Insights into protein-RNA complexes from computational analyses of iCLIP experiments

Nejc Haberman

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Department of Molecular Neuroscience

University College London

August 4, 2017

I, Nejc Haberman, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

RNA-binding proteins (RBPs) are the primary regulators of all aspects of post-transcriptional gene regulation. In order to understand how RBPs perform their function, it is important to identify their binding sites. Recently, new techniques have been developed to employ high-throughput sequencing to study protein-RNA interactions *in vivo*, including the individual-nucleotide resolution UV crosslinking and immunoprecipitation (iCLIP). iCLIP identifies sites of protein-RNA crosslinking with nucleotide resolution in a transcriptome-wide manner. It is composed of over 60 steps, which can be modified, but it is not clear how variations in the method affect the assignment of RNA binding sites. This is even more pertinent given that several variants of iCLIP have been developed. A central question of my research is how to correctly assign binding sites to RBPs using the data produced by iCLIP and similar techniques.

I first focused on the technical analyses and solutions for the iCLIP method. I examined cDNA deletions and crosslink-associated motifs to show that the starts of cDNAs are appropriate to assign the crosslink sites in all variants of CLIP, including iCLIP, eCLIP and irCLIP. I also showed that the non-coinciding cDNA-starts are caused by technical conditions in the iCLIP protocol that may lead to sequence constraints at cDNA-ends in the final cDNA library. I also demonstrated the importance of fully optimizing the RNase and purification conditions in iCLIP to avoid these cDNA-end constraints. Next, I developed CLIPo, a computational framework that assesses various features of iCLIP data to provide quality control standards which reveals how technical variations between experiments affect the specificity of assigned binding sites. I used CLIPo to compare multiple PTBP1 experiments produced by iCLIP, eCLIP and irCLIP, to reveal major effects of sequence constraints at cDNA-ends or starts, cDNA length distribution and non-specific contaminants. Moreover, I assessed how the variations between these methods influence the mechanistic conclusions. Thus, CLIPo presents the quality control standards for transcriptome-wide assignment of protein-RNA binding sites.

I continued with analyses of RBP complexes by using data from spliceosome-iCLIP. This method simultaneously detects crosslink sites of small nuclear ribonucleoproteins (snRNPs) and auxiliary splicing factors on pre-mRNAs. I demonstrated that the high resolution of spliceosome-iCLIP allows for distinction between multiple proximal RNA binding sites, which can be valuable for transcriptome-wide studies of large ribonucleoprotein complexes. Moreover, I showed that spliceosome-iCLIP can experimentally identify over 50,000 human branch points.

In summary, I detected technical biases from iCLIP data, and demonstrated how such biases can be avoided, so that cDNA-starts appropriately assign the RNA binding sites. CLIPo analysis proved a useful quality control tool that evaluates data specificity across different methods, and I applied it to iCLIP, irCLIP and ENCODE eCLIP datasets. I presented how spliceosome-iCLIP data can be used to study the splicing machinery on pre-mRNAs and how to use constrained cDNAs from spliceosome-iCLIP data to identify branch points on a genome-wide scale. Taken together, these studies provide new insights into the field of RNA biology and can be used for future studies of iCLIP and related methods.

Preface

This thesis describes my PhD work carried out at the UCL Institute of Neurology and the Francis Crick Institute, London, UK, between October 2013 and April 2017 under the supervision of Prof. Jernej Ule and Prof. Nicholas Luscombe.

All the experimental work in this thesis was performed by other members of the Ule group and collaborators. The experimental work described in chapter 3 was done by Dr Julian König, Dr Zhen Wang, Dr Jan Attig and Dr Ina Huppertz. The modified iCLIP protocol to identify 'readthrough cDNAs' in chapter 3 was developed and performed by Dr Ina Huppertz. The spliceosomal iCLIP protocol in chapter 5 was developed by Dr Michael Briese, with additional spliceosome-iCLIP and RNA-seq data produced by Dr Christopher Sibley. All the experiments and methods were done under the supervision of Prof. Jernej Ule.

Most of the work in chapter 3 have been published in the following article: Nejc Haberman*, Ina Huppertz*, Jan Attig, Julian König, Zhen Wang, Christian Hauer, Matthias W. Hentze, Andreas E. Kulozik, Herve Le Hir, Tomaz Curk, Christopher R. Sibley, Kathi Zarnack and Jernej Ule (2017), Insights into the design and interpretation of iCLIP experiments. *Genome Biology*, DOI: 10.1186/s13059-016-1130-x (See the Appendix).

Acknowledgements

I would first like to thank Professor Jernej Ule, my supervisor, for his help and guidance during my PhD, who has been a tremendous mentor for me. His enthusiasm and passion for research encouraged me to grow as a research scientist. I also want to thank him for all his brilliant comments and suggestions, and for allowing me to be a part of his group.

I would also like to thank my second supervisor, Professor Nicholas Luscombe. He gave me support and the opportunity to work with his group in the bioinformatic environment at The Francis Crick Institute in my final year. Working there enabled me to broaden my knowledge and experiences in bioinformatics.

I wish to thank all current and past members of the Ule group. Working and spending time with them was precious and made my past years the most enjoyable. I am also very grateful to Dr Chris Sibley, Dr Kathi Zarnack and Dr Jan Attig for the valuable feedback and support during my studies. Additionally I would like to express my gratitude to Dr Anob Chakrabarti and Dr Federico Agostini from Nicholas Luscombe's group for the valuable feedback in my final year.

I would like to acknowledge the Grand Challenge Programme who funded my PhD.

Finally, I would like to thank my family and friends, especially my parents for encouraging me to go studying in the first place and giving me all the support that I needed. Thank you!

List of Abbreviations

4SU	4-thiouridine
ALS	Amyotrophic Lateral Sclerosis
ATP	Adenosine triphosphate
BAM/SAM	Binary/Sequence alignment format
BP	Branch point
CCD	Charge-coupled device
cDNA	Complementary DNA
ChIP	Chromatin immunoprecipitation
CL	Cross-linking
CLIPo	CLIP optimisation tool
CLIP	UV cross-linking and immunoprecipitation
CMOS	Complementary metal-oxide-semiconductor
CRT	Cyclic reversible termination
CSV	Comma-separated values
DNA	Deoxyribonucleic acid
eCLIP	Enhanced CLIP
eIF4	Eukaryotic initiation factor
EJC	Exon junction complex
ENCODE	Encyclopedia of DNA elements
EST	Expression sequence tags
FDR	False discovery rate
FPKM	Fragments per kilobase of transcript per million mapped reads
FTD	Frontotemporal dementia
HITS-CLIP	High-throughput sequencing CLIP
HT-seq	High-throughput sequencing
hnRNP C	Heterogeneous nuclear ribonucleoproteins C1/C2
hnRNP	Heterogeneous nuclear ribonucleoprotein
iCLIP	Individual-nucleotide resolution CLIP
irCLIP	Infrared CLIP
mRNA	Messenger RNA

ncRNA	Non-coding RNA
NMD	Nonsense-mediated decay
NOVA	Neuro-oncological ventral antigen
PAR-CLIP	Photoactivatable-ribonucleoside-enhanced CLIP
PCR	Polymerase chain reaction
pH	Potential of hydrogen
PKR	Protein kinase R
PNK	Polynucleotide kinase
pre-mRNA	precursor mRNA
PTBP1	Polypyrimidine tract-binding protein 1
PTC	Premature termination codon
qPCR	Quantitative PCR
RBD	RNA binding domain
RBP	RNA-binding protein
RIP	RNA immunoprecipitation
RNA	Ribonucleic acid
RNase	Ribonuclease
RNP	Ribonucleoprotein
RRM	RNA recognition motif
rRNA	Ribosomal RNA
RT-HT-seq	Real-time high-throughput sequencing
RT	Reverse transcription
SDS	Sodium-dodecyl-sulfate
SF	Splicing factor
snRNA	Small nuclear RNA
snRNP	Small nuclear ribonucleoprotein
SBL	Sequencing by ligation
SBS	Sequencing-by-synthesis
STAU1	Staufen 1
SVM	Support vector machine
SVR	Vector regression

SS	Splice site
tRNA	Transfer RNA
U2AF	U2 auxiliary factor
UMI	Unique molecular identifier
UTR	Untranslated region
UV	Ultraviolet light

”See first, think later, then test. But always see first. Otherwise, you will only see what you were expecting. Most scientists forget that.”

- Douglas Adams

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this thesis work I will present an overview of CLIP-type data analysis combining previously available computational pipelines, together with new approaches that I have developed during my studies. The goal of this thesis is to present computational methods that will allow to better interpret iCLIP data, while extracting new insights from iCLIP experiments. This will allow to better address the challenge in correctly assigning binding sites for RNA-binding proteins (RBPs), a process that is crucial for RBP characterisation. In this first chapter, I will give a short introduction about RBPs and a quick overview of available computational methods to study protein-RNA interactions. I will then provide a brief summary of the process of cellular splicing and of the RBPs that I have been working on during my studies.

One of Britain's greatest scientists, Francis Crick, was the first to propose the term 'Central Dogma' in 1958 in reference to what is now a keystone of molecular biology: the process flow from genetic information to a functional protein. He went into greater detail by claiming that this process flow is possible in both directions between DNA and RNA, but it is not possible from the protein to nucleic acid or to another protein [1].

Gene expression is a cellular process that allows the genetic information stored in DNA to be synthetized into a functional gene product, where a part of DNA is transcribed to RNA to then produce a protein. Protein production in eukaryotes starts with the transcription of DNA to a precursor mRNA (pre-mRNA), which undergoes extensive co- and post-transcriptional processing. It is then exported to the cytoplasm and translated into a protein. The complexity of RNA is much greater than of DNA as it can form complex structures, controlling the sequence information together with trans-acting factors. The several steps in regulation of gene expression are mainly controlled by multiple interactions between RNA cis-elements and trans-acting factors such as RBPs.

1.1 RNA-binding proteins

RBPs bind to RNA molecules in the nucleus and/or the cytoplasm. RNA/RBP complexes form important ribonucleoprotein (RNPs) complexes involved in polyadenylation, RNA modification, pre-mRNA splicing, mRNA export from the nucleus to the cytoplasm, localisation, translation, control of mRNA stability and transcript degradation [2, 3, 4, 5]. In order to understand the post-transcriptional regulatory mechanisms and their functions, it is important to identify the exact binding site of RBPs on endogenous transcripts.

Genome sequencing technologies such as next-generation sequencing and protein mass spectrometry have led to fast research progress in the field of RNA processing over the past decade. These technologies allowed discovery of many new proteins that bind thousands of transcripts via specific binding sites [6]. There are over 1,500 catalogued genes encoding for RBPs in the human genome [7, 8] but for most of the RBPs we still have an incomplete understanding of their specificity and their involvement in cellular processes [9]. Many RBPs do not bind only a simple contiguous RNA motif, but rather assemble on RNA into a complex with other RBPs or bind clusters of short motifs that can be dispersed over dozens of nucleotides [10]. Since they have a major role in gene regulation, it is unsurprising that perturbations of RBP activity are strongly linked to disease. Mutations within RBP binding sites or the RBPs themselves can cause misregulation at different levels of RNA processing, such as gene expression or alternative splicing [11]. RNA regulatory dysfunction or loss is associated with many diseases [9] including cancer [12] and has become increasingly recognized as a central component in neurological disorders [13, 14, 15, 16, 17, 18]. Consequently, studies on the interactions of RBPs with the transcriptome are becoming increasingly popular to understand cellular function and disease [19, 20, 21].

The discovery of heterogeneous nuclear ribonucleoproteins (hnRNPs) and other pre-mRNA/mRNA-binding proteins led to the identification of the first amino acid motifs and functional domains that confer RNA-binding properties [22]. Most RBP-RNA interactions occur through a variety of single or multiple protein domains, including RNA recognition motifs (RRMs), K homology (KH) domains and small structural motifs such as zinc fingers [23]. Importantly, cellular RNA is not naked but it is associated with multiple proteins in RNP complexes, which are involved in RNA regulation [24]. Early biochemical studies found that the pre-mRNA transcripts are packaged into heterogeneous nuclear ribonucleoprotein (hnRNP) particles by a group of RBPs termed hnRNP proteins [25, 26]. *In vitro* analysis is insufficient to understand the characteristics of RBP targets because *in vitro* methods do not take into account the effect of the complex cellular environment in which protein-RNA interactions normally take place. It has been shown in fact that the most highly enriched RBP motifs detected *in vitro* are not necessarily the preferred binding site motifs *in vivo*, or *vice versa* [24]. Therefore, it is important to understand what the characteristics of RNA targets of RBPs are and how their *in vivo* binding specificity is achieved. Due to the fact that RNA is single-stranded, it is able to fold back on itself and form elaborate secondary and even tertiary structures. These structures have important regulatory roles in RBP-RNA interaction as, for example, they can mask RBP binding motifs [27]. Although most RBPs prefer to bind single-stranded regions, there are also proteins that bind structured regions such as Staufen 1 (STAU1), which is known for binding to double-stranded RNA duplexes [28]. There are numerous protein domains contained across the genome that have the potential to bind RNA among other factors, such as secondary structure and interactions with other complexes, and we may not be fully aware of all of the protein domains that can mediate interaction. For example, STAU1 protein can bind in an especially long-range form of RNA duplexes that can only be detected experimentally [29], so it is important to first study RBP specificity through experimental data before using any computational prediction models.

In order to understand these processes, many new methods have been developed by using techniques involving protein-RNA crosslinking and immunoprecipitation of RBPs (CLIP, iCLIP, eCLIP, PAR-CLIP, irCLIP) (see subsection 1.6.1). All these methods revealed the high coverage of protein-RNA interactions across all transcripts, which can be precisely defined by the crosslink clusters. But there are still many unanswered questions about RBP specificity and function in post-transcriptional regulation. Before we can start asking these questions, it is fundamental to first correctly assign the position of the RBPs' RNA binding sites. So far, most progress in this direction has been made with combination of *in vivo* experiments and computational methods [30, 31, 32, 33]. There are many available computational tools to predict RBPs binding sites based on RBP binding motifs [30], structured regions and other features [34, 35, 36], but currently they only work for a limited number of RBPs with strong motif characteristics and only a small fraction of predicted sites are in fact shown experimentally to be occupied by an RBP [37]. For example, NOVA protein was one of the first RBPs studied with the CLIP technology, where it was characterised as a splicing regulator in the mouse brain that binds to intronic YCAY clusters [38]. Another group was able to follow up this study by modelling the alternative splicing events that are regulated by the neuron-specific factor NOVA in the mouse brain [39]. It is certainly a great challenge to model these interactions but it is important to first learn about each RBP through experimental data. That is why it is so important that we first correctly analyse experimental data to precisely map the binding sites, which then allows us to learn more about RBP specificity and its function.

1.2 Pre-mRNA splicing

During the expression of a gene in eukaryotes, a stretch of DNA is transcribed into pre-mRNA. The process of creating mature mRNA from pre-mRNA is called splicing, by which introns are removed and exons are spliced together, before the RNA can be used to produce a correct protein during translation. Richard J. Roberts and

Phillip A. Sharp first discovered pre-mRNA splicing in the 1970s, and were awarded the Nobel Prize for their work in 1977. They first noticed unusually long RNA in the nucleus of vertebrate cells compared with the shorter mRNA that emerged in the cytoplasm. Gayle Knapp and his colleagues later sequenced a series of tRNAs from yeast and noticed additional nucleotide sequences within the middle of the gene that are not present in the mature tRNA. One of the main observations of their study was that there must be an activity of multiple enzymes that removes these intervening sequences to produce mature-sized RNA [40]. This process broadens the diversity of the transcriptome via alternative splicing (see subsection 1.3).

For most eukaryotic genes, splicing occurs in several steps, catalysed by the spliceosome, a complex and dynamic molecular machinery composed of small nuclear ribonucleoproteins (snRNPs). The splicing process requires multiple snRNPs to contact distant regions of the pre-mRNA, and involves a multitude of remodelling steps (Figure 1.1 - [41]). One of the key steps is the recognition of the 3' and 5' splice sites, which are located upstream and downstream of exons, respectively. Splice sites are highly conserved across species with strong sequence motifs that contribute to the splice site strength. Splice sites can be classified as weak or strong according to their similarity to consensus sequence [42]. The splice site strength plays a crucial role in the splicing efficiency by interacting with snRNPs that coordinate and catalyse the splicing reaction. In the early splicing complex (complex E), U1 and U2 snRNPs bind to the 5' and 3' splice site on the pre-mRNA, respectively. A number of auxiliary factors accompany the binding of these snRNPs including cytotoxic granule-associated RBPs TIA1 and TIAL1 at the 5' splice site, and the U2 auxiliary factor (U2AF) with its subunits U2AF35 (U2AF1) and U2AF65 at the 3' splice site [43, 44, 45] (Figure 1.1 - [41]). Within this machinery, U2AF forms multiple RNA-protein interaction together with other RBPs and RBP complexes during spliceosome assembly [46]. In an early stage of this process, the branch point (BP) is recognised by a dynamic complex comprising splicing factor 1 (SF1) in the region upstream of the 3' splice site, where U2AF35 identifies the AG dinucleotide at the end of an intron and U2AF65 binds the polypyrimidine tract downstream of the BP and upstream of U2AF35 [47, 48] (Figure 1.1 - [41]). U2AF is not only a crucial factor of the splicing machinery, but it is also a splicing regulator which can be superseded by other competitive RBPs, particularly around the weak splice sites in alternative splicing [49]. This competition is usually performed by RBPs with similar binding motifs such as hnRNPC [50], TIA-1 [51], YB-1 [52] and PTBP1 [53]. In the later complexes (complex B and C), three further snRNPs (U4, U6 and U5) finish the splicing process in a series of ATP-dependent steps to form a mature RNA [54].

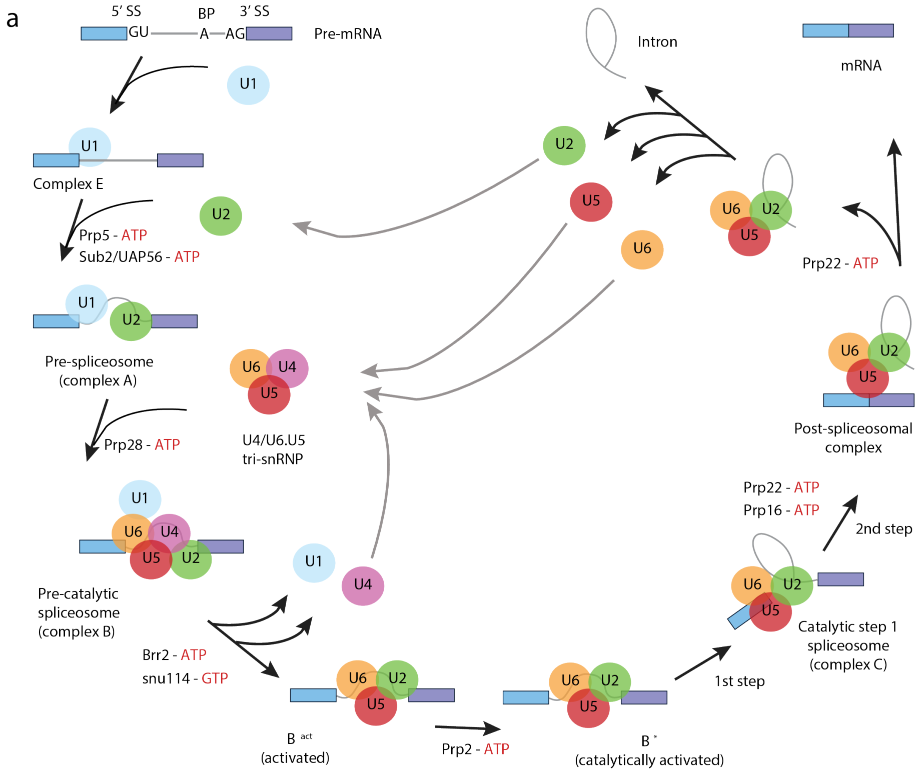


Figure 1.1: The schematic representation of pre-mRNA splicing in eukaryotes [41].

At the beginning of the splicing process pre-mRNA contains intronic and exonic elements. Introns are spliced out and exons are joined together to form a mature mRNA. The process involves a stepwise binding of small nuclear ribonucleoprotein (snRNP) complexes formed by small nuclear RNAs (snRNAs) interacting with proteins. First, the U1 snRNP binds to the pre-mRNA on the 5’ splice site, while the U2 snRNP binds close to the 3’ splice site on the branch point (BP) sequence encompassing the BP. U2 binding results in bulging out of the unpaired BP adenosine increasing its propensity to effectuate a nucleophilic attack on the 5’ splice site during the next catalytic step. Subsequently, the U4/U5/U6 tri-snRNP binds the intron and leads to the displacement of U1 and U4 and the formation of a catalytic spliceosome. This complex catalyzes the transesterification of the BP adenosine 2’-OH group to the guanosine phosphate of the 5’ splice site resulting in a lasso-shaped intron (lariat) that is spliced away thereby joining two exons together. The same process can occur with different combinations of 3’ and 5’ splice sites resulting in the formation of different isoforms that are regulated by the splicing machinery [55].

1.3 Alternative splicing and its regulators

Alternative splicing is a highly regulated mechanism that allows a single gene to codify for multiple protein variants as well as to regulate gene expression [56]. It has been shown that the majority of mRNA isoforms in humans are the result of alternative splicing. About 92 to 94% of human genes are alternatively spliced, of which 85% have a minor isoform frequency of at least 15% [57]. Alternative splicing plays an important role in development and physiology and is also associated with several diseases [58, 59]. A number of different RBPs identified as splicing factors (also called SF proteins) are involved in alternative splicing regulation, coordinating intron removal and the decision of including an alternative exon in the mature transcript. Some of these SF proteins can either enhance or repress exon inclusion in a position-dependent manner by assembling at different sites of the intron or the exon, either on the splice sites themselves or on other motifs known as splicing silencers or enhancers [60]. For example, the RBP PTBP1 mainly works as a splicing repressor by binding the upstream region of its target alternative exons (see subsection 1.3.1) [61]. A similar mechanism was first observed for the RBP NOVA, which mediates regulation of spliceosome assembly and alternative splicing of a subset of exons in neurons [33]. Later, another mode of alternative splicing regulation was observed for the RBP TDP-43 by discovering deep intronic regions where the protein binds to repress exon inclusion [62].

1.3.1 Polypyrimidine tract-binding protein 1

Polypyrimidine tract-binding protein 1 (PTBP1) is an RBP that regulates inclusion of a defined set of alternative exons. It has been extensively studied as a paradigm for its mechanism of splicing regulation and as an example of a tissue-specific alternative splicing regulator. PTBP1 is a 57-kDa protein that binds to CU-rich sequences [63]. It contains four RNA-binding domains, each of which can bind a pyrimidine-rich (Y-rich) motif to facilitate interactions [64, 65, 66]. Moreover, PTBP1 proteins can work in clusters to form higher-order complexes when bound to RNA, and up to eight PTBP1 proteins were observed on a long RNA binding site [67, 68]. All four RNA-binding domains of PTBP1 are capable of interacting with

RNA, and its three main isoforms are the result of alternative splicing [69]. PTBP1 is also one of the best-studied RBPs, for which the sequence-specific RNA binding is understood on an atomic level [66]. One of the main reasons why PTBP1 was one of the first discovered splicing regulators is because of the large number of its RNA targets and its efficient ultraviolet light (UV) crosslinking to the polypyrimidine tract [70]. One example of the well-studied mechanism of PTBP1 as a splicing repressor is its competitive assembly on the binding site of U2AF65 at the 3’ splice site (Figure 1.2a, b), [71, 72, 73]. In contrast, PTBP1 can also work as a splicing enhancer by assembling at a downstream region away from the alternative exon (Figure 1.2c) [61]. Besides being a splicing regulator, it is also known to function in a large number of additional cellular processes such as polyadenylation, mRNA stability and translation initiation [74].

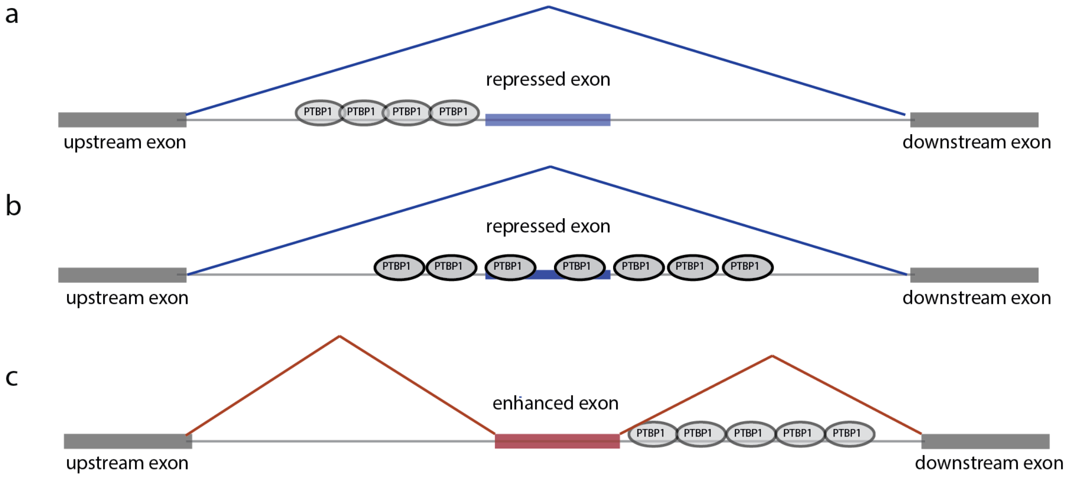


Figure 1.2: A schematic explaining different scenarios how PTBP1 regulates splicing of alternative exons.

If PTBP1 binds across a) the 3’ splice site or b) expands over the whole exon, it represses exon inclusion. PTBP1 can also enhance exon inclusion by binding over the 5’ splice site c) of an alternative exon.

1.4 Biological importance of alternative splicing

In order to respond quickly to physiological changes and external stimuli, cells undergo large-scale changes in gene expression that must be coordinated in a precise spatio-temporal fashion. Alternative splicing provides a fast, reliable and dynamic tool to tackle some of these major rearrangements. Comparison of expression sequence tags (ESTs) data revealed similar levels of alternative splicing in evolutionarily distinct species, emphasizing the importance of alternative splicing throughout evolution [75, 76]. Dynamic regulation of different isoforms enables complex cellular responses, which are essential in cellular responses such as regulation of cell viability, differentiation and apoptosis in response to environmental cues [77, 78]. Alternative splicing isoforms are most prevalent in brain cells compared to other tissues [79, 80]. Also, numerous diseases have been associated with changes in alternative splicing [81].

Here, it has been shown that AS mediated by the *Sex-lethal* (Sxl) RBP, regulates protein products *Sex-lethal* itself, *transformer*, and male specific *lethal-2* genes that are needed for sex determination in Drosophila melanogaster [82, 83, 84]. Prominent examples where alternative splicing plays a pivotal role are in cell differentiation, lineage commitment in neuronal progenitors [85] and immune response [86], suggesting direct involvement in tissue-identity acquisition and organ development [87].

RBP expression levels are tightly regulated in mammalian neuronal differentiation, to form different splicing products via regulation of alternative splicing. Several brain-specific factors such as PTBP1, PTBP2, NOVA1, NOVA2 and SRRM4 have been identified as important regulators during brain development [87, 88, 89]. For example, a switch between alternative-spliced PTBP1 and PTBP2 proteins is one of the most important mechanisms in neuronal differentiation [90]. This switch happens in neural progenitor cells, where PTBP1 protein represses the inclusion of exon 10 in PTBP2, which in turn leads to exon skipping, forming a transcript with a premature termination codon (PTC), targeting the transcript to be degraded by nonsense-mediated decay (NMD) [91, 90]. In contrast, SRRM4 protein enhances the PTBP2 exon 10 inclusion when cells exit the cell cycle, allowing the induction of PTBP2 to promote neuronal development and tissue maintenance [88, 89]. Another example of brain-specific regulator is the neuronal splicing factor Nova, which regulates neuronal pre-mRNA alternative splicing by binding to RNA in a sequence specific manner [92, 93], and has been identified as a target antigen in patients with paraneoplastic opsoclonus-myoclonus ataxia (POMA), a human neurological syndrome characterized by motor and cognitive deficits [94, 95, 96].

1.5 Next-generation sequencing

Next-generation sequencing (NGS), also known as high-throughput sequencing, is a series of modern sequencing technologies that allow sequencing of large portions of DNA and RNA fragments. NGS technology brings faster and low cost sequencing since it can work with small amounts of material in comparison with the previously used Sanger sequencing, developed by Frederick Sanger and colleagues in 1977 [97]. NGS technology has revolutionised the study of genomics and molecular biology, and it has been incorporated into multiple methods such as chromatin immunoprecipitation sequencing (ChIP-seq), RNA sequencing (RNA-seq), crosslinking immunoprecipitation (CLIP), whole genome sequencing, *de novo* genome assemby, genome wide structural variation, mutation detection, sequencing of mitochondrial genomes and personal genomics. This has also empowered researchers to detect, among others, causative changes in inherited disorders and complex human diseases [98, 99].

NGS technology is based on DNA or RNA fragmentation into smaller sequences, where millions or billions of them can be processed and sequenced in parallel. The length of these fragments can vary among different sequencing platforms. For example, Illumina HiSeq 2500 supports read lengths of 50, 100 and 150 base pairs (bp) for single or paired-end reads and Illumina MiSeq supports even longer reads up to 250 and 300 bp. As new technologies appeared, a number of sequencing companies emerged. Each developed their own methods and had variable impacts upon what type of experiments are more feasible. These includes Illumina (Solexa), Roche 454, Ion Torrent (Life Technologies product) and SOLiD sequencing [100]. These sequencing platforms can be classified as ’sequencing-by-synthesis’ (SBS) or ’sequencing by ligation’ (SBL), where the former uses DNA polymerase and the latter DNA ligase [101].

- Roche 454

  The 454 sequencer is one of the first next-generation sequencing technologies that was introduced by the Roche company in 2005 [102, 103]. It is a SBS pyrosequencing based sequencer, which relies on generation of light after nucleotides are incorporated into a growing chain of DNA [104]. The sequencing preparation starts with ligation of specific sequencing adapters to DNA fragments. These fragments are then captured in an aqueous droplet, along with a bead covered with millions of oligomers that are complementary to ligated adapters. Emulsion-PCR is then used to make multiple copies of each adapter-ligated DNA fragment, resulting in a chip containing individual micro wells with beads in which each well contains many cloned copies of the same DNA fragment [102, 101]. As a dNTP is incorporated into a strand, an enzymatic cascade occurs resulting in a luminescent signal that triggers pyrophosphate release, which produces flashes of light that are detected by a charge-coupled device (CCD). This signal is recorded as a series of light peaks that can be translated into genomic sequence [105, 102].

- Illumina (Solexa)

  This system covers the largest market for sequencing instruments compared to other available platforms and it was also used for sequencing HT-seq data presented in this thesis, including RNA-seq and iCLIP related methods. Similar to Roche 454, Illumina sequencers require a pre-amplification step, which involves the ligation of specific adapters to DNA fragments on either end. The surface of a glass flow cell is washed, together with oligos attached that hybridize to the ends of the fragments [106, 102]. Illumina does not involve pyrosequencing but it uses the cyclic reversible termination (CRT) instead, which sequences the template strand one nucleotide at a time through fragment replication of base incorporation, imaging, washing and cleavage [106]. This approach uses fluorescently labeled 3'-O-azidomethyl-dNTPs to pause the polymerization reaction of a single nucleotide per cycle and fluorescent imaging with CCD camera to identify the added nucleotide [107]. After each imaging cycle, 3'-O-azidomethyl-dNTPs are removed and the molecules are washed away so the process can be repeated [106].

- Ion Torrent

  Similar to 454 sequencer, Ion Torrent also uses a chip containing individual micro wells with beads to which DNA fragments are attached. However, unlike 454 which is based on CRT and SNA methods, Ion Torrent approach relies on a single signal marking the incorporation of a dNTP into an elongating strand. As a consequence, each of the four nucleotides must be added iteratively to a sequencing reaction to ensure only one dNTP is responsible for the signal [105]. In addition, the Ion Torrent platform does not use an enzymatic cascade to generate a signal, but uses a pH sensitive approach that relies on an electrochemical detection system called an ion-sensitive field-effect transistor (ISFET), using complementary metal-oxide-semiconductor (CMOS) microdetectors to detect small changes in pH. These changes are the result of a hydrogen ion (or proton) release when a nucleotide base is added to a growing DNA strand, which causes a slight pH change that can be detected by a CMOS sensor [105, 105].

- SOLiD

  SOLiD stands for Small Oligonucleotide Ligation and Detection System which was developed for parallel sequencing by stepwise ligation process. The process starts with an emulsion PCR step of DNA fragmented library of flanked ligated adapters in a similar way as used by 454, but the sequencing part is entirely different from the previously described sequencing platforms [102]. This one uses SBL approach to utilize DNA ligase, instead of 'sequencing-by-synthesis' [108]. Each cycle of sequencing involves the ligation of octamer probes, where the first two nucleotides represent 16 dinucleotide combinations including a fluorescent label. There are four different dinucleotide combinations with a fluorescent tag, enabling labelling of all 16 dinucleotide tags. When a probe anneals adjacent to the adapter, the primer strands are ligated and fluorescence is captured, corresponding to the ligated probe. Multiple cycles of ligation, imaging and cleavage are performed using two ligation events per base, determining the eventual read length, which significantly decreases the error rates [109].

Due to its cost-effectiveness Illumina is one of the most popular platforms. Nevertheless, for some applications such as whole genome assembly, other platforms with larger read lengths or lower error rate can be advantageous. All these sequencing methods use sensitive detectors (CCD, CMOS) and processing systems that produce raw sequencing data in a textual format known as FASTQ format. This format contains sequences also known as ’reads’, where each read information is composed of ID tag, base sequenced, and quality scores for each base [110].

1.6 UV crosslinking technologies to study RBPs in the context of RNA processing on a genome-wide scale

Crosslinking with ultraviolet (UV) light is commonly used to create a covalent bond between proteins and nucleic acid, which can be used to determine contact points of any nucleobase within DNA or RNA. Absorption of UV light by a molecule introduces energy sufficient to break or reorganize most covalent bonds. When low intensity light is used, an electron of the nucleobase can absorb a single photon, which is promoted to the first excited singlet state (Si) [111]. The excited nucleobase can either react from singlet state (Si), relax to the ground state, or enhance intersystem crossing to the first excited triplet state (Ti), which can also react or decay to the ground state. Crosslinking can happen in both states (Si or Ti) [111]. Different reactions within the protein and the nucleic acid can take place with the use of continuous UV irradiation. This reaction can occur as amino acid modifications (destruction of tryptophan), modifications to the DNA (strand breaks) or reactions between nucleobases (formation of cyclobutane dimers) [111, 112]. After the UV light crosslinking, many amino acids can react within a macromolecular complex. It has been shown that poly A, poly T, poly G, poly C and poly U, with cysteine, lysine, phenylalanine, tryptophan and tyrosine amino acids are the most reactive and have the high potential to crosslink to DNA molecules among all common amino acids [113, 114, 115]. On the individual nucleobases, uridine-cysteine [116], thymine-lysine [117], 5-methylcytosine-serine, 5-methylcytosine-threonine [118], and thymidine-tyrosine [119] had the highest crosslinking reactions, which can also lead to potential biases in the UV crosslinking based methods such as CLIP [120].

In recent years, there has been great progress and an increase in the number of methods using UV crosslinking to study protein-RNA interactions. These methods were first introduced by Gideon Dreyfuss and his group, who used UV light irradiation to crosslink direct contacts between RNA and proteins that form *in vivo* [121]. In order to understand how the binding of RBPs instructs their function, it is important to identify their binding sites on endogenous transcripts. Many RBPs bind clusters of short sequence or structural RNA motifs that can be dispersed over dozens of nucleotides and are therefore difficult to predict computationally [10]. Therefore, experimental methods for transcriptome-wide mapping of protein-RNA interactions have been developed [6].

For high-throughput sequencing of isolated RNAs it is important to use the optimal wavelength to efficiently crosslink proteins to RNA and not DNA. For example, in iCLIP and other CLIP related methods, it is important that the cells or tissue samples are irradiated with a ’low energy’ wavelength of 254 nm, and for PAR-CLIP a wavelength of 365 nm is needed for efficient crosslinking with UV-A light (see subsection 1.6.1 - PAR-CLIP) [6]. CLIP methods are also based on ’zero-length’ crosslinking, which allows direct evaluation of contact interactions between peptides to directly crosslink atoms of the protein to RNA [122].

1.6.1 Diversity of different techniques to study RNA-protein interactions, gene expression and alternative splicing

- **CLIP** (also known as HITS-CLIP or CLIP-seq)

  The UV crosslinking and immunoprecipitation (CLIP) [38] method was developed to identify positions of protein-RNA interactions *in vivo*. CLIP uses UV light exposure to form a covalent bond between the protein and RNA. On cell lysis, the protein-RNA complex is immunoprecipitated with an antibody for the RBP of interest [38, 6]. The co-purified RNA molecules are reverse-transcribed and amplified with the aid of 5' and 3' adapters. During library preparation, the crosslinked RBP is removed through proteinase K digestion, leaving a small peptide on the crosslink site, which impairs reverse transcription and commonly leads to truncation of cDNAs at the crosslink site. The original CLIP protocol only amplifies those cDNAs that readthrough the crosslinked peptide [38] and it was first demonstrated with NOVA proteins by sequencing cDNA clones with Sanger sequencing [38]. The CLIP method was later optimised with a more efficient protocol for cDNA amplification and ligation of barcoded adapters, which enabled amplification of the low concentration of isolated RNA and sequencing of multiplexed libraries. Moreover, CLIP was combined with high-throughput sequencing (HITS-CLIP), which allows sequencing of millions of cDNAs in a single run [123] and was later optimised to a single nucleotide resolution method known as iCLIP [124].

- **iCLIP**

  One disadvantage of the previously mentioned CLIP method is that only those cDNAs that readthrough the crosslink sites can be amplified. These account for merely $\sim$10% of cDNAs [120]. This loss of cDNAs results in less quantitative information in the resulting cDNA libraries. To overcome this limitation, individual-nucleotide resolution CLIP (iCLIP) was developed. Reverse transcriptase can be caused by a crosslinked protein, but also to UV-induced strand breaks in the RNA, intra-RNA crosslinks, or stuttering of the reverse transcriptase [125]. The CLIP method purifies short RNA fragments that are crosslinked to RBPs, and due to the low amount of UV light used and low efficiency of UV crosslinking, it is unlikely that additional types of intra-RNA crosslinks form in the same fragment. Also, given the short size of the RNA fragments and reverse transcription with efficient transcriptase at 55 degrees, it is unlikely that stuttering of the reverse transcriptase would lead to many truncation events. Therefore, cDNAs truncated at the crosslink site, referred to as 'truncated cDNAs', are expected to dominate the resulting cDNA library. The nucleotide on the genome that precedes the mapped cDNAs is thus expected to correspond to the crosslink site (Figure 1.3). iCLIP also solved another problem of earlier CLIP approaches, i.e. that PCR over-amplification of low amount of the sequencing material can generate a high number of cDNA duplicates. Partly this can be solved by collapsing identical cDNA reads, but in the more recent version (iCLIP and later) an additional step was added to the protocol. This allows a better quantification of cDNA molecules, by including unique molecular identifiers (UMIs) as a random barcode to discriminate unique cDNA products from PCR duplicates [124]. This technique has now been optimised [126] and broadly accepted for transcriptome-wide studies of protein-RNA interactions.

- **PAR-CLIP**

  A variant of the CLIP method, Photoactivatable-Ribonucleoside-Enhanced CLIP (PAR-CLIP), uses point mutations and deletions to identify crosslink sites of RBPs after the 4SU incubation. Exact crosslink sites are identified by thymidine-to-cytidine transitions on the cDNAs prepared from immunopurified RNPs of 4-thiouridine-treated cells [127] (Figure 1.3). In theory, the method improves the resolution problem compared to the CLIP method, but it is limited to cultured cells that are able to incorporate the required ribonucleoside analogs. This method was later optimised for *in vivo* experiments known as iPAR-CLIP (*in vivo* PAR-CLIP) [128]. Among the resolution improvements, PAR-CLIP also uses different crosslinking conditions, i.e. UV-A instead of UV-C light radiation [129] (Figure 1.3).

- **irCLIP**

  There are many technical challenges that can be improved in available CLIP-based protocols [130, 131]. One of the main challenges is the standard radioactive labelling of RBP-RNA complexes that occurs on the 5' ends of crosslinked RNA molecules. This can be an obstacle to setup iCLIP in institutions with restriction on radiation use. Also, the decay of radioactive reagents can interfere with the signal across experiments [132]. To improve these limitations, infrared-CLIP (irCLIP) was developed in which an infrared-dye-conjugated and biotinylated ligation adaptor allows rapid and quantitative analysis of *in vivo* captured protein-RNA interactions. This step keeps the same efficiency in ligation reactions as a standard adaptor ligation and reduces the time required for protein-RNA complex visualisation [132]. irCLIP also provides a more stream-lined protocol with a shortened cDNA isolation process allowing a certain degree of automatisation.

- **eCLIP**

  Enhanced CLIP (eCLIP) was developed to simplify and improve certain technical steps from the original iCLIP protocol, such as ligation efficiency in library preparation of RNA fragments and over-amplification of cDNAs. Similar to irCLIP, eCLIP skips the radioactivity step and adds two ligation reactions to improve ligation efficiency. In the first step, an indexed 3' RNA adapter is ligated to the crosslinked RNA fragment, still on the immunoprecipitation beads, and in the second step, a 3' single-stranded DNA adapter is ligated after reverse transcription [133]. The next advantage of the eCLIP method is that it decreases the background noise by generating size-matched input controls called mock eCLIP [134]. A pre-immunoprecipitation RNase-treated lysate control is prepared in parallel with the main experiment, following the same isolation procedure like for the protein of interest. This control serves as a non-specific background signal that can be used as a normalisation step in RBP binding site identification [133]. Another difference compared to other methods is that the eCLIP library is sequenced with paired-end cDNA reads. This method was used for the datasets submitted to the Encyclopedia Of DNA Elements (ENCODE), where eCLIP is systematically applied to all known RBPs in human HepG2 and K562 cells. The method is quite recent and therefore there are few publicly available analyses yet.

1.6.2 Methods to study gene expression and alternative splicing

To better understand the mechanism of RBPs, it is important to measure transcript abundance and alternative splicing.

- **Microarrays**

  Microarrays were first designed to measure gene expression by targeting already known isoforms through preparation of DNA oligonucleotide probes that correspond to a specific gene region [135]. This approach did not give any alternative splicing information: this came later with the annotate method, which aims to discover new alternative exons [58] and gene annotation of different isoforms [136]. The limitations of microarrays include their poor quantification of lowly and highly expressed genes, and the fact they can only detect a sequence that the array was designed to detect. This means that unannotated genes will not be detected by microarrays because there is no complementary sequence on the array [137].

- **RNA-seq**

  The RNA-seq brings robust analysis of gene expression across the transcriptome, which enables us to detect novel features of RNA expression at a high resolution [109]. RNA-seq describes the full length of known and novel transcripts to investigate all the RNAs present in a sample, including messenger RNAs (mRNAs), and long noncoding RNAs, along with other untranslated regions [138]. The method has a number of advantages over microarrays, such as increased specificity and sensitivity, discovery of single nucleotide variants (SNVs), detection of different transcript isoforms and RNA splicing events, and since it does not require reference genome assembly it can be used for any organism.

The basic workflow of the RNA-seq protocol involves a collection of the sample of interest, RNAs isolation, cDNA fragmentation, size selection, adapter ligation and RT-HT-sequencing, which is followed by computational analysis. In general, RNA is first isolated from tissue and can either be kept as a whole, including ribosomal-RNA (rRNA), or filtered by poly-dT hybridization for 3' polyadenylated RNA (poly(A)+ RNA). Poly(A) selection requires a high proportion of mRNA, where the 3' ends of transcripts are protected from degradation, leading to a higher number of reads at the end of the transcript. In this case rRNAs, which represents over 90% of the RNA in a cell, and also noncoding RNAs are depleted in the sample [139]. In the next step, fragmentation and size selection are performed to reverse transcribe the sample into a library of cDNA fragments with ligated adaptors. The library is then amplified and sequenced with the NGS to obtain short sequences from one end (single-end sequencing) or both ends (paired-end sequencing). The size of the read can vary in length from short ($\sim$30 nt) to hundreds of nucleotides, depending on the sequencing platform (see section 1.5) and library preparation [138, 140, 141].

Another essential factor in designing an RNA-seq experiment is the number of replicates. The number of technical and biological replicates in the experiment also depends on the amount of technical and biological variability, as well the desired type of statistical analysis [142]. Increasing the number of replicates minimizes the false positives and usually leads to more robust outcomes, ensuring meaningful biological interpretation of the results [143, 144].

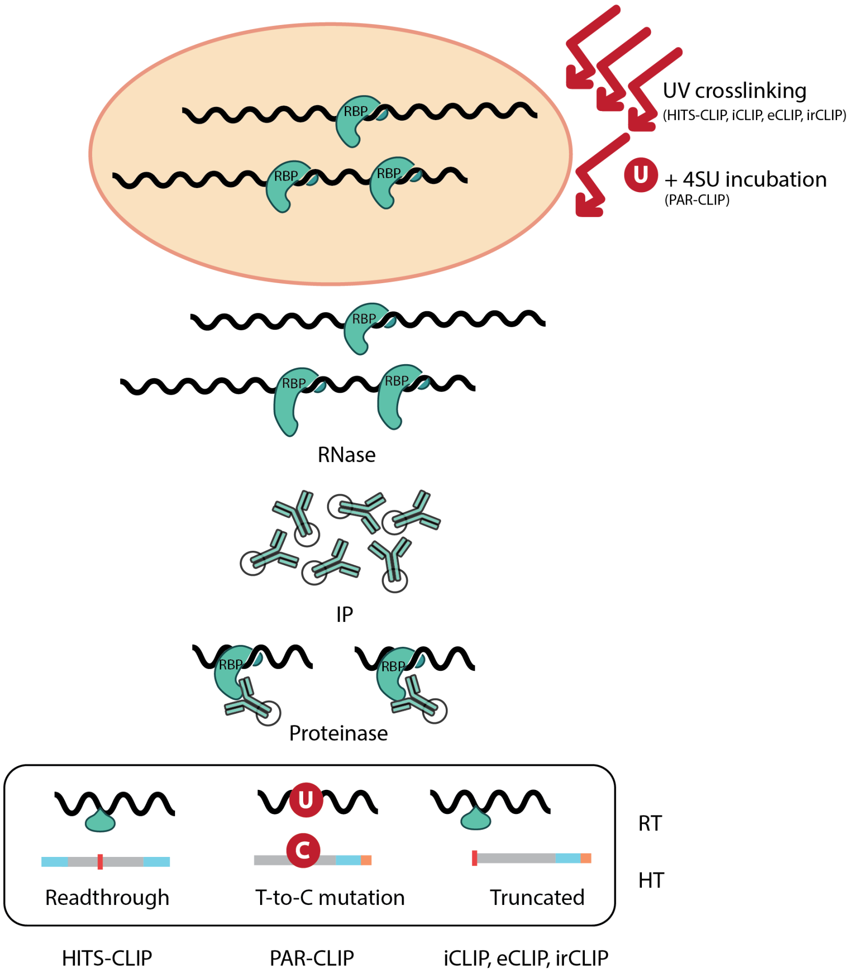


Figure 1.3: Schematic description of the CLIP-related protocols (HITS-CLIP, PAR-CLIP and iCLIP/eCLIP/irCLIP).

All protocols are based on crosslinking with ultraviolet light (UV), with an additional incubation with 4-thiouridine (4SU) for PAR-CLIP. After RNase digestion and immunoprecipitation (IP), RNA-protein complex is digested with proteinase K, and the RNA is reverse transcribed with a primer including barcode. Different types of cDNAs are generated: readthrough in HITS-CLIP, readthrough with T-to-C mutations in PAR-CLIP and truncated cDNAs in iCLIP/eCLIP/irCLIP. All cDNAs are then amplified with PCR and sequenced on a high-throughput sequencing platform.

1.6.3 Analysis of CLIP-related data

Over the years, CLIP has become a state-of-the-art method to study RNA-protein interactions. There are certain protocol differences between the CLIP related methods but the main workflow of data analysis is very similar in all these methods (Figure 1.4). After the initial quality check and filtering of sequenced cDNAs, they are pre-processed for adapter removal and trimming of low quality cDNA reads if necessary. Then, pre-processed cDNAs are aligned to the reference genome followed by the peak-calling step to identify enriched binding sites associated with the RBP of interest. This may be followed by motif analysis or prediction models to improve the genome-wide coverage of binding sites [145]. In this section, I will focus on the data analysis and specifications of CLIP-related methods. There are many different tools available for each step of analysis, but I will give a short overview of the most recent ones and focus on those that I used in this thesis.

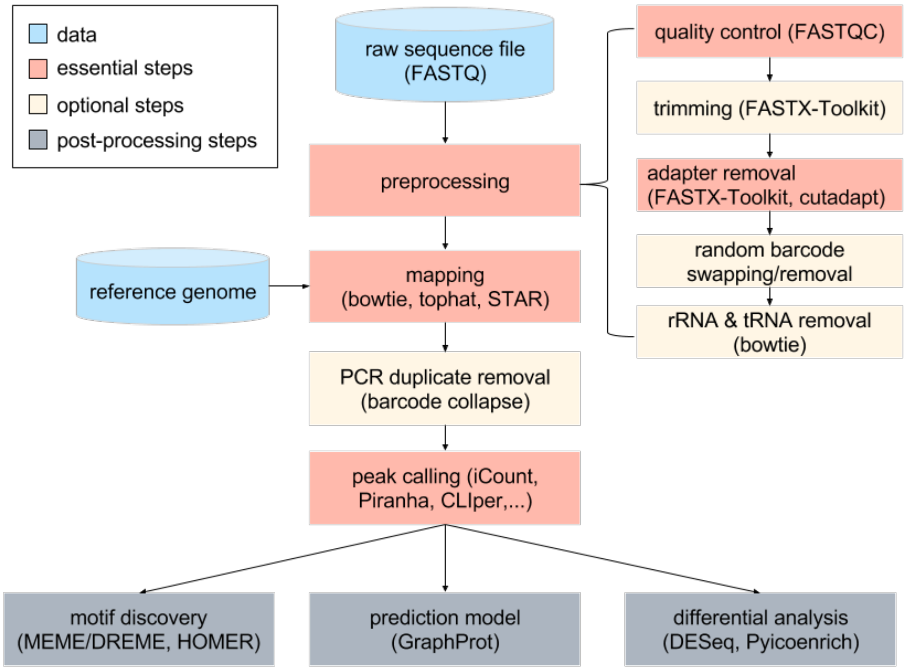

data
essential steps
optional steps
post-processing steps
raw sequence file (FASTQ)
preprocessing
quality control (FASTQC)
trimming (FASTX-Toolkit)
adapter removal (FASTX-Toolkit, cutadapt)
random barcode swapping/removal
rRNA & tRNA removal (bowtie)
reference genome
mapping (bowtie, tophat, STAR)
PCR duplicate removal (barcode collapse)
peak calling (iCount, Piranha, CLIper,...)
motif discovery (MEME/DREME, HOMER)
prediction model (GraphProt)
differential analysis (DESeq, Pyicoenrich)

Figure 1.4: Overview of the general analysis workflow for CLIP related methods.

The workflow starts with the unprocessed input data (coloured in blue) and is followed by the essential steps (red boxes) of analysis, together with optional processing steps (yellow boxes) that are not part of every CLIP related method. The final step of processed peak calling data can be followed by additional analysis (grey boxes) such as motif discovery, prediction models or differential analysis of comparing two datasets with different conditions.

1.6.4 Sequence quality control

In the first step of data analysis, we want to make sure that the quality of sequencing data is suitable for further analysis. One of the most popular tools that is used for all types of high-throughput sequencing data is the FASTQC tool [146]. The FASTQC tool can be used as a desktop application, or it can be integrated as a part of the bash script pipeline. It provides simple control steps to check whether the sequencing data has any errors before further analysis. For example, there can be some sequence quality drop at the end of sequencing cycles: this can be resolved with additional trimming steps in the pre-processing part of the analysis. Here are the main reports from the FASTQC quality check, which are the essential part of every high-throughput sequencing pipeline.

- **Per Base Sequence Quality** shows a quality overview across all cDNA sequences for each position that comes from a raw sequencing FASTQ file.

- **Per Sequence Quality Scores** checks if there is a subset of sequences that have lower quality scores compared to other subsets.

- **Per Base Sequence Content** shows the proportion of bases for each position across all sequences to see if there are any general sequence biases in the library.

- **Per Base GC Content** plots the GC content for each position compared to the overall GC content calculated from the observed data. An unusually shaped distribution of GC content at certain positions could indicate that there is a contamination of an over-represented sequence or a systematic problem during the sequencing of the library.

- **Per Sequence GC Content** measures the total GC content across each sequence and compares it with the observed GC distribution. A subset of sequences with higher or lower GC content could indicate that there is a contamination from another species.

- **Per Base N Content** represents the proportion of unidentified bases (N, any base) coming from the sequencer at each position across the library.

- **Sequence Length Distribution** can help to evaluate a length distribution after trimming or to determine a trimming cut-off if needed depending on the sequencer, as some generate sequences of the same length and others a variety of different lengths.

- **Duplicate Sequences** shows the level of duplicated sequences. If this is high, it could be a result of PCR over-amplification.

- **Overrepresented Sequences** allows us to detect if there is a contamination in our library. This function will also look for matches across a database for the most common contaminants.

- **Overrepresented Kmers** will report a positional enrichment of kmers across the library.

1.6.5 Pre-processing

- **Trimming**

  There are cases in which we have additional adapter sequences at the end or beginning of cDNAs or when the sequencing quality significantly drops towards the end of the cDNAs. With trimming, we can remove those parts if they are consistent across the library or read-wise trimming to increase the number of mapped cDNA reads. This can be done with (FASTX-Toolkit) or by any sequence trimming tool.

- **Adapter removal**

  In the third step of the iCLIP protocol there is a on-bead ligation of the 3' adapter. This sequence needs to be removed before mapping. The adapter sequence can be found at the end of the FASTA sequence and it can differ between CLIP-related methods. Two of the most popular tools for adapter removal are the FASTX-Toolkit adapter removal and Cutadapt [147]. FASTX-Toolkit comes with a lot of functions such as trimming, format converting, duplicate removal and even quality control, whereas Cutadapt is strictly for adapter removal and has many other features such as multi-adapter removal from 5' and 3' in a single run. Cutadapt also supports 454, Illumina and SOLiD (colour space) datasets and it is also a part of the ENCODE project pipelines for all high-throughput sequencing datasets [147].

- **Random and experimental barcode sequence swap/removal**

  The 3' adapter ligation in the standard iCLIP protocol also introduces an additional random barcode sequence and an experimental barcode. The experimental barcode is used to demultiplex samples from a single sequencing lane, and the random barcode is used to identify PCR amplification duplicates. In the iCLIP protocol, there is a random barcode positioned at the positions 1-4 nt and 9-12 nt and the experimental barcode is positioned at 5-8 nt. Processing of this barcode can be done by using a custom script (see Methods) that removes the experimental barcodes and adds the random barcodes into a header of the FASTQ file format.

- **Ribosomal RNA and transfer RNA removal**

  Ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs) are types of RNA molecules that are part of the translation processes from mRNA molecules to proteins. During the immunoprecipitation (IP) step there are abundant RNAs such as rRNAs and tRNAs that can bind non-specifically to the Dynabead-antibody-protein-RNA complexes [148]. In the CLIP and iCLIP protocol, only RNAs that are crosslinked to the RBP are selected, but there can still be some contamination of small RNAs such as rRNAs and tRNAs. These can be filtered out by mapping the cDNA reads directly to the rRNA and tRNA sequences that can be found in UCSC Table Browser, and removing hits without any mismatches. Preferably, this is done before the mapping to the remainder of the reference genome (Figure 1.4).

1.6.6 Mapping to reference genome

Most of the CLIP-related experiments are with well-annotated organisms, such as humans or mice, and the mapping process can sometimes be the most time-consuming step for the processing power. There are several mapping tools available for all kinds of high-throughput sequencing data. One of the most popular and fast tools is the Bowtie alignment software [149], which allows fast mapping to the genome or transcriptome. One disadvantage of Bowtie is that it cannot align cDNA sequences to splicing junctions. If we know that the targeted protein binds in the nucleus to pre-mRNA, Bowtie will probably be the tool of choice. Meanwhile, the TopHat alignment tool [150] overcomes this limitation by re-mapping all unmapped cDNAs from Bowtie to the spliced junctions, which is the most time-consuming step. Recently, the STAR alignment tool became a method of choice for RNA-seq data [151], and also across other high-throughput datasets that are also part of the ENCODE project. The STAR alignment tool guarantees the fastest and most accurate mapping across splice junctions [151, 152]. The only disadvantage of STAR is that the genome needs to be pre-built. This consumes a lot of time and memory and needs an enormous amount of pre-allocated memory for mapping compared to other alignment tools.

Besides using the most suitable tool, it is important to carefully adjust parameters to our needs. For example, in my pipeline (see Methods) and most of the publicly available pipelines, multiple hits are allowed but only one is selected by using score counts for the best fit. In cases of multiple hits with the same score, a random one is selected. Some protocols such as eCLIP are also designed with pair-end cDNA sequencing, so it is important that the alignment tool supports pair-end mapping.

1.6.7 Removal of PCR duplicates

After the mapping, it is important to remove PCR duplicates resulting from PCR amplification by collapsing cDNAs that have the same random barcode and map to the same genomic position. If we swapped the random barcode with the header of the FASTQ file, we can simply use a bash command from a mapped BED file format: `cat mapped.cDNAs.BED | sort -k1,1 -k2,2n -k5,5 -k6,6 | uniq`. This command will sort mapped cDNAs by genomic positions and random barcodes that are in the 5th column of the BED file, and every duplicate will be removed. There are other available tools with extra features, such as considering sequencing errors occurring in the region where the random barcode is positioned. But this is not the essential part of the analysis, since the random barcodes are positioned at the beginning of each sequence, whereas the sequencing errors are increased towards the end of the cDNA sequence [153]. For the ENCODE eCLIP pipeline there is a python script 'barcode_collapse_pe.py' available on GitHub. This script will remove PCR duplicates from the BAM file after mapping, but it is important that random barcodes are included at the end of the FASTQ header separated with the ':' character. Another popular tool is FastUniq which is a fast *de novo* tool for PCR duplicate removal from pair-end cDNA reads. It identifies duplicates by comparing sequences between read pairs before the data is mapped to the reference genome. However, it has been shown that the majority of duplicates (70-80%) are un-mappable or come from the same genomic positions [154].

1.6.8 Assignment of protein binding sites

How to correctly assign a protein-binding site is one of the most important and challenging parts of the CLIP data analysis. Firstly, there is no standardisation for a negative control for the CLIP-related methods. For comparison, some other traditional RNA immunoprecipitation and sequencing methods such as RIP-seq use a non-targeted approach to measure the background noise as control. A similar approach is taken for chromatin immunoprecipitation and DNA sequencing (ChIP-seq) methods where the affinity of the RBP is independent from the transcript abundance because they look at DNA binding of RBPs. However, normalisation is crucial to separate specific and unspecific bindings from the high signal to noise ratio data [155]. This can also be problematic in the CLIP methods since we do not know if a high enrichment comes from strong RBP-RNA interactions or from artefacts such as PCR over-amplification or from high gene expression which means that binding sites with low occupancy can outnumber highly occupied binding sites in terms of sequencing reads. One option to overcome this limitation of detecting binding sites with low occupancy on highly expressed transcripts, is by using RNA-seq data in parallel to correct for the transcript abundance or by performing an unspecific control experiment such as mock eCLIP in the eCLIP protocol [133].

Thus, a large proportion of CLIP cDNAs still represent protein-RNA interaction sites, and highly occupied binding sites appear as clusters of crosslink events if the CLIP library is of sufficient complexity, which makes it easier to identify true binding events [6]. Approaches to identify such clusters are based on peak calling algorithms to discriminate high-affinity binding sites from unspecific binding sites in a genome-wide manner. This is done after pre-processing the raw data, where we want to first discard unspecific peaks from the background noise and cluster the enriched ones into significant clustered regions over a certain threshold. This signal-to-noise ratio can be improved by avoiding PCR amplification artefacts [6], normalising to input RNA or RNA-seq [156, 6, 134], using different controls such as non-crosslinked control samples for background RNA [155], and by increasing the number of biological replicates.

There are different types of publicly available tools for CLIP/iCLIP/eCLIP/irCLIP or PAR-CLIP data analyses with different peak calling approaches (see Table 1.1):

- **Piranha**

  Piranha is a method for binding site identification that can be applied to all CLIP-related methods as well as RIP-seq data. It supports a nucleotide resolution by using cDNA-starts as crosslinking positions with external covariant data support of measured transcript abundances, which can improve the peak identification process [157]. Piranha assumes the majority of sites to be noise, so the sum of all sites can be used to fit a background model [155]. An appropriate binning window size needs to be defined by the user, into which Piranha adds cDNA counts. After the binning of cDNA counts, it models the cDNA count within bins by using a zero-truncated negative binomial distribution or a zero-truncated negative binomial regression model if external covariant data such as RNA-seq are included [157]. It can also perform a statistical analysis to measure p-values without any control data but it does not support any statistical measurements together with biological replicates.

- **CLIPper**
  CLIPper was first developed to identify clusters representing binding sites for Rbfox1 and Rbfox2 [158] and later became a part of the ENCODE standard pipeline. The program requires genome annotation to correctly separate thresholds on a gene level. The peak significance is defined by the number of cDNAs and gene length, relative to the number of other cDNAs from the same gene. There is one additional feature to improve significance: the user can pre-define whether the targeted RBP binds to mRNAs or pre-mRNAs. For this purpose, there are pre-compiled regions of mRNAs and pre-mRNAs from Ensembl annotation for mouse and human [134, 159]. Like Piranha, CLIPper supports a statistical approach on a single dataset but there is no option to incorporate biological replicates.

- **JAMM**
  JAMM (Joint Analysis of NGS replicates via Mixture Model clustering) is a universal peak calling bash script implemented by R and Perl. It can integrate information from multiple replicates in order to find consensus peaks, determine accurate peak widths and resolve neighbouring narrow peaks [160]. JAMM contains six peak finding steps: Extended cDNA reads count, Estimate Optimum Bin Size, Scan Chromosome in Non-overlapping Bins, Merge Enriched Bins into Enriched Windows, Determine Peak Width, Peak Scoring and Filtering. The main finding step searches for enriched windows compared to the background noise clusters. Clusters are normalised locally and can adapt peaks to different lengths by using the cost function that optimises the bin width [161]. For the clustering, it uses multivariate Gaussian models [162]. These models support peak calling across biological replicates [160] but do not support differential analysis across different conditions and do not take into account mappability-related features, such as GC content. JAMM was originally designed for the ChIP-Seq method and was tested on ENCODE

data among other available tools [160]. It can be applied to any other related datasets or protocols with the additional step of separating the data into two datasets by their strand orientation, and then running it separately for each strand.

- **Pyicoclip**
  Pyicoclip is another implementation of a peak calling tool that uses a False Discovery Rate algorithm (FDR) to discover significant clusters from mapped cDNA reads across the genome [163]. Originally the method was designed for CLIP data, but it can be applied to any CLIP-related method. One disadvantage is that it uses whole cDNA coverage for peak discovery, which lowers the resolution for methods in which RBP interactions are determined by truncated cDNAs such as iCLIP. The peak calling algorithm also supports custom genomic regions from the input file and comes with a pyicoregion function which generates exploratory regions from current annotation into intergenic, intragenic, exonic, intronic and TSS sides of any genome that is in GFF format by Sanger Institute standards [164].

- **ASPeak (an abundance sensitive peak detection algorithm)**
  ASPeak (Abundance Sensitive Peak Detection Algorithm) is a peak calling pipeline implemented in Perl to identify binding sites [165]. The algorithm is sensitive to differential expression datasets that uses RNA-seq data for the expression measure as an additional input which increases the sensitivity of peak detection from low-abundance transcripts. Originally it was tested to successfully detect binding sites for the exon junction complexes in the predicted -24 nt position upstream from exon-exon junctions [166]. It uses genomic intervals that can come from a custom BED format annotation or additional source such as RefSeq to regions separated into coding exons, 5' untranslated region (5' UTR), 3' untranslated region (3' UTR) and introns. For each genomic interval a negative binomial distribution is used to detect significant binding sites with dynamic window sizes. For peak calling without RNA-seq input the algorithm uses a local window approach, the size of which can be defined

by users. All the input data needs to be in BED or BAM format from any variant of CLIP, RIP-seeq or RNA-seq datasets. ASPeak can run locally or it can be parallelized to multiple cores or cluster computers for advanced users [165].

- **PIPE-CLIP**

  PIPE-CLIP was built on a Galaxy online framework as part of the complete pipeline to reliably identify binding sites and process the raw CLIP, PAR-CLIP and iCLIP data. It can be used just as a peak finding tool that only accepts the BAM file format. PCR duplicate removal can also be applied in the peak finding step that uses a zero-truncated negative binomial model for identifying the significantly enriched peaks [167].

- **CLIP Tool Kit (CTK)**

  CTK is a software package of tools for CLIP data analysis from pre-processing raw reads including PCR duplicate collapse. However, its main function is to define clusters and peaks from all variants of the CLIP method and it works on single-nucleotide resolution. It also supports the PAR-CLIP method that detects the crosslinking position by T-to-C transitions across cDNAs. For the peak calling step, it uses a 'valley seeking' algorithm. The first stage of the algorithm looks for the local maximum peaks of overlapping cDNAs. Two local peaks are considered to be significant only when they are separated by a valley of depth `d=h-v`, where `h` is the smallest peak and `v` is the cDNA coverage at the valley position [168]. The threshold for the valley depth can be set by the user to adjust the stringency of peak discovery.

- **iCount**

  iCount is a Python module and works as command-line interface and as a web-based platform. It provides a large number of functions to process the iCLIP data: demultiplexing and adapter removal, mapping to a reference genome, identifying protein-RNA crosslink sites by using a False Discovery Rate (FDR) algorithm to discover significant peaks and merging them into crosslink clusters (Figure 1.5), and grouping of individual experiments into large datasets. As downstream analyses, it offers RNA-map visualisation showing the positional distribution of crosslink sites relative to genomic landmarks and kmer enrichment analysis. It supports two types of normalisation across the genome, one at the transcript level and the other by genomic regions. The window size for peak calling and clustering needs to be set in advance, and iCount does not support differential analysis between conditions [169].

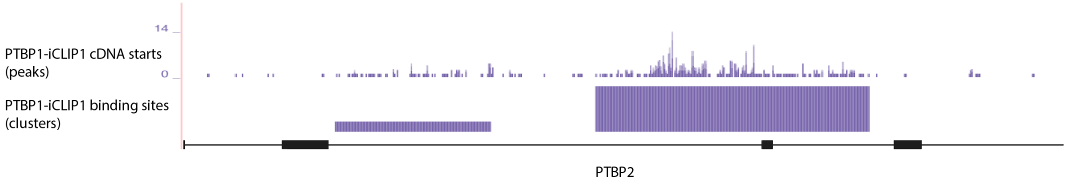


Figure 1.5: Example of PTBP1 iCLIP cDNA-starts (peaks) and binding sites (clusters) of crosslink positions on PTBP2 transcript defined by iCount in UCSC Genome Browser.

The crosslinking positions were identified by iCount with peak calling tool using 0.05 FDR threshold within 15 nt window size and were merged into crosslink clusters within 15 nt window size. The number on y-axis represents a maximum number of raw cDNA-starts.

Peak finding tool	Method	Replicate support	Normalisation by custom annotation (transcript, UTR, intron, exon)	Supported data format	Supported Methods	Normalisation on transcript abundance	Pros	Cons
Piranha	Zero-truncated negative binomial distribution, a zero-truncated negative binomial regression model.	no	no	BAM, BED	CLIP, RIP-CLIP, iCLIP variations	yes	Fast and easy to use. It corrects the cDNA reads dependence on transcript abundance with additional control data.	Does not support normalisation based on genomic region.
PIPE-CLIP	False discovery rate with local randomisation with Fisher's method.	no	yes	BAM	CLIP, PAR-CLIP, iCLIP variations	no	Detects differential binding regions by comparing two CLIP experiments.	Difficult to install and just BAM file support that needs to be in a strict format (BAM format from STAR alignment is not supported).
CLIPper	False discovery rate with local randomisation.	no	yes	BAM	CLIP, iCLIP variation	yes	Supports normalisation by genomic region . and additional input dataset.	Low resolution, more suitable for CLIP data.
JAMM	Single or multivariate Gaussian mixture model with MannWhitney U non-parametric test to compare it with background noise.	yes	no	BED	ChIP-seq, (flexible for others)	no	Fast and easy to use.	More suitable for ChIP-seq (no strand support).
Pyioclip	False discovery rate with local randomisation.	no	yes	BAM, BED	CLIP	no	It supports different types of reginal normalisations.	No nucleotide resolution support.
ASPeak	Negative binomial distribution.	no	yes	BAM, BED	RIP-seq, CLIP	yes	Fast and it can be parallelised. Local normalisation by genomic regions together with RNA-seq.	For the single nucleotide resolution CLIP variants the data needs to be pre-processed for the correct crosslinking positions.
CLIP Tool Kit	'Valley seeking' algorithm using statistics if replicates and genomic regions are included.	yes	yes	BED	CLIP, PAR-CLIP, iCLIP variations	no	It also supports crosslink identifications by T-to-C transitions.	It supports on local normalisation.
iCount	False discovery rate with local randomisation.	no	yes	BED	CLIP, iCLIP variation	no	It comes with a number of useful features such as RNA-maps, kmers analysis, dataset comparisons, grouping, etc.	Slow because of the randomisation step.

Table 1.1: Overview of available peak calling methods.

1.6.9 Validation of identified clusters

It is very difficult to say which of the peak calling approaches is the most suitable since there are no standard ways to validate the results. Mostly they are validated by the number of significantly detected clusters or by reproducibility between replicates. But these approaches lack biological validation, instead they are telling us which statistical model better fits the data. Another way to validate the data is to include a motif enrichment across identified clusters [167, 134, 168] but first we need to know what the binding motifs of our target RBPs are.

1.6.10 Motif discovery

After the binding site assignment, it is important to understand the specificity of the RBP. One way is by first discovering motifs that are specifically recognised by the protein of interest. This step is very challenging since it is highly dependent on the previous peak calling step, and so far only about 15% of motifs from known RBPs have been discovered [170]. The number of discovered motifs is even lower in other studied organisms [156].

- **DREME**
  DREME (Discriminative Regular Expression Motif Elicitation) is a software package and also a web-based platform to discover motifs from FASTA sequences. It is part of the MEME Suite, which collects software packages for all kind of motif-based sequence analysis. The basic input is a positive (target) and negative (control) set of sequences, but it can also work without controls by shuffling the target set to provide a control set. For the statistics, it uses Fisher's Exact Test, which determines significance of each discovered motif from the positive set compared to the control set with the significance threshold that can be set by user. It was designed to discover short motifs (up to 8 nt) in a very short time. When we analyse the RNA sequence data, it is important that we use the -norc option, which searches the given primary sequences in a single direction and also to set the minimum and maximum lengths of motifs to optimise the processing time [171].

- **Homer**

  HOMER (Hypergeometric Optimization of Motif Enrichment) is a software package for next-generation sequencing or microarray data analysis from genome-wide experiments, such as ChIP-Seq, GRO-Seq, RNA-Seq, DNase-Seq, Hi-C, CLIP, iCLIP and more. It is based on a Unix platform written in Perl and C++, and its primary function is to discover *de novo* motifs from large datasets. The motif discovery algorithm uses two types of inputs. The first is a target sequence of interest, and the second is a background sequence. The background sequence can also be produced by HOMER from a random collection of nucleotides, but this can be very biased since the real distribution in the genome/transcriptome is not random and will find motifs that may not be significant. Therefore, the compilation of an appropriate set of background sequences is very important and can be either done with other tools or using another dataset as control data [172]. HOMER supports a discovery of extremely long motifs (up to 20 nt) but to optimise its processing power for our needs, it is better to specify the maximum length in advance. The motif enrichment is measured by the cumulative hypergeometric distribution or the cumulative binomial distribution when using a large dataset. Originally it was designed for DNA analysis but it can also be applied for RNA motifs. It has been used to determine miRNA seeds from mRNAs and RNA binding motifs from CLIP data. For the CLIP data, it can search for motifs directly from the RNA sequences by using '-rna' option which results in strand-specific analysis and replaces 'T' with 'U' in the motif results [173].

- **Zagros**

  Zagros is a motif discovery software that was designed to characterize the RBP binding sites from CLIP-related methods. Previous studies showed that RNA secondary structure plays an important role in RBP binding site selection [174, 175]. In comparison to other motif discovery methods, Zagros uses secondary structure from input sequences to improve the accuracy of motif discovery. It has been shown that RBP binding sites show less structural constraints for RBPs with highly specific sequence motifs compared to RBPs with less specific motifs [176]. Zagros is using McCaskill's algorithm [177] to measure pairing probability around crosslink sites. These predictions of pairing probabilities are then included into a final model together with crosslink events to perform the expectation maximization algorithm for motif discovery. So far it has been tested on CLIP, iCLIP and PAR-CLIP methods for 40 RBPs from human and mouse [176].

1.6.11 Prediction of protein binding sites

- **GraphProt**

  GraphProt is a computational framework that uses machine learning models to identify RBP binding sites from experimental data. Like any other machine learning tool, it first needs to build a training set from the input data. But it also includes the RNA sequence and structure characteristics into a graph-kernel strategy to obtain a large set of features from the dataset [156]. The core of the model is to extend a similarity of kmer motifs and structures into graphs that can be applied to Support Vector Machine (SVM) [178] and Support Vector Regression (SVR) [179] for classification and regression analysis [34]. For the correct classification of binding site predictions, the tool requires two sets of training data: one needs to be the positive dataset of bound sites and the negative of truly unbound sites [156]. This method has been tested on CLIP, PAR-CLIP and iCLIP for several RBPs including PTBP1.

1.6.12 Differential analysis

- **DESeq/DESeq2**

  DESeq/DESeq2 is an R package from Bioconductor (open source software for bioinformatics) for the analysis and comprehension of high-throughput genomics data (bioconductor.org) [180]. DESeq estimates the variance mean dependence that needs to be provided by the count dataset from high-throughput sequencing assays [155]. It was originally designed to test for differential expression based on a model using the negative binomial distribution from RNA-seq data, but the same model can also be applied to any type of CLIP data. This can be done with a few additional steps to prepare the CLIP data in the right format. The input count table can be generated with the htseq-count software that uses crosslink positions as the input and clusters of binding sites as genomic regions. The essential part of differential analysis is also to have multiple biological replicates in order to measure significant changes [180].

- **Pyicoenrich**

  Pyicoenrich is a part of Pyicoteo software package and supports enrichment analysis on any type of sequencing data from two conditions. Pyicoenrich will report basic scores between conditions and the significance of the difference by comparing the overlap of sequences between enriched regions. It also supports multiple replicates and presents the data conditions with MA-plots with a log2 ratio of normalised cDNAs in a similar way as DESeq (Pyicoenrich online reference).

1.7 Aims of the Thesis

In recent years CLIP methods have become a state-of-the-art technique for transcriptome-wide studies of protein-RNA interactions. It is thus crucial to understand the technical aspects of the method that need to be taken into account when interpreting the data. The main focus of my thesis is to first explore variants between different CLIP and iCLIP techniques to better understand technical biases behind the methods. My aim is to provide a comprehensive assessment of the intricacies of a 'good' CLIP dataset, and stream lined computational analysis, which are required to correctly assign protein-RNA binding sites.

In iCLIP, the start positions of cDNAs identify the protein-RNA crosslink sites, but the correct assignment of binding sites through iCLIP remains challenging [29, 181, 182]. A recent study found that the positions of cDNA-starts depend on cDNA length in several iCLIP datasets and proposed two alternative interpretations: the first recommended use of cDNA-centres due to a hypothetical prevalence of readthrough cDNAs, while the second proposed that non-coinciding cDNA-starts still correctly assign the crosslink sites [181]. The second hypothesis was mentioned in the discussion but was not examined, and the study focused on the first interpretation. Since the previous analyses show that non-coinciding starts are present in most, if not all of datasets produced by iCLIP and its variant methods (CITS-CLIP, eCLIP, FAST-iCLIP), and thus it was unclear if cDNA-starts or cDNA-centres should be used for analysing resulting data, I decided to investigate the second hypothesis in depth. Previous studies used a limited analysis of cDNA deletions to support the use of cDNA-starts [29]. In this thesis, I undertook a more systematic comparative analysis of many different features in cDNA libraries, which provided more thorough evidence that non-coinciding sites are valid crosslinking positions. Through this detailed analysis, I developed experimental and computational solutions to improve binding site assignment from iCLIP data. I focus on polypyrimidine tract-binding protein 1 (PTBP1), and the exon junction complex protein eIF4A3, but the findings are relevant to iCLIP studies of all RNA-binding proteins. The two studied proteins are particularly challenging, as pointed out by a recent study showing that the start positions of long and short iCLIP cDNAs do not always coincide in their iCLIP data [181].

1. I assessed the mechanisms for this alignment discrepancy, and presented the following findings:

1.1 I used a new a modified iCLIP protocol developed in our lab to experimentally identify the readthrough cDNAs, which shows that only a minor portion of our PTBP1 iCLIP cDNAs read through the crosslink site, and therefore they do not explain the cause of the non-coinciding cDNA-start sites. I further used the insights gained from analysis of PTBP1 data to examine the binding regions of the exon-junction complex (EJC) by using four different publicly available and newly generated iCLIP or CLIP datasets.

1.2 I examined the ends of cDNAs to discover that the non-coinciding cDNA-starts are caused by constrained cDNA-ends, which result from the RNA sequence and structure constraints of RNase cleavage. Therefore two experimental aspects of iCLIP are crucial for correct assignment of binding sites: broad distribution of RNase cleavage sites, and a cDNA library containing a broad range of cDNA sizes. I discussed how this is achieved in iCLIP, and showed that by following these iCLIP guidelines, a more comprehensive set of crosslink sites of eIF4A3 and PTBP1 is identified.

1.3 I analysed exon-junction complex (EJC) that binds upstream of exon-exon junctions with nucleotide precision, by using two independent methods (CLIP and iCLIP) to find precise crosslinking peaks that align to exon-exon junctions across all exons. I then demonstrated that EJC also forms additional crosslinks over a broader surrounding region.

2. I developed CLIPo, a computational tool for quality control of iCLIP, which reveals how technical variations between experiments affect the specificity of assigned binding sites. I examined multiple datasets for proteins produced by different variants of CLIP or iCLIP to reveal major effects of sequence constraints at cDNA ends or starts, cDNA length distribution and non-specific contaminants.

2.1 I tested whether CLIPo can detect constraints across different experiments that were already identified through previous analysis. For this purpose, I developed new computational pipelines to visualise the impact of these features on the sequenced cDNA libraries, which helps to correctly interpret the assigned binding sites.

2.2 The assignment of the RBP binding site is the crucial step to characterize RBPs and their functions. Here, I focused on optimal peak calling algorithm window sizes by using motif enrichment and RNA-map approach to measure specificity for PTBP1.

2.3 In addition to the technical insights, I provided mechanistic insights into PTBP1-dependent splicing regulation. Past studies examined PTBP1 binding at narrowly defined positions when attempting to explain its position-dependent mechanisms of splicing regulation [53, 61]. I now find that PTBP1 most often regulates splicing in three distinct modes of its position-dependent activity.

3. To investigate the spliceosomal interactions with other RBPs on pre-mRNA, I examined a new dataset from spliceosome-iCLIP method.

3.1 I hypothesise that cDNAs that end at the end of introns are truncating at the branch point (BP) position during splicing.

3.2 I developed a new pipeline to detect BPs genome wide and validate these results with comparison of predicted BP, sequence motifs, RNA-maps with other RBPs.

3.3 I applied a large dataset of eCLIP data to identify known and novel RBP targets that are interacting with BPs.

To summarise, I present several technical advances that aid the assignment of RNA binding regions of RBPs from CLIP and iCLIP-related data, and described how such binding regions can provide insights into the function of PTBP1. Application of these approaches will be particularly useful for studies of RBPs that regulate splicing. I also demonstrate the importance of conditions of library preparation that can lead to constrained cDNA-ends, which result in non-coinciding cDNA-starts. More importantly, I demonstrated how these effects can be minimised by optimising iCLIP conditions, and how they should be taken into account during computational analysis to ensure correct assignment of binding sites. Finally, I examined a spliceosome-iCLIP method to gain new insight of splicing machinery.

Chapter 2

Methods

All the source codes used for the analyses in this thesis are available at the following GitHub repository: *https://github.com/nebo56/PhD-Thesis*.

2.1 Computational tools and working environment for data analyses

All the data analyses were mainly performed by the usage of following software and packages.

- Programming languages: R (3.1.0), Python (2.7), Unix based bash scripting
- Software: Samtools, Bedtools, Bowtie, Tophat, Cufflinks, STAR, Cutadapt, FASTXtools, FASTQC
- Bioconductor packages: DESeq, DEXseq, JunctionSeq
- Other R packages: ggplot2, smoother, heatmaps2
- Integrated development environments: Aptana, R-studio
- Operating System: Ubuntu Linux (16.04 LTS), CentOS Linux (release 6.5)

2.2 Mapping and pre-processing of high-throughput sequencing data

2.2.1 CLIP, iCLIP, irCLIP, eCLIP and spliceosome-iCLIP

The data produced by modified iCLIP protocol with additional 5' marker for PTBP1 and eIF4A3 experiment in chapter 3 was analysed separately. I separated cDNA reads containing 'CAGTCCGACGATC' sequence from Illumina 5' adapter at the beginning of each read. These sequences were marked as 'readthrough cDNAs' and were analysed separately.

1. **Trimming of the adapter sequences**

   Before mapping the cDNAs, I removed unique molecular identifiers (UMIs) and trimmed the 3' Solexa adapter sequence. Adapter sequences were trimmed with the FASTX-Toolkit (version 0.0.13) adapter removal software, using the following parameters: `fastx_clipper -Q 33 -a AGATCGGAAG -c -n -l 26 -i INFILE -o OUTFILE`. For reads that did not contain parts of the adapter sequence (incomplete cDNAs), the '-C' parameter was used, and they were analysed separately. All sequences shorter than 26 nt (17nt sequence + 4 nt experimental barcode + 5 nt random barcode) were isolated from the analysis to avoid multiple mapping bias.

2. **Random barcode removal**

   I used a custom python script that removes experimental barcodes and includes the random barcodes into the read names within the FASTQ file (see GitHub).

3. **Mapping with Bowtie alignment software**

   To map CLIP, iCLIP, irCLIP and eCLIP sequence data for PTBP1 and U2AF65 (in all chapters), I used the UCSC hg19/GRCh37 genome assembly and the Bowtie2 (version 2.1) alignment software with default settings. More than 80% of all cDNAs from the published and newly generated iCLIP data mapped uniquely to a single genomic position. The first 9 nt of the sequenced iCLIP read correspond to the barcode. This is composed of the experimental identifier, which allows to separate experimental replicates, and the unique molecular identifiers (UMIs), which allow the avoidance of artefacts caused by PCR over-amplification of different cDNAs [124]. I used these UMIs to quantify the number of unique cDNAs that mapped to each position in the genome or transcriptome (for eIF4A3 iCLIP dataset) by collapsing cDNAs with the same UMI that mapped to the same starting position to a single cDNA. For the analysis described in chapter 3, I have separated cDNAs into four different length classes to retrieve a similar number of cDNAs per each class: <30 nt, 30-34 nt, 35-39 nt or >40 long after adapter trimming.

4. **Mapping with STAR alignment software**

   In chapter 5, to map eCLIP sequencing data for all RBPs, I used GENCODE (GRCh38.p7) genome assembly and the STAR alignment (version 2.4.2a) with the following parameters taken from the ENCODE pipelines section:

   ```
   STAR --runThreadN 8 --runMode alignReads --genomeDir GRCh38_Gencode_v25 --genomeLoad LoadAndKeep --readFilesIn read1, read2, --readFilesCommand zcat --outSAMunmapped Within --outFilterMultimapNmax 1 --outFilterMultimapScoreRange 1 --outSAMattributes All --outSAMtype BAM Unsorted --outFilterType BySJout --outFilterScoreMin 10 --alignEndsType EndToEnd --outFileNamePrefix outfile
   ```
   .

5. **Custom mapping to CDS transcripts**

   To improve mapping of iCLIP sequence data for eIF4A3 that binds on mRNA exon-exon junctions, I extracted all the transcripts CDS sequences from Ensembl Genes 79 BioMart. Then, I compiled a set of the longest mRNA sequence available for each multi-exon gene. I mapped the eIF4A3 pre-processed iCLIP sequences with the Bowtie2 (version 2.1) alignment software directly to the longest mRNA sequence, allowing a maximum of two mismatches. To keep the gene information and position of each exon-exon junction I added information to the FASTA header for each transcript sequence separated with `'|'` symbol.

2.2.2 ENCODE and customised pipeline for eCLIP data analysis

In chapter 4, I used 'narrow peaks' for PTBP1 eCLIP data that is available online at ENCODE eCLIP data section. The pipeline uses Cutadapt tool for adapter removal, RepBase database to remove consensus sequences of repetitive elements and peak calling CLIPper tool to define final peaks, together with control mock-eCLIP data. In the second part of analysis, I used a modified custom pipeline (Figure 2.1), where -1 nt upstream from the cDNA-start position were considered as crosslinking position and used as input for the iCount clustering algorithm.

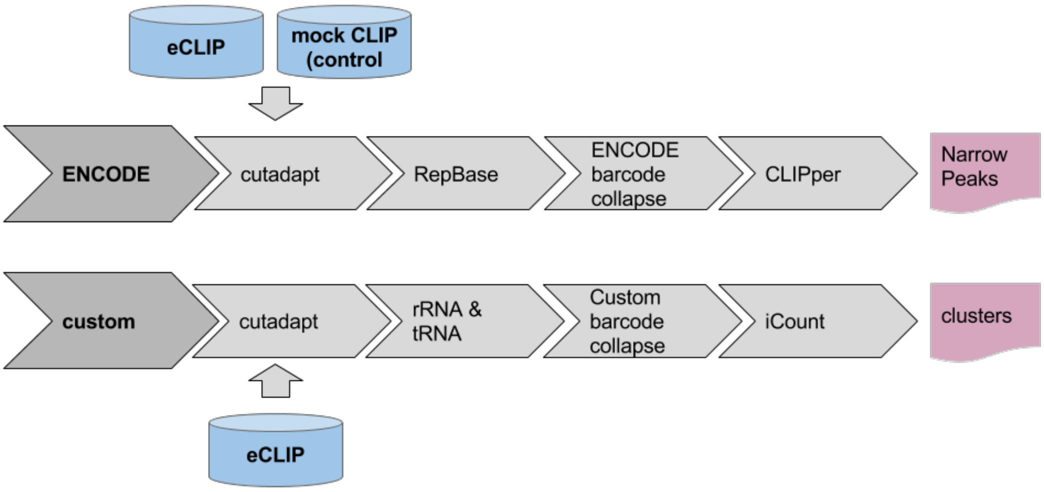


Figure 2.1: Overview of the workflow for eCLIP data analysis workflow.

ENCODE pipeline uses control mock-eCLIP data for normalisation of final narrow peaks and RepBase database to remove consensus sequences of repetitive elements. The custom pipeline does not use any additional data for normalisation and keeps all repetitive elements. It uses a bash command for PCR duplicates removal and iCount peak calling algorithm to define final clusters.

2.2.3 RNA-seq

The following pipeline was used for RNA-seq data of hnRNPC knock down (KD) in chapter 4.

1. **Trimming of the adapter sequences**

   Before mapping the cDNAs, I removed the adapter sequence by using Cutadapt tool (version 2.7), with following command: `-cutadapt -a GATCGGAAGAGCACACGTCTGAACTCC -m 17`. Sequences shorter than 17 nucleotides were removed from the downstream analysis to increase the number of uniquely mapped cDNAs.

2. **Mapping with STAR**

   RNA-seq from ENCODE presented in chapter 4 for hnRNPC KD and control samples were mapped to GENCODE (GRCh38.p7) genome assembly, using the STAR alignment (version 2.4.2a), with the following command: `STAR --runThreadN 8 --genomeDir genomeDir --readFilesIn RNA-seqFile.fq --winAnchorMultimapNmax 101 --outFilterMultimapNmax 100 --outFileNamePrefix path --outSAMtype BAM SortedByCoordinate --outWigType wiggle --quantMode GeneCounts`.

3. **Count tables for differential analysis**

   Count tables for differential analysis of mapped RNA-seq data were generated by using QoRTs (Quality of RNA-seq Tool-Set), with the following command: `java -Xmx8G -jar QoRTs.jar QC --minMAPQ 255 --stranded --singleEnded --runFunctions writeKnownSplices, writeNovelSplices, writeSpliceExon Aligned.sortedByCoord.out.bam GRCh38.gtf QoRTs-results`.

4. **Differential analysis of regulated exons**

   Exons that are regualted by hnRNPC were identified by JunctionSeq R package by using two biological replicates of hnRNPC KD and control data. Regulated exons were selected with a log2 fold change +1.0 for down and -1.0 for up regulated exons, and adjusted p-value lower than 0.01. The control exons were selected with p-adjusted value higher than 1.0.

The following pipeline for RNA-seq data was used in chapter 5 for filtering genes with low expression. This data was performed in parallel with spliceosome-iCLIP experiment.

1. **Trimming of the adapter sequences**

   Adapter sequences were trimmed with the FASTX-Toolkit (version 0.0.13) adapter removal software, using the following command: `fastx_clipper -Q 33 -a ATCTCGTATGCCGTCTTCTGCTTG -n 17 -i INFILE -o OUTFILE`.

2. **Mapping with Tophat alignment software**

   To map RNA-seq from the spliceosomal experiment in chapter 5, I used the UCSC hg19/GRCh37 genome assembly and the Tophat2 (version 2.0.9) alignment software allowing a maximum of 2 mismatches and uniquely mapped reads.

3. **Measuring gene expression**

   For the introns selection of branch point (BP) analysis from chapter 5, I used only introns coming from expressed genes with a median Fragments Per Kilobase of transcript per Million mapped reads (FPKM) higher than 10 across all 4 replicates. The threshold of 10 FPKM was set by the visually examining the density distribution (data not shown) and selecting a reasonable number of expressed introns (35,056 introns). FPKM values were generated with cufflinks version 2.1.1 with default settings and the same annotation that was used for mapping.

2.3 Genomic annotations

2.3.1 Definition of Y-tracts

I obtained genomic positions of all TC-rich and T-rich low complexity sequences in the human genome using the UCSC table browser.

2.3.2 Genomic lift over

For characterisation of spliceosomal interactions from upstream region (-50 to -10 nt) of identified branch points in chapter 4, I used UCSC lift over tool to convert genomic positions from hg19 to hg38 genome builds.

2.4 Post-mapping analysis

2.4.1 Identification of crosslink clusters from CLIP, eCLIP and iCLIP datasets

The crosslink clusters were identified by a False Discovery Rate (FDR) peak finding algorithm implemented in iCount, which considers the crosslink sites as significant, with minimum half-window spacing of 3 nt (iCount2 default settings), by assessing the significance of cDNA enrichment under the FDR <0.05 threshold compared to shuffled data [183]. Then, all the significant crosslink sites also known as peaks, are merged into final clusters separated by 3 nt, 15 nt, 25 nt, 50 nt and maximum 100 nt (see Chapter 6). In iCLIP, eCLIP and irCLIP data I used the -1 position from cDNA-start as a crosslinking position and the middle position for the CLIP data.

2.4.2 Classification of cDNA length

Only cDNAs that mapped to a unique genomic position were considered. These were separated into cDNAs that either did or did not contain parts of the 3' Solexa primer adapter. For libraries sequenced with the 50 cycle Illumina kit, the cDNAs with the adapter sequence were further separated into length groups of <30 nt, 30-34 nt or 35-39 nt after trimming. The length groups were defined by the cDNA length distribution of all cDNAs to keep the similar number of cDNAs per group. For other libraries with more than 50 cycles and untrimmed cDNAs were considered as a group of >40 nt.

2.4.3 Definition of crosslink-associated motifs

I reasoned that sequence motifs enriched directly at the cDNA-starts of the mock-eCLIP cDNAs could uncover preferences of UV crosslinking, since they are thought to represent a mixture of crosslink sites for many different RBPs and they should not reflect sequence specificity of any specific RBP [134]. I therefore examined occurrence of tetramers that overlapped with the nucleotide preceding the cDNA-starts (position 1 nt) in comparison with the ones overlapping with the 10th nucleotide preceding the cDNA-starts (position 10 nt) in PTBP1 mock input eCLIP. I excluded the TTTT tetramer from further analyses, since it is often part of longer tracts of uridines (Ts), and its inclusion results in a decrease of the resolution of analysis. Thus, the tetramers that are enriched over 1.5-fold at position -1 compared to -10 to select the top ten tetramers including TTTG, TTTC, TTGG, TTTA, ATTG, ATTT, TCGT, TTGA, TTCT and CTTT, and were considered for all analyses of 'CL-motifs'.

2.4.4 Definition of Y-rich motifs

To identify the Y-rich motifs bound by PTBP1, I searched for pentamers enriched in the 10 nt region around the cDNA-start peaks identified in each crosslink cluster defined by PTBP1-iCLIP2. 69 pentamers showed enrichment z-score >299 and were used as PTBP1-target pentamers for the analyses. Their sequences are: TCTTT, CTTTC, TCTTC, CTTCT, TCTCT, CTCTC, TTTCT, TTCTC, TTCTT, TTTTC, TCCTT, CTCTT, ATTTC, TTCCT, CTTCC, TTTCC, CCTTT, CTTTT, CCTTC, TCTGT, TTCTG, TCCTC, CTTCA, ATCTT, TGTCT, TCTGC, CTCCT, CCTCT, GTCTT, TCTAT, TCTCC, ATTCC, TTCTA, CTTTG, TATCT, ACTTC, TTATC, CTTAT, CTATT, TTCAT, TTCCA, TCTTG, TTGTC, TTGCT, CTCTA, CTCTG, TATTT, TCCCT, TCATT, TTCCC, CATTT, ATTCT, TTTAC, GTTCT, CTATC, TCATC, CTTTA, TGTTC, TATTC, CATCT, TACTT, CTGTT, CTTGC, ACCTT, TTTCA, TTTGT, TGTTT, CTTGT, ACTTT. All of these pentamers are enriched in pyrimidines, which is in agreement with the known preference of PTBP1 for UC-rich binding motifs [66]. Therefore, these pentamers were also referred to as Y-rich motifs.

2.4.5 Assignment of the cDNA-end peak in eIF4A3 iCLIP

For cDNA-end peak assignment in eIFA3 iCLIP data, I used exons longer than 100 nt to avoid the enrichment of neighbouring exons. I used the top 50% of the distribution of exons based on cDNA coverage to avoid transcripts with low expression. This ensured that sufficient cDNAs were available for assignment of the putative binding sites. I then summarised all cDNA-end positions in the 20 nt upstream and 25 nt downstream region around exon-exon junctions, where cDNA-ends are highly enriched and selected the position with the maximum cDNA count as the 'cDNA-end peak'. The putative eIF4A3 RNA binding regions were then defined by using the region between the cDNA-end peak and the start of the longest cDNA that ends precisely at the cDNA-end peak.

2.4.6 Analysis of pairing probability

All computational predictions of the secondary structure were performed by RNAfold (Vienna Package) software with default parameters and no post scripting ('–noPS' option) to shorten computational power and processing time [184]. I used the following command:

```
RNAfold -noPS < input.fasta > output.RNAfold.fasta.
```

The RNAfold results are provided in a customised format, where brackets are representing the double stranded region on the RNA and dots are used for unpaired nucleotides. I measured the density of pairing probability by implementing a custom python script 'RNAfold-sum.py' (see GitHub). This script creates an array with the size of the FASTA sequence, where brackets are counted as one and dots as zeroes representing a density of double stranded nucleotides from the fasta sequence file. Final density plots were generated in R, using ggplot2 package.

2.4.7 Normalisation of cDNA-starts/ends for the density graphs

All normalisations were performed in R (version 3.1.0) by using the 'ggplot2' and the 'smoother' package for the final graphical output. For the analysis of eIFA3 iCLIP, each density graph shows a distribution of cDNA-starts and cDNA-ends relative to positions of exon-exon junctions or end peaks in mRNAs. To avoid any border effects, I examined only exon-exon junctions within coding regions, excluding the first or the last junction. The number of cDNAs starting or ending at each position on the graph was normalised by the number of all cDNAs mapped to representative mRNAs, the mRNA length, and the number of examined exon-exon junction positions, as described below:

$$RNAmap[n] = \frac{(cDNAs[n]/sum(cDNAs)) * length(mRNAs)}{count(\text{exon junctions})}, \tag{2.1}$$

where [n] stands for a specific position on the density graph.

To draw the graph, I then used the Gaussian method for smoothing with a 5 nt sliding window. For the analysis of PTBP1 iCLIP and CLIP, each density graph shows a distribution of cDNA-starts and cDNA-ends relative to positions of its binding sites, which were defined using the position of Y-tracts. I obtained genomic positions of all TC-rich and T-rich low complexity sequences that are present in introns inside protein-coding genes on the human genome using the UCSC table browser. To avoid the effects of variable abundance of intronic RNAs (and occasional presence of highly abundant non-coding transcripts, such as snoRNAs), I normalised cDNA-starts at each position by the density of cDNAs in the same region. For this purpose, I examined the binding region, as well as 120 nt flanking regions, to find the nucleotide with the largest count of cDNA-starts or ends (according to whether starts or ends were plotted on the graph), which is referred to as 'MaxCount'. I thus obtained 'MaxCount-normalised cDNA counts' at each position (which were between 0 and 1). For drawing RNA density maps, I examined the enrichment of cDNA counts within binding sites compared to nearby regions outside of binding sites. I therefore calculated the average 'MaxCount-normalised cDNA counts' at each position across the evaluated binding sites, and divided each position by the average 'MaxCount-normalised cDNA counts' in the region between 50 and 100 nt downstream of the binding site, as described in the formula below:

$$RNAmap[n] = \frac{\text{average normalised cDNAs}[n]}{\text{average normalised cDNAs}[\text{50 to 100 nt downstream of the binding site}]}, \tag{2.2}$$

where [n] stands for a specific position on the density graph.

2.4.8 Visualisation of RNA-maps

This method takes into account predefined crosslink cluster positions, exonic positions that are regulated by the RBP of our interest, together with unregulated control exons and motif sequences. The pipeline can also be used with non-regulated control exons or with out motifs (RBP kmers), which are used for visualisation purposes. In the first step it employs bedtools intersect function to select all the neighbouring clusters in 300 nt flanking region around regulated exons, and then sort them by the distance from exon starts. In the second step, it extracts the genomic sequences around the splice site regions and creates a matrix based on cluster positions and motif enrichment (Y-rich motifs were used for PTBP1 and hnRNPC). This part of the process is written in Python and it takes each nucleotide position around regulated exons to set a value based on their cluster position and motif enrichment: -1 value is for exon start or end position, value 1 is each position that overlaps with any of the motifs, 2 is for cluster positions, 3 is for the motif coverage that is inside of a cluster region and every other position is set to 0 value. These values are stored as a matrix in a comma-separated values (CSV) format and are then visualised with R script, where the matrix is plotted as a heatmap, a density plot of cluster enrichment and a table (Figure 2.2). In the heatmap, every row represents a regulated exon with the cluster position and the motif coverage in surrounding region 300 nt upstream and downstream from the regulated exons and 50 nt inside of an exon. The second plot is a density plot of clusters enrichment compared to the enrichment in control exons. The last result of this pipeline is a table with crosslink cluster enrichments with distances and ratios between control exons and regulated exons in 3' splice site region, inside of exons and 5' splice site region (Figure 2.2).

Analysed RBPs:

- **PTBP1**

  Dataset of exons regulated by PTBP1 was identified by the previous micro array study [185]. Three subsets of exons were used: 6419 control exons, 359 exons that are enhanced by PTBP1 and 419 silenced exons.

- **hnRNPC**

  I used publicly available RNA-seq data from the ENCODE project to identify exons that are regulated by hnRNPC.

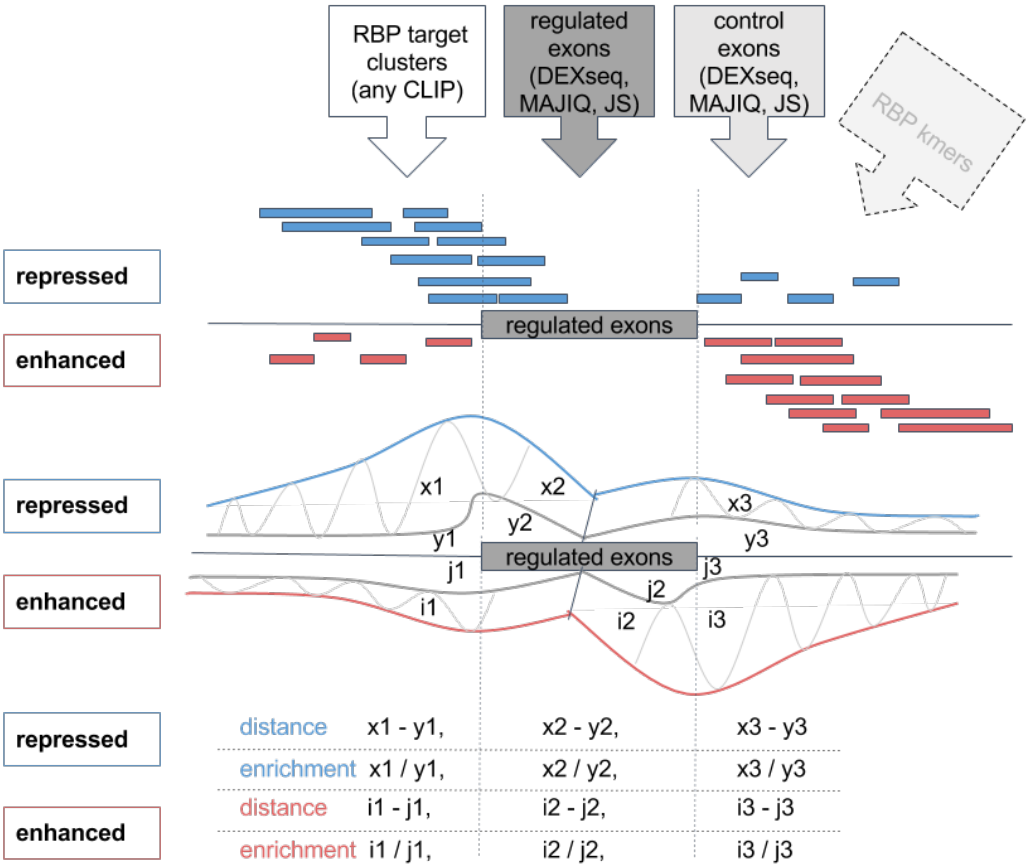

RBP target clusters (any CLIP)
regulated exons (DEXseq, MAJIQ, JS)
control exons (DEXseq, MAJIQ, JS)
RBP kmers
repressed
enhanced
regulated exons
repressed
enhanced
x1
x2
x3
y1
y2
y3
j1
j2
j3
i1
i2
i3
regulated exons
repressed
enhanced
distance x1 - y1, x2 - y2, x3 - y3
enrichment x1 / y1, x2 / y2, x3 / y3
distance i1 - j1, i2 - j2, i3 - j3
enrichment i1 / j1, i2 / j2, i3 / j3

Figure 2.2: Schematic visualisation of RNA-map pipeline.

The first part of the diagram shows the input datasets for the RNA-map pipeline. It is a schematic figure of the output which creates a heatmap of clusters around regulated exons (repressed and enhanced), density figure of crosslink clusters compared to control exons and a table of ratios and distances of crosslink clusters between regulated and control exons in the surrounding region.

2.4.9 Visualisation of crosslink positions around splice sites in the form of RNA-maps

RNA-maps were produced by summarising the cDNA-start counts at significant crosslink sites at all exon/intron and intron/exon boundaries and branch points (BPs) on pre-mRNAs. The definition of intronic start and end positions was based on Ensembl annotation version 75. Only introns longer than 300 nt were used to draw RNA-maps, since they enabled best normalisation of data by the intronic abundance with the following procedure:

1. For each intron, calculate the total count of cDNAs that identify crosslink sites in the deep intronic region (from 50 nt downstream from exon-intron junction to 100 nt upstream of exon-intron junctions).

2. If the count is more than 10, then I proceed with the analysis of this intron. The average cDNA count per nt in the deep intronic regions is used as a normalisation factor.

3. Divide the counts at each position in the intron and flanking exons by the normalisation factor of this intron.

4. Sum up the normalised values from step 2 for each position relative to splice site across the examined introns, and divide this value by the number of examined introns. For all analyses in Cal51 cells from Spliceosomal iCLIP, I only assessed protein-coding genes with FPKM more than 10 in RNA-seq data.

2.4.10 Identification of branch points

For the branch point (BP) identification, I used the spliceosome-iCLIP data produced under mild and medium stringency conditions from Cal51 cell line. In the first step of analysis, I used the spliceosome-iCLIP cDNAs that ended precisely at the ends of introns (I considered only introns terminating with an AG dinucleotide) after removal of the 3' adapter sequence and defined the position where these cDNAs started. The nucleotide preceding the cDNA-start corresponds to the position where cDNAs truncated during the reverse transcription, and I selected the adenine ribonucleotide that had the highest number of truncated cDNAs as the best BP candidate. If two positions with equal number of cDNAs were found, I selected the one closer to the 3' splice sites. This identified 35,056 intronic positions within genes with FPKM higher than 10. The sequence composition around these positions corresponded to the previously reported sequence around BPs. I then proceeded to the second step of the analysis, where I considered all cDNAs (regardless where they ended), but including trimming of the first nucleotide if there was a mismatch within the the adenine. I then overlapped cDNA truncation sites with computationally predicted BPs in the last 100 nt of introns [186]. If this analysis identified a position with a higher cDNA-start count than the initial analysis (or if the initial analysis did not identify any BPs in the same intron), then the newly identified position was assigned as the BP. For introns without any BPs identified by either first or second steps in the analysis, I assessed computationally predicted BPs located further than 100 nt from the 3' splice sites, and if any of these overlapped with a truncating cDNA, I assigned the position closest to the 3' splice sites as the BP. As a result, I identified 50,812 BPs within genes with FPKM higher than 10. These BPs were used for all analyses in the manuscript, and their coordinates were used for BP positioning in RNA-maps. I additionally identified 13,496 BPs in introns of lowly expressed genes, but these were not used for the analyses.

Computational prediction of the secondary structure around BPs (see Chapter 6) was performed using the RNAfold program with default parameters [184], as described previously but considering 40 nt flanking region around identified BPs. I also examined the overlap with BPs identified by previous studies [187, 188], using bedtools intersect tool to compare it with BP coordinates identified by spliceosome-iCLIP data.

2.4.11 Analysis of cDNA C to T mutations

Mapped cDNAs in BAM format from Tophat/Bowtie mapping tool were used as the pipeline input. In the first step I used the 'valmd' function from samtools software with '-e' parameter. This function compares each cDNA sequence with genomic sequence, where only mutations are reported and every match is replaced by '='. In the next step I use a custom script *get_positions_of_transitions_density.py* that creates a numeric vector of summed mutations that were transitioned from C to T across all cDNAs. Final results were plotted as a density graph by using costume R script *coverage_of_C_to_T_transitions.R* and ggplot2 package (see GitHub).

2.5 CLIPo analysis

Results from the CLIPo table in chapter 4 were calculated with the following analysis.

2.5.1 Data complexity

Library size of uniquely mapped cDNAs after PCR duplicate removal.

2.5.2 cDNA-end constrains

I only focus on cDNAs that had a full length sequenced by looking at those which are less than 40 nt long, which is the longest sequencing length for most datasets used in this thesis.

- **Length constraints**

  For the narrow distribution of cDNA length constraints, I used a sliding window approach to detect the most enriched cDNA length density in a 10 nt window frame. This value tells us if there are some strong cDNA length constraints such as narrow cDNA lengths group across the library.

- **Sequence constraints at cDNA-ends**

  In the same way as for definition of crosslink-associated motifs, I used the ratio of the top 10 tetramers that are positioned around cDNA-ends (from less than 40 nts long cDNAs) compared to the top 10 tetramers from central region between -15 to -5 nt upstream from cDNA end.

$$R = \frac{(\sum top10 \; at \; cDNA.end) * 6}{(\sum top10 \; at \; cDNA.centre)} \tag{2.3}$$

- **Structure constraints at cDNA-ends**

  Ratio of single strandedness around cDNA-ends (from less than 40 nt long cDNAs) compared to the central region between -15 to -5 nt upstream from cDNA end. Single strandedness was measured same as above in a 30 nt surrounding region (see subsection 2.4.6).

2.5.3 Specificity of binding sites

- **Number of crosslink clusters**

  The number of crosslink clusters was obtained with iCount peak calling tool. The clusters were identified as described earlier, by considering all crosslink sites that were significant with a FDR lower than 0.05 but with a maximum spacing of 15 nt between crosslink sites.

- **Percentage of cDNA-starts in the clusters**

  Measured percentage of cDNA-starts that are inside of identified crosslink clusters.

2.5.4 Motif enrichment inside the clusters

Enrichment of tetramer coverage between identified clusters and 300 nt control regions downstream from the clusters. Top 10 tetramers were selected from those identified around cDNA-starts in 10 nt surrounding region.

2.5.5 Identification of cDNA-start peaks and tetramer enrichment

I processed each mapped PTBP1 iCLIP, eCLIP and mock-eCLIP dataset with the iCount pipeline to define crosslink clusters with 3 nt spanning window, 20 nt cluster merging and with the threshold lower than 0.05 FDR to identify significant crosslink clusters at the high resolution. I only selected cDNAs that were inside of those clusters and then I selected position with the highest cDNA-start count for each cluster and defined it as a cDNA-start peak for further analysis. Next, I ranked all tetramers that were enriched in 20 nt flanking region around the maximum peaks. The enrichment of each tetramer was measured in comparison with the control frequency of tetramers from non-overlapping region of 200 to 300 nt downstream from cDNA-start peaks. I used the same peaks with the same surrounding region and controls to measure the enrichment of pairing probability using RNAfold software and a python script as described before. For the correlation between tetramer enrichments I used Pearson correlation. I calculated the individual upper and lower quartile of cDNA-start peaks for the most common tetramers and used them for further analysis. The same conditions were used for the pairing probability analysis.

2.5.6 Heatmap of tetramer enrichment around cDNA-start peaks

For each PTBP1 iCLIP, eCLIP and mock-eCLIP dataset, I used a top quartile tetramers from the PTBP1-eCLIP experiment that were enriched around cDNA-start peaks. For each tetramer, I plotted a heatmap of the enrichment across a 50 nt flanking region around cDNA-start peaks. The rows represent the tetramer entries, and they are ranked by enrichment from top to bottom and normalised by the maximum enrichment score. The tetramer sequence is reported on the right side.

Chapter 3

Assessing potential biases in protein-RNA binding site assignment with iCLIP

Individual-nucleotide resolution ultraviolet crosslinking and immunoprecipitation (iCLIP) identifies the RNA crosslink sites of RNA-binding proteins (RBPs) with the use of cDNA-starts. However, a recent study found that positions of cDNA-starts depend on cDNA length in several iCLIP datasets and proposed two alternative solutions: the first recommends use of middle position of the cDNA due to a hypothetical prevalence of readthrough cDNAs, while the second suggests that non-coinciding cDNA-starts are caused by constrained cDNA-ends and thus it was unclear if cDNA-starts or cDNA-centres (middle positions) should be used for analysing resulting data [181]. Here I present in-depth computational comparisons of multiple experiments performed with CLIP, iCLIP and eCLIP methods for three different RNA-binding proteins (PTBP1, U2AF65 and eIF4A3) to determine which of these two solutions is more appropriate.

3.1 Introduction

CLIP is principally composed of eight experimental and two computational steps, each of which can affect the assignment of protein-RNA binding sites (Figure 3.1 - protocol). The first step relies on irradiation of cells or tissues with ultraviolet light

(UV), which creates a covalent bond between proteins and RNAs that are in direct contact. After cell lysis, RNA is then fragmented with the use of RNase to remove the bound RNA excess (Figure 3.1, step 2). The time of incubation and concentration of RNase affect the length of RNA fragments, and can thereby affect the cDNA length distribution in the final cDNA library. The crosslinked RNA fragments are co-immunoprecipitated with antibodies targeting the RBP of interest and ligated to an oligonucleotide adapter at their 3' end (Figure 3.1, step 3). Due to sequence preferences of some RNases and of the RNA ligase, steps 2 and 3 can lead to sequence biases at the 3' end of RNA fragments also known as sequence constraints [189] at cDNA-ends in the final cDNA library. The main difference between CLIP and other immunoprecipitation-based methods, such as chromatin immunoprecipitation (ChIP) [190] or RNA immunoprecipitation (RIP) [191], is that the immunoprecipitated complexes are separated and visualised by SDS-PAGE (Figure 3.1, step 4) [38, 192]. With the use of appropriate controls, this crucial experimental quality control step ensures high specificity, while incomplete optimisation enhances the background signal. When this step is fully optimised, non-specific background is absent, ensuring that only the RNA fragments crosslinked to the purified RBP are obtained. The crosslinked RNA affects the migration of the RBP on SDS-PAGE, and therefore visualisation of the protein-RNA complex also ensures that conditions of RNA fragmentation resulted in an appropriately broad RNA length distribution [126, 192]. The protein-RNA complex is then isolated in a size-specific manner, and the RBP is removed through proteinase K digestion, leaving a small peptide at the crosslink site (Figure 3.1, step 5). This peptide impairs reverse transcription and commonly leads to truncation of cDNAs at the crosslinked peptide. Changes in the conditions of proteinase digestion and reverse transcription could affect the ratio between cDNAs that truncate or readthrough the crosslink site [181, 120].

Therefore, individual-nucleotide resolution CLIP (iCLIP) was developed to amplify the truncated cDNAs, which can identify the protein-RNA crosslink sites with nucleotide resolution [193]. Thus, the main difference between CLIP and iCLIP (and its variants such as FAST-iCLIP, HITS-CLIP or eCLIP) is that iCLIP amplifies truncated cDNAs in addition to the readthrough cDNAs that are amplified in CLIP [194]. The primer used for reverse transcription contains a unique molecular identifier (UMI, also referred to as randomer or random barcode), which can separate unique cDNAs from artefacts of PCR amplification (Figure 3.1, step 8). After mapping the sequenced library to the genome, the number of different UMIs among cDNAs that map to the same genomic position is considered as the 'cDNA count' (Figure 3.1, step 9). Analysis of these cDNA counts within transcripts is then examined to identify crosslink clusters, which are used to assign the RNA binding sites (Figure 3.1, step 10). It has been shown that the positions of crosslink sites needs to be interpreted with caution during binding site assignment [195, 65] (see Chapter 1). Analysis of cDNA libraries of several RBPs indicated that UV crosslinking has a slight uridine preference, which can affect the efficiency of crosslinking at different positions within the RNA binding site [195, 65]. Amino acids can also vary in their crosslinking efficiencies, which together can restrict crosslinking to specific positions within the protein-RNA binding sites. For instance, the binding sites of the Unkempt protein consist of a UAG triplet followed by a U-rich motif and CPSF-160/CPSF-30 proteins bind to AAUAAA flanked by U-rich motifs, but only the U-rich motifs crosslink to both proteins [196, 197]. Recently, new variants of CLIP that rely on amplification of truncated cDNAs were developed, including BrdU-CLIP [198], eCLIP [134] and irCLIP [199]. Therefore, understanding the characteristics of readthrough cDNAs in these methods is essential, since their presence could erroneously shift the boundaries of predicted RBP binding sites to positions upstream of the true binding sites.

Moreover, a recent study observed that the starts of long and short iCLIP cDNAs do not fully coincide across iCLIP libraries for several RBPs, including PTBP1 and eIF4A3 [181]. This non-coinciding cDNA-starts can now be detected by the iCLIPro tool [181]. Accordingly, it is important to fully understand the technical aspects of iCLIP that need to be taken into account during binding site assignment. To further investigate this phenomenon, I focused on experiments produced on the polypyrimidine tract binding protein 1 (PTBP1), the eukaryotic initiation factor 4A-

III (eIF4A3), which is a component of the exon junction complex (EJC) that binds the region between 20 and 24 nt upstream of the exon-exon junction [200, 201, 166], and the splicing factor U2 auxiliary factor 65 kDa subunit (U2AF2), which represents an example of non-coinciding or coinciding cDNA-starts in introns or exons.

The Exon Junction Complex (EJC) is a set of proteins forming a complex that is deposited on mRNAs at the junction sites between exons after they have been joined together during the splicing process. The EJC is involved in various cellular processes such as nucleo-cytoplasmic mRNA export, subcellular localisation, quality control and translation [202, 203]. There are four core proteins: eIF4A3 (DDX48), MAGOH, Y14(RBM8A) and Barsentz (BTZ, CAC3 or MLN51) [204, 203, 205]. The RNA sequence plays only a marginal role in defining the position of EJC binding, and therefore it is not yet possible to computationally predict the binding sites without experimental data. Rather than sequence specificity, EJC binding is established by distance from the exon-exon junction. This interpretation was presented in the first CLIP study of the EJC complex [200], where the authors used eIF4A3 protein to identify EJC binding sites across the human genome. It has been previously shown that the EJC complex protects the region between 20 and 24 nt upstream of the exon-exon junction from cleavage by RNase H [200, 201, 166]. Further studies also showed that the sequence and structure of a nascent mRNA can shift EJC deposition as far as 10 nt away from this expected site [206]. The iCLIP study of eIF4A3 found that the non-coinciding iCLIP cDNA-starts mapped upstream of this expected region (-24 to -20 nt), while cDNA-centres were located closer to this region. This suggests that non-coinciding cDNA-starts might reflect a high prevalence of readthrough cDNAs. An alternative hypothesis, however, proposed that non-coinciding cDNA-starts are unrelated to readthrough cDNAs [181]. Understanding the cause of non-coinciding starts is therefore crucial to ensure correct analysis and interpretation of iCLIP and related protocols that can amplify truncated cDNAs, such as BrdU-CLIP, eCLIP and irCLIP [198, 134, 199].

To identify the characteristics of readthrough cDNAs and their potential influence on the iCLIP data, we first developed a modified iCLIP protocol that enables direct identification of such cDNAs. This showed that sequence features at the starts of readthrough cDNAs are different from the majority of cDNAs in PTBP1 and eIF4A3 iCLIP. To further argue against predominance of readthrough cDNAs in iCLIP, I analysed the iCLIP data with high frequency of non-coinciding cDNA-starts, where I first examined the position and prevalence of crosslink-induced mutations. In agreement with previous findings, I showed that crosslink-induced mutations are generally more than 5-fold less common within iCLIP than in 'readthrough' CLIP cDNAs, regardless of the presence of non-coinciding cDNA-starts [120].

Next, I identified RBP motifs that are commonly associated with crosslink sites and found them highly enriched at cDNA deletions in CLIP, and cDNA-starts in iCLIP, eCLIP and irCLIP method. This enrichment was observed in coinciding and non-coinciding cDNA-starts. Interestingly, I observed that in a modified iCLIP protocol where the UV radiation was performed with the photoactivatable 4-thiouridine (4SU)-based crosslinking, the motifs were more highly enriched at cDNA-starts than at T-to-C transitions. These results demonstrate that the cDNA-starts can reliably be used to determine crosslink sites in iCLIP, regardless of the crosslinking method.

By continuing with my research, I discovered that the non-coinciding cDNA-starts result from sequence and structural constraints that are present at cDNA-ends. To follow up on that finding, more experimental iCLIP data was produced for PTBP1 and eIF4A3 to demonstrate that the prevalence of the non-coinciding cDNA-starts is directly correlated with the extent of cDNA-end constraints. Finally, I demonstrated that the broad size range of iCLIP cDNAs in these new experiments allows the cDNA-starts to assign binding sites that align with the expected binding motifs (for PTBP1) or binding regions (for eIF4A3). I conclude that the use of the iCLIP cDNA-starts is appropriate to assign the protein-RNA crosslink sites in iCLIP and other related methods.

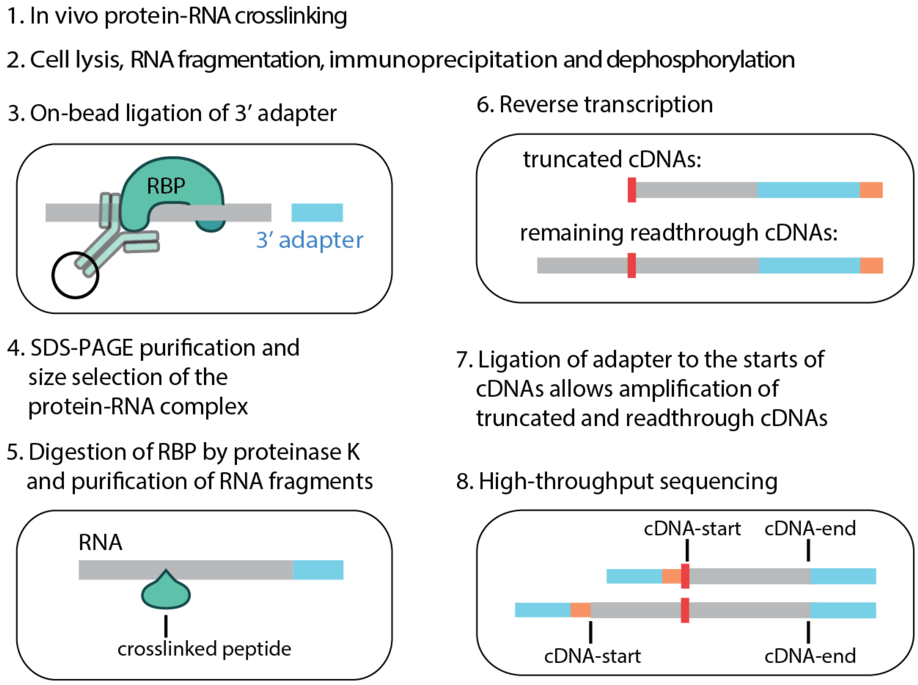

1. In vivo protein-RNA crosslinking
2. Cell lysis, RNA fragmentation, immunoprecipitation and dephosphorylation
3. On-bead ligation of 3′ adapter
RBP
3′ adapter
4. SDS-PAGE purification and size selection of the protein-RNA complex
5. Digestion of RBP by proteinase K and purification of RNA fragments
RNA
crosslinked peptide
6. Reverse transcription
truncated cDNAs:
remaining readthrough cDNAs:
7. Ligation of adapter to the starts of cDNAs allows amplification of truncated and readthrough cDNAs
8. High-throughput sequencing
cDNA-start
cDNA-end
cDNA-start
cDNA-end

Figure 3.1: Schematic representation of the iCLIP protocol with truncated and readthrough cDNAs.

First, cells or tissues are irradiated with UV light, which creates covalent bonds between proteins and RNAs that are in direct contact (step 1). After lysis, the crosslinked RNA is fragmented by limited concentration of RNase I, and RNA fragments are then co-immunoprecipitated with the RBP (step 2), followed by ligation of a 3' adapter (step 3). After SDS-PAGE purification (step 4), the crosslinked RBP is removed through proteinase K digestion and RNA fragments are purified; since the ligation reaction is not 100% efficient, only a subset of the fragments contain the 3' adapter (step 5). Reverse transcription is performed with a primer that includes a barcode (orange) containing both an experimental identifier and a unique molecular identifier (UMI) (step 6). The peptide that is on the crosslink site impairs reverse transcription and commonly leads to truncation of cDNAs at the crosslink site. Therefore, two types of cDNAs are generated: truncated cDNAs and readthrough cDNAs. In iCLIP, the cDNA library is prepared in such a way that both truncated and readthrough cDNAs are amplified (step 7). After PCR amplification and sequencing (step 8), both truncated and readthrough cDNAs are present.

3.2 Crosslink sites are identified by cDNA-starts in iCLIP

There are eight primary experimental steps in the original iCLIP protocol (Figure 3.1). In the first step, cells are exposed to ultraviolet light irradiation, which can create a covalent bond between RBPs and RNA. Cell lysates are then treated with RNAse, and the crosslinked RNA fragments are co-immunoprecipitated with the RBP. In the third step, an oligonucleotide adapter is ligated to the 3' end of the RNA fragments. The immunoprecipitated complexes are then separated and visualised by SDS-PAGE, the protein-RNA complex is isolated in a size-specific manner, and the RBP is removed through proteinase K digestion, leaving a small peptide at the crosslink site. This peptide stops the reverse transcription that commonly leads to the truncation of cDNAs at the crosslinked peptide. Therefore, iCLIP cDNA-start positions are at the nucleotide downstream of the crosslinked peptide and the cDNA-ends at the site of RNase cleavage.

Protein	Method and experiment number	(PMID)	Source	Cell line	Total number of unique cDNAs
eIF4A3	CLIP	23085716	GSM1001330	HeLa	11,690,349
eIF4A3	iCLIP1	26260686	E-MTAB-2599	HeLa	7,148,538
eIF4A3	iCLIP2	new	E-MTAB-3618	HEK293	14,454,772
eIF4A3	iCLIP3	new	E-MTAB-4000	HeLa	11,935,788
PTBP1	CLIP	23313552	GSE19323	HeLa	1,779,318
PTBP1	iCLIP1	25599992	E-MTAB-3108	HeLa	8,447,229
PTBP1	iCLIP2	new, 4SU-crosslinked	E-MTAB-5027	HEK293	9,211,541
PTBP1	iCLIP3	new, -3'deP	E-MTAB-5026	HeLa	3,275,592
PTBP1	iCLIP4	new	unpublished	HeLa	90,098
PTBP1	eCLIP	new, Encode	ENCSR981WKN	K562	6,060,266
mock	eCLIP	new, Encode	ENCSR445FZX	K562	5,669,907
PTBP1	irCLIP	27111506	CSR981WKN	HeLa	65,593,070
U2AF2	CLIP	25326705	GSM1509288	HeLa	4,702,278
U2AF2	iCLIP	23374342	E-MTAB-1371	HeLa	116,771,612

Table 3.1: Overview of methods and experiments in chapter 3.

To analyse how binding sites can be affected by variations in experimental conditions, I compared published and newly produced experimental data for eIF4A3, PTBP1 and U2AF2. For simplicity reasons I labelled each experiment with a unique number after the protein and methods name (see Table 3.1). For the eIF4A3 protein, eIF4A3-iCLIP1 refers to data produced in a previous study [181] and so does eIF4A3-CLIP [201], while eIF4A3-iCLIP2 and eIF4A3-iCLIP3 were newly produced by the Le Hir and Ule labs respectively. For the PTBP1 protein, the PTBP1-iCLIP1 also refers to data produced in a previous study [185], while PTBP1-iCLIP2, PTBP1-iCLIP3 and PTBP1-iCLIP4 were newly produced with the protocol modifications for the purpose of this study. Specifically, 4SU was used to induce crosslinking and RNase I conditions were adjusted in PTBP1-iCLIP2, while the 3' dephosphorylation step was omitted in PTBP1-iCLIP3. For eIF4A3-iCLIP3 and PTBP1-iCLIP4 experiments, the 5' ligation was added (Figure 3.4) to separate readthrough cDNAs from truncated ones. I compared all these datasets to the published PTBP1-CLIP [207], PTBP1-eCLIP [134] and PTBP1-irCLIP [199], together with U2AF2-CLIP [49] and U2AF2-iCLIP [50] datasets (Table 3.1).

A recent study proposed that the presence of non-coinciding cDNA-starts indicates that these cDNAs could be an outcome of readthrough from the crosslink site during reverse transcription [181]. One feature that can serve as an identifier of crosslink sites within readthrough cDNAs is the presence of deletions, which are often introduced into cDNAs at the position of the crosslink site during reverse transcription [120, 208]. I therefore compared the proportion of cDNAs with deletions in the eIF4A3 datasets used in the present study (Table 3.1). Some datasets were prepared by different sequencing protocols, where some of the libraries contained longer RNA fragments. For this reason, I only examined cDNAs shorter than 40 nt since the rate of sequencing errors increases with increasing cDNA length. Even though the proportion of deletions is lower in eIF4A3 iCLIP compared to CLIP, a bimodal distribution of deletions is apparent in all datasets, with one peak of deletions close to the cDNA-starts (5th to 8th nt) and the second one close to the cDNA-centres (22nd to 27th nt) (Figure 3.2a). Thus, the deletions present in iCLIP show the same features as in CLIP and likely inform on the presence of readthrough cDNAs. More importantly, the proportion of deletions is lower by a factor of 5 or more in all eIF4A3 iCLIP experiments compared to CLIP, indicating that readthrough cDNAs represent a minor proportion of iCLIP data (Figure 3.2a).

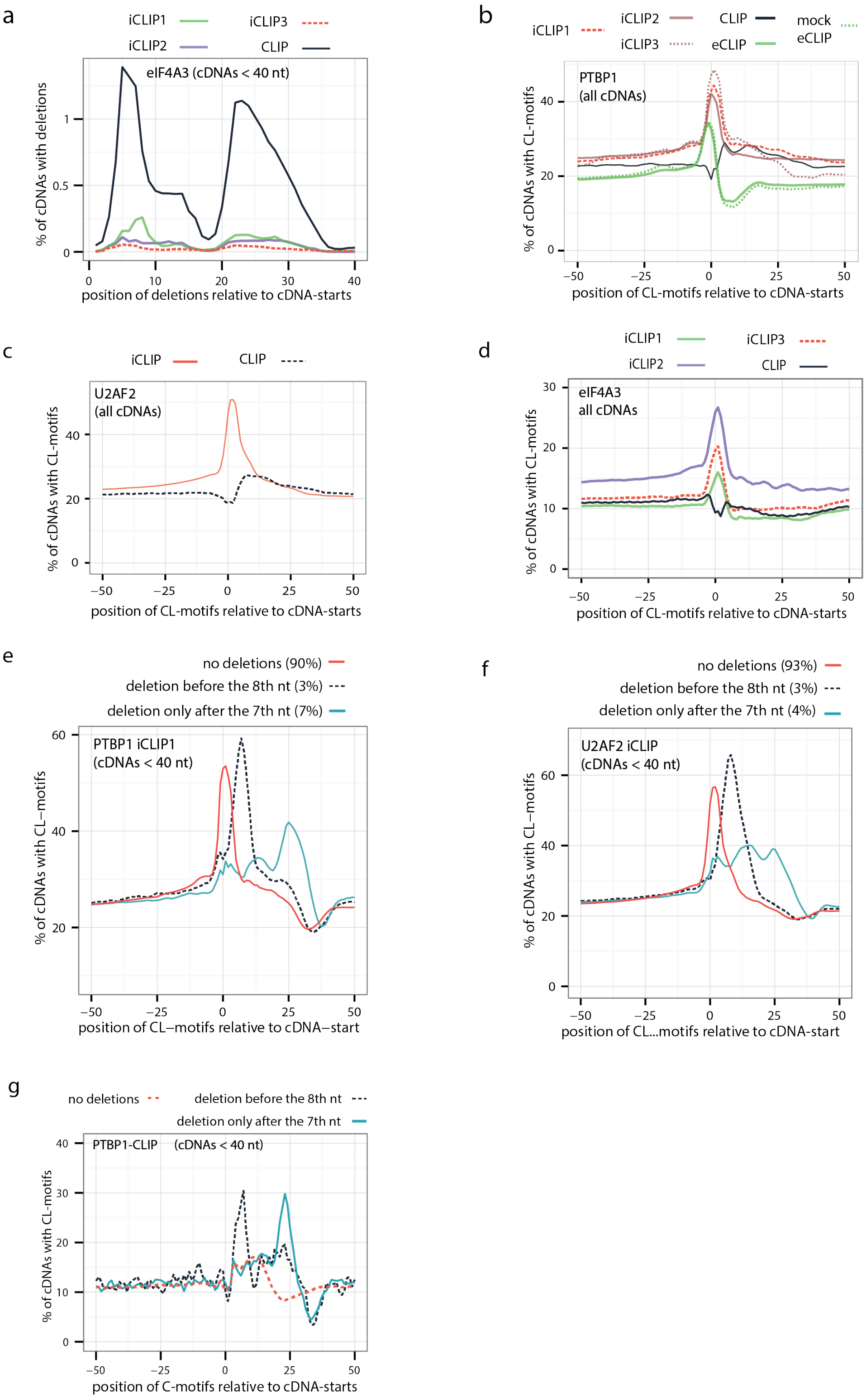


Figure 3.2: Distribution of deletions and crosslink-associated (CL)-motifs in CLIP and iCLIP experiments.

(continued)

Figure 3.2: a) Proportion of eIF4A3 cDNAs with deletion at each position relative to the cDNA-start. Only cDNAs shorter than 40 nt are examined.
b) Proportion of PTBP1 cDNAs from each experiment that overlap with a CL-motif at each position relative to the cDNA-start.
c) Proportion of U2AF2 cDNAs from each experiment that overlap with a CL-motif at each position relative to the cDNA-start.
d) Proportion of eIF4A3 cDNAs from each experiment that overlap with a CL-motif at each position relative to the cDNA-start. e) Proportion of PTBP1-iCLIP1 cDNAs that overlap with a CL-motif at each position relative to the cDNA-start. Only cDNAs shorter than 40 nt are examined; they are divided into those lacking deletions or containing a deletion within the first 7 nt or anywhere in the remaining portion of the cDNA.
f) Same as e), but for U2AF2-iCLIP.
g) Proportion of PTBP1-CLIP cDNAs that overlap with a CL-motif at each position relative to the cDNA-start. Only cDNAs shorter than 40 nt are examined; they are divided into those lacking deletions or containing a deletion within the first 7 nt or anywhere in the remaining portion of the cDNA.

A second feature that can serve as an identifier of crosslink sites is the potential nucleotide preference of crosslinking, and therefore I defined these sequence motifs based on analysis of eCLIP mock input data. The mock eCLIP co-purified RNA fragments that were crosslinked to a random mixture of RBPs, and thus the enrichment of motifs at cDNA-starts in this experiment indicates that these motifs represent crosslinking preferences common to most RBPs [134]. To identify these common motifs, I used the kmer analysis (see Methods 2.4.3), where I distinguished 10 tetramers that were enriched by 1.5 factor at cDNA-start positions compared to the region 10 nucleotides preceding the cDNA-starts. Thus, I assume that these tetramers represent preference of UV crosslinking motifs, and therefore I refer to them as 'CL-motifs' (for UV crosslink-associated motifs). All of these CL-motifs are rich in uridines (see Methods 2.4.4), in agreement with the previous finding that crosslinking tends to have preference for uridines [120]. The CL-motifs also contain polypyrimidine (polyY) sequences that are preferentially bound by PTBP1 and U2AF2, and thus it is expected that their enrichment should be even higher at the crosslink sites of these proteins. Next, I examined the occurrence of CL-motifs around the starts of all cDNAs in each experiment, including PTBP1 CLIP, iCLIP, eCLIP and its mock control and U2AF65 CLIP and iCLIP [134, 185, 50, 49, 207].

Notably, the CL-motifs are enriched at cDNA-starts of all eIF4A3, PTBP1 and U2AF2 iCLIP and eCLIP experiments (Figure 3.2b-d). This agrees with the presence of 'truncated cDNAs' in iCLIP and eCLIP but not CLIP. However, CL-motifs have an almost identical distribution around cDNA-starts in the PTBP1 eCLIP and mock eCLIP experiments, which is somewhat surprising (Figure 3.2b). While no further increase in CL-motif enrichment is seen at cDNA-starts of PTBP1-eCLIP, it is reassuring to find their increased enrichment at cDNA-starts of all PTBP1 and U2AF2 iCLIP experiments (Figure 3.2b, c). Taken together, analysis of CL-motifs indicates that the incidence of readthrough cDNAs is low, and that the majority of cDNAs truncate at crosslink sites in iCLIP and its variants, such as eCLIP.

Significant CL-motif enrichment at cDNA-starts is present in all eIF4A3 iCLIP experiments (Figure 3.2d, Figure 3.3a). This protein is not thought to bind RNA with sequence specificity according to biochemical and transcriptomic studies [200, 201, 166], and its sequence-independent interaction with RNA is consistent with the properties of DEAD-box proteins [209]. Moreover, I did not find any generic enrichment of CL-motifs at nucleotides 20 to 24 upstream of the exon-exon junctions, where EJC normally binds [201]. Thus, it is most likely that CL-motifs only reflect crosslinking preferences in the case of eIF4A3 iCLIP. In contrast to their enrichment at cDNA-starts of all iCLIP experiments, CL-motifs are depleted from the cDNA-starts of all CLIP experiments and instead they are enriched within the sequence of CLIP cDNAs (Figure 3.2b-d). This agrees with the expected prevalence of truncated cDNAs in iCLIP and readthrough cDNAs in CLIP.

To validate the CL-motifs, I exploited the bimodal distribution of deletions in cDNAs shorter than 40 nt. I separated these cDNAs into three classes: those with deletions in the first 7 nts, those with deletions elsewhere, and those with no deletions. In iCLIP and CLIP datasets of all examined RBPs, enrichment of CL-motifs closely followed the position of the deletions. If cDNAs contain a deletion in PTBP1 and U2AF2 iCLIP, CL-motifs are most highly enriched at the position of deletion, but not at cDNA-starts, which confirms that they represent readthrough cDNAs (Figure 3.2e, f). Analysis of cDNAs without deletions reveals a striking difference between CLIP and iCLIP in the positions of CL-motifs. These cDNAs contain CL-motif enrichment almost exclusively at cDNA-starts in iCLIP, and unlike the cDNAs with deletions, CL-motifs are not enriched downstream of cDNA-starts (Figure 3.2e, f and Figure 3.3a, e). Interestingly, even the cDNAs with deletions also contain some CL-motif enrichment at their starts, and this is most apparent in eIF4A3 iCLIP, indicating that the readthrough cDNAs often readthrough one crosslink site and truncate at another crosslink site (Figure 3.3b). In contrast, occurrence of CL-motifs decreases at the start of cDNAs lacking deletions in CLIP, and thus these cDNAs are similar to those containing deletions (Figure 3.2g), which is expected given that all cDNAs in CLIP are readthrough cDNAs. In conclusion, the overlap with the deletions confirms that the CL-motifs can be used to estimate the general distribution of crosslink sites within CLIP and iCLIP cDNAs.

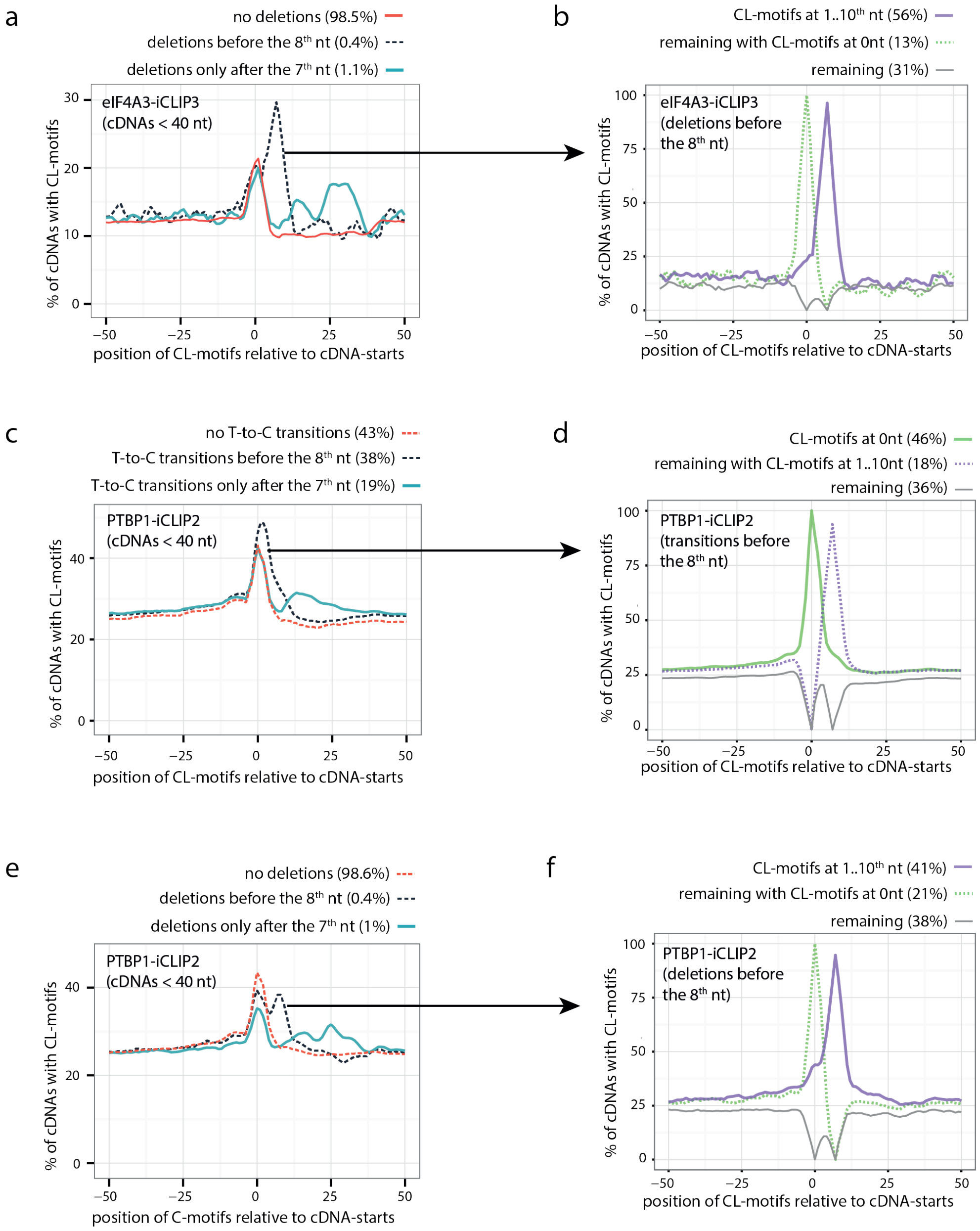


Figure 3.3: Crosslink-associated (CL)-motifs are enriched at cDNA deletions and cDNA-starts in iCLIP.

a) Proportion of eIF4A3-CLIP3 cDNAs that overlap with a CL-motif at each position relative to the cDNA-start. Only cDNAs shorter than 40 nt are examined; they are divided into those lacking deletions or containing a deletion within the first 7 nt or anywhere in the remaining portion of the cDNA.
b) The cDNAs of eIF4A3-CLIP3 containing a deletion within the first 7 nt are further sub-divided into three categories. First, cDNAs with CL-motifs between the 1st and 10th nucleotide of the cDNA. Second, the remaining cDNAs that contain CL-motifs at the position 0. And third, all remaining cDNAs. The proportion of cDNAs that overlap with a CL-motif at each position relative to the cDNA-start is then plotted for each sub-category.

(continued)

Figure 3.3: c) Proportion of PTBP1-iCLIP2 cDNAs that overlap with a CL-motif at each position relative to the cDNA-start. Only cDNAs shorter than 40 nt are examined and are divided into those lacking T-to-C transitions or containing a transition within the first 7 nt or anywhere in the remaining portion of the cDNA.
d) The cDNAs of PTBP1-iCLIP2 containing a T-to-C transition within the first 7 nt are further sub-divided into three categories. First, cDNAs with CL-motifs overlapping the position 0. Second, the remaining cDNAs that contain CL-motifs between the 1st and 10th nucleotide of the cDNA. And third, all remaining cDNAs. Visualisation as in (d).
e) Same as a), but for PTBP1-iCLIP2.
f) Same as b), but for PTBP1-iCLIP2.

3.3 cDNA-starts assign crosslink sites in iCLIP regardless of the crosslinking method

Even when cDNAs contain deletions in eIF4A3-iCLIP3 experiment, an enrichment of CL-motifs can still be seen at cDNA-starts (Figure 3.3a). The most likely explanation for this dual enrichment is that some of the deletions were introduced by sequencing errors from truncated cDNAs that originally lacked deletions. To follow up this explanation, I separated cDNAs with deletions in 8 nt upstream region from cDNA-starts and grouped them into three classes: cDNAs that have CL-motif overlapping with deletion (56%), cDNAs that have CL-motif at cDNA-start (13%) and remaining cDNAs (31%) (Figure 3.3b). Thus, the dual enrichment of CL-motifs indicates that the majority of the iCLIP reads with deletions and CL-motifs contains a mixture of truncated cDNAs, and that the other cDNAs are more likely a result of sequencing errors within truncated cDNAs, rather than readthrough cDNAs. This analysis of deletions and CL-motifs further indicates that the incidence of readthrough cDNAs is low, and the majority of cDNAs truncate at crosslink sites in iCLIP.

Since cDNAs with deletions are rare in iCLIP, I included another modified iCLIP experiment (PTBP1-iCLIP2) to examine the affect of 4SU incubation that is used to induce crosslinking. In PTBP-iCLIP2 experiment, cells were incubated with 4SU, crosslinking was induced with UV-A and RNase conditions were optimised. One characteristic of 4SU-mediated crosslinking is the presence of C-to-T (thymidine to cytidine) transitions that can be used to identify crosslinking positions and it was originally introduced in the PAR-CLIP method [210]. I used the same CL-motif analysis to examine the alignment of cDNA truncations and C-to-T transitions to crosslink sites across different experiments. Even though CL-motifs are similar to the known PTBP1 binding motifs [211], I expect that the CL-motifs correspond to the crosslinking sites of PTBP1, independently of the fact that they also overlap with its binding preferences. Interestingly, just like in PTBP1-iCLIP1 there is a similar pattern in PTBP1-iCLIP2 experiment, where 57% of cDNAs contained transitions, but CL-motifs were mainly enriched at cDNA-starts (Figure 3.3c, d). In 67% of cDNAs with transitions, the position of the transition mapped to the first few nucleotides close to the cDNA-start. To further explore these group of cDNAs in more detail, I used the same type of analyses as before by dividing them into three classes (Figure 3.3d): cDNAs that have CL-motif overlapping with deletion (46%), cDNAs that have CL-motif at cDNA-start (18%) and remaining cDNAs without any CL-motifs (36%). One potential explanation for this enrichment could be that a quarter of transitions correspond to the crosslink sites and the majority could be a result of some other causes. Taken together, the presence of transitions does not faithfully separate readthrough from truncated cDNAs in iCLIP experiments since CL-motifs are evenly enriched at cDNA-starts in cDNAs that contain or lack transitions.

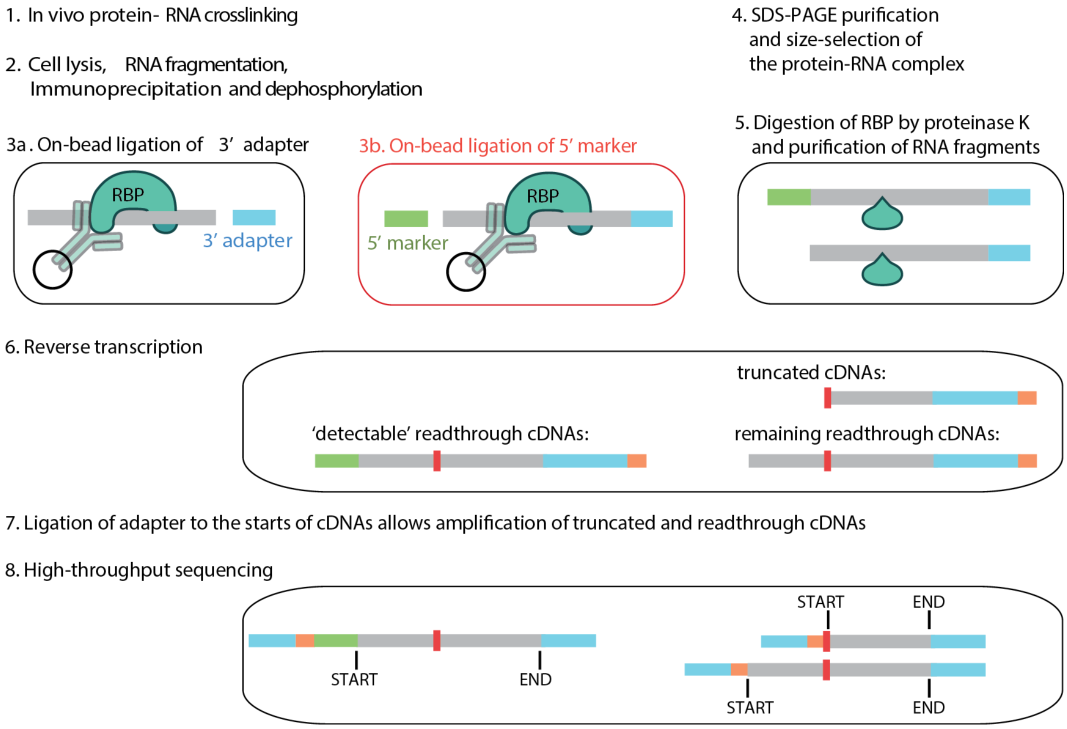
Interestingly, in PTBP1-iCLIP2 experiment with 4SU incubation there was only a small proportion (1.4%) of PTBP1-iCLIP2 cDNAs that contained deletions (Figure 3.2e), a greater proportion contain CL-motifs at the position of the deletion than at cDNA-starts (Figure 3.2f). This indicates that deletions are more reliable than transition to identify crosslink sites in readthrough cDNAs, even when 4SU is used for crosslinking in iCLIP. Taken together, the analysis of deletions, transitions and CL-motifs indicates that the incidence of readthrough cDNAs is generally low and that the majority of cDNAs truncate at crosslink sites in iCLIP regardless of the crosslinking method.

3.4 Defining the characteristics of readthrough cDNAs in iCLIP

Assignment of protein-RNA crosslink sites from iCLIP data relies on the start positions of truncated cDNAs. Previous computational comparisons of CLIP and iCLIP cDNAs estimated that over 80% of iCLIP cDNAs truncate at the crosslink sites of NOVA, TIA1, TIAL1, hnRNP C and TDP43, and 57% of cDNAs at the crosslink sites of the RBFOX protein [120]. In order to correctly assign protein-RNA binding sites from iCLIP, it is important to understand the contribution of readthrough and truncated cDNAs [120, 181]. However, it has not been possible to experimentally distinguish readthrough from truncated cDNAs.

Therefore, a new modified iCLIP protocol with an additional ligation step was designed in Ule lab in a manner that enables direct identification of readthrough cDNAs (Figure 3.4). This additional ligation step was included to add a 'marker' oligonucleotide to the 5' end of RNA fragments (Figure 3.4, step 3b, green). The 5' marker will become part of only those cDNAs that readthrough the crosslink site. The sequence of the 5' marker is not complementary to the PCR primers, and thus is present in the sequencing read. As a consequence, only the readthrough cDNAs will contain the sequence of the new 5' marker (Figure 3.4, step 6 and step 8, green).

The library was prepared with the low concentration of RNAse according to the iCLIP guidelines [126, 192] with cDNAs that were between 20 and 140 nt long. The sequenced reads were produced in size 150 nt using the Illumina HiSeq platform for eIF4A3-iCLIP2 and MiSeq platform for PTBP1-iCLIP4, which after removal of adapters obtained complete sequences for cDNAs up to a length of 140 nt. The first 9 nt of the sequenced iCLIP read correspond to the barcode, which contains the experimental identifier that allows to separate experimental replicates, and the UMI, which allows to avoid artefacts caused by variable PCR amplification of different cDNAs (Figure 4, step 6, orange) [124]. I used these UMIs to quantify the number of unique cDNAs that mapped to each position in the transcriptome [124].



1. In vivo protein- RNA crosslinking
2. Cell lysis, RNA fragmentation, Immunoprecipitation and dephosphorylation
3a. On-bead ligation of 3' adapter
RBP
3' adapter
3b. On-bead ligation of 5' marker
RBP
5' marker
4. SDS-PAGE purification and size-selection of the protein-RNA complex
5. Digestion of RBP by proteinase K and purification of RNA fragments
6. Reverse transcription
truncated cDNAs:
'detectable' readthrough cDNAs:
remaining readthrough cDNAs:
7. Ligation of adapter to the starts of cDNAs allows amplification of truncated and readthrough cDNAs
8. High-throughput sequencing
START
END
START
END
START
END

Figure 3.4: Schematic representation of the modified 5' marker iCLIP protocol.

A schematic description of the modified iCLIP protocol. Before, cells or tissues are irradiated with UV light, which creates a covalent bond between proteins and RNAs that are in direct contact (step 1). After lysis, the crosslinked RNA is fragmented by limited concentration of RNase I, and RNA fragments are then co-immunoprecipitated with the RBP (step 2). Ligation of a 3' adapter (step 3a) is followed by ligation of a 5' marker that is unique to the modified protocol (red balloon, step 3b). After SDS-PAGE purification (step 4), the crosslinked RBP is removed through proteinase K digestion and purification of RNA fragments; since the ligation reaction is not 100% efficient, only a subset of the fragments contain both the 3' adapter and the 5' marker (step 5). Reverse transcription is performed with a primer that includes a barcode (orange) containing both an experimental identifier and a unique molecular identifier (UMI) (step 6). The peptide that is on the crosslink site impairs reverse transcription and commonly leads to truncation of cDNAs at the crosslink site. Therefore two types of cDNAs are generated: truncated cDNAs (which never contain the 5' marker) and readthrough cDNAs (some of which contain the 5' marker). In iCLIP, the cDNA library is prepared in such a way that both truncated and readthrough cDNAs are amplified (step 7). After PCR amplification and sequencing (step 8), the 5' marker sequence is present only at the beginning of readthrough cDNAs.

Using this modified iCLIP protocol, I produced iCLIP datasets for PTBP1-iCLIP4 and eIF4A3-iCLIP3 (Table 3.1), 3.4% and 0.2% of the resulting reads contained the 5' marker at their start, respectively (Figure 3.5a, c). Since the efficiency of the 5' marker ligation is unknown, some readthrough cDNAs could also be present in the remaining pool of cDNAs, but I can be confident that those containing the 5' marker correspond to readthrough cDNAs. Nevertheless, the nucleotide composition at starts of readthrough cDNAs (Figure 3.5a, c) is quite different from the remaining cDNAs, which suggests that readthrough cDNAs represent a minor portion of the remaining cDNAs (Figure 3.5b, d).

The readthrough cDNAs most often contain adenosine as their first nucleotide. Since the start of readthrough cDNAs marks the position of RNase I cleavage, this suggests that RNase I may have a sequence preference for cutting upstream of adenosines. In contrast, the nucleotide before the start of the remaining cDNAs is enriched in thymidine (T), which likely reflects a combination of binding preference of the studied RBPs and the potential preference of UV crosslinking at uridines [120]. Importantly, the sequence characteristics at the starts of readthrough cDNAs are quite different from the remaining cDNA pool, which implies that they do not represent a major proportion of the iCLIP cDNA libraries.

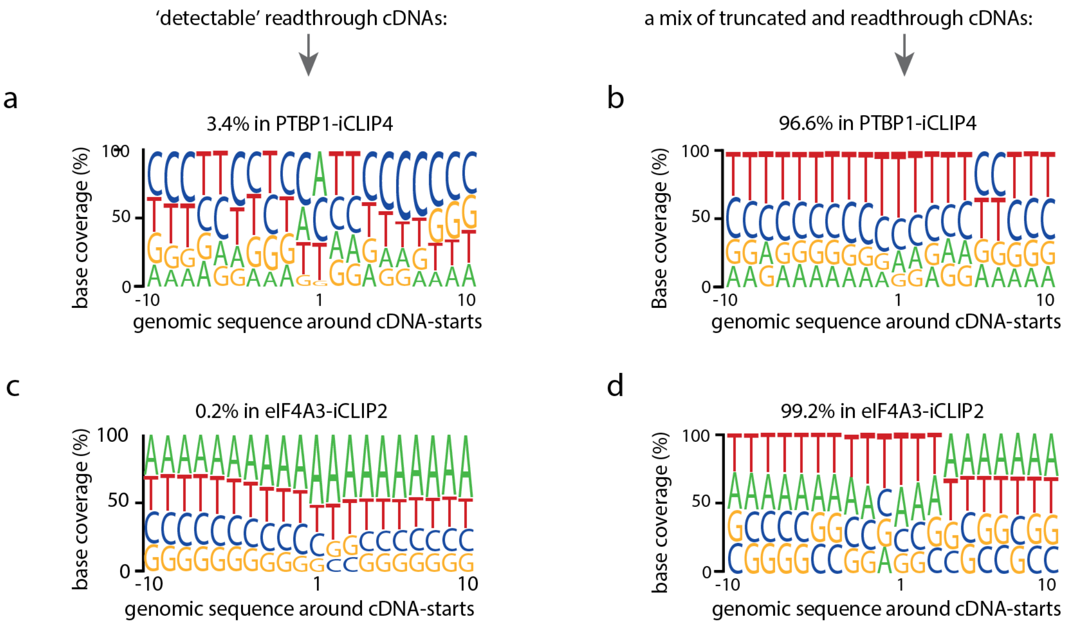


Figure 3.5: A modified 5'marker iCLIP protocol identifies readthrough cDNAs.

a-d) The composition of genomic nucleotides around iCLIP cDNA-starts that were generated using the modified protocol. These include 3.4% of the mapped PTBP1-iCLIP4 cDNAs a) and 0.2% of the mapped eIF4A3-iCLIP2 cDNAs c) that were preceded by the 5' marker (readthrough cDNAs), as well as 96.6% of the mapped PTBP1-iCLIP4 cDNAs b) and 99.8% of the mapped eIF4A3-iCLIP2 cDNAs d) that lacked the 5' marker sequence.

3.5 Non-coinciding cDNA-starts result from constrained cDNA-ends

The previous study [181] also discussed an alternative model in addition to the argued model based on readthrough cDNAs. In the alternative model the non-coinciding cDNA-starts could be derived from constraints on RNase cleavage, particularly when these are combined with the presence of long binding sites [181]. All previous analyses in this study showed that in spite of non-coinciding starts, the prevalence of readthrough cDNAs appears low in iCLIP, and therefore I decided to study this alternative model in more detail. First, I examined the prevalence of non-coinciding cDNA-starts with the CLIPro tool that was developed by the previous study together with PTBP1-iCLIP1 cDNA library which was also produced as part of the previous study [181]. CLIPro is a tool that compares different cDNA length groups by plotting heat maps of overlapping cDNA-start positions compared to long cDNAs. For this purpose, I examined cDNAs longer than 16 nt, which is the minimum cDNA length after trimming the 3' adapter sequence, and shorter than 40 nt, which is the longest cDNA length after the random and experimental barcode removal (see Methods 1). In the CLIPro analysis I compared trimmed cDNAs (between 17 nt and 39 nt) with the reference cDNAs longer than 39 nt since we do not know their exact length because of the maximum sequencing length. I was able to reproduce the same results (Figure 3.6a) of non-coinciding cDNA-starts as in the previous study [181]. Next, I focused on cluster regions where cDNA-starts are significantly enriched and performed the same CLIPro analysis on cDNAs coming from short (5 to 30 nt) and long (more than 30 nt) clusters to see if the non-coinciding cDNA-starts depends on the length of binding sites. Interestingly, the ratio of non-coinciding cDNA-starts increases in clusters that are longer than 5 and shorter than 30 nt but this ratio is even more visible in clusters that are longer than 30 nt (Figure 3.6c). This analysis reveals that the non-coinciding cDNA-starts originate mainly from the long binding sites.

To follow up the alternative model of RNase cleavage constraints, another modified iCLIP experiment (PTBP1-iCLIP2) was performed by Ule lab. In this modified PTBP1-iCLIP2 experiment, RNAse treatment conditions were optimised by including an inhibitor of endogenous RNases into the lysis buffer (antiRNase which does not inhibit RNase I), and by slightly increasing the concentration of RNase I compared to PTBP1-iCLIP1. This condition could potentially ensure that RNase I, which does not to have any sequence specificity, was responsible for fragmenting the RNAs. Notably, the ratio of non-coinciding cDNA-starts decreased in PTBP1-iCLIP2, especially in cDNAs coming from long clusters (Figure 3.6e, f). In addition to overlapping cDNA-starts, the cDNA-ends in PTBP1-iCLIP2 also often overlap with cDNA-starts (diagonal enrichment in Figure 3.6d-f). Both RNase I and UV crosslinking require single-stranded RNA, and thus their similar RNA structure preferences could increase their possibility of overlap. Taken together, our results show that the prevalence and position of non-coinciding cDNA-starts can vary greatly between different iCLIP experiments performed for the same RBP, and thus they are most likely a result of technical differences between these experiments.

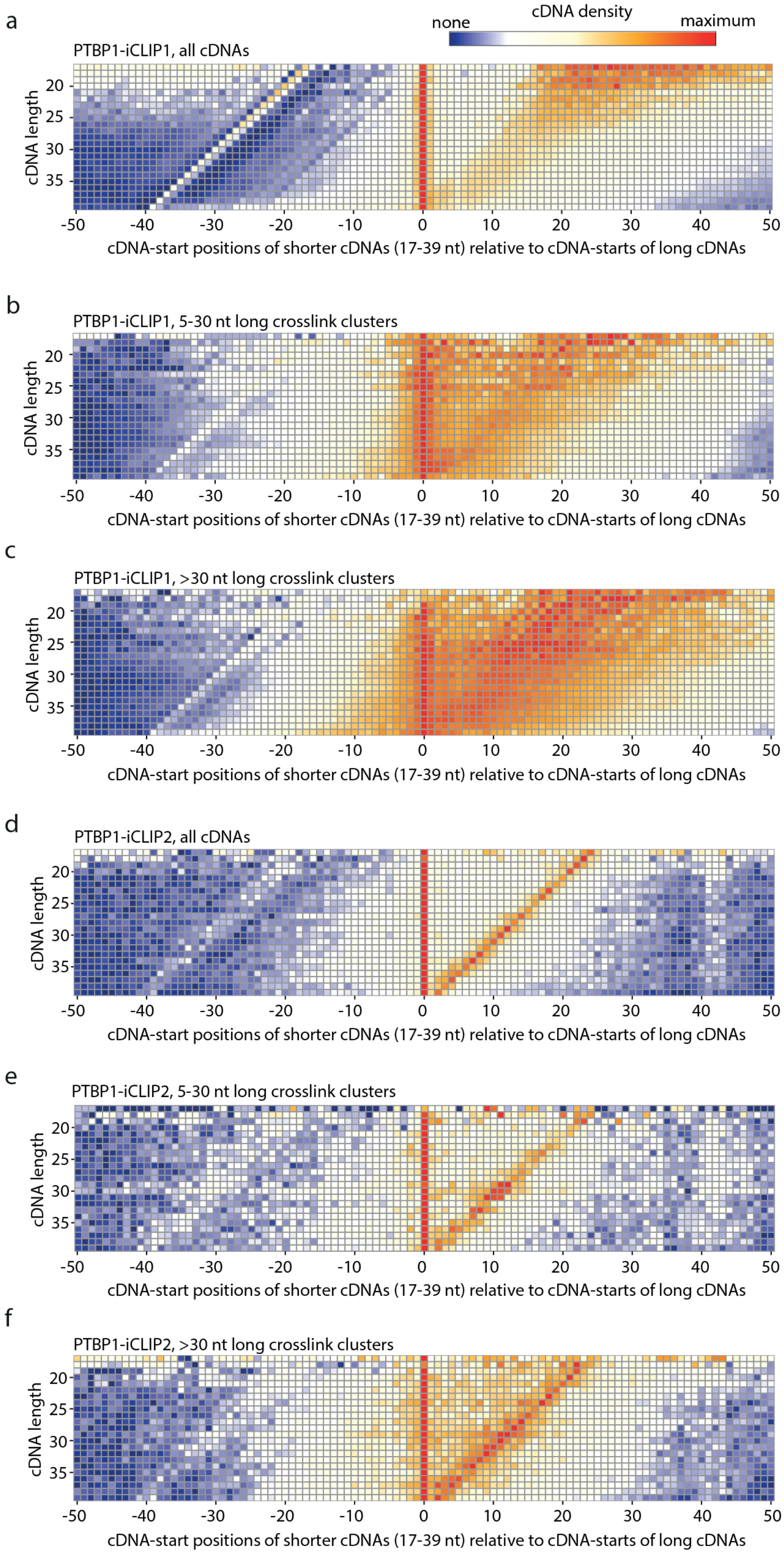


Figure 3.6: Proportion of non-coinciding cDNA-starts differs between PTBP1-iCLIP1 and PTBP1-iCLIP2 experiments.

(continued)

Figure 3.6: a) Heatmap for PTBP1-iCLIP1 generated using the previously developed software iCLIPro [181] to show the relative positioning of cDNA-starts of shorter iCLIP cDNAs (17-39 nt) compared to cDNA-starts of long cDNAs (longer than 39 nt).
b) As in a), but for cDNAs of PTBP1-iCLIP1 that overlap with 5-30 nt long crosslink clusters.
c) As in a), but for cDNAs of PTBP1-iCLIP1 that overlap with >30 nt long crosslink clusters.
d) As in a), but for PTBP1-iCLIP2.
e) As in a), but for cDNAs of PTBP1-iCLIP2 that overlap with 5-30 nt long crosslink clusters.
f) As in a), but for cDNAs of PTBP1-iCLIP2 that overlap with >30 nt long crosslink clusters.

In order to understand these technical differences, I focused on non-coinciding cDNA-starts in PTBP1-iCLIP1 and PTBP1-iCLIP2 experiments, particularly in long clusters (more than 30 nt), where cDNAs with non-coinciding cDNA-starts are the most dominant (Figure 3.6c, f). To ensure that I have high enough coverage for this analysis, I selected cDNAs from the top 1000 enriched clusters. First, I identified the maximum peak of cDNA-start (cDNA-start peak) and cDNA-end (cDNA-end peak) positions within each cluster. Then, I separated cDNAs into five length categories and compared their cDNA-starts and cDNA-ends density around cDNA-start (Figure 3.7a) and cDNA-end (Figure 3.7b) peaks in 50 nt surrounding region. In both experiments cDNA-starts are broadly distributed around cDNA-start peaks (Figure 3.7a, c) but more interesting, the distribution of different cDNA length categories has a much stronger effect on cDNA-start distribution in PTBP1-iCLIP1 (Figure 3.7a) than in PTBP1-iCLIP2 (Figure 3.7c) experiment. These differences of cDNA-start distributions around cDNA-start peaks between these two experiments is even more obvious by looking at measured empirical cumulative distribution (Figure 3.7e, f). Next, I examined distribution of cDNA-ends of different cDNA length categories around cDNA-end peaks. Strikingly, in PTBP1-iCLIP1 cDNA-ends overlaps exactly with cDNA-end peaks (Figure 3.7b), while they are much more evenly distributed in PTBP1-iCLIP2 (Figure 3.7d). The sharp peak of cDNA-ends at cDNA-end peaks in both experiments corresponds to cDNA-start peaks of different cDNA length categories in upstream region (Figure 3.7b, d), while they are much stronger in PTBP1-iCLIP1, where the fold change of cDNA-end is increased by two-fold (Figure 3.7g). Taken together, I conclude that reduced constraints of cDNA-ends decreases the presence of non-coinciding cDNA-starts in PTBP1-iCLIP2.

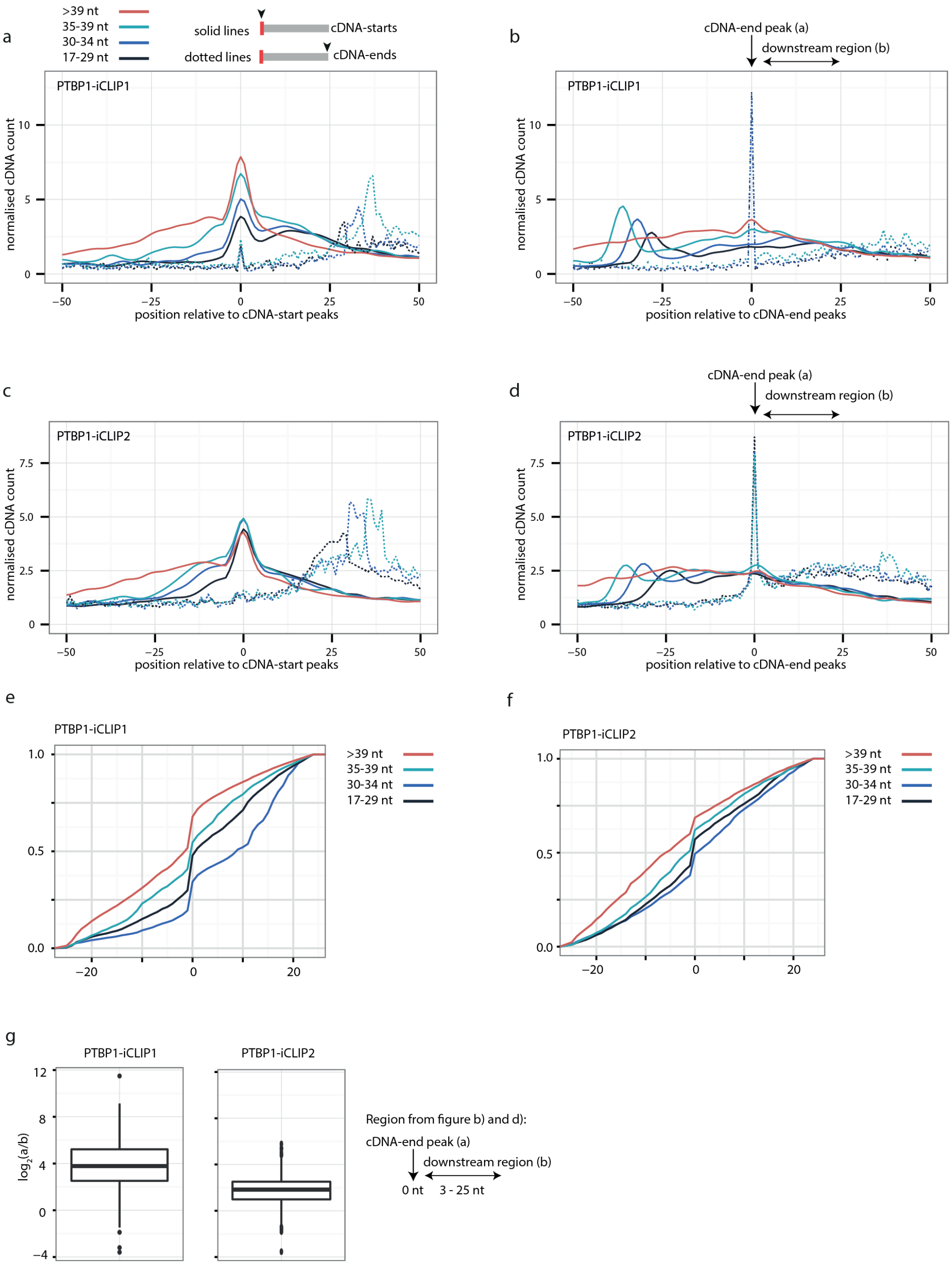


Figure 3.7: Non-coinciding cDNA-starts are a result of constrained cDNA-ends.

a) The cDNA-starts (solid lines) and cDNA-ends (dotted lines) of PTBP1-iCLIP1 are plotted around the cDNA-start peak that was identified within each of the 1000 clusters that have the highest total cDNA crosslink count and are more than 30 nt long. cDNAs are divided into four length categories: 17-29 nt, 30-34 nt, 35-39 nt and >39 nt.

(continued)

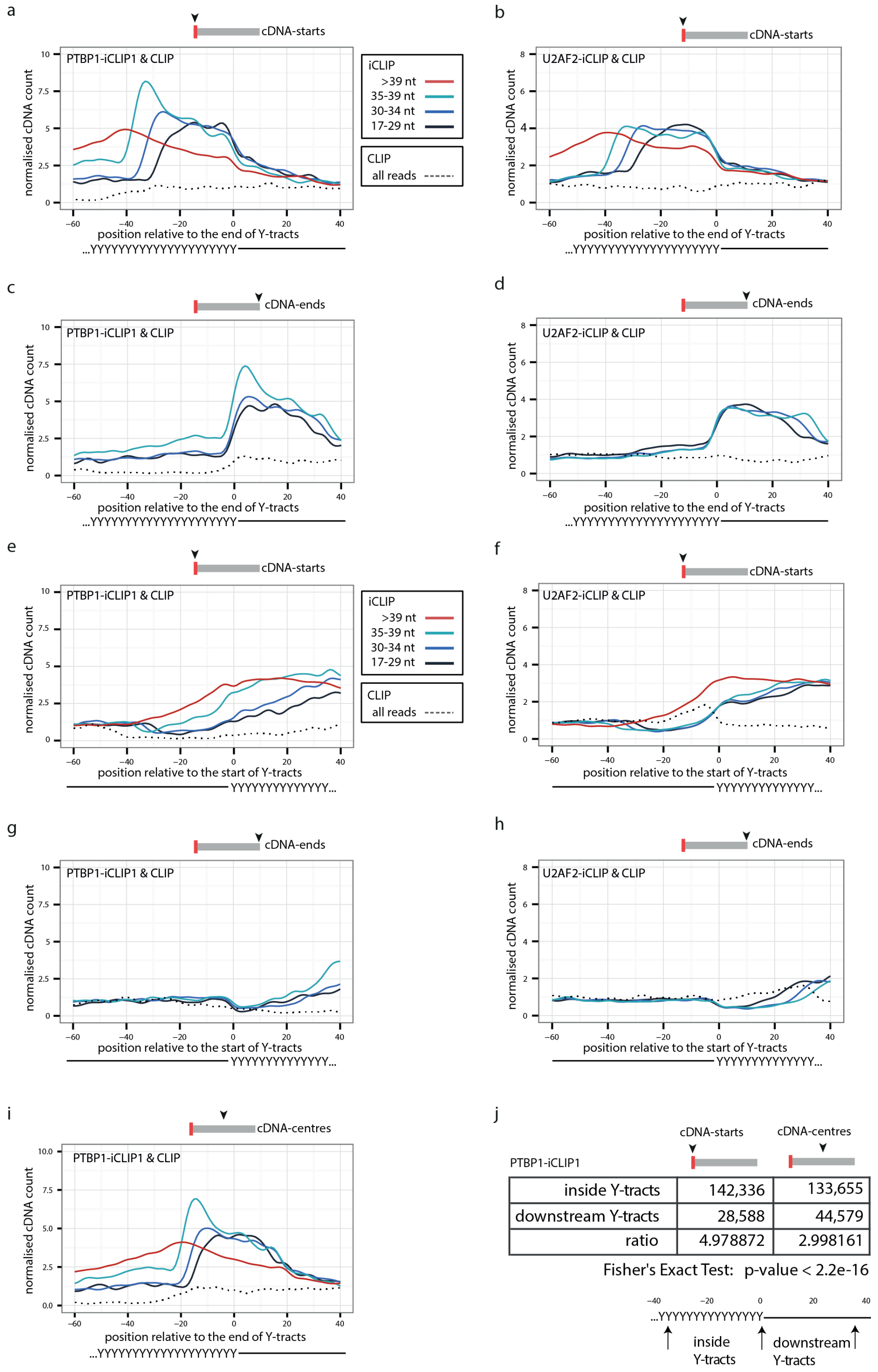
Figure 3.7: b) As in a), but plotted around the cDNA-end peak that was identified within the 30 nt downstream of each of the 1000 clusters that have the highest total cDNA crosslink count and are more than 30 nt long.
c) As in a), but for PTBP1-iCLIP2.
d) As in b), but for PTBP1-iCLIP2.
e) The empirical cumulative distribution from all four length categories in the region between -25 nt and 25 nt around cDNA-start peaks for PTBP1-iCLIP1.
f) as in a) but for PTBP1-iCLIP2.
g) The ratio of cDNA counts (log2) between PTBP1-iCLIP1 and PTBP1-iCLIP2 at the position 0 (overlapping with cDNA-end peak at Figure 3.7b, d) compared to the average count of cDNAs in the region from 5 nt to 25 nt downstream of the cDNA-end peak (marked by horizontal arrow).

3.6 PTBP1 binding sites can be assigned correctly despite non-coinciding cDNA-starts

For PTBP1-iCLIP1 and PTBP1-iCLIP2 I demonstrated that the non-coinciding cDNA-starts are the most dominant within long clusters. Interestingly, the previous study [181] did not detect non-coinciding cDNA-starts for U2AF65 protein with their iCLIPro tool.

To further investigate the phenomenon of non-coinciding cDNA-starts, I asked if their prevalence may depend on the location of cDNAs within transcripts. For this purpose, I examined CLIP and iCLIP data for PTBP1 and U2AF2, two RBPs that bind to polypyrimidine tracts (Y-tracts). To define the coordinates of potential PTBP1 and U2AF65 binding sites independently of iCLIP data, I made use of the Y-tracts that are annotated in human genome as T-rich or TC-rich 'low complexity sequences', and are located at multiple locations within transcripts [50, 49, 211]. While PTBP1 preferentially binds intronic Y-tracts further away from splice sites, U2AF2 primarily binds to Y-tracts at 3' splice sites (see subsection 1.2). Nevertheless, PTBP1 does also bind at some 3' splice sites to repress alternative splicing (see subsection 1.3.1), while U2AF2 also binds to deep intronic regions. In both experiments all cDNA-starts are enriched within the Y-tracts, which these proteins are known to bind (Figure 3.8a, b, e, f). Interestingly, the cDNA-starts of different cDNA length categories do not coincide towards the end of Y-tracks, the short iCLIP cDNAs identify the crosslink sites close to the 3' region of the Y-tracts, while longer cDNAs identify crosslink sites that are located further towards the 5' region (Figure 3.8a, b).

Since I noticed before that the non-coinciding cDNA-starts are the outcome of cDNA-end constraints, I was interested in the cDNA-end distribution around Y-tracks. As expected, in both PTBP1-iCLIP1 and U2AF54-iCLIP experiments the cDNA-ends are constrained to downstream positions of the Y-tracks (Figure 3.8c, d). Importantly, the cDNA-ends of PTBP1-iCLIP1 and U2AF65-iCLIP cDNAs of all length categories align at the end of the Y-tracts, demonstrating that RNase cleavage sites are constrained to positions just downstream of Y-tracks, where RNAse cleavage seems to be inefficient within Y-tracks (Figure 3.8c, d). In order to understand how the choice of method may affect the assignment of binding sites, I also examined the positioning of cDNA-starts and cDNA-ends identified by PTBP1-CLIP and U2AF65-CLIP experiments (see Table 3.1). Surprisingly, the enrichment of CLIP cDNAs is extremely low across Y-tracts for both experiments (Figure 3.8a-i). This agrees with the findings of a previous study, which showed that CLIP of NOVA proteins was not well suited for identifying the YCAY-tracts that represent high-affinity binding sites of NOVA [120]. As shown before, the presence of non-coinciding cDNA-starts in iCLIP experiments reflects constrained positions of cDNA-ends. To avoid the artefacts that could be caused by the RNase cleavage constraints, a broad range of cDNA sizes is required to identify crosslink sites across the full binding sites. In particular, the long cDNAs are most important to overcome these constraints, since they can truncate at crosslink sites that are located far from the site of RNase cleavage.



PTBP1-iCLIP1	cDNA-starts	cDNA-centres
inside Y-tracts	142,336	133,655
downstream Y-tracts	28,588	44,579
ratio	4.978872	2.998161

Fisher's Exact Test: p-value < 2.2e-16

Figure 3.8: Non-coinciding cDNA-starts are required to map the crosslink sites within Y-tracts.

(continued)

Figure 3.8: a) The cDNA-starts of PTBP1-iCLIP1 and CLIP experiments are plotted around the ends of >35 nt Y-tracts that are annotated as T-rich or TC-rich low-complexity sequence in the human genome (hg19). cDNAs of PTBP1-iCLIP1 are divided into four length categories: 17-29 nt, 30-34 nt, 35-39 nt and >39 nt.
b) Same as a), but using U2AF2-iCLIP and CLIP cDNAs.
c) Same as a), but showing the positions of cDNA-ends.
d) Same as b), but showing the positions of cDNA-ends.
e) The cDNA-starts of PTBP1-iCLIP1 and CLIP experiments are plotted around the starts of >35 nt Y-tracts that are annotated as T-rich or TC-rich low-complexity sequence in the human genome (hg19). cDNAs of PTBP1-iCLIP1 are divided into four length categories: 17-29 nt, 30-34 nt, 35-39 nt, and >39 nt.
f) Same as e), but using U2AF2-iCLIP and CLIP cDNAs.
g) Same as e), but showing the positions of cDNA-ends.
h) Same as b), but showing the positions of cDNA-ends.
i) Same as a), but for the cDNA-centres around the ends of Y-tracks.
j) Ratio of cDNA-starts and cDNA-centres that are inside of Y-tracks compared to the downstream region (schematic description at the bottom). Statistical test for cDNAstarts and cDNA-centres enrichment in Y-tracks region was done by Fisher's Exact Test with p-value <2.2e-16.

To examine the effect of non-coinciding cDNA-starts in assigned binding sites, I examined PTBP1 motif enrichment across crosslink clusters that were defined by the iCount peak finding tool (see Methods 2.4.1) for PTBP1-iCLIP1 and PTBP1-iCLIP2 experiments. The PTBP1 motifs were identified from PTBP1-iCLIP2 experiment by selecting the most highly enriched tetramers around cDNA-starts compared to the downstream control regions (see Methods 2.5.4). To visualise the enrichment of selected tetramers, I grouped clusters of similar lengths and plot them as a heatmap with the surrounding region for iCLIP, eCLIP or irCLIP experiments (Figure 3.9). In all three experiments, the enrichment correctly overlaps crosslink clusters, regardless of which type or which variant of library preparation protocol was used. Taken together, I conclude that the use of cDNA-starts is appropriate for the computational analysis of data produced by iCLIP or any related method that is capable of efficiently amplifying truncated cDNAs.

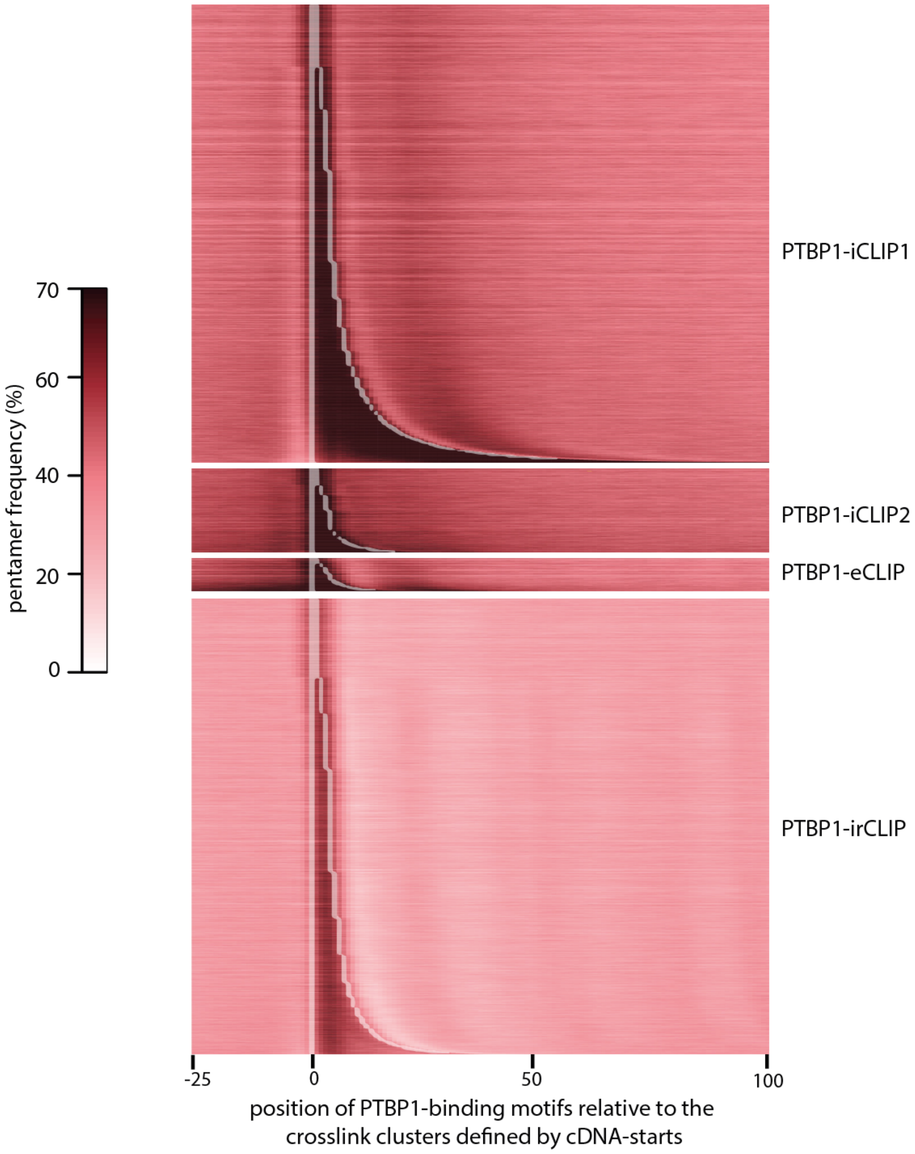


Figure 3.9: PTBP1-binding motif enrichment across PTBP1 crosslink clusters.

Heatmap showing the coverage of PTBP1-binding motifs at the PTBP1-iCLIP1, PTBP1-iCLIP2, PTBP1-eCLIP or PTBP1-irCLIP crosslink clusters that were defined with a 3 nt clustering window. Each row shows the average coverage for 300 clusters of similar length, sorted from shortest to longest clusters. The white line marks the nucleotide preceding the start and the nucleotide following the median end of all clusters that were combined in each row. A colour key for the coverage per nucleotide of the PTBP1-binding motifs is shown on the right.

3.7 Adenosine enrichment at RNase I cleavage sites in PTBP1 CLIP and iCLIP

The cDNA-ends of PTBP1 and U2AF65 CLIP and iCLIP experiments of all cDNA length categories align at the end of the Y-tracts, demonstrating that RNase cleavage sites are constrained to positions just downstream of their binding sites (Figure 3.8c, d, e, f). In order to understand how the choice of method may affect the assignment of binding sites, I also examined the positioning of cDNAs identified by PTBP1-CLIP and compare it to PTBP1-iCLIP1. PTBP1-CLIP cDNA-ends are slightly enriched in the last portion of Y-tracts (grey dotted line in Figure 3.8c). This agrees with the findings of a previous study, which showed that CLIP of NOVA proteins was not well suited for identifying the YCAY-tracts that represent high-affinity binding sites of NOVA [120]. To gain additional insight into the different classes of cDNAs, I further examined the enrichment of Y-rich motifs. Y-rich motifs are most enriched at the beginning of PTBP1-iCLIP1 cDNAs, consistent with these cDNAs truncating at crosslink sites (Figure 3.10a). In contrast, Y-rich motifs are most enriched in the centre of CLIP cDNAs, consistently with these being readthrough cDNAs that contain crosslink sites within them (Figure 3.10c). Y-rich motifs are also enriched within the iCLIP cDNAs, which is expected given that cDNAs end downstream of the Y-rich binding sites (Figure 3.10a). To understand the reasons for RNase cleavage downstream of PTBP1 binding sites, I examined the sequence composition around the cDNA-ends that are shorter than 40 nt. Interestingly, there is a high enrichment of adenosine at the RNase I cleavage sites (cDNA-ends) in PTBP1-iCLIP1 and PTBP1-iCLIP3 but not in PTBP1-iCLIP2, where different RNase conditions were used (Figure 3.10e, b, f). An even stronger enrichment of the ARRW motif is seen at cDNA-ends in PTBP1-CLIP data (Figure 3.10d). Whereas AA and ARRW motifs are enriched at the cDNA-ends of PTBP1-iCLIP1 and PTBP1-CLIP cDNAs, respectively, the Y-rich motifs recognised by PTBP1 are depleted from the cDNA-ends (Figure 3.10a, c). This suggests that lack of purines within the PTBP1 binding sites could prevent the different RNases from cleaving.

In conclusion, the RNase cleavage sites identified by the cDNA-ends of

PTBP1-CLIP and PTBP1-iCLIP1 cDNAs generally locate to a narrow region immediately downstream of the PTBP1 binding sites. To avoid artefacts that could be caused by these RNase cleavage constraints a broad range of cDNA sizes is required to identify crosslink sites across binding sites. In particular, the long cDNAs are most important to overcome these constraints, since they can be truncated at crosslink sites that are located far from the site of RNase cleavage.

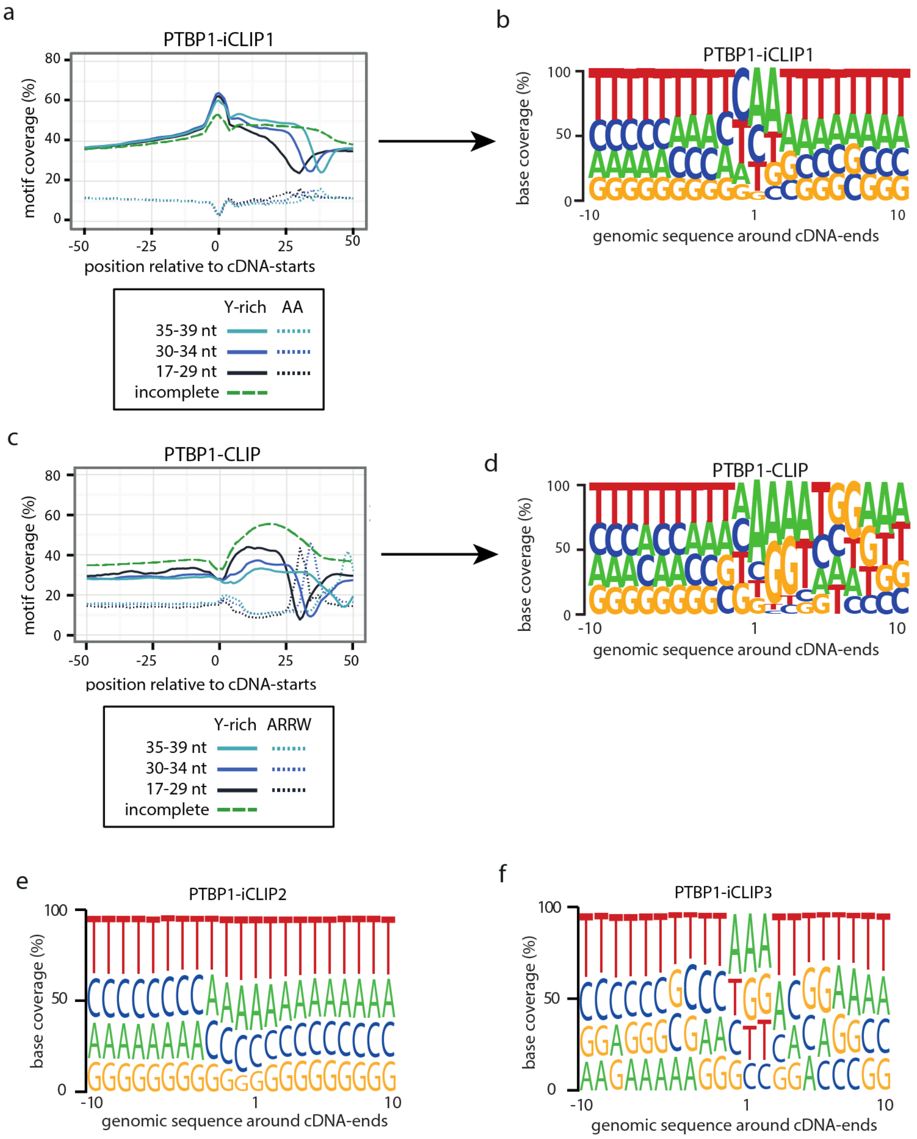


Figure 3.10: Adenosine enrichment at RNase I cleavage sites in PTBP1-CLIP and PTBP1-iCLIP1.

a) The percentage of cDNAs overlapping with Y-rich motifs at each nucleotide around cDNA-starts is plotted for different cDNA categories from published PTBP1 iCLIP data [185]: long cDNAs (dashed green line) as well as short (less than 40 nt) cDNAs that were divided into three length categories (35-39 nt , 30-34 nt, 17-29 nt; different shades of blue). In addition, the dotted lines show the positional frequency of adenosine dinucleotides (AA), demonstrating that their enrichment is mutually exclusive with Y-rich motifs in the region of cDNAs ends.
b) The composition of genomic nucleotides around the ends of short cDNAs from the published PTBP1-iCLIP1 data.

(continued)

Figure 3.10: c) The percentage of cDNAs overlapping with Y-rich motifs at each nucleotide around cDNA-starts is plotted for different cDNA categories from published PTBP1-CLIP data [207]. Length categories and colour labelling is as in c). In addition, the dotted lines show the positional frequency of ARRW motifs, demonstrating that their enrichment is mutually exclusive with Y-rich motifs in the region of cDNAs ends.
d) Same as b) but for PTBP1-CLIP data.
e) Same as b) but for PTBP1-iCLIP2 data.
f) Same as b) but for PTBP1-iCLIP3 data.

3.8 Efficient RNase I-mediated RNA fragmentation minimises the cDNA-end constraints

To further investigate the phenomenon of cDNA-end constraints, I decided to compare PTBP1-iCLIP1 and PTBP1-iCLIP2 experiments in more details. The cDNA-end positions correspond to the position of the RNA fragments that were cleaved by the RNase treatment (Figure 3.1). To ensure that RNase 1 is the primary cause of RNA framgentation, the RNase treatment conditions were fully optimised in the PTBP1-iCLIP2 experiment. The sharp cDNA-end peak in PTBP1-iCLIP1 and PTBP1-iCLIP2 (Figure 3.7b, d) indicates that the RNA fragmentation could be caused by some other factors. To further investigate possibilities of other factors, I wanted to know if the prevalence of cDNA-end constraints depends on the location of cDNAs within transcripts, for example relative to intron-exon junctions, since these are subject to endogenous RNA cleavage by the spliceosome. For this purpose, I examined CLIP and iCLIP data of PTBP1 and U2AF2, two RBPs that bind to polypyrimidine tracts (Y-tracts) at multiple locations within transcripts [50, 49, 211]. While PTBP1 preferentially binds intronic Y-tracts further away from splice sites, U2AF2 primarily binds to Y-tracts at 3’ splice sites. Nevertheless, PTBP1 does also bind at some 3’ splice sites to repress alternative splicing, while U2AF2 also binds to deep intronic regions. As observed previously [181], the cDNA-starts of all cDNA length classes mainly coincide at the 3’ splice sites for U2AF2-iCLIP, while they do not coincide well for PTBP1-iCLIP1 (Figure 3.11a, d). Strikingly, the analysis of cDNA-ends uncovered an unusually sharp peak at the last intronic nucleotide, which is more pronounced for the PTBP1-iCLIP1 data (Figure 3.6d), but is also visible for the U2AF2-iCLIP data, even though most cDNA-ends are in the exonic sequence (Figure 3.11a, b). Here, the cDNA-ends of all cDNA length categories directly overlap, indicating that cDNA-ends are often constrained to the last position of an intron. To test whether constrained cDNA-ends in the intron are indeed the cause of non-coincidng cDNA-starts, I separated the cDNAs in U2AF2-iCLIP into two classes depending on the position of their cDNA-end. When the cDNA-end is present in the intron, the cDNA-starts are non-coinciding (Figure 3.11b), while the cDNAs ending in the exon have fully coinciding cDNA-starts (Figure 3.11c). Thus, prevalence of non-coinciding cDNA-starts is not a generic feature of a specific iCLIP dataset, but instead it depends on the position in transcripts. Thus, the cDNAs in U2AF2-iCLIP and PTBP1-iCLIP1 have similar features; both proteins contain a mixture of cDNAs with coinciding and non-coinciding starts, and the proportion of those with non-coinciding cDNA-starts is higher at 3' splice sites in PTBP1 due to increased proportion of cDNAs ending within the intron.

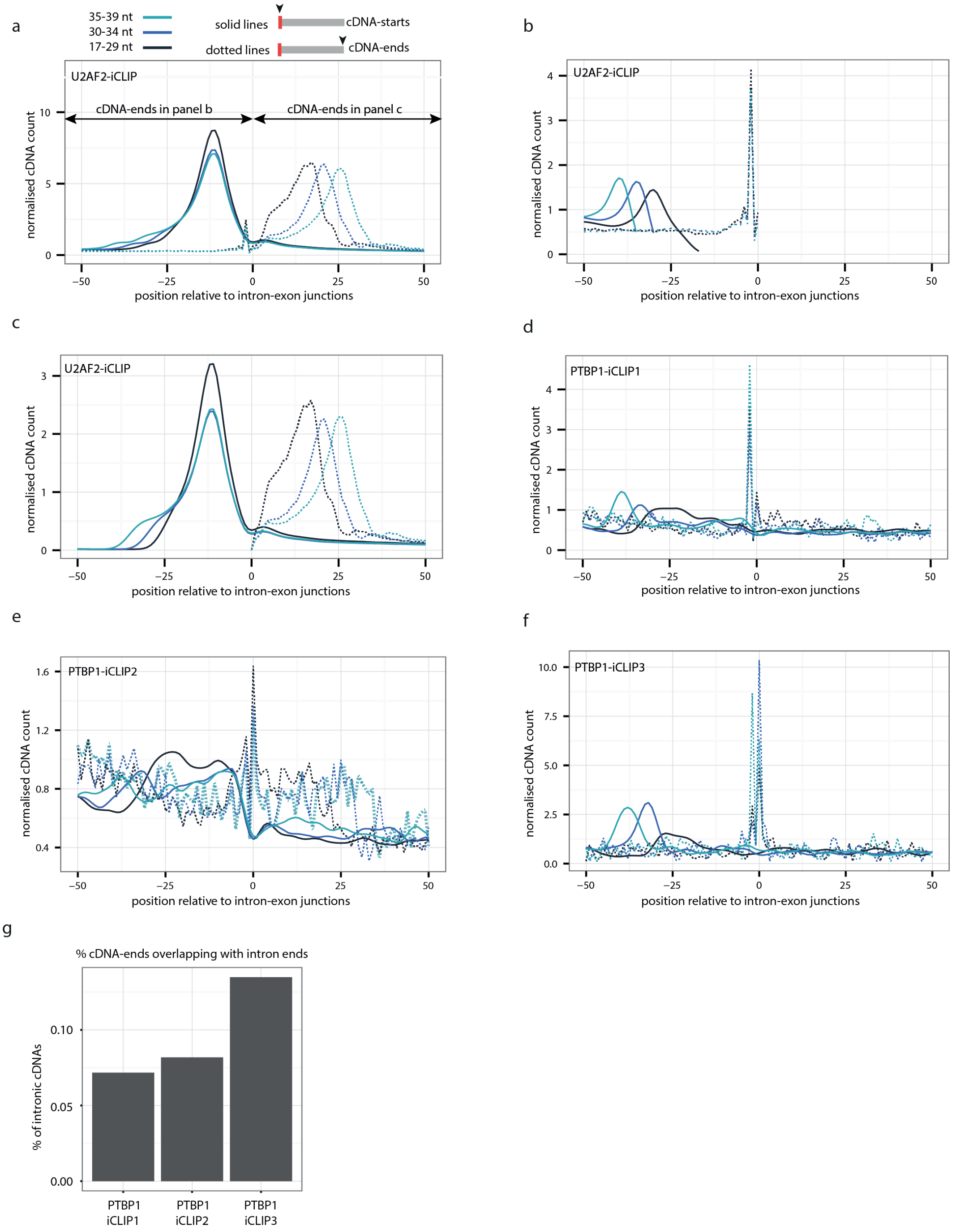


Figure 3.11: Constrained cDNA-ends affect the cDNA-starts at 3' splice sites.

a) The cDNA-starts (solid lines) and cDNA-ends (dotted lines) of U2AF2-iCLIP are plotted around intron-exon junctions (position 0 being the first nucleotide of the exon). cDNAs are divided into three length categories: 17-29 nt, 30-34 nt and 35-39 nt; the distribution of all cDNAs together is shown in grey.
b) Same as a), but using only cDNAs that end in the intron.
c) Same as a), but using only cDNAs that end in the exon.

(continued)

Figure 3.11: d) Same as a), but showing PTBP1-iCLIP1 cDNA-starts (full lines) and cDNA-ends (dotted lines).
e) Same as a), but showing PTBP1-iCLIP2 (using 4SU and optimised RNase conditions) cDNA-starts (full lines) and cDNA-ends (dotted lines).
f) Same as a), but showing PTBP1-iCLIP3 (omitting 3' dephosphorylation) cDNA-starts (full lines) and cDNA-ends (dotted lines).
g) Proportions of cDNAs that map to introns which contain cDNA-ends at positions overlapping the last two nucleotides of introns. PTBP1-iCLIP1 and PTBP1-iCLIP2 are compared to PTBP1-iCLIP3 iCLIP, which was performed without using PNK to dephosphorylate RNAs in step 2. This enriches for RNAs that contain a 3' OH, which can occur when they are cleaved at their 3' end independently of RNase I, such as the 3' ends of intron lariats.

I speculated that the cDNAs that end at the last intronic nucleotide were generated from RNA fragments that originated from the 3' end of intronic lariats, which are produced when introns are spliced out from pre-mRNAs. The stronger peak of cDNA-ends at the last intronic nucleotide would suggest that PTBP1 more commonly remains bound to the intron lariat, while U2AF2 is released before splicing is completed. To test this hypothesis another PTBP1 iCLIP experiment (PTBP1-iCLIP3) was performed by Ule lab, with the exploited fact that intron lariats lack a phosphate at their 3' end, and therefore no 3' dephosphorylation would be needed in the iCLIP protocol (Figure 3.1 - step 2). For that reason, the PTBP1-iCLIP3 was prepared with omitting dephosphorylation from step 2 and continuing directly to the ligation of the 3' adapter in step 3 (Figure 3.1) and therefore only those RNA fragments cleaved by other means were amplified in PTBP1-iCLIP3. Notably, both in PTBP1-iCLIP1 and PTBP1-iCLIP3, the cDNA-ends at 3' splice sites are strongly constrained at the introns end, while these constraints are minor in PTBP1-iCLIP2 (Figure 3.11d-f). Thus, non-coinciding cDNA-starts predominate at 3' splice sites in PTBP1-iCLIP1 and PTBP1-iCLIP3, while in PTBP1-iCLIP2 most cDNA-starts coincide in the region of 20 nt to 5 nt upstream of the intron-exon junction. This suggests that the RNAs overlapping with the 3' splice sites were fragmented by spliceosome-mediated cleavage in PTBP1-iCLIP1 and PTBP1-iCLIP3 and by RNase I in PTBP1-iCLIP2 and in U2AF2-iCLIP. It is this difference that appears to explain the variation in the prevalence of non-coinciding cDNA-starts at

3' splice sites.

To further compare the characteristics of cDNA-ends between the PTBP1 iCLIP experiments, I visualised the sequence composition of cDNA-ends. As mentioned before, I observed almost no sequence biases at cDNA-ends in PTBP1-iCLIP2 only (Figure 3.10e). This could be explained by the lack of sequence specificity of RNase I, since it was performed with the optimal RNase conditions. In contrast, a preference for adenosines was observed at the cDNA-ends in PTBP1-iCLIP1 and PTBP1-iCLIP3, suggesting that this preference results from an RNase I-independent fragmentation of RNAs (Figure 3.10b, d, e, f). Spliceosome-mediated RNA cleavage contributes to only about 0.1% of these fragments (Figure 3.11g) and therefore the primary cause of RNase I-independent fragmentation remains to be identified. Nevertheless, it is clear to conclude that the efficient use of RNase I avoids the constraints at cDNA-ends in iCLIP and this decreases the incidence of non-coinciding cDNA-starts.

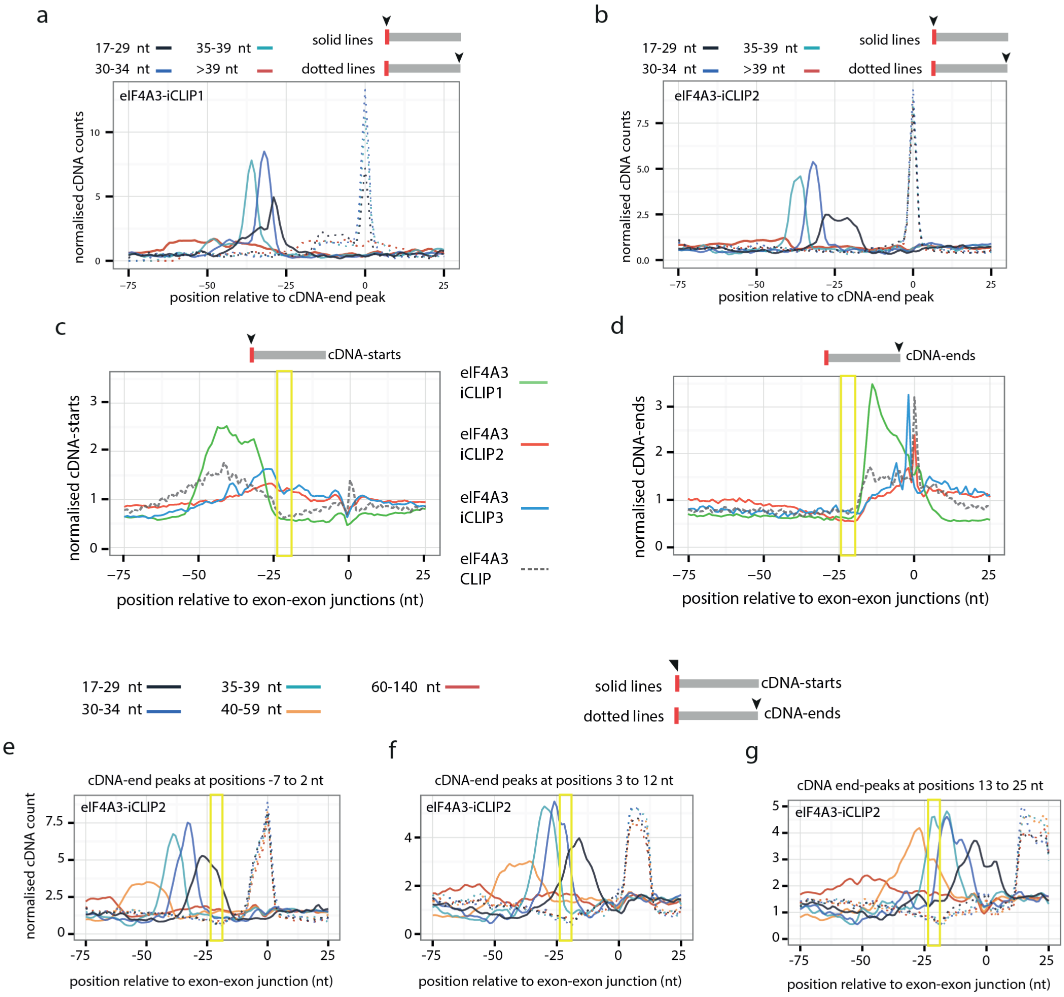


Figure 3.12: Constrained cDNA-ends in eIF4A3 iCLIP.

a) The distribution of cDNA-starts (solid lines) and ends (dotted lines) relative to the cDNA-end peaks that were identified at each exon-exon junction in eIF4A3-iCLIP1. cDNAs are divided into four length categories: 17-29 nt, 30-34 nt, 35-39 nt and more than 39 nt.
b) Same as a), but for eIF4A3-iCLIP2.
c) The cDNA-starts of eIF4A3 iCLIP and CLIP experiments are plotted around the 1000 exon-exon junctions with the highest number of cDNAs.
d) Same as c), but showing cDNA-ends.
e-g) Distribution of cDNA-starts (solid lines) and ends (dotted lines) in eIF4A3-iCLIP2 relative to exon-exon junctions. Junctions were divided into three different classes according to the position cDNA-end peaks at: -7 to 2 nt e), 3 to 12 nt f), or 13 to 25 nt g) relative to exon-exon junctions. cDNA length categories and labelling as shown on top.

Next, I examined three different iCLIP and one CLIP experiments produced for eIF4A3 protein (see Table 3.1). Surprisingly, the distribution of cDNA-starts varies considerably between eIF4A3 experiments (Figure 3.12c). As observed by the previous study, the cDNA-starts in eIF4A3-iCLIP1 are shifted to positions upstream of the expected EJC-binding region (yellow rectangle in Figure 3.12c) [181]. However, there is an overlap between eIF4A3-iCLIP2 and eIF4A3-iCLIP3 in the expected binding region of EJC (yellow rectangle in Figure 3.12c), where the cDNA-starts in the eIF4A3-CLIP experiment are shifted upstream from expected region, which agrees with the likely prevalence of truncated cDNAs in iCLIP and readthrough cDNAs in CLIP. To see if the non-coinciding cDNA-starts are also influenced by cDNA-end composition, I examined cDNA-ends across all exon-exon junctions. Interestingly, the distribution of cDNA-ends is highly enriched in a broad downstream region from exon-exon junction (-17 nt to 0 nt relative to exon-exon junction) only in eIF4A3-iCLIP1 (Figure 3.12d). Depending on the position of cDNA-ends, different lengths of cDNAs identify crosslink sites within the expected region. This is also evident if assessing cDNAs of multiple length categories at distinct classes of exon-exon junctions with different cDNA-end peak positions (Figure 3.12e-g). Even though the cDNA-starts are fully defined by the constrained position of cDNA-ends, at least one length category in eIF4A3-iCLIP2 has its cDNA-starts in the expected EJC-binding region in each class of junctions (yellow rectangle in Figure 3.12e-g).

To further understand how cDNA-ends are constrained, I grouped all exon-exon junctions that had the same distance between the cDNA-end peak and the junction. Then I classified all exon-exon junctions by the distance between the maximum cDNA-end peak and the junction position. Strikingly, this confirmed that each junction has a single dominant position of cDNA-ends, which differs from junction to junction (Figure 3.13a, c, e). In eIF4A3-iCLIP2, I observe a second peak of cDNA-ends precisely at the end of the exon (Figure 3.13c), which probably reflects the co-splicing deposition of eIF4A3 on the splicing intermediate, in agreement with previous reports [212, 213, 214]. Moreover, the cDNA-end peak in eIF4A3-iCLIP2 and less in eIF4A3-iCLIP3 coincides with a strong decrease in pairing probability (Figure 3.13b, d, f). This is consistent with the preference of RNase I to cleave single-stranded RNA. Absence of these features from eIF4A3-iCLIP1 suggests a difference in the RNase conditions between the experiments, which remains to be fully understood. All three eIF4A3 iCLIP experiments have enriched adenosine at the position following the cDNA-end peak, but with stronger enrichment in eIF4A3-iCLIP2 and eIF4A3-iCLIP3 (position 1 in Figure 3.13g, h, i). Moreover, a third eIF4A3-iCLIP3 experiment that was conducted in the Le Hir laboratory (eIF4A3-iCLIP3 in Table 3.1), had a similar sequence enrichment of adenosine after the cDNA-end peak as eIF4A3-iCLIP2 (Figure 3.13i). This preference for adenosine after the cleavage site might be an indication that RNase I has a nucleotide bias. Interestingly, the sequence signature at cDNA-end peaks is different in the published eIF4A3-CLIP data [201], which used RNase T1 rather than RNase I, which is used in iCLIP protocol (Figure 3.13j). It is known that RNase T1 preferentially cuts after guanosine, and guanosine is indeed strongly enriched at the position preceding the cDNA-end peaks in this dataset (position 0 in Figure 3.13j). This suggests that the preference of the corresponding RNase for specific RNA structure and sequence motifs explains why cDNA-ends favour a specific position around exon-exon junctions. In conclusion, I find that the non-coinciding cDNA-starts in eIF3A3 iCLIP datasets result from the constrained cDNA-ends, arising from technical features such as RNase sequence and structure preferences.

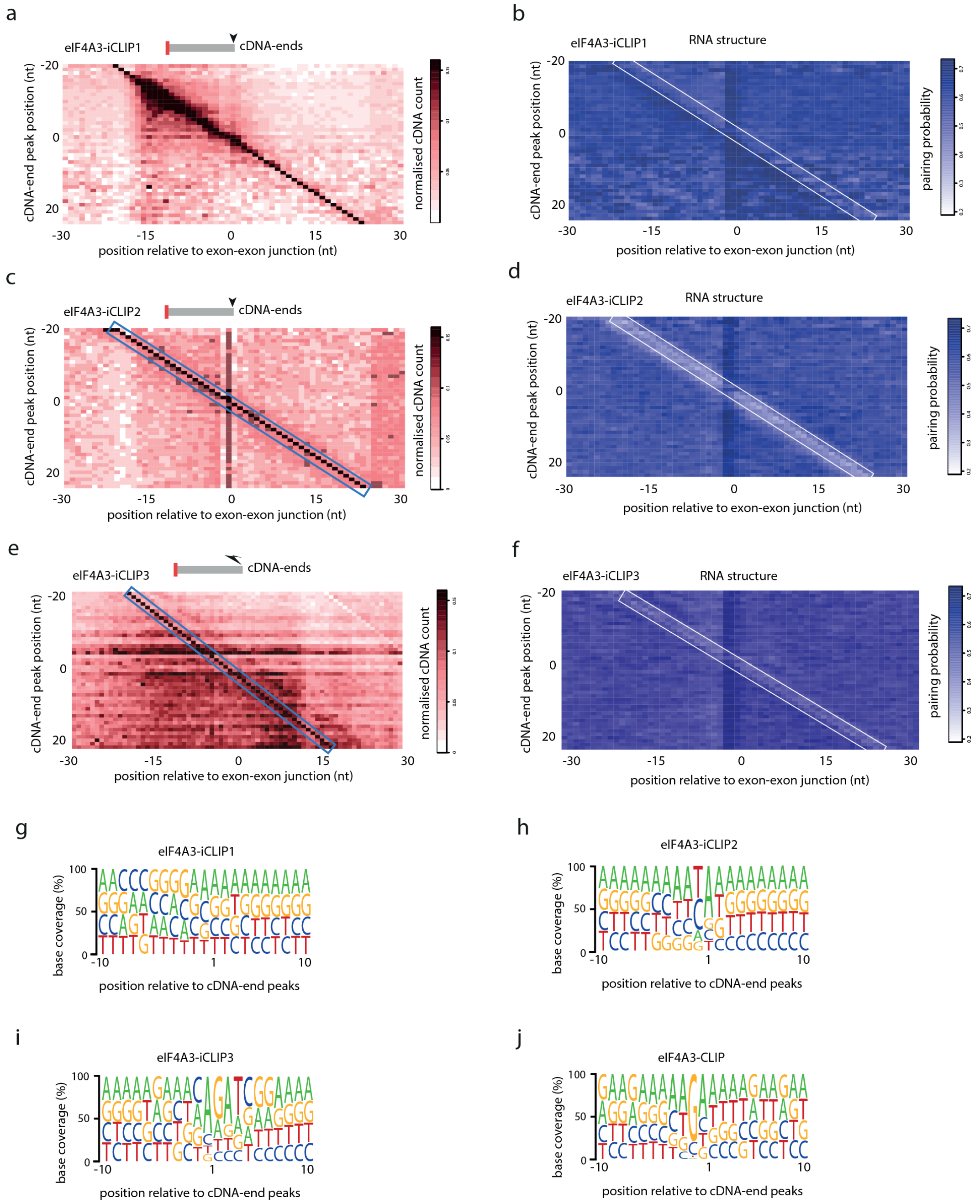


Figure 3.13: Affect of sequence and structure constraints at cDNA-ends in eIF4A3 iCLIP and CLIP.

a) Heatmap showing the position of cDNA-ends around exon-exon junctions in eIF4A3-iCLIP1. Junctions are sorted according to their cDNA-end peak position. Each row shows the average of cDNA counts at all junctions with a cDNA-end peak at the indicated position. The values are normalised against the maximum value across all rows.

(continued)

Figure 3.13: b) Heatmap of summarised pairing probability around exon-exon junctions in eIF4A3-iCLIP1. Junctions are sorted according to their cDNA-end peak position. Each row shows the average pairing probability of all junctions with a cDNA-end peak at the indicated position. The positions of cDNAs end peak coincide with decreased pairing probability, as indicated by the arrow.
c) Same as a), but for eIF4A3-iCLIP2.
d) Same as b), but for eIF4A3-iCLIP2.
e) Same as a), but for eIF4A3-iCLIP3.
f) Same as b), but for eIF4A3-iCLIP3.
g) Genomic nucleotide composition around cDNA-end peaks in eIF4A3-iCLIP1.
h) Same as g), but for eIF4A3-iCLIP2.
i) Same as g), but for eIF4A3-iCLIP3.
j) Same as g), but for eIF4A3-CLIP.

To understand the constraints at cDNA-ends in more detail, I examined the exon-exon junctions with highest coverage of cDNAs in greater detail for eIF4A3-iCLIP1 and eIF4A3-iCLIP3, where I first identified the maximum cDNA-end peaks (Figure 3.14a, b) in a similar way as I did for PTBP1-iCLIP1 and PTBP1-iCLIP2 (see Methods 2.4.5). For this purpose, I focused on the 1000 junctions with the highest cDNA count to minimise the noise of genome wide analysis. This demonstrates that the cDNA-ends are largely restricted to a single position in the eIF4A3-iCLIP3 experiment, while they are more variable in eIF4A3-iCLIP1 (Figure 3.14a, b). Then, I selected 3 individual examples of eIF4A3-iCLIP1 and eIF4A3-iCLIP3 from the top 15 exon-exon junctions with the highest cDNA count, to see the constraints at cDNA-ends at the level of individual exon-exon junction. As a result, the cDNA-starts often coincide in eIF4A3-iCLIP1, but are fully non-coinciding in eIF4A3-iCLIP3 (Figure 3.14c-h). This again demonstrates that the cDNA-end constraints are the primary cause of non-coinciding cDNA-starts in iCLIP. These constraints therefore need to be considered when interpreting the position of binding sites assigned by iCLIP and related methods.

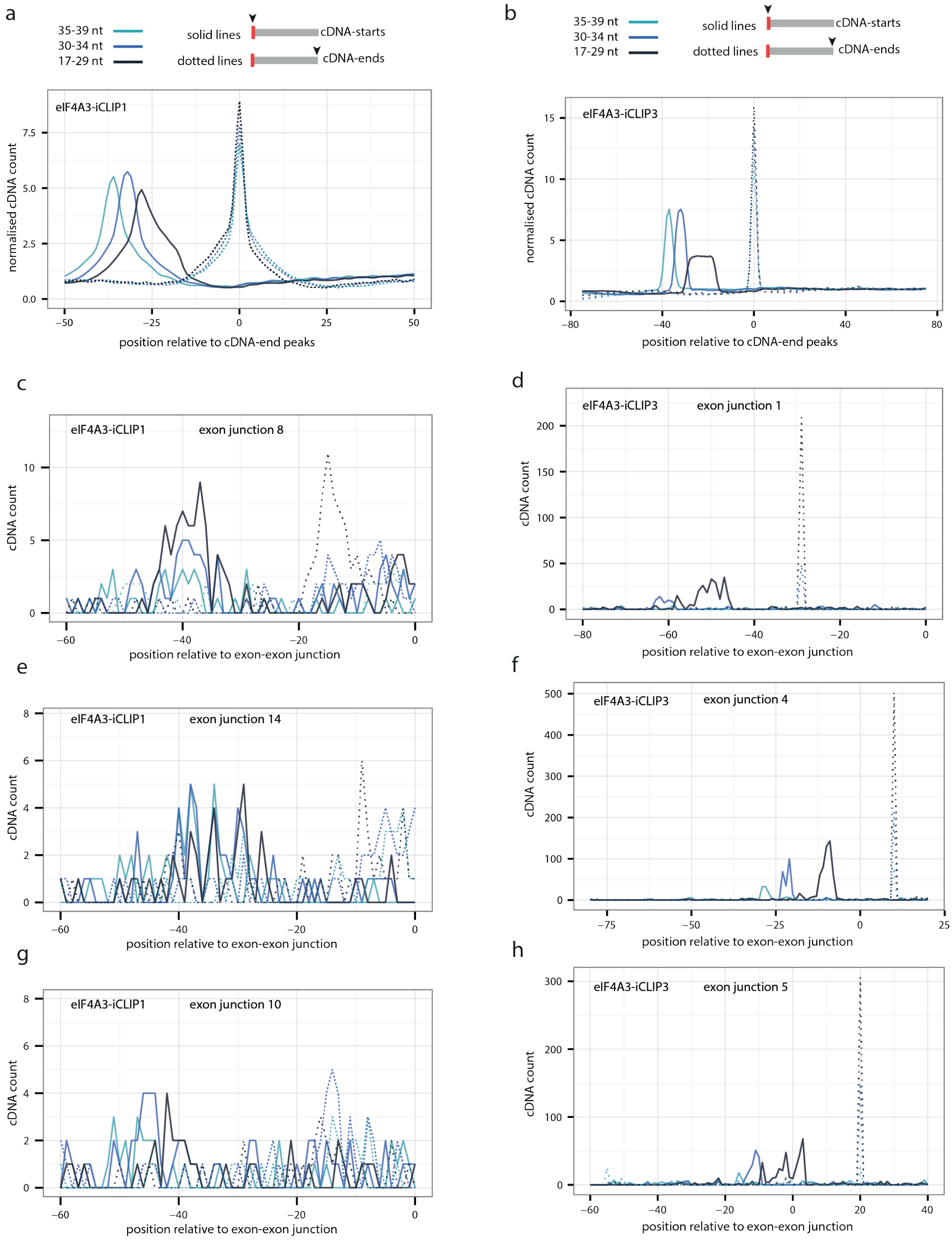


Figure 3.14: The impact of cDNA-end constraints on cDNA-starts in eIF4A3 iCLIP.

a) The distribution of cDNA-starts (solid lines) and ends (dotted lines) relative to the cDNA-end peaks that were identified at top 1,000 exon-exon junctions contain the highest number of cDNAs in eIF4A3-iCLIP1. cDNAs are divided into three length categories: 17-29 nt, 30-34 nt, and 35-39 nt. b) Same as a), but for eIF4A3-iCLIP3.

c) Same as a), but for the junction that ranks 8th by the number of cDNAs in eIF4A3-iCLIP1.

(continued)

Figure 3.14: d) Same as b), but for the junction that ranks 1st by the number of cDNAs in eIF4A3-iCLIP3.
e) Same as a), but for the junction that ranks 14th by the number of cDNAs in eIF4A3-iCLIP1.
f) Same as b), but for the junction that ranks 4th by the number of cDNAs in eIF4A3-iCLIP3.
g) Same as a), but for the junction that ranks 10th by the number of cDNAs in eIF4A3-iCLIP1.
h) Same as b), but for the junction that ranks 5th by the number of cDNAs in eIF4A3-iCLIP3.

3.9 A broad range of cDNA lengths compensates for the constrained cDNA-ends

To understand how the cDNA-end constraints effect the cDNA-start positions in eIF4A3, I grouped all exon-exon junctions that had the same distance between the maximum cDNA-end peak and the junction position, focused on the 1000 junctions with the highest cDNA count. Next, I visualised the density of cDNA-start positions for all the groups in eIF4A3-iCLIP1, eIF4A3-iCLIP2 and eIF4A3-iCLIP3 (Figure 3.15a, b, c). As expected, the cDNA-start enrichment has a strong effect on the cDNA-end composition (marked as blue rectangle in Figure 3.15a, b, c). Particularly, this effect is even stronger for eIF4A3-iCLIP1, which can be seen by the narrow enrichment of cDNA-starts within the same distance relative to cDNA-end peaks (Figure 3.15a). To investigate the difference between eIF4A3 iCLIP libraries, I analysed the cDNA length distribution for the examined experiments. Notably, the analysis of cDNA length distribution shows that of the examined experiments, eIF4A3-iCLIP1 has the largest proportion (58%) of cDNAs that are shorter than 39 nts, in comparison to eIF4A3-iCLIP3 that has only 36% and 46% in eIF4A3-iCLIP2 (Figure 3.16a). In addition to the large proportion of short cDNAs in eIF4A3-iCLIP1, there is also a dominant range of cDNAs that are between 27 and 38 nt long (Figure 3.16a) in comparison to other libraries with more even size distribution (Figure 3.16b, c). The narrow range of cDNA-starts in eIF4A3-iCLIP1 rarely identify crosslink sites within the expected EJC-binding region (marked by the yellow rectangle in Figure 3.15a).

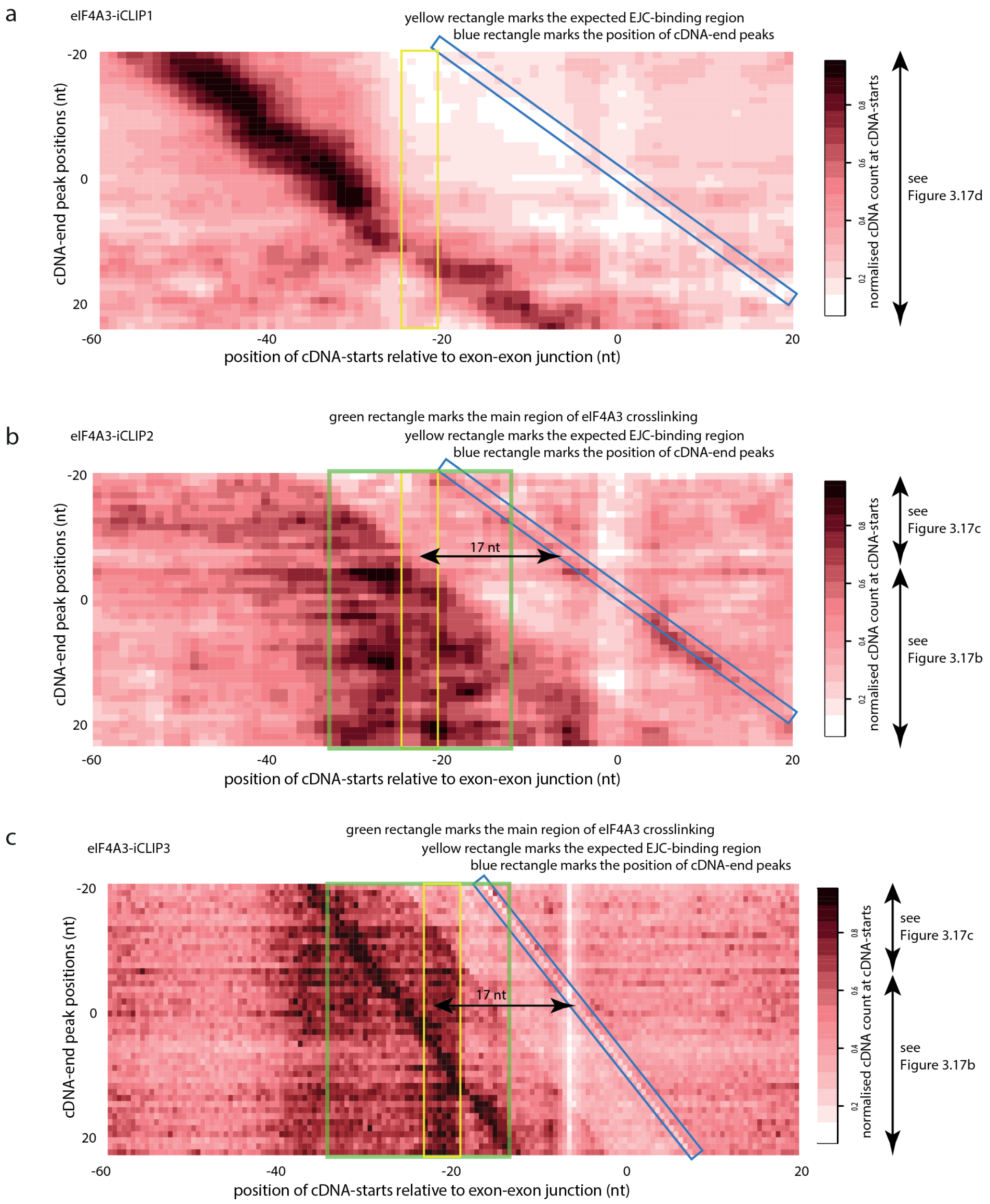


Figure 3.15: A broad cDNA length range ameliorates the effects of constrained cDNA-ends.

a) Heatmap showing the position of cDNA-starts in eIF4A3-iCLIP1 around the 1000 exon-exon junctions with the highest number of cDNAs. Junctions are sorted according to their cDNA-end peak position. Each row shows the average of cDNA counts at all junctions with a cDNA-end peak at the indicated position. The values are normalised against the maximum value across all rows. On the right, the arrows mark parts of the figure in which binding site assignment corresponds to the schematic shown in (Figure 3.17d). Coloured rectangles mark the main region of eIF4A3 crosslinking (green), the expected EJC-binding region (yellow) and the position of the cDNA-end peak (blue).

(continued)

Figure 3.15: b) Same as a), but for eIF4A3-iCLIP2. The arrow in the figure marks the 17 nt minimal distance between cDNA-starts and the expected EJC-binding region that is required for cDNA-starts to be able to identify crosslink sites within the binding site. On the right, the arrows mark sections that correspond to the schematics shown in (Figure 3.17c, b).
c) Same as a), but for eIF4A3-iCLIP3. The arrow in the figure marks the 17 nt minimal distance between cDNA-starts and the expected EJC-binding region that is required for cDNA-starts to be able to identify crosslink sites within the binding site. On the right, the arrows mark sections that correspond to the schematics shown in (Figure 3.17c, b).

In comparison with eIF4A3-iCLIP1 experiment, eIF4A3-iCLIP2 and eIF4A3-iCLIP3 have a broad range of cDNA lengths, where majority of cDNAs are longer than 40 nts of their sequencing length (Figure 3.16a). The broad range of cDNA lengths identifies a broad area of crosslink positions upstream from cDNA-end peaks, including the expected 24 nt EJC-binding upstream region relative to exon-exon junction in both eIF4A3-iCLIP2 and eIF4A3-iCLIP3 (marked as yellow and green rectangle in Figure 3.15b, c). Notably, there is a 17 nt distance of low crosslink enrichment between cDNA-end peaks and crosslinks, that can be explained by the iCLIP procedure. In the computational pipeline of iCLIP analysis (see Methods 1), cDNAs shorter than 17 nt are removed from the pipeline, since they rarely map to a unique genomic position. For this reason, the cDNA-ends should ideally be at least 17 nt away from the RBP binding region.

The majority of cDNA-ends are present more than 17 nt downstream of the expected EJC-binding region in eIF4A3-iCLIP2 and eIF4A3-iCLIP3 (Figure 3.15b, c), which decreased the cDNA-end constraints. Another potential explanation for better overlap of crosslink sites over the EJC binding region in eIF4A3-iCLIP2 and eIF4A3-iCLIP3 could be the broad range of cDNA lengths (Figure 3.16a). Indeed, most crosslinking in eIF4A3-iCLIP2 and eIF4A3-iCLIP3 is seen within the expected EJC-binding region, as well as approximately 10 nt on each side of this region (marked with green rectangle in Figure 3.15b, c). In conclusion, the broad range of cDNA lengths can overcome the cDNA-end constraints by producing the non-coinciding cDNA-starts that can more comprehensively identify crosslink sites (marked by the yellow rectangle in Figure 3.15a).

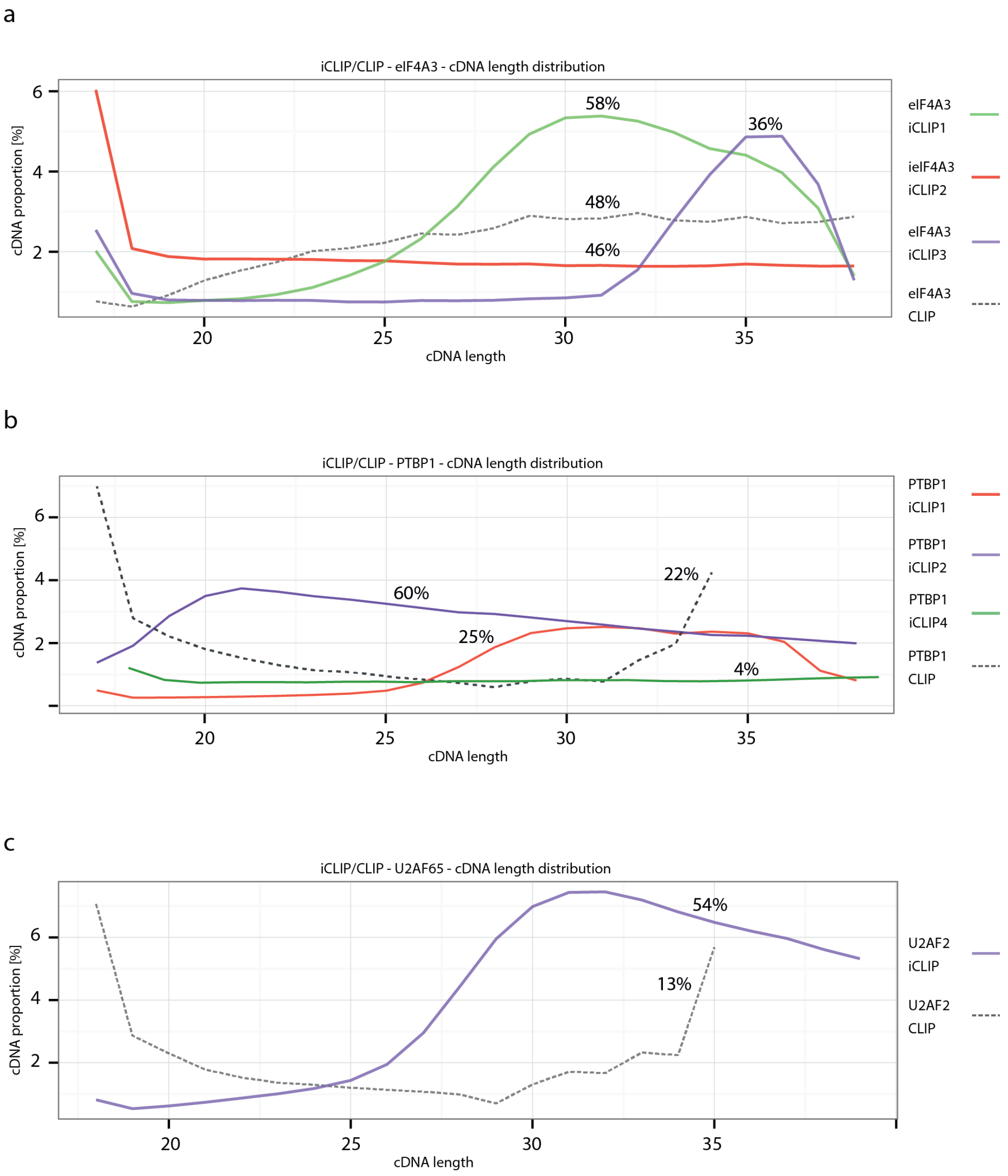


Figure 3.16: Distribution of cDNA sizes in the studied experiments.

a) Distribution of cDNA sizes in eIF4A3 CLIP and iCLIP experiments of cDNAs that are shorter than 39 nt. The number above the lines reports the % of cDNAs shorter than 39 nt. For longer cDNAs, it is not possible to draw the distribution as their precise lengths are unknown due to the limited length of sequencing. Thus, both the distribution and the % needs to be taken into account to estimate if there is a narrow distribution of cDNA sizes. For example, the distribution shows preferred lengths for both eIF4A3-iCLIP1 and eIF4A3-iCLIP3, but in case of eIF4A3-iCLIP3 only 36% of cDNAs are shorter than 39 nt, while in eIF4A3-iCLIP1 approximately 50% of cDNAs are in the length range of 27-37 nt. Thus, only eIF4A3-iCLIP1 has a strong potential for the cDNA distribution to affect binding site assignment.

(continued)

Figure 3.16: b) Same as a), but for PTBP1 CLIP experiments, showing the % of cDNAs shorter than 34 nt due to the shorter sequencing length.
c) Same as b), but for U2AF2-iCLIP which shows a trend for shorter cDNA size distribution, with 54% of cDNAs <39 nt. However, this is not a major problem due to the lesser cDNA-end constraints in this experiment.

3.10 Discussion

This chapter demonstrates that use of iCLIP cDNA-starts is appropriate for assignment of crosslink sites. Moreover, it shows that assessing cDNA lengths and cDNA-ends can help to understand any biases that can limit assignment of crosslink sites to specific regions of binding sites, particularly in long binding sites (Figure 3.17a). I present the computational approaches to visualise these technical features in the sequenced cDNA libraries. I find that cDNA-ends are often constrained in CLIP and iCLIP libraries, most likely a result of preferred RNase cleavage, which leads to the non-coinciding cDNA-starts (Figure 3.12a, b). For example, pre-mRNAs are cleaved during the splicing process within cells, which explains the peak of cDNA-ends at exon-exon junctions in eIF4A3-iCLIP2, eIF4A3-iCLIP3 (Figure 3.12d), and at intron-exon junctions in PTBP1 iCLIP experiments (Figure 3.12b, d, e, f). I also show that the RNases used in iCLIP and CLIP can have preference for single-stranded RNA or for specific sequence motifs, which can also lead to cDNA-end constraints (Figure 3.13b, d, f, g-j). When cDNA-ends are located at least 17 nt (minimum cDNA length for mapping after trimming, see Methods 1) downstream of the binding site, then a broad distribution of cDNA lengths can compensate for the cDNA-end constraints to ensure that the assigned binding sites are correctly assigned by crosslink cDNAs (Figure 3.16d and Figure 3.17b). However, if an iCLIP library contains a narrow distribution of cDNA sizes, the cDNA-end constraints can lead to an overly narrow assignment of binding sites (Figure 3.15a and Figure 3.17d). Similarly, only the 5' region of the binding sites is assigned if cDNA-ends are constrained to positions very close to the binding sites (Figure 3.16b, c and Figure 3.17c). Since cDNA fragments can also be too long to be isolated from the gel (Figure 3.17b in grey) or too short (Figure 3.17c in grey) to be uniquely mapped to genomic position, it would be very challenging to identify crosslink sites closer than 17 nt from the cDNA-ends, which is the minimum length for mapping iCLIP cDNA reads (see Methods 1).

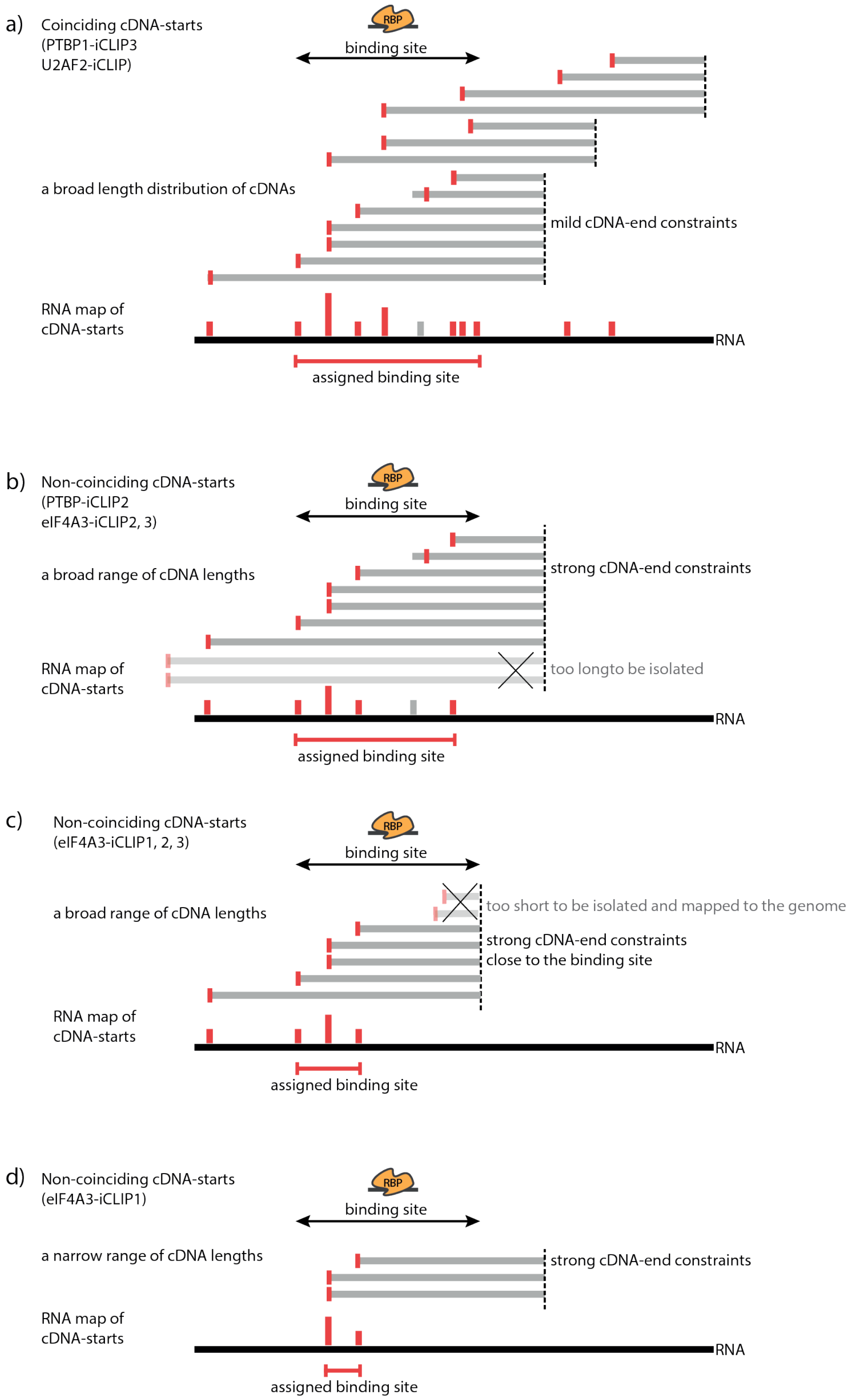

a)
Coinciding cDNA-starts
(PTBP1-iCLIP3
U2AF2-iCLIP)
RBP
binding site
a broad length distribution of cDNAs
mild cDNA-end constraints
RNA map of
cDNA-starts
RNA
assigned binding site
b)
Non-coinciding cDNA-starts
(PTBP-iCLIP2
eIF4A3-iCLIP2, 3)
RBP
binding site
a broad range of cDNA lengths
strong cDNA-end constraints
RNA map of
cDNA-starts
too longto be isolated
RNA
assigned binding site
c)
Non-coinciding cDNA-starts
(eIF4A3-iCLIP1, 2, 3)
RBP
binding site
a broad range of cDNA lengths
too short to be isolated and mapped to the genome
strong cDNA-end constraints
close to the binding site
RNA map of
cDNA-starts
RNA
assigned binding site
d)
Non-coinciding cDNA-starts
(eIF4A3-iCLIP1)
RBP
binding site
a narrow range of cDNA lengths
strong cDNA-end constraints
RNA map of
cDNA-starts
RNA
assigned binding site

Figure 3.17: A schematic explaining how different extents of cDNA-end constraints affect binding site assignment.

(continued)

Figure 3.17: a) If the iCLIP library contains a broad range of cDNA lengths and unconstrained positions of cDNA-ends, then crosslink sites are identified in an unbiased manner, allowing assignment of the full binding site (RNA-map at the bottom). The crosslink sites assigned by cDNA-starts are marked in red bars and a grey bar marks a crosslink site that is incorrectly assigned by a readthrough cDNA.
b) If cDNA-ends are constrained, most likely as a result of biased RNase cleavage, then the resulting cDNA-starts do not coincide. Nevertheless, if a broad distribution of cDNA lengths is available and the cDNA-ends are placed far enough from the binding site, then crosslink sites can still be identified across the full binding site, allowing correct assignment, as was seen in the case of eIF4A3-iCLIP2 (Figure 3.15b). If a broad distribution of cDNA lengths is used there can also be too long fragments to be isolated from the SDS gel (long cDNAs in grey).
c) If cDNA-ends are constrained to a position very close to the binding site, then those cDNAs that truncate at crosslink sites in the 3' region of the binding site are too short to be isolated and mapped to the genome (short cDNAs in grey). Therefore, crosslink sites are identified only in the 5' region of the binding site, leading to an overly narrow assignment of binding sites, as was seen in some of the sites identified by eIF4A3-iCLIP1, eIF4A3-iCLIP2 and eIF4A3-iCLIP3 (Figure 3.15a, b, c).
d) If cDNA-ends are constrained and an iCLIP library contains a narrow distribution of cDNA sizes, then cDNA-end constraints lead to an overly narrow assignment of binding regions, as was seen in the case of eIF4A3-iCLIP1 (Figure 3.15a).

I provide several pieces of evidence to argue against the previous hypothesis that the non-coinciding cDNA-starts reflect a high prevalence of readthrough cDNAs. First, I analyse the iCLIP library from a modified iCLIP protocol to identify readthrough cDNAs with the use of an additional 5' marker ligation step (Figure 3.4). The purpose of this method is not to define the precise proportion of readthrough cDNAs; that would require a more specialised protocol. Instead, it examines the sequence characteristics at the starts of readthrough cDNAs as part of the standard iCLIP protocol, and without loss of any cDNAs. This confirmed that the detected readthrough cDNAs have distinct characteristics from most other cDNAs in iCLIP (Figure 3.5a-d). Second, I observe lower proportion of crosslink-induced deletions in eIF4A3 iCLIP compared to CLIP, in agreement with the previous study [120]. Third, I show that in spite of non-coinciding cDNA-starts, CL-motifs are enriched mainly at cDNA-starts in iCLIP, but not in CLIP and this also applies to the PTBP1-iCLIP2 experiment in which 4SU was used for crosslinking. Moreover, I show that readthrough and truncation at crosslink sites are not mutually exclusive, since many readthrough cDNAs in iCLIP appear to also truncate at crosslink sites, as evident by CL-motif enrichment both at position of deletions and cDNA-starts (Figure 3.2g). This could occur when two separate crosslinking events are present in a single RNA fragment, allowing the cDNA to readthrough the first event but truncate at the second. Fourth, I show that non-coinciding cDNA-starts in iCLIP are caused by constrained cDNA-ends, which can be caused by the RNase used for RNA fragmentation, or by *in vivo* RNA cleavage, for example during production of intron lariats (Figure 3.11). Fifth, while cDNA-starts of readthrough cDNAs could lead to spurious assignment of crosslink sites upstream of the expected binding regions, I find that the expected EJC binding region locates at the centre of cDNA-starts in eIF4A3-iCLIP2 and eIF4A3-iCLIP3 (Figure 3.15b, c), and similarly the Y-tracts overlap well with cDNA-starts in PTBP1 and U2AF2 (Figure 3.8a, b, e, f).

Collectively, I believe the presented evidence is sufficient to reject the hypothesis that non-coinciding starts are caused by a high prevalence of readthrough cDNAs as it was proposed by the previous study [181]. Instead, I find that crosslink sites are correctly assigned by cDNA-starts even if non-coinciding cDNA-starts are present, and instead the non-coinciding cDNA-starts are a result of cDNA-end constraints that can be explained by multiple causes (Figure 3.17).

Based on the readthrough hypothesis, non-coinciding cDNA-starts served as an argument for using cDNA-centres instead of cDNA-starts, since use of cDNA-centres corrected the shift in the EJC-binding sites assigned by eIF4A3-iCLIP1 [181]. I now show that eIF4A3-iCLIP1 experiment is unique due to its narrow position of cDNA-ends immediately next to the expected EJC-binding region (Figure 3.12d) and its relatively narrow cDNA size distribution (Figure 3.16a), both of which lead to assignment of overly narrow binding sites (Figure 3.15a and Figure 3.17d). The newly-generated eIF4A3-iCLIP2 and eIF4A3-iCLIP3 experiments demonstrate that these problems can be addressed experimentally, and therefore experimental optimisations of iCLIP would be more appropriate instead of use of cDNA-centres (Figure 3.8i, j).

My findings and computational approaches will help users to optimise iCLIP conditions towards a broad range of cDNA lengths and unconstrained positions of cDNA-ends. They demonstrate the importance of optimised conditions in iCLIP to avoid cDNA-end constraints such as 3' dephosphorylation of RNA fragments needs to be efficient (Figure 3.1, step 2), since this is necessary for efficient 3' adapter ligation to the RNA 3' ends produced by RNase I (Figure 3.1, step 3). Ideally, most RNA fragments would be ligated to the 3' adapter, which minimises potential biases. Another important aspect of iCLIP protocol optimisation is the purification of cDNAs, that should be performed in a way that maintains a broad range of cDNA lengths in the final amplified library. This should ideally include isolation of both short and long cDNAs to maximise mapping of crosslink sites that are located either close or far from the site of RNase cleavage, respectively (Figure 3.17a). Moreover, it indicates that special procedures for genomic mapping of short cDNAs may be beneficial; for example, due to the problem that short cDNA reads often map at multiple genomic positions, mapping of short cDNAs could be narrowed down to the genomic regions where longer cDNAs map. Taken together, it is important to ensure that RNase I is the primary source of RNA fragmentation, that 3' dephosphorylation of RNA fragments is efficient and that the cDNA library has a broad range of cDNA sizes.

This chapter provides insights into the design, analysis and interpretation of iCLIP data. It demonstrates the importance of a broad cDNA length distribution and optimised RNase fragmentation conditions, according to published guidelines [126, 124, 192]. These analysis confirms that cDNA-starts are the correct input for the computational analysis of iCLIP data, even if non-coinciding cDNA-starts are present. Finally, it informs about the interpretation of binding sites that are assigned with cDNA-starts. For example, even though the non-coinciding cDNA-starts in iCLIP identify contacts with 10 nt on each side of the expected EJC-binding region, this is compatible with the finding that the sequence and structure of a nascent mRNA can shift EJC deposition as far as 10 nt away from this expected site [206].

It also agrees with finding of previous transcriptomic studies, which have shown that the precise position of EJC binding can vary between different junctions, and can be influenced by RNA structure or by other RNA-binding proteins that bind in the vicinity [201, 166].

I conclude that non-coinciding cDNA-starts are not a cause for concern in iCLIP, and instead they reflect the capacity of broad cDNA length distribution to compensate for constrained cDNA-ends (Figure 3.17b). This allows the cDNA-starts in iCLIP to identify crosslink sites across the complete RNA binding sites of RBPs, ensuring that the binding sites are correctly assigned.

Chapter 4

CLIPo: a tool to identify the features underlying protein-RNA interactions from CLIP data

4.1 Introduction

Analysis of endogenous RNA binding sites of RBPs has been aided by the development of UV crosslinking and immunoprecipitation (CLIP) [38], and its successor, individual nucleotide resolution crosslinking and immunoprecipitation (iCLIP) [124]. iCLIP employs UV crosslinking and immunoprecipitation to identify sites of protein-RNA crosslinking with nucleotide resolution in a transcriptome-wide manner. Moreover, many other variations of CLIP method have been developed, including eCLIP and irCLIP, each of which modify multiple enzymatic steps from the CLIP protocol that can affect the quality of the resulting cDNA library [120] and binding site assignment. However, the ways that variations in the method affect the assignment of RNA binding sites are unclear.

Tools such as FASTQC exist that can examine the sequenced library to assess the potential effects of poor sequencing quality, but no tools are available to examine the quality of preceding steps in the preparation of CLIP cDNA libraries. The development of such a tool was hampered by a lack of high-quality cDNA libraries produced for the same RBP in different laboratories using different protocols. Recently, such libraries have become available for several proteins, which provides the basis for comparing the technical features affecting each cDNA library systematically.

In this chapter, I will present CLIPo (CLIP optimisation tool), a newly developed computational pipeline that assesses features that can affect the correct identification of binding sites (see GitHub). CLIPo examines three general features of cDNA libraries that can inform on the quality of the steps during library preparation. As a proof of principle, I examined these features in multiple datasets for PTBP1 produced by iCLIP, eCLIP and irCLIP to demonstrate the impact of several biases, such as sequence constraints, narrow cDNA length distribution and background noise that can affect the assignment of binding sites. CLIPo can assess technical features in all variants of iCLIP, and thus offers the quality control standards for the transcriptome-wide assignment of protein-RNA binding sites.

In the first part of this chapter, I will focus on the effects of technical features that differ between CLIP methods and assess in detail the binding pattern of PTBP1, which binds to well-defined polypyrimidine-rich RNA motifs [66, 47]. In addition to providing novel technical insights such as the optimal clustering conditions for PTBP1, I will give new mechanistic insight into PTBP1 splicing regulation. With a new RNA-map pipeline, which is a part of the CLIPo package, I will demonstrate that PTBP1 regulates splicing mostly in a position-dependent manner.

In this chapter, I will refer to local accumulations of significantly enriched crosslink sites as peaks, and to clusters as regions of merged peaks. Both clusters and peaks are identified by iCount, such that the user can adjust the window size for their assignment.

4.2 CLIPo reports on the quality and specificity of CLIP experiments

In the previous chapter, I showed that iCLIP data contain non-coinciding cDNA-starts, which are caused by constrained cDNA-ends resulting from the RNA sequence and structure constraints of RNase cleavage. I provided new computational approaches to visualise the impact of these features on the sequenced cDNA libraries to interpret the assigned binding sites correctly. Here, I will focus on the quality controls for the PTBP1 dataset from the eCLIP, iCLIP and irCLIP methods, and show how the constraints can be measured based on the findings of cDNA-end constraints from the previous chapter (see Chapter 3).

CLIPo is a computational pipeline that examines three general features of sequencing libraries (see Methods 2.5) that inform on the quality of the principal library preparation steps in CLIP (Figure 3.1). First, it examines the cDNA complexity, or the total number of unique cDNAs in the library. This cDNA count depends on the sequencing depth and reflects the complexity of the library, since the cDNAs mapping at the same genomic position need to have distinct UMIs and hence are removed as PCR duplicates. High cDNA complexity shows that a sufficient amount of RNA was co-purified with the RBP, which depends on the amount of starting material (cells or tissue), the abundance of the RBP, and the crosslinking and immunoprecipitation efficiency. Moreover, it reflects the efficiency of linker ligation, cDNA circularisation and reverse transcription, and the loss of RNAs or cDNAs at each step in the protocol.

Secondly, CLIPo examines the level of cDNA-end constraints by analysing the cDNA length distribution and the sequence at cDNA-starts and cDNA-ends. The distribution of cDNA lengths informs on the conditions of RNA fragmentation (Figure 3.11, step 2), size-selection of protein-RNA complexes (Figure 3.1, step 4) and cDNA library preparation and amplification (Figure 3.1, step 7). Sequence constraints at cDNA-starts can result from the sequence preferences of UV crosslinking (Figure 3.1, step 1) and from the conditions of adapter ligation to cDNA-starts (Figure 3.11, step 7). To examine length distribution, CLIPo reports the proportion of cDNAs that are shorter than 40 nt (see Table 4.1) and the highest percentage of cDNA lengths that are in the 10 nt length window are marked as length constraints (see Table 4.1). To examine the sequence constraints, CLIPo examines a sequence composition around cDNA-ends to measure the enrichment of all tetramers. It calculates a numerical value that estimates the constraints (see Table 4.1) by comparing the enrichment of the 10 most highly enriched tetramers directly at cDNA-ends relative to a region 10 nt upstream. A high enrichment of these motifs shows that the sequence of cDNA-ends is different from the surrounding sequence, and thus indicates high sequence constraints (Figure 3.10b, d, f).

Thirdly, CLIPo examines the specificity of cDNAs (see Table 4.1). To examine the capacity of the cDNA library to assign binding sites, CLIPo reports the number of significant crosslink clusters. Since it applies the same clustering algorithm for all data (except the the so-called 'narrow peaks' for eCLIP-NarrowPeaks, which are predefined clusters by ENCODE) by using iCLIP, irCLIP and eCLIP cDNA-starts as an estimated position of crosslink sites; this method allows comparative analysis. To examine the specificity for these clusters, it reports the proportion of cDNAs that identify crosslink sites within the clusters. Sequence specificity of the assigned binding sites is analysed through identification of the ten most enriched tetramers within the clusters and the enrichment of these tetramers is compared to the genomic sequence preceding the clusters. These specificity features can also reflect the nature of protein-RNA interaction: in case of RBPs that do not bind RNA with high specificity, values will be low regardless of the quality of cDNA library.

In order to understand how variations in specificity features of different experiments influence binding site assignment, I examined these features for PTBP1 across different methods. These experiments have similar cDNA complexity between 6 and 10 million uniquely mapped cDNAs, and thus they are among the most complex CLIP cDNA libraries published to date (see Table 4.1). As expected, I detected several cDNA constraints, some of which have been highlighted in the previous chapter (see Chapter 3). For example, there are pronounced sequence constraints at cDNA-ends in PTBP1-CLIP, PTBP1-iCLIP1 and PTBP1-iCLIP3 that have been investigated in the previous chapter (see Chapter 3), in which I demonstrated the adenosine enrichment at cDNA-end positions. There are also length constraints in the PTBP1-iCLIP5 experiment, similar to the ones discussed in the previous chapter 3, in which I show that the eIF4A34-iCLIP1 experiment had a narrow distribution of cDNA lengths (Figure 3.16a). However, enrichment in clusters from the top 10 most enriched tetramers is lower in PTBP1-iCLIP5 and PTBP1-NarrowPeaks compared to other experiments (see Table 4.1). Moreover, the number of assigned clusters and the percentage of cDNAs within them is very low in the PTBP1-eCLIP data compared to other experiments (see Table 4.1). The PTBP1-irCLIP experiment has ten times more uniquely mapped cDNAs and identified clusters compared to other experiments, but there is a much lower motif enrichment inside those clusters (see Table 4.1), suggesting that this experiment might have a large proportion of background noise. There are also strong structure constraints at cDNA-ends in PTBP1-irCLIP that could be related to the low specificity of the data. Interestingly, the PTBP1-eCLIP experiment shows much less structure constraints around cDNA-ends (see Table 4.1), which could be explained by the difference in the library preparation compared to the standard iCLIP, as eCLIP uses long fragments and paired-end read sequencing.

In summary, CLIPo predicts that PTBP1-iCLIP1 and PTBP1-iCLIP2 have the highest specificity compared to other experiments (see Table 4.1). More generally, I conclude that specificity features vary between experiments performed with the same RBP, suggesting that they provide information on the quality of the different cDNA libraries (see Table 4.1). This tool will now be a useful as a quality control of RBPs from different variants of CLIP method.

Protein	Method	Experiment number and Pubmed ID	Data complexity (unique cDNA number)	% of length constraint in 10 nt window	Sequence constraints at cDNA-ends	Structure constraints at cDNA-ends	Number of crosslink clusters	% of cDNAs in the clusters	Motif enrichment inside the clusters
PTBP1	iCLIP1	25599992	8,160,912	25.361	5.821	0.865	262026	36.573	3.471
PTBP1	iCLIP2	new, 4SU, RNase	9,211,541	59.741	1.615	0.815	299865	30.761	2.977
PTBP1	iCLIP5	26260686	10,580,744	61.444	2.904	0.897	114654	38.909	1.956
PTBP1	eCLIP	27018577	6,060,266	18.644	1.871	0.779	129995	35.925	2.675
PTBP1	eCLIP-NarrowPeaks	27018577					60625	20.551	2.164
PTBP1	irCLIP	27111506	60,921,471	52.758	2.775	0.938	1704218	30.085	0.958

Table 4.1: CLIPo report table for PTBP1 produced by iCLIP, eCLIP, and irCLIP method.

4.3 Exploring the PTBP1 ENCODE eCLIP data

The ENCODE project chose eCLIP as the method of choice for a study of 300 RBPs. Since January 2016, all the datasets are available online, including mapped cDNAs and final clusters known as 'narrow peaks' produced with the peak calling tool CLIPper (see subsection 1.6.8). I first analysed the ENCODE processed data by using the narrow peaks from CLIPper, which have been normalised with the mock eCLIP controls and filtered as significantly enriched crosslink clusters (see PTBP1-Narrow Peaks in Table 4.1). I focused my study on the PTBP1 protein since it has been the main example used in this thesis (see Chapter 3). Surprisingly, the CLIPo quality measures of the protein specificity, such as coverage of cDNA-starts and motif enrichment, seem to be particularly low in narrow peaks compared to other experiments including the total number of detected crosslink clusters (see Number of crosslink clusters in PTBP1-NarrowPeaks in Table 4.1). A simple approach to explore the RBP specificity is to look at the motif enrichment across identified clusters and their surrounding region. For this purpose, I stratified clusters into length categories and produced a heatmap of motif enrichment including the surrounding region (see Methods 2.5.4).

Interestingly, there is a significant enrichment of PTBP1 motifs in the region upstream of detected clusters (Figure 4.1a). To further assess the validity of the PTBP1-eCLIP data, I looked at the distribution of cDNA-starts for the same cluster groups. As expected, the cDNA-start positions show an upstream shift away from the narrow peak clusters with a similar pattern as the heatmap of motif enrichment (Figure 4.1a, b). This shift in the coverage of cDNA-starts and the motif enrichment could be explained by the use of complete cDNAs, rather than cDNA-starts, in the ENCODE narrow peak calling.

The eCLIP protocol [134] proposed that eCLIP data maintains the single-nucleotide resolution, and so it is important that we use the appropriate software for peak analysis that exploits such resolution by the use of cDNA-starts. Therefore, I analysed the PTBP1-eCLIP data with the iCount peak calling algorithm (see Methods 2.4.1) that I also used for all the iCLIP and irCLIP experiments. I analysed the data using a 3 nt window size setting and FDR threshold lower than 0.05 to merge peaks that are within 15 nt surrounding region into crosslink clusters, which is also the default setting of the CLIPper tool. This comparison of two different pipelines showed a clear difference in the CLIPo results table (see Table 4.1), with twice as many clusters identified in PTBP-eCLIP and an over 2-fold increase of motif enrichment compared to the PTBP1-NarrowPeaks. The low number of identified clusters by eCLIP is unexpected, since the number of eCLIP cDNAs is similar to iCLIP, and it might reflect increased noise of non-specific cDNAs in eCLIP. Importantly, in both iCLIP and eCLIP, motifs and cDNA-starts are enriched in the expected region of the clusters defined by the iCount pipeline, demonstrating that analysis of cDNA-starts is more appropriate than the use of complete cDNAs by CLIPper (Figure 4.1).

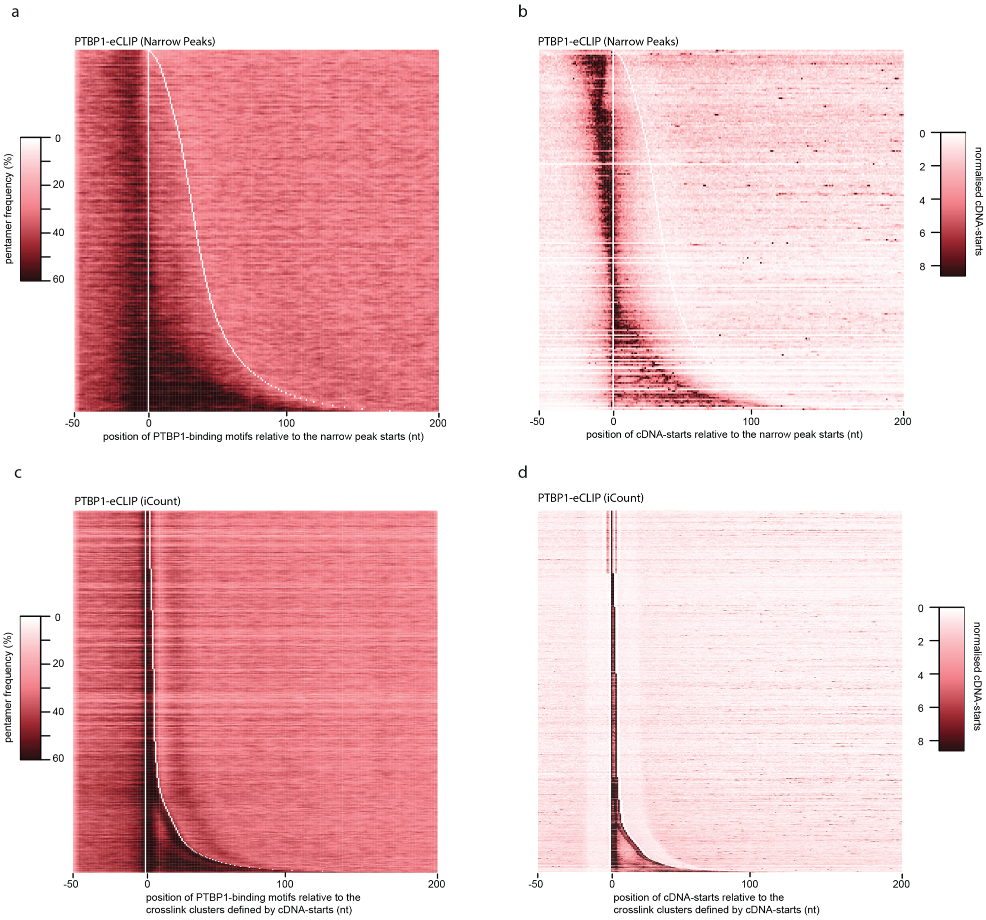


Figure 4.1: Heatmaps of PTBP1 motifs and cDNA-starts for comparing ENCODE narrow peaks and the iCount peak calling pipeline.

a) Heatmap showing the coverage of PTBP1-binding motifs at the PTBP1-eCLIP1 narrow peak clusters that were downloaded from the ENCODE website. Each row shows the average coverage for 300 clusters of similar length, sorted from the shortest to longest clusters. The white line marks the nucleotide preceding the start and the nucleotide following the median end of all clusters that were combined in each row. A colour key for the coverage per nucleotide of the PTBP1-binding motifs is shown on the top left.
b) Heatmap showing the density of normalised cDNA-starts around PTBP1-eCLIP1 narrow peak clusters that were downloaded from ENCODE website. Each row shows the average coverage for 300 clusters of similar length, sorted from shortest to longest clusters. The white line marks the nucleotide preceding the start and the nucleotide following the median end of all clusters that were combined in each row. A colour key for the coverage per nucleotide of the PTBP1-binding motifs is shown on the top left.
c) Same as a) but showing the motif coverage of PTBP1-eCLIP1 clusters identified by iCount.
d) Same as a) but showing the normalised cDNA-starts of PTBP1-eCLIP in PTBP1-eCLIP1 clusters identified by iCount.

4.4 Differentiation of biological and technical features between iCLIP and eCLIP data

One of the differences between the iCLIP and eCLIP methods is that eCLIP comes with an additional mock-eCLIP control experiment, which represents a mixture of crosslink sites for many different RBPs. Thus this mock-eCLIP should not reflect the sequence specificity of any specific RBP [134]. CLIPo analysis showed a similar specificity for the PTBP1-eCLIP and PTBP1-iCLIP2 experiments, when I used the same pipeline for identification of crosslink binding clusters without any additional normalisation. To explore differences in the data further, I investigated whether there were any technical or biological differences between these two experiments by analysing tetramer enrichment around cDNA-start peaks together with the control mock-eCLIP experiment. I measured the tetramer enrichment for each experiment compared to the controlled region (see Methods 2.5.5) and compared the enrichment between experiments (Figure 4.2a, b, c). Indeed, there is a high correlation of tetramer enrichment between PTBP1-iCLIP2 and PTBP1-eCLIP (Figure 4.2c) and very low correlation between PTBP1-iCLIP2 and control mock-eCLIP (Figure 4.2a). Interestingly, there is much higher correlation between PTBP1-eCLIP and control mock-eCLIP compared to PTBP1-iCLIP2 and control mock-eCLIP (Figure 4.2a, b), suggesting that PTBP1-iCLIP2 identified more protein-specific motifs compared to PTBP1-eCLIP. Next, I looked at how tetramers are enriched and positioned relative to the cDNA-start peaks. First, I identified the position within each crosslink cluster with the highest count of cDNA-starts for all three experiments and visualised the enriched tetramers, which were sorted by the PTBP1-eCLIP tetramer enrichment from top to bottom with its sequence on the right side (see Figure 4.2). I noticed that the tetramer enrichment from all three experiments is positioned between -2 and 10 nt relative to the cDNA-start peak, which agrees with my previous findings (see Chapter 3), where I concluded that cDNA-start positions should be used for the data analysis of iCLIP and other related methods.

Next, I noticed a high enrichment of GACG, ACGG, CGGA tetramers in PTBP1-eCLIP and also in mock-eCLIP but not in PTBP1-iCLIP2 (orange rectangle in Figure 4.3). These motifs are not specific for PTBP1 protein which preferentially binds to pyrimidine tracts (Y-tracts) [66, 211]. Since these motifs are also enriched in the control mock-eCLIP experiment, I assumed these to be parts of technical features of the eCLIP method and that they can potentially be minimised or filtered with the mock eCLIP normalisation pipeline [134]. Taken together, there is no need for additional controls such as the mock eCLIP experiment for the iCLIP method if the quality of the library is high enough and prepared by following the recommended protocol [126]. More importantly, even though the PTBP1-eCLIP data has a much lower complexity (number of uniquely mapped cDNAs, number of identified clusters in Table 4.1), it can still detect a correct set of motifs that correlate with PTBP1-iCLIP2 experiment. Overall, it is important to distinguish technical from biological features and this type of tetramer visualisation around cDNA-start peaks (Figure 4.3, see Methods 2.5.6) could be used to interpret these features between different CLIP related methods.

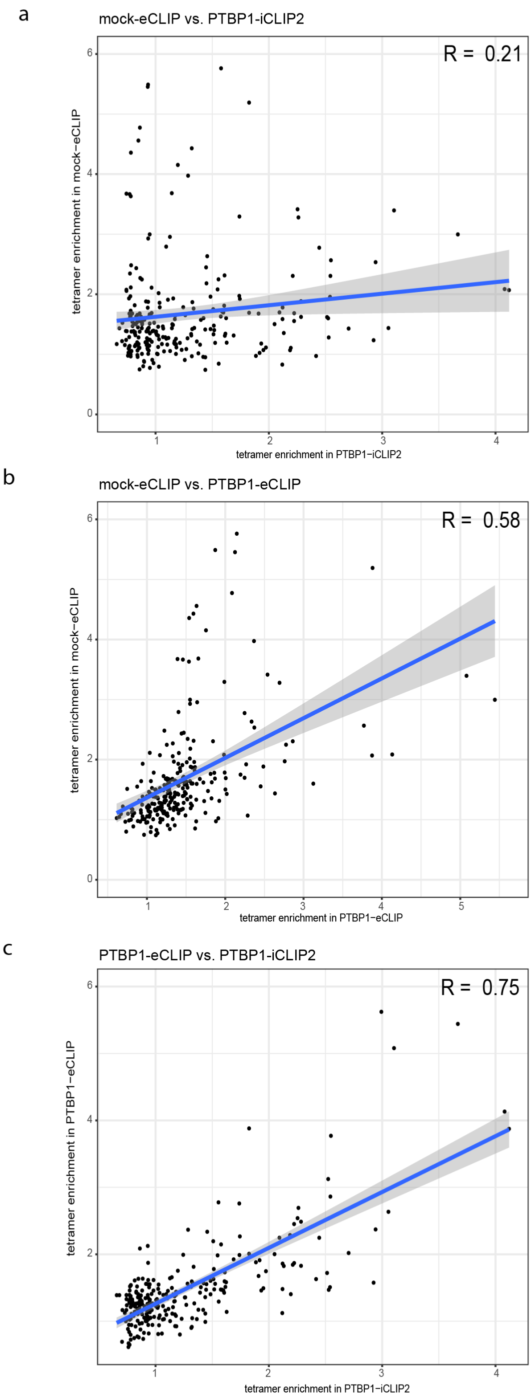


Figure 4.2: Scatter plot of tetramer enrichment between PTBP1-eCLIP, mock-eCLIP and PTBP1-iCLIP2 experiments.

a) Tetramer enrichment compared to the control between mock-eCLIP and PTBP1-iCLIP2 experiments.
b) Same as a) but for mock-eCLIP and PTBP1-eCLIP.
c) Same as a) but for PTBP1-iCLIP2 and PTBP1-eCLIP.

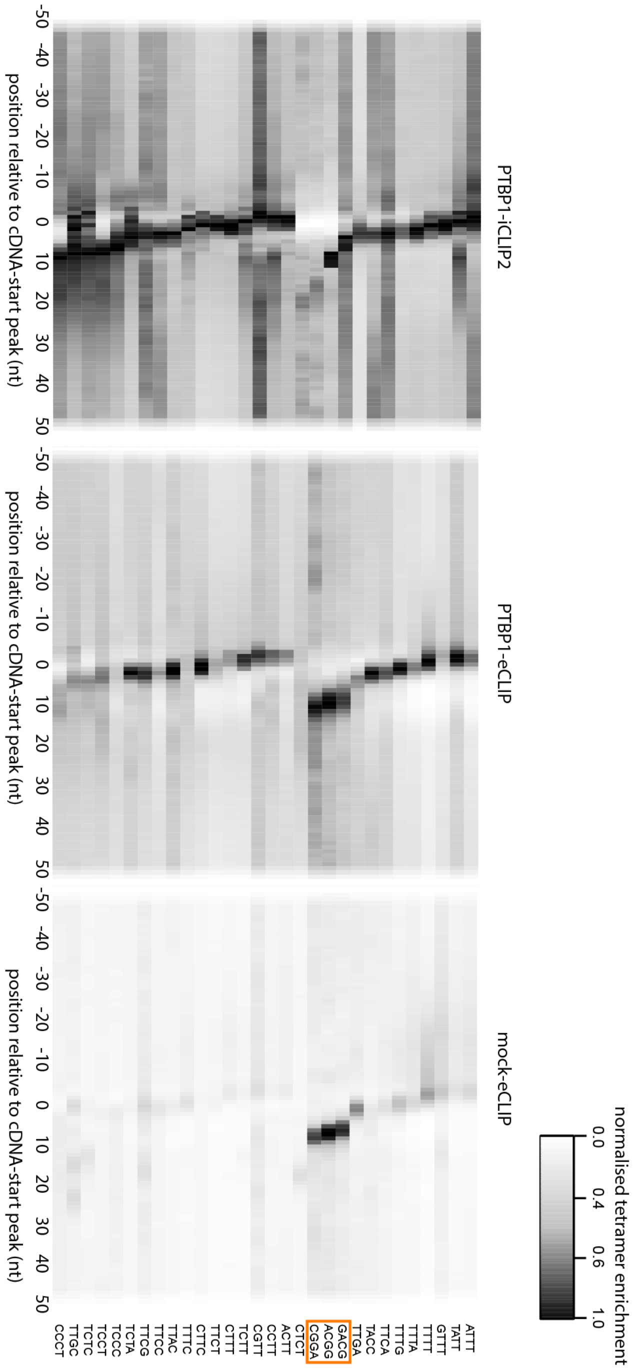


Figure 4.3: Heatmap of tetramer enrichment around cDNA-start peaks.

Each row shows the normalised tetramer enrichment relative to cDNA-start peak in region 50 nt surrounding region from left to right: PTBP-iCLIP2, PTBP-eCLIP and mock-eCLIP. All tetramers are sorted by PTBP1-eCLIP enrichment from top to bottom.

4.5 Optimal peak calling settings for PTBP1 binding sites in iCLIP data

We noticed that the distribution of cDNA-starts in clusters that were identified by iCount shows a drop of cDNA-starts in the 15 nt surrounding the cluster region (Figure 4.1 - Heatmap of eCLIP-PTPB1). This can be explained by the use of 15 nt peak merging conditions, which lead to the peaks being positioned at the boundaries of clusters. To understand how the window size options affect the features of assigned crosslink clusters better, I decided to redefine PTBP1 crosslink clusters to find the most optimal peak calling and clustering windows size. For this purpose, I grouped PTBP1-iCLIP1 and PTBP1-iCLIP2 experiments (here referred as PTBP1-iCLIP1-2), since both experiments showed a similar specificity in the CLIPo analysis (see Table 4.1). I first employed the standard binding site pipeline (see Methods 2.4.1) of 5 different groups with 20 nt clustering windows and 3 nt, 10 nt, 25 nt, 50 nt and 100 nt peak calling window sizes and by the significance of cDNA enrichment compared to the shuffled data [183] of 0.05 FDR (False Discovery Rate) threshold.

To examine if the newly defined clusters represent *bona fide* PTBP1 binding sites, I first assessed the sequence of clusters that were identified by different peak calling window sizes. PTBP1 motifs (see Methods 2.4.4) are highly enriched across the full length of these clusters but with decreased motif enrichment in long cluster regions (Figure 4.4). The drop in motif enrichment can also be seen by looking at the PTBP1 motif enrichment across all clusters for each peak window size group (see Table 4.2: Optimal peak calling window size for PTBP1-iCLIP1-2), suggesting that the short 3 nt peak calling window size is more precise in obtaining the full coverage of PTBP1 motifs across clusters in the PTBP1-iCLIP1-2 data. The enrichment of PTBP1 motifs is restricted to the region inside the 3 nt peak calling clusters, whereas a slight enrichment of PTBP1 motifs continues in the surrounding region of other cluster groups (10 nt, 25 nt, 50 nt, 100 nt in Figure 4.4).

Thus, the enrichment of PTBP1 motifs in the surrounding region of crosslink clusters might be explained by the presence of binding sites of a closely neighbouring PTBP1. Indeed, many RBPs contain multiple domains or form homomultimers, which allows them to simultaneously bind at multiple sites over longer RNA regions [64, 65, 33, 62]. For example, the two RRM domains of TIA1 were found to promote its cooperative binding, and thereby increase the avidity of interaction with longer binding regions [215]. Similarly, PTBP1 contains four RNA-binding domains, each of which binds a pyrimidine-rich (Y-rich) motif to facilitate interactions with long RNA regions [64, 66, 65]. Moreover, PTBP1 can form higher-order complexes when bound to RNA; binding of up to eight PTBP1 proteins was observed at a regulated exon [67, 68]. Nevertheless, it is not known how precisely iCLIP data can assign the position of full RBP binding sites on endogenous transcripts, and therefore the length of these binding sites has not yet been systematically assessed.

Since the short 3 nt peak calling window is the most suitable for this type of analysis, I was next interested in the optimal clustering window size of surrounding peaks. I used the same window size groups as I did for peak calling and draw the PTBP1 motif enrichment for all 5 clustering groups (3 nt, 10 nt, 25 nt, 50 nt and 100 nt in Figure 4.5). Similar to the peak window comparison I observed the same decreased trend of motif enrichment in the long clusters of large clustering window sizes (Figure 4.5). To ensure that the surrounding region is considered in the motif enrichment analysis (see Table 4.3 - PTBP1 motif enrichment inside clusters), I used the motif enrichment of mirrored clusters from the upstream region to deduce the inside motif enrichment across clusters (Table 4.3: PTBP1 motif-corrected enrichment). I found that the clustering with the approximate window size of 10 nt has the highest motif enrichment compared to other clustering groups (see Table 4.3). In essence, I found that the PTBP1 crosslink clusters from the iCLIP data are optimally classified under the short window sizes. In my example, this is demonstrated by the 3 nt peak calling window size and 10 nt clustering window size for merging the neighbouring peaks for PTBP1-iCLIP1-2 data.

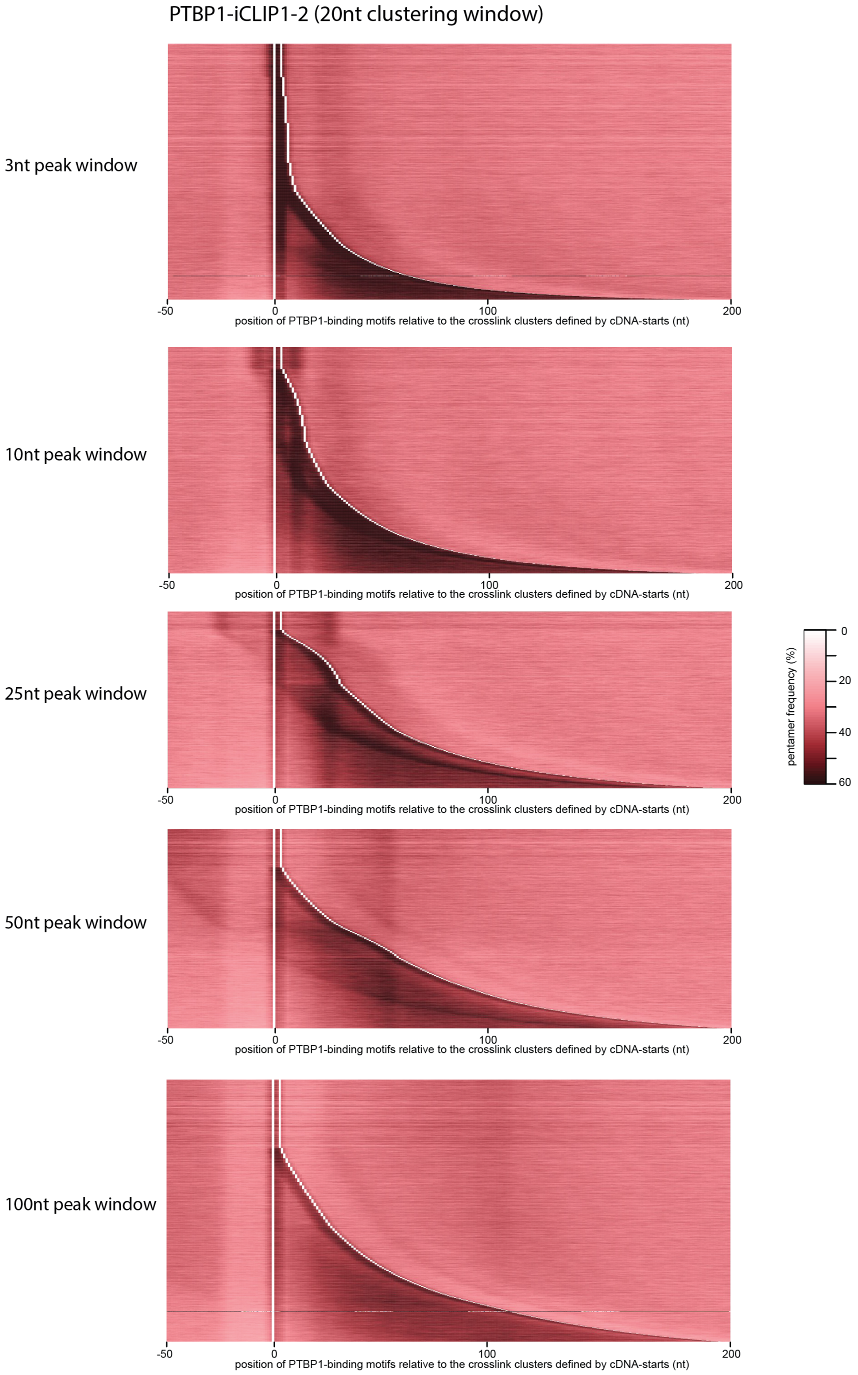


Figure 4.4: PTBP1-binding motif enrichment across PTBP1 crosslink clusters with different peak calling window sizes.

Heatmap showing the coverage of PTBP1-binding motifs at the PTBP1-iCLIP1-2 crosslink clusters that were defined with a 20 nt clustering window and 3nt, 10 nt, 25 nt, 50 nt, 100 nt peak calling window sizes. Each row shows the average coverage for 300 clusters of similar length, sorted from shortest to longest clusters. The white line marks the nucleotide preceding the start and the nucleotide following the median end of all clusters that were combined in each row. A colour key for the coverage per nucleotide of the PTBP1-binding motifs is shown on the right.

Peak calling windows size	PTBP1 motif enrichment inside clusters
3 nt	0.223
10 nt	0.222
25 nt	0.198
50 nt	0.181
100 nt	0.167

Table 4.2: Optimal peak calling window size for PTBP1-iCLIP1-2.

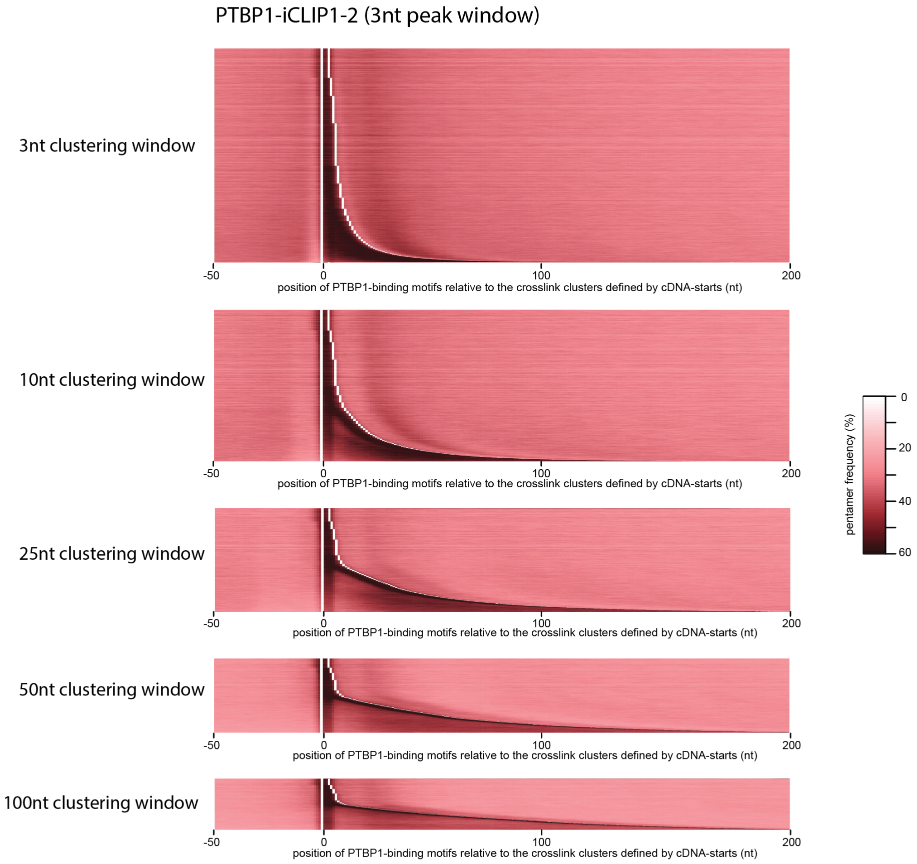


Figure 4.5: PTBP1-binding motif enrichment across PTBP1 crosslink clusters with different clustering window sizes.

Heatmap showing the coverage of PTBP1-binding motifs at the PTBP1-iCLIP1-2 crosslink clusters that were defined with a 3nt peak window and 3 nt, 10 nt, 25 nt, 50 nt, 100 nt clustering window sizes. Each row shows the average coverage for 300 clusters of similar length, sorted from shortest to longest clusters. The white line marks the nucleotide preceding the start and the nucleotide following the median end of all clusters that were combined in each row. A colour key for the coverage per nucleotide of the PTBP1-binding motifs is shown on the right.

Clustering window size	PTBP1 motif enrichment inside clusters	PTBP1 motif surrounding enrichment upstream	PTBP1 motif corrected enrichment
3nt	0.228	0.103	0.125
10nt	0.235	0.105	0.131
25nt	0.222	0.099	0.123
50nt	0.203	0.091	0.112
100nt	0.180	0.083	0.098

Table 4.3: Optimal clustering size for PTBP1-iCLIP1-2.

4.6 Assessing the position-dependent principles of splicing regulation with RNA-maps

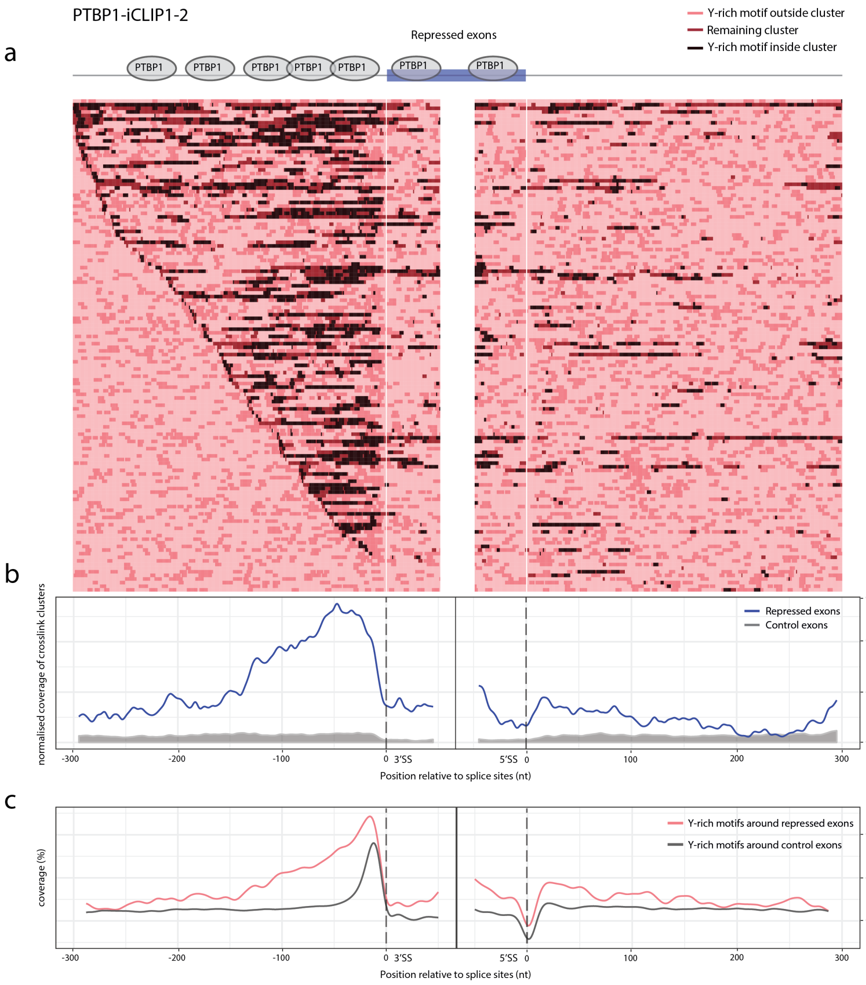
Similar to many other RNA-binding proteins, PTBP1 regulates splicing in a position-dependent manner [216]. The initial CLIP study concluded that PTBP1 binding close to an alternative exon generally causes skipping, whereas binding near a flanking exon induces inclusion [53]. Furthermore, later studies showed that PTBP1 promotes either skipping or inclusion by binding close to an alternative exon in a similar manner to NOVA proteins [33], such that binding upstream or inside the exon causes skipping, whereas binding downstream of the exon causes its inclusion [61, 30]. After the optimised binding site identification conditions for PTBP1-iCLIP1-2 data, I was interested in how PTBP1 binding contributed to position-dependent splicing regulation.

I examined PTBP1 positioning within 300 nt of PTBP1-regulated exons that were identified by the previous splice-junction microarray study [185]. The RNA-maps reveal a clear position-dependent regulatory outcome for PTBP1 (Figure 4.6 and Figure 4.7). There is a strong enrichment of crosslink clusters at the repressed exons that overlap with the 3’ splice site or the exon (normalised coverage of crosslink enrichment in Figure 4.6b, enrichment in left table of Figure 4.6d). In contrast, crosslink clusters at the enhanced exons locate downstream of the 5’ splice site, with low coverage in the 3’ splice site or the exon (normalised coverage of crosslink enrichment in Figure 4.7b, enrichment in left table of Figure 4.7d).

Using the RNA-map approach, I found the two main arrangements of PTBP1 complexes around regulated exons: the clusters either extend upstream (Figure 4.6a, b) or downstream of the 3’ splice site (Figure 4.7a, b). Moreover, in repressed exons the cluster can also extend over the exons (Figure 4.6a, blue line in Figure 4.6b). As expected, the position of crosslink clusters appears to be dictated by the enrichment of Y-rich motifs around regulated exons in comparison to the control exons; repressed exons with upstream clusters contain a Y-rich motif enrichment that extends upstream of the 5’ splice site (coverage of Y-rich motifs in Figure 4.6c), whereas with the downstream clusters of enhanced exons, the enrichment of Y-rich motifs is lower compared to repressed exons and is enriched only downstream of the 5’ splice sites where crosslink clusters are enriched (Figure 4.7a, c).

Several features explain the activity of PTBP1 exons with the upstream PTBP1 bindings. Firstly, it has been shown that PTBP1 complexes repress splicing when binding directly at 3’ splice sites [66, 211, 53, 61]. This analysis shows that these repressed exons are preceded by the long coverage of PTBP1 crosslink clusters, and with strong enrichment of Y-rich motifs that extends upstream of the canonical position of the poly-Y tract (Figure 4.6a, b, c). In contrast there is almost no Y-rich motif enrichment compared to the control exons in the 3’ splice site and exonic region (Figure 4.6a, b, c), showing that PTBP1 binding upstream of the exon can decrease accessibility of the 3’ splice site and/or the branch point that is needed for the U2AF2 and U2 snRNP complexes to repress or enhance splicing, respectively.

Moreover, at the repressed exons, Y-rich motif enrichment is present at the 3’ splice site, as well as within and downstream of the exon (Figure 4.6c). This agrees with previous studies showing that PTBP1 complexes repress splicing when binding across the exon [66, 211, 53, 61]. In contrast, the motifs are enriched only downstream of the enhanced exons (Figure 4.7c). This enhancing effect of downstream binding also agrees with the previous study [61] (Figure 4.7a, b, c). Notably, at enhanced exons, enrichment starts downstream of the 5’ splice site and continues as far as 300 nt away from the exon (Figure 4.7a, b). Taken together, PTBP1 binding upstream or downstream of the regulated exons appears to play a central role in its position-dependent manner of splicing regulation.



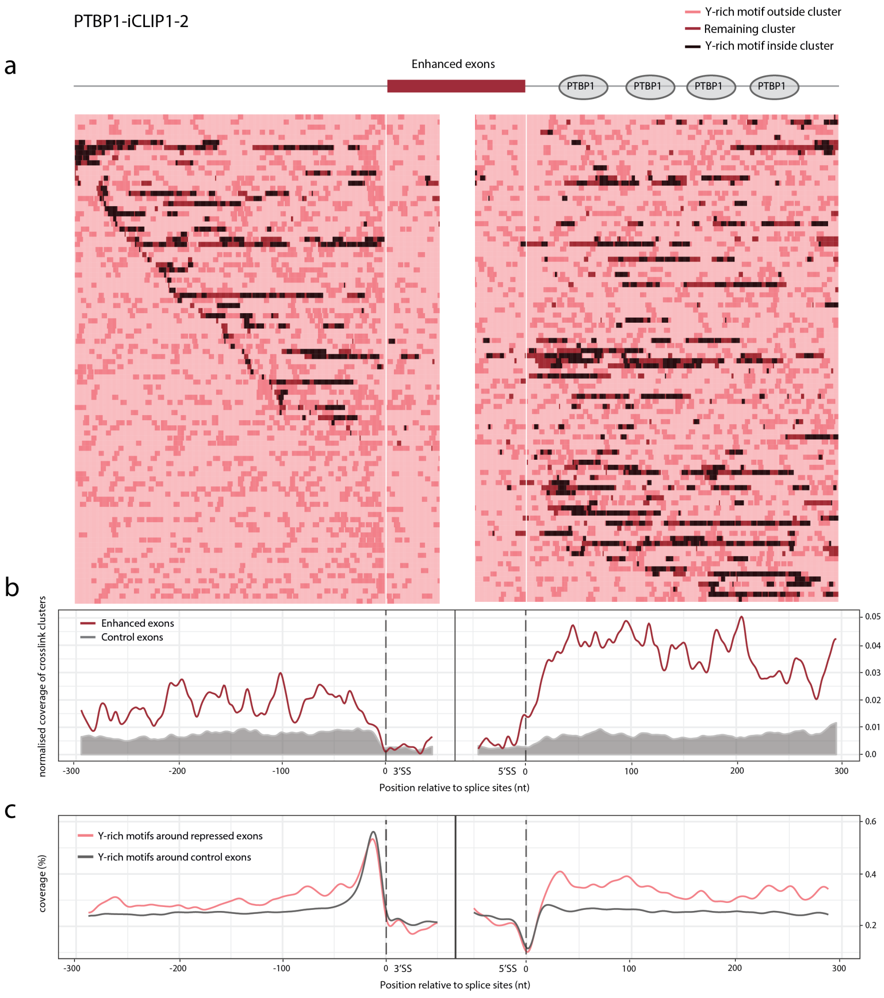
d

	reg.exons.cov	control.exons.cov	enrichment	distance
3’ss	19.29	2.19	8.79	17.1
exon	3.38	0.26	13.13	3.13
5’ss	5.62	1.71	3.28	3.9

	regulated	control
3’ss vs. 5’ss	3.44	1.28
exon vs. 3’ss	0.18	0.12
exon vs. 5’ss	0.6	0.15
5’ss vs. 3’ss	0.29	0.78

Figure 4.6: RNA-map for PTBP1 repressed exons.

a) Heatmap showing the positioning of PTBP1-iCLIP1-2 crosslink clusters and Y-rich motifs around PTBP1-repressed exons. All PTBP1 exons, were identified with splice-junction microarrays [185], and their flanking regions were aligned to the 3’ (left) and 5’ (right) splice sites (vertical white lines). The positions of the PTBP1-iCLIP1-2 crosslink clusters are indicated by dark shading, where the Y-rich motifs are shown as black or light red rectangles inside or outside of the clusters, respectively.
b) Density plot showing the normalised coverage of PTBP1-iCLIP1-2 crosslink clusters around repressed (blue) and control exons (grey).
c) Density plot showing the average coverage of Y-rich motifs around repressed (light red) and control exons (grey).
d) Table on the left shows PTBP1-iCLIP1-2 crosslink cluster enrichment and distance between repressed and control exons for the 3’ splice site, the 5’ splice site and exonic region. Table on the right shows the ratio between the 3’ splice site, the 5’ splice site and exonic regions for repressed and control exons.



d

	reg.exons.cov	control.exons.cov	enrichment	distance
3'ss	5.35	2.19	2.44	3.16
exon	0.43	0.26	1.65	0.17
5'ss	9.3	1.71	5.44	7.59

	regulated	control
3'ss vs. 5'ss	0.58	1.28
exon vs. 3'ss	0.08	0.12
exon vs. 5'ss	0.05	0.15
5'ss vs. 3'ss	1.74	0.78

Figure 4.7: RNA-map for PTBP1 enhanced exons.

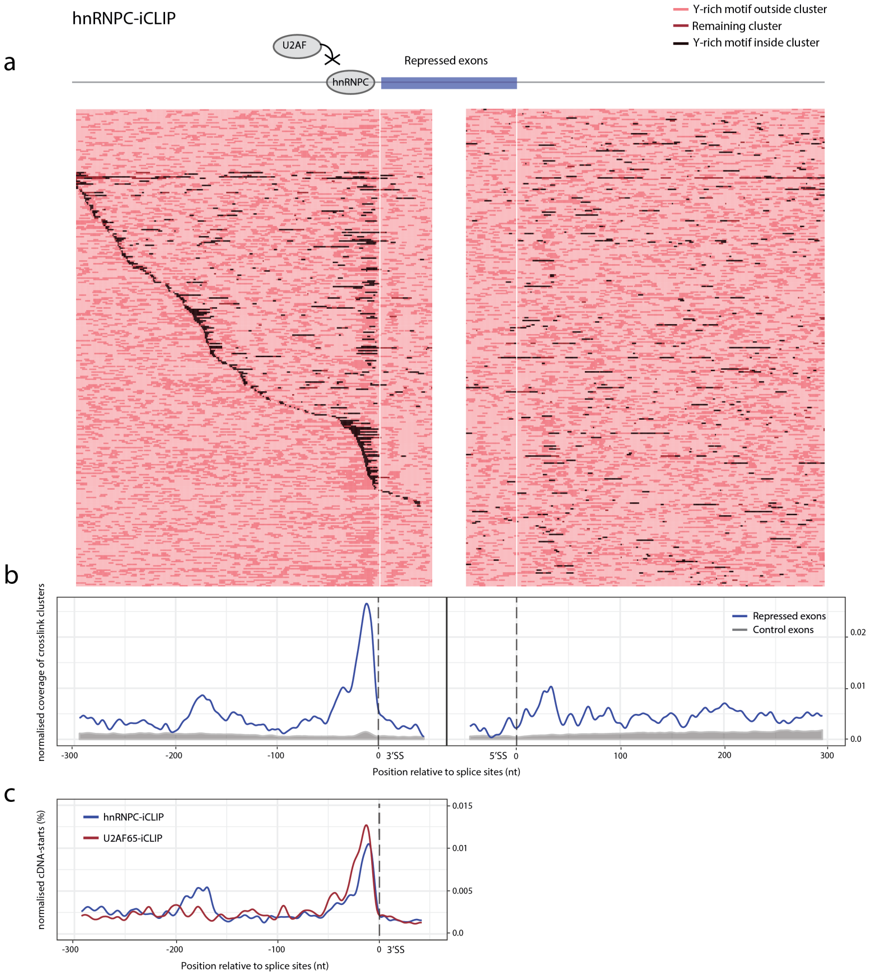
a) Heatmap showing the positioning of PTBP1-iCLIP1-2 crosslink clusters and Y-rich motifs around PTBP1-enhanced exons. All PTBP1 exons, were identified with splice-junction microarrays [185], and their flanking regions were aligned to the 3' (left) and 5' (right) splice sites (vertical white lines). The positions of the PTBP1-iCLIP1-2 crosslink clusters are indicated by dark shading, where the Y-rich motifs are shown as black or light red rectangles inside or outside of the clusters, respectively.

b) Density plot showing the normalised coverage of PTBP1-iCLIP1-2 crosslink clusters around enhanced exon (blue) and control exons (grey).

c) Density plot showing the average coverage of Y-rich motifs around enhanced (light red) and control exons (grey).

d) Table on the left shows PTBP1-iCLIP1-2 crosslink cluster enrichment and distance between enhanced and control exons for the 3' splice site, the 5' splice site and exonic region. Table on the right shows the ratio between the 3' splice site, the 5' splice site and exonic regions for enhanced and control exons.

In addition to PTBP1, I was interested in whether the RNA-map would also show splicing regulation for another protein such as hnRNPC. To answer this question, I examined published hnRNPC iCLIP data [124] with 552,440 significantly enriched clusters that I identified with the same pipeline and peak calling parameters that I used for PTBP1-iCLIP1-2 (3 nt peak calling and 10 nt clustering window size). First, I examined hnRNPC positioning within 300 nt of hnRNPC-regulated exons that were identified by JunctionSeq by using publicly available ENCODE RNA-seq hnRNPC KD data for the K562 cell line. Regulated exons were selected by having an adjusted p-value lower than 0.01 and log2 fold-change more than 1.0 (see Methods 2.2.3). Indeed, hnRNPC clusters are highly enriched upstream from the 3' splice site (Figure 4.8a, b, d), overlapping exactly over the the upstream position next to 3' splice site, where U2AF2 protein binds (Figure 4.8c) as a part of U2 snRNP and is needed for the splicing process (see subsection 1.2). Since hnRNPC mainly works as a repressor [217, 218, 124, 50], I was also interested in the hnRNPC cluster composition around enhanced exons (Figure 4.9). I found that there was no significant enrichment in the downstream region relative to the 5' splice site, as for PTBP1 (Figure 4.8a, b, d), but there is a clear drop in hnRNPC binding at the upstream region from 3' splice site where the control data shows the enrichment (Figure 4.8b). This drop could be explained by previous studies [50, 219], which demonstrated competition between U2AF65 and hnRNPC protein through a knock down of hnRNPC, which revealed a large inclusion of alu-exons that would otherwise be repressed by the hnRNPC [50, 219]. Taken together, RNA-maps can be used to understand splicing regulation of alternative spliced exons by RBPs that regulate splicing in the position-dependent manner.



d

	reg.exons.cov	control.exons.cov	enrichment	distance
3ss	1.59	0.28	5.7	1.32
exon	0.25	0.06	4.11	0.19
5ss	1.23	0.3	4.13	0.93

	regulated	control
3ss vs. 5ss	1.3	0.94
exon vs. 3ss	0.15	0.21
exon vs. 5ss	0.2	0.2
5ss vs. 3ss	0.77	1.06

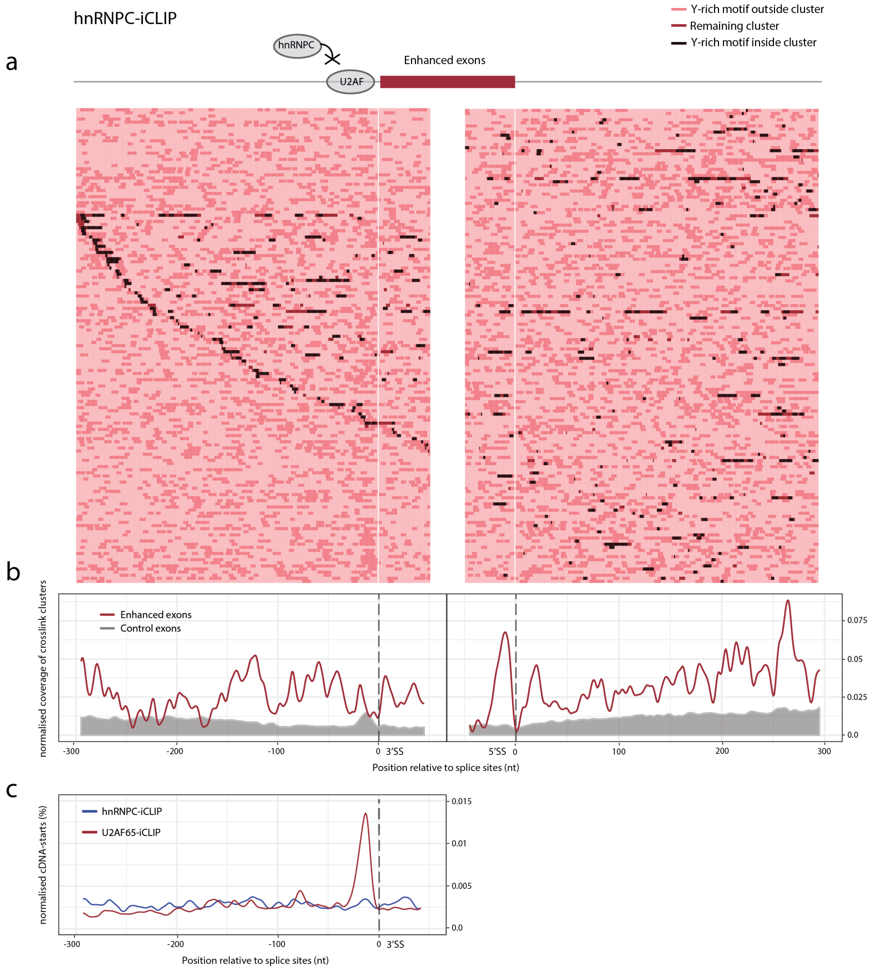
Figure 4.8: RNA-map for hnRNPC repressed exons.

a) Heatmap showing the positioning of hnRNPC-iCLIP crosslink clusters and Y-rich motifs around hnRNPC-repressed exons. All hnRNPC exons, were identified with JunctionSeq tool from RNA-seq hnRNPC KD data that is publicly available on ENCODE website. The exon flanking regions were aligned to the 3' (left) and 5' (right) splice sites (vertical white lines). The positions of the hnRNPC-iCLIP crosslink clusters are indicated by dark shading, where the Y-rich motifs are shown as black or light red rectangles inside or outside of the clusters, respectively.

b) Density plot showing the normalised coverage of hnRNPC-iCLIP crosslink clusters around repressed (blue) and control exons (grey).

c) Density plot showing the coverage of normalised cDNA-starts in the 300nt upstream region from 3' splice site for hnRNPC-iCLIP and U2AF65 protein.

d) Table on the left shows hnRNPC-iCLIP crosslink cluster enrichment and distance between repressed and control exons for the 3' splice site, the 5' splice site and exonic region. Table on the right shows the ratio between the 3' splice site, the 5' splice site and exonic regions for repressed and control exons.



d

	reg.exons.cov	control.exons.cov	enrichment	distance
3ss	0.78	0.28	2.78	0.5
exon	0.26	0.06	4.32	0.2
5ss	0.8	0.3	2.71	0.51

	regulated	control
3ss vs. 5ss	0.97	0.94
exon vs. 3ss	0.33	0.21
exon vs. 5ss	0.32	0.2
5ss vs. 3ss	1.04	1.06

Figure 4.9: RNA-map for hnRNPC enhanced exons.

a) Heatmap showing the positioning of hnRNPC-iCLIP crosslink clusters and Y-rich motifs around hnRNPC-enhanced exons. All hnRNPC exons, were identified with JunctionSeq tool from RNA-seq hnRNPC knockdown data that is publicly available on ENCODE website. The exon flanking regions were aligned to the 3’ (left) and 5’ (right) splice sites (vertical white lines). The positions of the hnRNPC-iCLIP crosslink clusters are indicated by dark shading, where the Y-rich motifs are shown as black or light red rectangles inside or outside of the clusters, respectively.
b) Density plot showing the normalised coverage of hnRNPC-iCLIP crosslink clusters around enhanced (blue) and control exons (grey).
c) Density plot showing the coverage of normalised cDNA-starts in the 300 nt upstream region from 3’ splice site for hnRNPC-iCLIP and U2AF65 protein.
d) Table on the left shows hnRNPC-iCLIP crosslink cluster enrichment and distance between enhanced and control exons for the 3’ splice site, the 5’ splice site and exonic region. Table on the right shows the ratio between the 3’ splice site, the 5’ splice site and exonic regions for enhanced and control exons.

4.7 Discussion

Methods such as iCLIP are composed of many stages, which can influence the resulting data. Here, I have examined the computational aspects of biological and technical features from different PTBP1 datasets produced by the CLIP, iCLIP, eCLIP and irCLIP method. In the past, the quality of CLIP libraries was examined by ranking the enriched motifs by their significance and comparing the top motifs with evidence from *in vitro* binding assays. I now find that even poor quality CLIP libraries can detect correct motifs, and thus this analysis is not sufficient evidence for high-quality data.

While optimised experimental conditions are essential to produce high quality protein-RNA interaction data, to some extent the quality of data can also be improved using computational filtering of non-reliable sequence reads, and statistical solutions. Removing false positives from the dataset can improve RBP specificity but also reduces the sensitivity [220]. For example, sensitivity of iCLIP data from RBPs that bind to mRNA can be increased by mapping cDNAs directly to transcriptome which in turn would improve the mapping of cDNAs that are too short to detect exon-exon junctions. However, this could also decrease its specificity by limiting the analysis only to known transcripts and missing the important bindings of other RBP targets [221].

From chapter 3, it is clear that the conditions of RNA fragmentation can affect studies of protein-RNA interactions, just as the conditions of DNA fragmentation affect the studies of protein-DNA interactions [222]. Here, I demonstrate how these constraints can be taken into account through the integrated analysis of quality checks to detect cDNAs that have strong cDNA-end constraints. The background signal is not an inherent property of the original CLIP method, but such background can be increased in methods that do not exploit the quality control step of visualising the protein-RNA complex by SDS-PAGE. Therefore, background subtraction methods developed as part of the eCLIP protocol use a mock eCLIP, potentially improving the enhanced background by removing non-specific bindings (orange rectangle in Figure 4.3).

Here, I have showed how variations in specificity features influence the binding site assignment. CLIPo evaluates different features such as complexity of the data, cDNA-end constraints, and specificity of the data as quality measures for the cDNA libraries. Complexity of the data is measured by uniquely mapped cDNAs to a single genomic position after PCR duplicate removal (see Methods 2.5). Previously I demonstrated how cDNA-end constraints can have an effect on non-coinciding cDNA-starts. CLIPo can detect these constraints if they are part of the sequence, structure constraints or cDNA length constraints. More importantly, CLIPo can measure specificity of the data by looking at the number of identified clusters and how enriched they are in top 10 sequence motifs around cDNA-starts, as well as how enriched they are for cDNA-starts. For example, the ratio of cDNA-start coverage across clusters can tell us how noisy the dataset is by using the FDR stringency of peak calling algorithms. These features from CLIPo analysis are very useful to compare the different variants of CLIP methods. I demonstrated this for PTBP1, where most of the methods showed similar specificity except the PTBP1-NarrowPeaks, which can be explained by the different ENCODE pipeline approach (Figure 2.1, see Methods 2.2.2). Another example is PTBP1-iCLIP5 data (see Table 4.1), which showed low enrichment of motifs and high cDNA-length constraints (Table 4.1). This data had been previously published as an example of a library with strong non-coinciding cDNA-starts [181]. In conclusion, CLIPo provides insights into experimental and computational features of all variants of CLIP, which need to be optimised to ensure comprehensive and unbiased assignment of the protein-RNA binding sites.

The next important step after mapping is using optimal peak calling settings (see subsection 1.6.8) to remove non-specific bindings of false positives. Optimising parameters such as peak calling window size and FDR threshold can reduce the false positive rate and provide a set of high affinity binding sites. Many cDNA reads are discarded by peak calling algorithms to increase specificity of RBP targets, though at the cost of reduced sensitivity. For example, cDNAs discarded by peak calling algorithms could also be result of lowly expressed genes. This type of correction can be improved by integrating complementary data, such as RNA-seq, [223] or additional controls for background correction such as mock eCLIP [134].

In this chapter, I demonstrated how the published ENCODE Narrow Peaks from the eCLIP method can be improved with a customised pipeline using an iCount peak calling tool that uses cDNA-start positions as crosslink sites. Moreover, most of the peak calling algorithms use a fixed window size for peak calling and clustering. Here, I examined the optimal window size by changing clustering and peak calling parameters for binding site assignment in PTBP1-iCLIP1-2 data. Measuring specificity between different CLIP-related methods or even replicates for the same method still presents a significant challenge. In part, this is a problem relating to experimental conditions, such as different cell lines or other variation in expression profiles [221]. In the CLIPo analysis I focused on PTBP1 protein, which is known to bind polypyrimidine tracks. This enabled me to use Y-rich motifs as a measurement of RBP specificity between different methods and peak calling conditions. I validated my results by PTBP1 motif enrichment, and I found that short windows spanning upstream and downstream of approximately 3 nt, together with longer windows of approximately 10 nt for clustering are the most appropriate. These parameters could be improved by observation at a single nucleotide resolution, but this is a good starting point for further analysis.

Most of the available quality control measures rely on data complexity, the number of clusters identified using a certain FDR threshold and motif discovery or the reproducibility of biological replicates [220, 224, 225]. These quality controls are important but they do not exploit known biological functions to validate specific enrichments. One way to demonstrate CLIP library specificity by biological function is using the RNA-map approach for RBPs involved in splicing regulation [33]. For this purpose, I developed a new pipeline that can generate the cluster composition around regulated exons, and calculated their enrichment compared to the unregulated control exons (see Methods 2.4.8). Using this, I could better understand how specific RBPs can regulate splicing in a position-dependent manner. Next, I was interested in whether PTBP1 clusters from optimised peak calling conditions

(PTBP1-iCLIP1-2 dataset) could reveal biological insights by using the RNA-maps approach on regulated exons that were identified by another microarray study [185]. Indeed, PTBP1 crosslink clusters show that binding upstream or downstream of the regulated exons appears to play a central role in its position-dependent manner of splicing regulation. In greater detail, I demonstrated how PTBP1 binding upstream from the 3' splice site represses the exon inclusion that can also span across the exonic region (Figure 4.6a, b). In contrast, the mechanisms of PTBP1-mediated enhancing are far less understood [61]. The PTBP1 RNA-maps also indicate that PTBP1 can directly enhance splicing when binding over the downstream region of the alternative exon, but not in the upstream or exonic region (Figure 4.7a, b). It remains to be seen whether PTBP1 assembles into higher-order complexes, potentially in combination with other RBPs, in order to regulate splicing via these binding regions. This could be done by assessing other cooperative RBPs or RBP-complexes to this study.

In a similar way, I was interested in whether RNA-maps could demonstrate how the hnRNPC protein represses exon inclusion by binding to a U-rich track at the 3' splice site and blocking U2AF65 or *vice versa* [217, 218, 124, 50]. For this purpose, I used hnRNPC-iCLIP data from the original iCLIP study [124], and analysed hnRNPC knockdown RNA-seq data from the ENCODE project to identify exons that are regulated by the hnRNPC protein. The RNA-maps showed a clear position-dependent splicing regulation of these proteins, even though the iCLIP data and controlled exons were prepared by different lab groups using methods with different cell lines (Figure 4.8 and Figure 4.9). These RNA-maps could be even clearer if the RNA-seq or microarray experiments to detect regulated exons were prepared in parallel together with the iCLIP method, or at least through using the same cell line. For example, in the hnRNPC RNA-map of repressed exons, there are still several positions lacking hnRNPC crosslink clusters at the 3' splice site: however, there is a clear enrichment of Y-rich motifs to which hnRNPC binds (Figure 4.8 and Figure 4.9) [124]. Experimental conditions, such as expression differences between cell lines, could be one of the potential explanations. Interestingly, RNA-maps showed position-dependent regulation of RNA splicing for PTBP1 and hnRNPC proteins overlapping with Y-rich motifs, especially in PTBP1-repressed exons (Figure 3.6a, c). These mechanistic insights could be interesting for predicting the regulated exons of such proteins. For example, exons with a similar pattern of Y-rich motif enrichment in the long upstream region around splice sites (Figure 4.6a, c) could classify exons as potential targets that are regulated by PTBP1, in a similar way to what has been done previously for the NOVA, hnRNPC, PTBP1 and TDP43 [39, 30]. Another example would be for hnRNPC, where I demonstrated how it can repress exon inclusion when it fully overlaps with U2AF65, binding at the 3' splice site (Figure 4.8). These are all important insights that are RBP-specific and can be used for modelling in future studies. More importantly, the RNA-maps indicate that PTBP1 or hnRNPC binding at its regulated exons represses splicing when competing with U2AF2 binding, or when it incorporates the alternative exon within a binding region, agreeing with findings from past studies [211].

At present, CLIPo mostly examines quality control measures, but in future I could include more functions to measure the specificity of RBPs. In chapter 3, I discovered that the cDNAs that end at the last intronic nucleotide are generated from RNA fragments that originated from the 3' end of intronic lariats, which are produced when introns are spliced out from pre-mRNAs. The cDNA-end peak enrichment could potentially be used as a measurement of lariat binding frequency. For example, the stronger cDNA-end peak at the last intronic nucleotide could suggest that PTBP1 more commonly remains bound to the intron lariat, while U2AF65 is released before splicing is completed (Figure 3.11a, d, e, f). This also agrees with previous studies that showed U2AF65 binds only temporarily to the pre-mRNA when the spliceosome is formed [226, 227, 228]. Another useful feature would be the ability to filter cDNAs that have strong cDNA-end constraints, by removing cDNAs that have a sharper peak at cDNA-ends rather than cDNA-starts. This would ensure that these cDNAs do not cause any misleading results in the final protein binding site assignment.

In summary, I present several technical approaches that aid the assignment of

RNA binding regions of RBPs from iCLIP data, and describe how such binding regions can provide insights into the function of PTBP1. Application of these approaches will be particularly useful for studies of RBPs from different CLIP variants and RBPs that likely play important roles in the regulation of splicing, as well as other modes of post-transcriptional regulation.

Chapter 5

Assignment of RNA binding sites for higher-order proteins complexes

5.1 Introduction

The vast majority of mammalian transcripts undergo splicing, a process through which introns are excised and respective exons adjoined (see subsection 1.2). Splicing is integral to gene expression, and via alternative splicing, it broadens the diversity of the transcriptome. An early step of the splicing process is branch point (BP) selection by the spliceosome, which defines the 3' splice site and leads to inclusion of the downstream exon in the mRNA [229]. Point mutations are commonly assumed to affect the encoded proteins in the coding gene regions, whereas mutations at BP nucleotides can result in exon skipping and aberrant splicing, which can result in disease [65]. In recent years, methods employing high-throughput sequencing have enabled high-resolution studies of the positioning of RNA Polymerase II or the ribosome in a transcriptome-wide manner [230, 231]. However, such a method is not yet available for high-resolution studies of endogenous pre-mRNA splicing reactions due to the challenges caused by the dynamic remodelling of spliceosomal interactions.

In order to understand these spliceosomal interactions, the iCLIP technique was adapted for the study of multiple proteins by applying it to the spliceosome in a so-called 'spliceosome-iCLIP'. In this protocol, the SmB protein was used as a bait to target other proteins involved in splicing. The SmB protein is a part of the Spliceosome Sm ring, a collection of 7 proteins (E, F, G, D1, D2, D3 and B/B) that form a stable assembly around the core of snRNAs that is common to all of them except for the U6 snRNP [232, 233]. Accordingly, SmB-iCLIP can pull out the different snRNPs located at different positions along an RNA transcript. By changing the lysis and wash buffers in iCLIP it is possible to capture closely associated proteins from the spliceosome to simultaneously analyse their bound RNA targets with iCLIP. Through my analysis of spliceosome-iCLIP data, I show that spliceosome-iCLIP identifies strongly enriched crosslinking at specific positions around splice sites, which can be used to distinguish interactions of multiple spliceosomal components.

Another major challenge to understanding the endogenous splicing reaction is the difficulty in assigning the position of BPs. Most of the high-throughput methods rely on lariat-spanning reads that cross from the 5' portion of the intron past the BP to the 3' portion of the intron [187, 188]. Even though such cDNA reads are present in RNA-seq data, they are very rare, and have so far not been completely identified in humans. Methods that enhance the proportion of lariats in RNA-seq can increase this number, but it requires a very deep sequencing due to the great length and low abundance of some introns [234]. CLIP-based methods are unique because they freeze interactions at the point of cross-linking. This has previously allowed rare events such as NMD exons, which are normally rapidly degraded, to be discovered. Here, I found that spliceosome-iCLIP cDNAs that end at the last dinucleotide of the intron are the ones that truncate at the BP position. I show that the medium and mild purification conditions in the iCLIP method are optimal to identify cDNAs that truncate at BPs. For this purpose, I developed a computational pipeline to detect these hybrid cDNAs which allows us to identify the BPs within most introns in expressed genes. Computational BP predictions, on the other hand, most often predict multiple potential BPs in most introns. It is unknown if the top scoring predictions represent the most commonly used BPs [235, 236].

5.2 iCLIP identifies interactions between spliceosomal proteins and snRNAs

SmB/B' proteins are part of the collection of 7 proteins which form around the core of the highly stable Sm core common to all spliceosomal snRNPs except U6 [237], making them suitable candidates for enriching snRNPs via immunopurification. The standard iCLIP protocol employs a high concentration of detergents in the lysis buffer, followed by washing buffer, which are not compatible with many protein-protein interactions, except stable complexes such as snRNPs (Figure 5.1a). In order to adapt iCLIP for the study of RBP complexes like the spliceosome, the SmB/B' proteins were immunopurified under different conditions. Therefore, a modified purification was established in the Ule lab with decreased concentration of detergents in the lysis buffer in the washing buffer (Figure 5.1a, — mild: 0.1% Igepal CA-630, 0.01% SDS, 0.05% Na-Deoxycholate — medium: 1% Igepal CA-630, 0.1% SDS, 0.5% Na-Deoxycholate — stringent: 6 Urea in M, 1% SDS). For spliceosome-iCLIP from UV-crosslinked lysates, a broad size distribution of protein-RNA complexes was isolated in order to recover the greatest possible diversity of spliceosomal protein-RNA interactions. For each purification condition of spliceosome-iCLIP, cDNA libraries with two biological replicates were prepared.

As in previous iCLIP studies [124], cDNA-starts were considered as the crosslink site and cDNAs at each crosslink site were summed as cDNA counts. The replicate datasets were highly reproducible as indicated by the observation that more than 80% of crosslinks with cDNA counts of five or more were present in all three replicates of the medium and mild experimental condition (Figure 5.1b). Under all conditions, highly enriched crosslinking was identified on all major and minor Sm-class spliceosomal snRNAs (Figure 5.1c, d). When stringent conditions were used, more than 75% of spliceosomal crosslinking mapped to snRNAs (Figure 5.1c). However, under medium or mild conditions, the proportion of snRNAs decreased to approximately 10%. For comparison, less than 0.5% of the auxiliary splicing factors TIA1 or U2AF65 crosslinked to snRNAs [183, 50]. U1cU5 snRNAs were most highly enriched by spliceosome-iCLIP, as expected by their high abundance (Figure 5.1d). In comparison to stringent spliceosome-iCLIP, the medium and mild conditions identified additional crosslink sites outside the Sm sites which indicates that mild and medium conditions of spliceosome-iCLIP can identify binding sites of multiple snRNP-associated proteins (Figure 5.1d).

On unspliced substrate RNA, spliceosomal crosslinking was observed at several positions in the vicinity of splice sites, which I marked in numerical order to simplify the analysis (Figure 5.2b, peak 1-6). To characterise the binding sites of these proteins, I plotted the density of crosslinks for each replicate (Figure 5.2b, green lines) from medium and mild conditions around splice sites for spliceosome-iCLIP, together with TIA1/TIAL1 and U2AF65 (Figure 5.2b, blue lines) produced from HeLa cells in previous studies [183, 50]. Their binding exactly corresponds to the positioning of peaks 2 and 5, respectively (Figure 5.2b), but the exact identity of other peaks remains undetermined. In conclusion, I find that we can identify spliceosomal interaction together with other associated proteins that can be distinguished using the nucleotide resolution of the iCLIP method from higher-order protein complexes.

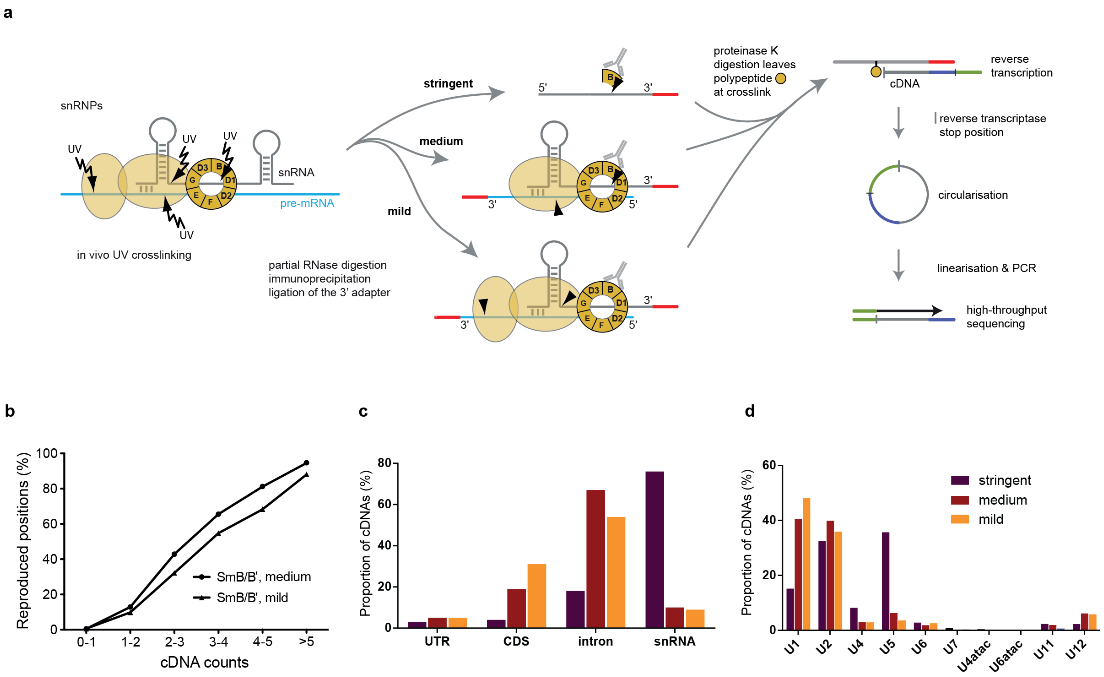


Figure 5.1: Spliceosome-iCLIP identifies the known protein-snRNA interactions.

a) Schematic representation of the spliceosome-iCLIP method performed under conditions of varying lysis stringency.
b) Percentage of crosslinks with a cDNA count value within the indicated range that were reproducible in all three replicates of the Sm iCLIP experimental groups indicated. Both single and multiple hits were considered.
c) Genomic distribution of spliceosome-iCLIP cDNAs.
d) Distribution of spliceosome-iCLIP cDNAs between different snRNAs.

5.3 Spliceosome-iCLIP can identify branch point positions genome-wide

In the chapter 3 (section about Non-coinciding cDNA-starts can result from locus-specific cDNA-end constraints), I hypothesised that the sharp peak at the last intronic nucleotide (Figure 3.11b, d, e, f) could be the result of cDNAs that were generated from RNA fragments that originated from the 3' end of intronic lariats. To test this hypothesis I investigated whether peak A and peak B (Figure 5.2b) corresponded to cDNAs truncating at the 5' splice site and the BP (Figure 5.2a). Since I found that the less stringent purification conditions were the most suitable to identify intronic cDNAs (Figure 5.1c), I used spliceosome-iCLIP under mild and medium purification conditions from Cal51 cells and grouped them to continue the investigation. According to the model showing that cDNAs truncating at peak B may also originate from intron lariats, these cDNA reads should overlap with the 3' ends of introns (Figure 5.2a) as seen before in the PTBP1 and U2AF65 experiment (Figure 3.11b, d, e, f). Moreover, rather than these cDNAs terminating at protein cross-link sites, the model implies cDNA reads could instead terminate at the 3',5'-phosphodiester linkages (Figure 5.2a). In agreement with this model, I discovered that cDNAs that terminate at the last nucleotide of endogenous introns truncate at the known sequence consensus of BPs (Figure 5.3b). In contrast, the remaining spliceosome-iCLIP cDNAs preferentially truncate at uridines (Figure 5.3c), which agrees with the high propensity of uridines for protein-RNA crosslinking [120] as well as findings from chapter 3. I used the cDNAs that end at the last nucleotide of endogenous introns to identify adenines in 35,056 introns that putatively act as BPs. Additionally, to identify more distally located candidates whose reads would not terminate at intron ends due to our 39 nt cDNA-length limit, I overlapped cDNA truncation sites with computationally predicted BPs [186] and selected the position with the highest number of truncated cDNAs as the experimentally identified BP (Figure 5.3a). This identified BP candidates in a further 15,756 introns. Thus, this collectively identified candidate BPs in 50,812 introns of the most highly expressed genes (FPKM >10 as determined by RNA-seq in Cal51 cells). Since these genes in total contain 78,894 annotated introns, I was able to identify putative BPs in 64% of introns in expressed genes.

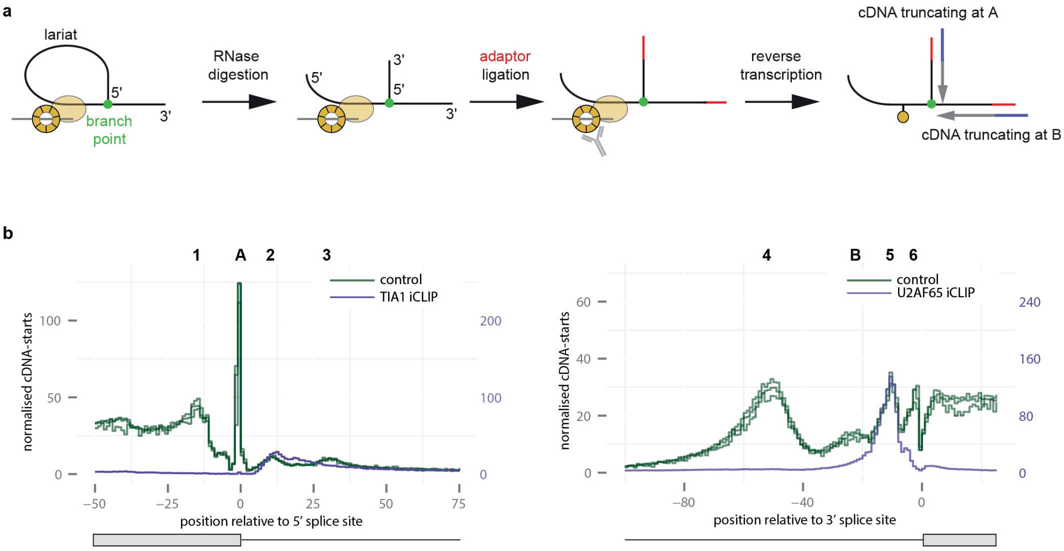


Figure 5.2: Analysis of splicesomal interactions with pre-mRNAs *in vivo*.

a) Schematic description of the three-way junctions of intron lariats, which would produce cDNAs that truncate at peak A or B on the RNA maps a) and b). These cDNAs initiate from the end of the intron and truncate at the BP (peak B), or downstream of the 5' splice site and truncate at the first nucleotide of the intron (peak A). The three-way junction is produced after limited RNase I digestion of intron lariats, followed by ligation of the L3 adaptor to the two 3' ends of the three-way junction. This leads to cDNAs corresponding to peaks A and B that do not truncate at sites of protein-RNA crosslinking, but rather at the three-way junction of intron lariats.

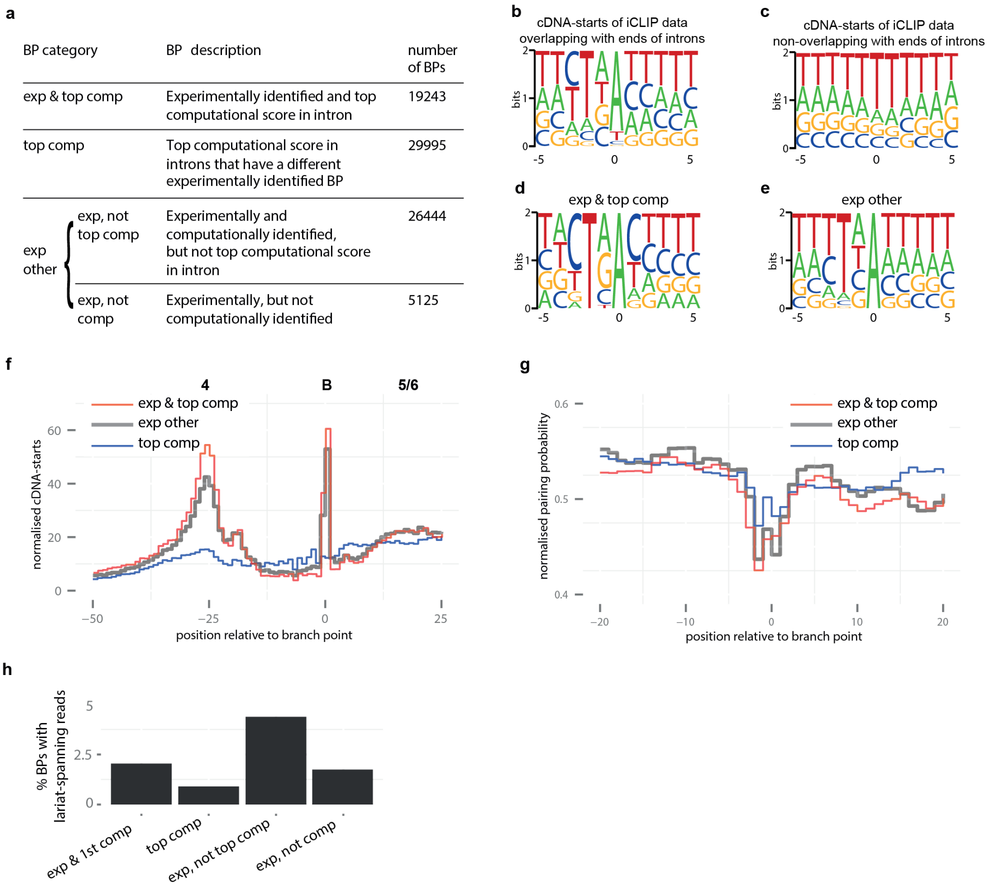
b) RNA-map of summarised crosslinking of spliceosome-iCLIP from mild and medium conditions of 4 replicates (green lines) at all exon-intron and intron-exon boundaries of spliceosome-iCLIP from Cal51 cells. For comparison, crosslinking of TIA1 or U2AF65 is also shown, and the scale for the normalised TIA1 or U2AF65 data is shown on the right of the respective panel.

To facilitate further analysis, I broke defined different categories of BP based on their presence in experimental and/or computational data. These categories were: experimentally identified and top computational score, top computational score that have a different experimentally identified BP, and two other categories that are neither part of top experimentally or computationally BPs (Figure 5.3a). Surprisingly, only 38% of the experimentally identified BPs overlapped with the top-scoring computationally predicted BPs in the same introns [235] (exp & top comp, 19,243 BPs, Figure 5.3a). The remaining experimental BPs (exp other, 31,569 BPs, Figure 5.3a) had smaller enrichment of C at the -3 and +1 position, and T at the -2 position (Figure 5.3e). Furthermore, a subset of these sites lacked uridine at the position two nucleotides upstream of the BP, and therefore these did not overlap with any computational BP (exp, not comp, 5,125 BPs, Figure 5.3a). This represents a potentially new and sizeable category of BPs that computational approaches have not yet identified.

I was also interested in whether I could validate BP groups, by using RNA-seq data and spliceosome-iCLIP cDNAs in a manner that can identify lariat-spanning reads that cross the BP from the 5' to the 3' region of the intron. This type of analysis has previously been successful in RNA-seq data but the lariat-spanning reads are very rare, and have so far identified less than 1000 BPs in humans [238, 239]. I found very few lariat-spanning cDNAs present in RNA-seq, but many such cDNAs present in spliceosome-iCLIP. In all categories of experimentally defined BPs, more than 2% of BPs could also be identified by the lariat-spanning reads, regardless of whether they overlapped with a computational BP or not (Figure 5.3h). In contrast, less than 1% of other computationally top-scoring BPs were identified by lariat-spanning reads (Figure 5.3h - top comp).

Next, I examined spliceosome-iCLIP crosslinking around the different BPs categories (Figure 5.3f) where I found that the crosslinking peak 4 (Figure 5.3b) is aligned to position 25 nt upstream of the BP that were experimentally identified (Figure 5.3f). Since this peak was not used to identify BPs, it could be used for independent validation of different BP categories. Only experimental, but not computationally predicted BPs, are preceded by peak 4 (Figure 5.3f). This suggests that spliceosomal complexes bind at the experimentally determined BPs, but not at other computationally top-scoring BPs. Together, these results demonstrate that splicesome iCLIP can identify valid BPs in most introns of expressed genes.

In order to understand the discrepancy between the experimental and computationally predicted BPs better, I evaluated the structure (see Methods 2.4.6) around the putative BPs. Experimental BPs had a low intramolecular pairing probability regardless of their computational score, whereas computationally top-scoring BPs had higher pairing probability, demonstrating that experimental BPs are single-stranded and therefore accessible (Figure 5.3g). Strikingly, the majority of computationally top-scoring BPs that do not overlap with the experimental BPs (top comp) appear to be in structurally inaccessible conformation. This indicates that the accuracy of computational predictions could be increased by taking into account the analysis of RNA secondary structure at BPs.



BP category		BP description	number of BPs
exp & top comp		Experimentally identified and top computational score in intron	19243
top comp		Top computational score in introns that have a different experimentally identified BP	29995
exp other	exp, not top comp	Experimentally and computationally identified, but not top computational score in intron	26444
	exp, not comp	Experimentally, but not computationally identified	5125

Figure 5.3: Comparison of experimentally and computationally identified BPs.

a) A table explaining the different categories of BPs, and the total number of these BPs identified in highly expressed genes (FPKM >10).
b) The composition of genomic nucleotides around the nucleotide preceding all spliceosome-iCLIP cDNA-starts that overlap with ends of introns.
c) The composition of genomic nucleotides around the nucleotide preceding all spliceosome-iCLIP reads that do not overlap with ends of introns.
d) The composition of genomic nucleotides of BPs that were experimentally identified, and also have the highest score of all computationally identified BPs in the same intron (exp & top comp).
e) The composition of genomic nucleotides of BPs that were experimentally identified, and do not have the highest score of all computationally identified BPs in the same intron (exp).
f) RNA-map of summarised crosslinking around the three categories of BPs, as defined in a).
g) Pairing probability was calculated at each nucleotide around BP using RNAfold program with the default parameters. Average pairing probability was then calculated for the two groups described in a) the computationally top-scoring BPs that were not experimentally identified (comp).
h) Percentage of the above BPs categories which have accompanying lariat spanning reads in spliceosome-iCLIP and RNA-seq.

5.4 The effect of branch point position on spliceosomal interactions

It is not yet fully understood how the position of BPs on endogenous introns determines spliceosomal interactions. I examined the effect of BP position by dividing 3' splice sites into three categories, depending on the distance between the BP and the 3' splice site (Figure 5.4 - distance groups of 17-21 nt, 22-35 nt, 36-100 nt). The spliceosome-iCLIP replicates from mild and medium conditions were highly reproducible (Figure 5.1b, Figure 5.2b, green lines), therefore they were grouped for the remaining analysis in order to maintain high crosslink coverage for each distance group. U2AF65 crosslinking peaks were at similar positions upstream of the 3' splice site at all three categories, with a slight shift upstream for the more distally located BPs, demonstrating that U2AF65 binding is generally independent of the BP position (Figure 5.4c, d, f). Nevertheless, introns with the distal BPs had broader U2AF65 crosslinking (Figure 5.4f). More importantly, the position of peak 4 is precisely determined by the position of the BP: regardless of how far the BP is positioned from the 3' splice site, peak 4 is invariably present 25 nt upstream of the BP, with an additional position with enriched crosslinking at 19 nt upstream from the BP (Figure 5.4a, c, e). This indicates that peak 4 may reflect crosslinking of two proteins or two domains of a protein. I used the DREME motif discovery tool to examine potential sequence-specificity of proteins that bind at these positions by comparing the sequence of the region between 30 and 20 nt upstream of the BP to the sequence of the deep intronic region further than 100 nt from the splice sites, but I was not able to identify any enriched sequence motifs that could potentially identify the specificity of the 25 nt upstream peak (data not shown). This demonstrates that RNA contacts at peak 4 upstream of the BP are determined solely by the position of the BP, rather than the sequence specificity that could be masked by secondary structure or interact with other spliceosomal complexes. Taken together, the BP-dependent position of peak 4 further indicates that it likely represents contacts of late spliceosomal components that are involved in BP recognition, although its identity remains unresolved.

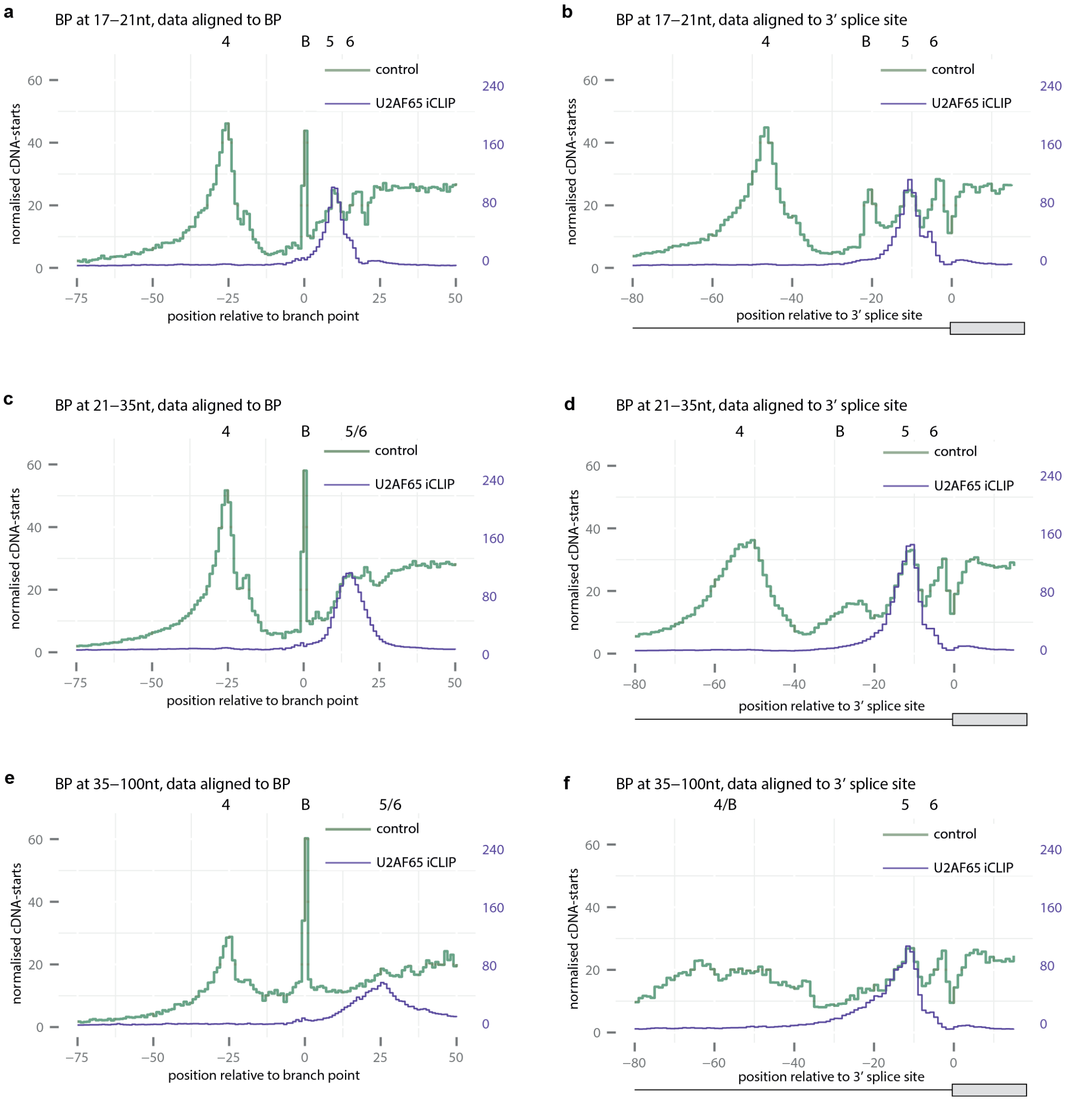


Figure 5.4: The effect of BP position on spliceosomal interactions.

RNA-map of summarised cDNA-starts around BPs a), c), e) and intron-exon junctions b), d), f). BPs were divided into the following categories based on their distance from the 3’ splice site: a), b) 17-21 nt, c), d) 22-35 nt, e), f) 36-100 nt. The scale for the spliceosome-iCLIP data is on the left, and U2AF65-iCLIP data is shown on the right of each graph.

5.5 Identification of 25 nt upstream peak relative to branch points by using ENODE eCLIP dataset

By analysing spliceosome-iCLIP data, I was able to identify over 50,000 BPs genome-wide in expressed genes and more importantly, I identified that the position of peak 4 (Figure 5.2b) is precisely determined by the position of the BP in the surrounding region of 25 nt upstream from the identified BPs (Figure 5.4a, c, e). First, I tried to identify the protein that might interact with that region by using publicly available iCLIP data and the DREME tool for motif discovery (data not shown). With both approaches I could not identify any protein enrichment or a motif that would be significantly enriched in that region. To continue with the investigation, I decided to use the large eCLIP dataset from the ENCODE project to see if any of their proteins are enriched in that region. I systematically analysed the eCLIP data of 140 samples from 70 different proteins in the HepG2 cell line and 178 samples from 89 different proteins in the K562 cell line by using customised pipeline (Figure 2.1, see Methods 2.2.2). Next, I intersected cDNA-starts from each sample to the -50 to -10 nt upstream region away from the BPs, where the 4th peak showed the highest enrichment of cDNA-starts in the spliceosome-iCLIP (Figure 5.4a, c, e). Surprisingly, the eCLIP dataset revealed a candidate, with a more than 2-fold enrichment of the SF3B4 protein overlapping peak 4 from both replicates and in both cell lines (orange in Figure 5.5a, b). SF3B4 is a subunit of the splicing factor 3b protein complex, which is a multi-protein complex that forms the U2 snRNP together with other units and is essential for the splicing process [240]. It is also known that the SF3b subunit binds to the pre-mRNA near the BP to reinforce U2 snRNP [241, 242, 243, 244] and plays a key role in BP recognition during constitutive and alternative splicing [245, 246, 247]. There is also enrichment of the SF3A3 protein (red in Figure 5.5a) in the HepG2 cell line, which is another subunit of splicing factor 3 that interacts with the same U2 snRNP. Besides the splicing factor associated proteins, another enriched protein SMNDC1 was detected in the K562 cell line (purple in Figure 5.5b), known as the survival motor neuron domain-containing 1 protein, associated with autosomal recessive proximal spinal muscular atrophy and already identified as a constituent part of the spliceosome complex [248, 249]. Non-consistent enrichment of SMNDC1 protein between these two cell lines could be because of their differences in expression level. Taken together, this demonstrated how the eCLIP dataset can be used to identify targets genome-wide, and also revealed proteins from peak 4 that are already known to interact with BPs. However, this approach is another confirmation that the identified BPs are valid.

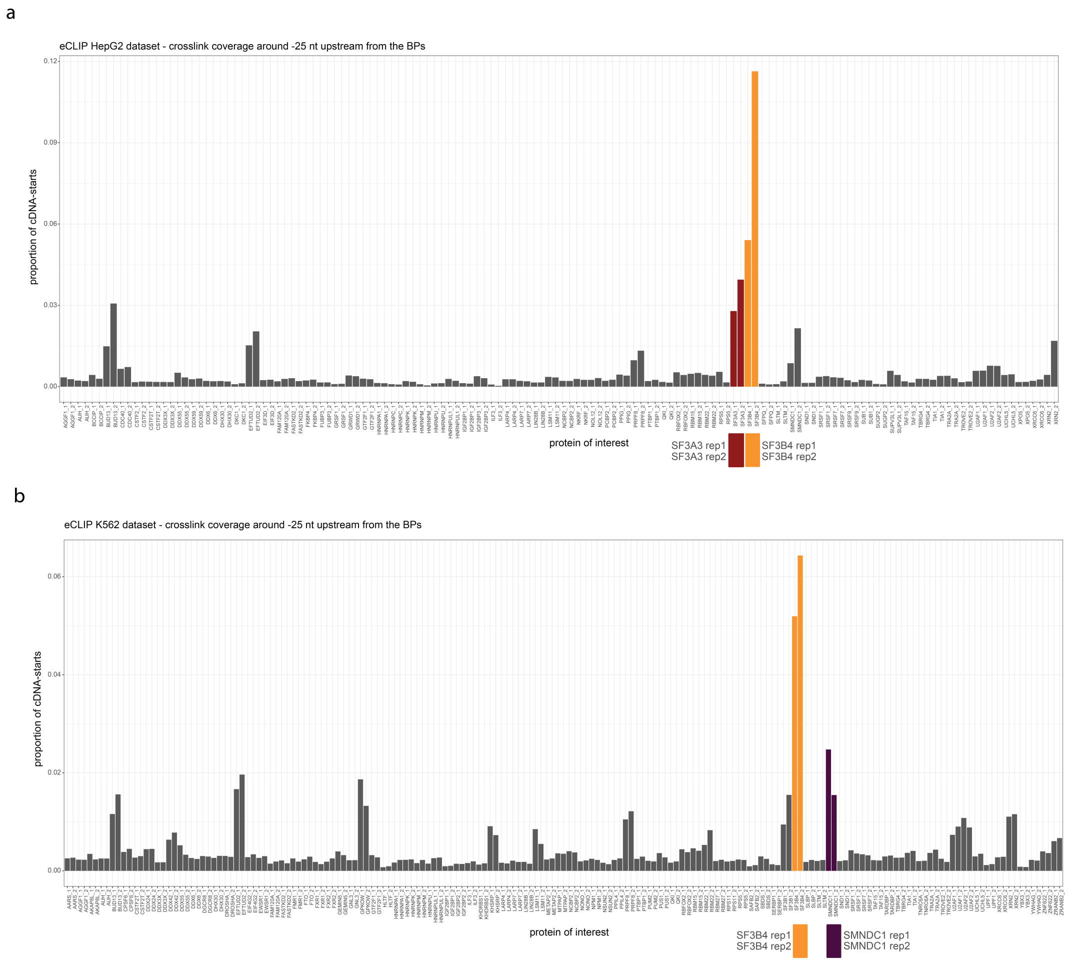


Figure 5.5: Proportions of cDNA-starts from eCLIP dataset around 25 nt peak upstream from the BPs genome-wide.

a) Each barplot represents its own replicate of 70 different proteins from the eCLIP HepG2 dataset that intersect with -10 to -50 nt upstream region away from BPs. Both replicates from the SF3B4 protein (in orange) are showing the highest enrichment followed by SF3A3 (in red) protein replicates. b) Each barplot represents its own replicate of 89 different proteins from the eCLIP K562 dataset that intersect with -10 to -50 nt upstream region away from BPs. Both replicates from the SF3B4 protein (in orange) are showing the highest enrichment followed by SMNDC1 (in purple) protein replicates.

5.6 Discussion

In this chapter, I assessed the features of the spliceosome-iCLIP method that was developed in the Ule lab. Spliceosome-iCLIP uses Sm core proteins as a bait to purify endogenous snRNPs and associated proteins. This was possible due to the high strength of protein-protein interactions within snRNPs as well as the mild purification conditions that were used, which preserved interactions between snRNPs and other associated proteins. Spliceosome-iCLIP identified known and novel crosslink sites on snRNAs, both within and outside the Sm site, indicating that the method can be used to study snRNAs and potentially other non-coding RNAs that may play a role in splicing. In addition, enriched crosslinking was identified in pre-mRNAs at defined positions around the splice sites and BPs, which correspond to the positions where snRNPs and associated splicing factors typically bind. I characterised two major peaks that overlapped with TIA1/TIAL1 and U2AF65 iCLIP data (Figure 5.2b peak 2 and 5). The identity of the remaining peaks is presently unclear. However, it is possible that peaks 2, 5 and 6 corresponds to auxiliary factors that can be co-purified with U1 and U2 snRNP in the complex E [226, 250, 251].

I next presented how cDNA-starts from spliceosome-iCLIP data that overlap with the ends of introns can be used as a new feature to identify BP positions in most human introns. Both the number of identified BPs, and their strong consensus sequence, compare favourably to previous methods that were based on analysis of lariat reads in RNA-seq. For identified BPs, I used different methods to show that their validity was comparable to the BPs that overlapped with the computationally top-scoring prediction.

It is also known that the majority of BPs are positioned in the region between 19 and 35 nt upstream from the 3' splice site [239]. One limitation of the current spliceosome-iCLIP experiment is that only 50 cycles of Illumina sequencing were used, which limits the BP detection in the region between 17 nt (the minimum cDNA length used for mapping) and 40 nt (the longest cDNA length after adapter and barcode removal, see Methods 1). To overcome the problem of identifying BPs that are located near 3' splice sites, I could lower the minimum length of the 17 nt threshold for mapping sequences and increase the number of multiple-hits from the current single-hit approach (see Methods 3). It would be important to analyse these short sequences separately, as they often map to multiple genomic positions. The BP discovery method should prove a useful starting point for future studies of splicing machinery.

Here, I focused on the spliceosome-iCLIP data for BP identification but this method could also be applied to other RBPs; for example PTBP1, where the cDNA-end peak at the 3' splice site (Figure 3.11d, e, f) could be a part of lariat cDNAs. This suggests that other datasets of relevant RBPs could also be applied to BP discovery genome-wide. Furthermore, including datasets with a better representation of long cDNAs, such as eCLIP dataset from ENCODE, which is generated using paired-end sequencing which allows up to 120 nt long cDNAs [134]. These long cDNAs could potentially identify more distal BPs which are known to be involved in exon skipping events [235, 239, 252].

It is not surprising that, given the broad importance of splicing regulation for cellular differentiation, a number of diseases are caused by mutations in components of the splicing machinery [253, 254]. Another challenge for the future would be to explore genomic variations around BPs systematically for a better understanding of how splicing-associated mutations can lead to disease.

During the BP analysis, I additionally identified a strong peak of spliceosomal crosslinking at 25 nt upstream (Figure 5.4a, c, e - peak 4) of experimentally derived BPs, which could be a part of the SF3 proteins that are known to bind at the 'anchoring site' upstream of the BPs [241]. To follow up this hypothesis I used the large ENCODE eCLIP dataset to show that peak 4 (25 nt upstream from the BPs) is interacting with SF3 proteins. I validated these results by using two different cell lines from the eCLIP data and presented it as a useful approach for future studies to identify potential RBP targets.

iCLIP has only been previously used for the study of single proteins. Recently, interactome capture was developed to study all RBPs interacting with mRNAs [255]. Since this approach isolates the whole compendium of RBPs, the sequence reads distribute over the whole length of mRNAs, and crosslink patterns of specific proteins cannot be easily distinguished [256]. Accordingly, interactome capture has been used primarily to identify RBPs, rather than their binding sites [255, 8, 257]. Nevertheless, high-throughput analysis of RNA interactions of multi-protein complexes is crucial in order to understand the dynamic assembly of such complexes on target RNAs. As a proof-of-principle method for the study of multi-protein complexes, I showed that spliceosome-iCLIP can delineate the crosslinking positions of spliceosomal complexes on endogenous transcripts at high resolution and in a transcriptome-wide manner. As others have shown, it is possible to distinguish differences between 5' and 3' splice sites binding in different conditions [258]. The ability of spliceosome-iCLIP to monitor the concerted pre-mRNA binding of spliceosomal proteins indicates that the method could also be readily applied to the study of other multi-protein RNA-binding complexes.

Chapter 6

Conclusion

In the first part of this thesis, I explored how variations in CLIP and iCLIP-related methods affect the assignment of protein-RNA binding sites. I reject the previously postulated hypothesis that highly prevalent readthrough cDNAs would explain the presence of non-coinciding cDNA-starts in iCLIP cDNA libraries, and show that the use of cDNA-starts is appropriate. Moreover, I found that the non-coinciding cDNA-starts are caused by constrained cDNA-ends, which result from the RNA sequence and structure constraints of RNase cleavage sites. These have an effect in particular on the assignment of long binding sites. These constraints can be overcome by optimizing iCLIP conditions and library preparation by RNase fragmentation, RNA ligation and cDNA purification, and by ensuring the recommended purification of protein-RNA complexes and cDNAs [126].

While showing that cDNA-starts are appropriate to assign crosslink sites in iCLIP, I also found that cDNA-ends can be informative for specific purposes. I exploited analysis of cDNA-ends in order to capture novel insights in RNA processing, showing how the sequence and structure of a nascent mRNA can lead to preferential sites of RNase fragmentation, which can lead to suboptimal assignment of the protein-RNA binding site. I developed computational approaches to visualise the impact of these features on the sequenced cDNA libraries, which helps to interpret the assigned binding sites correctly. These considerations apply to all protocols that amplify truncated cDNAs, including iCLIP, eCLIP and irCLIP, and they ensure that cDNA-starts comprehensively identify protein-RNA crosslink sites across the transcriptome.

In the second part of my thesis I developed a set of pipelines named ’CLIPo’, which use the findings of the third chapter to perform quality control analyses. The main function of CLIPo is to establish cDNA-end constraints, together with protein specificity and library complexity from the data produced by iCLIP and other related methods. These cDNA-end constraints can be recognized by analysis of their secondary structure with motif analysis as a quality control measure. CLIPo complements the findings from chapter 3, which I demonstrated by focusing on the PTBP1 protein. This approach can now be applied to other RBPs and any iCLIP-related methods.

In the last part of my thesis, I presented a new spliceosome-iCLIP method developed in the Ule lab, that identifies the positioning of spliceosomal complexes at nucleotide resolution in a transcriptome-wide manner. Furthermore, I demonstrated how cDNA-end constraints that are the result of intron-exon cleavage sites can be used as a new feature to identify BP positions in most human introns.

6.1 Future directions for integrating quality control into machine learning algorithms

How the RBPs recognize the target RNAs and why they bind to specific positions is still unclear. Taking all of these quality controls into consideration, machine learning could have great potential to aid the prediction of RBP interactions. Progress has been made in the prediction of RBP binding sites for several RBPs using computational prediction models. To date this has only really been possible for RBPs with strong binding patterns, such as sequence and structure specificity. Recent studies, have now shown great potential in the field through the usage of matrix factorization and multiple-kernel learning, which encodes multiple features to predict RBPs binding characteristics [259, 34, 35, 36]. This type of modelling allows integration of multiple factors in order to identify discriminative non-overlapping, class-specific RNA binding patterns of different strengths [35]. With the integration of multiple biological features across large datasets, such as ENCODE, these prediction models could be rapidly improved.

Meta-analysis and prediction models to study RBPs are becoming more and more popular, as they integrate large datasets of CLIP related methods and incorporate available sources of information such as RNA sequence and structure to model protein-RNA interactions [260, 261, 262]. These types of analysis rarely lead to specific biological insights; a major reason being the limited effort to disentangle the effects of technical noise from biological information in CLIP data. Filtering or more careful assessment of samples that are of a low quality, as well as better understanding of the variation between them, is crucial for modelling. However, computational approaches for quality control of CLIP data that is used for the meta-analyses have not yet been developed and implemented. Therefore, I have assessed the value of assessing cDNA-end structure and sequence constraints, cDNA-length constraints, data complexity, motif enrichment, and noise measurement for quality controls, as well as their validating specificity of the data with the use of RNA-maps.

In future studies, these quality controls could be included when using larger datasets to better understand other technical and biological insights. For example, how much might the choice of cell line affect the composition of binding sites? The ENCODE consortium has already produced a large number of eCLIP data for two cell lines. It would be important to include RBP clusters from different cell lines to determine missing binding sites that are results of lowly or not expressed genes in a certain cell type. Other technical artefacts that should be considered include technical batch effects of each experiment. A method such as iCLIP is a complex multi-step technique, and there are variations in how each laboratory and person produces data each time as shown in the datasets analysed in chapters 3 and 4 (eIF4A3-iCLIP1 in Figure 3.16a and PTBP1-iCLIP5 in Table 4.1). Variations in experiments produced by the same lab, suggest that these cDNA length constraints could result from a technical batch effect due to their library preparation. Another way to test this is by analysing RBPs produced by CLIP-related methods that cluster together, despite these RBPs having a very distinct motif specificity. This would indicate that the batch effect is a stronger determinant of binding variation than biological recognition of the same binding sites by RBPs. To assess this batch effect more systematically, it is valuable to define clusters of specific RBPs based on multiple datasets, especially those produced by different methods. However, there are additional types of technical or biological artefacts across CLIP-related experiments, such as variations in gene expression or batch effects that are produced by a specific lab. Therefore, it is crucial to consider these artefacts before we can start modelling RBP interactions.

6.2 Cooperative binding of RBPs to non-optimal binding sites

Crosslinking studies indicate that over 1,500 potential RBPs might be encoded by the human genome [7, 8], and most of these RBPs cooperate or compete with each other for their binding sites [263]. Studying the interactions and cooperative bindings between RBPs using an experimental approach is very expensive and time consuming, since one needs to design and integrate multiple conditions of pre-selected targets. Computational analysis and prediction of these interactions is therefore critical to gain a comprehensive understanding of RBP functions [264]. More importantly, these modelling methods could potentially discover cooperative interactions between RBPs at the protein level, so as to identify targets that are using non-optimal binding sites to stabilise their proximal binding site [265].

An advantage of cooperative binding to non-optimal binding sites could be to increase efficiency and specificity of RBPs [265]. The full functional potential will only be achieved with both binding partners present, decreasing the 'off-target effects' of both proteins on RNAs that contain similar sequences where no functional binding events are needed. It is possible that the cooperative binding between RBPs is used to refine their target spectrum in healthy physiology, but is detrimental in the context of disease, since it predisposes them for aggregation with each other. For example, factors that are associated with ALS accumulate in distinct foci that are phenotypically similar to stress granules formed by multiple RBPs [266]. There are groups of RBPs that have a major role in neurological diseases such as ALS and

FTD. Both have been associated with a group of RBPs including TDP43, Matr3, and hnRNPA2. These RBPs have already been associated with common diseases, and they also correlate to changes in gene expression, where they bind in proximity to one another. By applying machine learning towards cooperative binding of RBPs to non-optimal binding sites, we could improve understanding of their function. Methods such as spliceosome-iCLIP also have a great potential to be applied to the study of other RBP-complexes to better understand cooperative bindings of RBPs. Alongside an experimental approach, there is a great advantage of identifying RBP targets with large datasets in a similar way as I demonstrated in chapter 4, where I overlaid over 150 eCLIP samples to identify new targets that are involved in spliceosomal interactions. This approach could be used across other RBP studies, especially for discovery of new cooperative targets.

6.3 Final thoughts

In recent years, several variants of iCLIP methods have been developed, including FAST-iCLIP, eCLIP, miCLIP, hiCLIP and irCLIP. Each of these variants has features that have not yet been systematically analysed. At the moment there are no tools available to measure the quality of the data in depth. It is therefore crucial to first understand technical differences between the methods and experimental conditions, in order to separate technical and biological variability. The tools I have generated in this thesis will be influential in this regard, as I have shown how it can comprehensively quality control CLIP datasets from various techniques to identify suitable datasets for further analysis. Beyond this, there is still a great need for additional experimental and computational methods to validate results produced by these methods. Developing methods to generate RNA-maps is a good example of integrating different sets of experimental data with computational methods to learn about splicing regulation. In this regard, I have developed new tools for summarisation of CLIP data in RNA maps which have been used throughout this thesis and which will be useful to CLIP biologists in future. Lastly, due to the nucleotide resolution of iCLIP, I demonstrated here that it can also be used to study a complex of proteins that binds to distinct positions on pre-mRNAs by using spliceosome-iCLIP. This provided a new transcriptome-wide view of the spliceosome in action which led to new insights into branch point usage. All the computational methods for data analysis and data visualisation applied in this thesis are now available on the GitHub repository and can be applied to different datasets in the future. Taken together, this thesis will therefore enable us to better understand iCLIP data, and should be useful for future studies of any other iCLIP-related method.

Bibliography

[1] Francis Crick. Central Dogma of Molecular Biology. *Nature*, 227(5258):561–563, aug 1970.

[2] Patricia Hilleren, Terri McCarthy, Michael Rosbash, Roy Parker, and Torben Heick Jensen. Quality control of mRNA 3′-end processing is linked to the nuclear exosome. *Nature*, 413(6855):538–542, oct 2001.

[3] Andrea Kyburz, Arno Friedlein, Hanno Langen, and Walter Keller. Direct Interactions between Subunits of CPSF and the U2 snRNP Contribute to the Coupling of Pre-mRNA 3′ End Processing and Splicing. *Molecular Cell*, 23(2):195–205, jul 2006.

[4] Stefania Millevoi, Clarisse Loulergue, Sabine Dettwiler, Sarah Zeneb Karaa, Walter Keller, Michael Antoniou, and Stéphan Vagner. An interaction between U2AF 65 and CF Im links the splicing and 3′ end processing machineries. *The EMBO Journal*, 25(20):4854–4864, oct 2006.

[5] Frank Rigo and Harold G. Martinson. Functional Coupling of Last-Intron Splicing and 3'-End Processing to Transcription In Vitro: the Poly(A) Signal Couples to Splicing before Committing to Cleavage. *Molecular and Cellular Biology*, 28(2):849–862, oct 2007.

[6] Julian König, Kathi Zarnack, Nicholas M. Luscombe, and Jernej Ule. Protein–RNA interactions: new genomic technologies and perspectives. *Nature Reviews Genetics*, jan 2012.

[7] Yaseswini Neelamraju, Seyedsasan Hashemikhabir, and Sarath Chandra Janga. The human RBPome: From genes and proteins to human disease. *Journal of Proteomics*, 127:61–70, sep 2015.

[8] Alfredo Castello, Bernd Fischer, Katrin Eichelbaum, Rastislav Horos, Benedikt M Beckmann, Claudia Strein, Norman E Davey, David T Humphreys, Thomas Preiss, Lars M Steinmetz, Jeroen Krijgsveld, and Matthias W Hentze. Insights into RNA Biology from an Atlas of Mammalian mRNA-Binding Proteins. *Cell*, 149(6):1393–1406, jun 2012.

[9] Stefanie Gerstberger, Markus Hafner, and Thomas Tuschl. A census of human RNA-binding proteins. *Nature Reviews Genetics*, 15(12):829–845, nov 2014.

[10] S. D. Auweter, F. C. Oberstrass, and F. H.-T. Allain. Sequence-specific binding of single-stranded RNA: is there a code for recognition? *Nucleic Acids Research*, 34(17):4943–4959, sep 2006.

[11] Maria Gutierrez-Arcelus, Halit Ongen, Tuuli Lappalainen, Stephen B. Montgomery, Alfonso Buil, Alisa Yurovsky, Julien Bryois, Ismael Padioleau, Luciana Romano, Alexandra Planchon, Emilie Falconnet, Deborah Bielser, Maryline Gagnebin, Thomas Giger, Christelle Borel, Audrey Letourneau, Periklis Makrythanasis, Michel Guipponi, Corinne Gehrig, Stylianos E. Antonarakis, and Emmanouil T. Dermitzakis. Tissue-Specific Effects of Genetic and Epigenetic Variation on Gene Regulation and Splicing. *PLoS Genet*, 11(1):e1004958, jan 2015.

[12] Mee-Young Kim, Jung Hur, and Sun-Joo Jeong. Emerging roles of RNA and RNA-binding protein network in cancer cells. *BMB Reports*, 42(3):125–130, mar 2009.

[13] Claudia Racca, Alejandra Gardiol, Taesun Eom, Jernej Ule, Antoine Triller, and Robert B. Darnell. The Neuronal Splicing Factor Nova Co-Localizes with Target RNAs in the Dendrite. *Front Neural Circuits*, 4:5, Mar 2010.

[14] Weirui Guo, Yanbo Chen, Xiaohong Zhou, Amar Kar, Payal Ray, Xiaoping Chen, Elizabeth J Rao, Mengxue Yang, Haihong Ye, Li Zhu, Jianghong Liu, Meng Xu, Yanlian Yang, Chen Wang, David Zhang, Eileen H Bigio, Marsel Mesulam, Yan Shen, Qi Xu, Kazuo Fushimi, and Jane Y Wu. An ALS-associated mutation affecting TDP-43 enhances protein aggregation fibril formation and neurotoxicity. *Nature Structural & Molecular Biology*, 18(7):822–830, jun 2011.

[15] Hua Lin Zhou, Marie Mangelsdorf, JiangHong Liu, Li Zhu, and Jane Y Wu. RNA-binding proteins in neurological diseases. *Science China Life Sciences*, 57(4):432–444, mar 2014.

[16] Julia K. Nussbacher, Ranjan Batra, Clotilde Lagier-Tourenne, and Gene W. Yeo. RNA-binding proteins in neurodegeneration: Seq and you shall receive. *Trends in Neurosciences*, 38(4):226–236, apr 2015.

[17] Katannya Kapeli and Gene W. Yeo. Genome-Wide Approaches to Dissect the Roles of RNA Binding Proteins in Translational Control: Implications for Neurological Diseases. *Frontiers in Neuroscience*, 6, 2012.

[18] Epaminondas Doxakis. RNA binding proteins: a common denominator of neuronal function and dysfunction. *Neuroscience Bulletin*, 30(4):610–626, jun 2014.

[19] Kiven E. Lukong, Kai wei Chang, Edouard W. Khandjian, and Stéphane Richard. RNA-binding proteins in human genetic disease. *Trends in Genetics*, 24(8):416–425, aug 2008.

[20] Alger Fredericks, Kamil Cygan, Brian Brown, and William Fairbrother. RNA-Binding Proteins: Splicing Factors and Disease. *Biomolecules*, 5(2):893–909, may 2015.

[21] Alfredo Castello, Bernd Fischer, Matthias W. Hentze, and Thomas Preiss. RNA-binding proteins in Mendelian disease. *Trends in Genetics*, 29(5):318–327, may 2013.

[22] C. Burd and G Dreyfuss. Conserved structures and diversity of functions of RNA-binding proteins. *Science*, 265(5172):615–621, jul 1994.

[23] Ivica Letunic, Tobias Doerks, and Peer Bork. SMART 6: recent updates and new developments. *Nucleic Acids Research*, 37(Database):D229–D232, jan 2009.

[24] Minna-Liisa Änkö and Karla M. Neugebauer. RNA–protein interactions in vivo: global gets specific. *Trends in Biochemical Sciences*, 37(7):255–262, jul 2012.

[25] Ann L. Beyer, Mark E. Christensen, Barbara W. Walker, and Wallace M. LeStourgeon. Identification and characterization of the packaging proteins of core 40S hnRNP particles. *Cell*, 11:127–38, May 1977.

[26] Piol-Roma S and Dreyfuss G. hnRNP proteins: localization and transport between the nucleus and the cytoplasm. *Trends Cell Biol*, 3:151–5, May 1993.

[27] Emad Bahrami-Samani, Luiz O.F. Penalva, Andrew D. Smith, and Philip J. Uren. Leveraging cross-link modification events in CLIP-seq for motif discovery. *Nucleic Acids Res*, 43:95–103, Jan 2015.

[28] Glen N. Barber. The NFAR's (Nuclear Factors Associated with dsRNA): Evolutionarily conserved members of the dsRNA binding protein family. *RNA Biology*, 6(1):35–39, jan 2009.

[29] Yoichiro Sugimoto, Alessandra Vigilante, Elodie Darbo, Alexandra Zirra, Cristina Militti, Andrea D'Ambrogio, Nicholas M. Luscombe, and Jernej Ule. hiCLIP reveals the in vivo atlas of mRNA secondary structures recognized by Staufen 1. *Nature*, 519(7544):491–494, mar 2015.

[30] Matteo Cereda, Uberto Pozzoli, Gregor Rot, Peter Juvan, Anthony Schweitzer, Tyson Clark, and Jernej Ule. RNAmotifs: prediction of mul-

tivalent RNA motifs that control alternative splicing. *Genome Biology*, 15(1):R20, 2014.

[31] I. Paz, I. Kosti, M. Ares, M. Cline, and Y. Mandel-Gutfreund. RBPmap: a web server for mapping binding sites of RNA-binding proteins. *Nucleic Acids Research*, 42(W1):W361–W367, may 2014.

[32] C. Zhang, K.-Y. Lee, M. S. Swanson, and R. B. Darnell. Prediction of clustered RNA-binding protein motif sites in the mammalian genome. *Nucleic Acids Research*, 41(14):6793–6807, may 2013.

[33] Jernej Ule, Giovanni Stefani, Aldo Mele, Matteo Ruggiu, Xuning Wang, Bahar Taneri, Terry Gaasterland, Benjamin J. Blencowe, and Robert B. Darnell. An RNA map predicting Nova-dependent splicing regulation. *Nature*, 444(7119):580–586, oct 2006.

[34] Daniel Maticzka, Sita J Lange, Fabrizio Costa, and Rolf Backofen. GraphProt: modeling binding preferences of RNA-binding proteins. *Genome Biology*, 15(1):R17, 2014.

[35] Martin Stražar, Marinka Žitnik, Blaž Zupan, Jernej Ule, and Tomaž Curk. Orthogonal matrix factorization enables integrative analysis of multiple RNA binding proteins. *Bioinformatics*, 32(10):1527–1535, jan 2016.

[36] Xiaoyong Pan and Hong-Bin Shen. RNA-protein binding motifs mining with a new hybrid deep learning based cross-domain knowledge integration approach. *BMC Bioinformatics*, 18(1), feb 2017.

[37] Jeremy R. Sanford, Xin Wang, Matthew Mort, Natalia VanDuyn, David N. Cooper, Sean D. Mooney, Howard J. Edenberg, and Yunlong Liu. Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts. *Genome Research*, 19(3):381–394, dec 2008.

[38] Jernej Ule, Kirk B. Jensen, Matteo Ruggiu, Aldo Mele, Aljaz Ule, and Robert B. Darnell. CLIP identifies Nova-regulated RNA networks in the brain. *Science*, 302:1212–5, Nov 2003.

[39] Chaolin Zhang, Maria A. Frias, Aldo Mele, Matteo Ruggiu, Taesun Eom, Christina B. Marney, Huidong Wang, Donny D. Licatalosi, John J. Fak, and Robert B. Darnell. Integrative Modeling Defines the Nova Splicing-Regulatory Network and Its Combinatorial Controls. *Science*, 329(5990):439–443, jun 2010.

[40] Gayle Knapp, Jacques S.Beckmann, Peter F.Johnson, Shella A.Fuhrman, and John Abelson. Transcription and processing of intervening sequences in yeast tRNA genes. *Cell*, 14:221–36, Jun 1978.

[41] Cindy L. Will and Reinhard Lhrmann. Spliceosome Structure and Function. *Cold Spring Harbor Perspectives in Biology*, 3(7):a003707–a003707, dec 2010.

[42] Angela N. Brooks, Julie L. Aspden, Anna I. Podgornaia, Donald C. Rio, and Steven E. Brenner. Identification and experimental validation of splicing regulatory elements in Drosophila melanogaster reveals functionally conserved splicing enhancers in metazoans. *RNA*, 17(10):1884–1894, aug 2011.

[43] Maria Bennett, Susan Michaud, Joy Kingston, , and Robin Reed. Protein components specifically associated with prespliceosome and spliceosome complexes. *Genes Dev*, 6:1986–2000, Oct 1992.

[44] Rita Das, Zhaolan Zhou, and Robin Reed. Functional association of U2 snRNP with the ATP-independent spliceosomal complex E. *Mol Cell*, 5:779–87, May 2000.

[45] Patrik Förch, Oscar Puig, Nancy Kedersha, Concepcion Martinez, Sander Granneman, Bertrand Seraphin, Paul Anderson, and Juan Valcarcel. The apoptosis-promoting factor TIA-1 is a regulator of alternative pre-mRNA splicing. *Mol Cell*, 6:1089–98, Nov 2000.

[46] Justin R. Prigge, Sonya V. Iverson, Ashley M. Siders, and Edward E. Schmidt. Interactome for auxiliary splicing factor U2AF65 suggests diverse roles. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1789(6-8):487–492, jun 2009.

[47] Singh R, Valcarcel J, and Green MR. Distinct binding specificities and functions of higher eukaryotic polypyrimidine tract-binding proteins. *Science*, 268(5214):1173–1176, may 1995.

[48] J. Valcarcel, R. K. Gaur, R. Singh, and M. R. Green. Interaction of U2AF65 RS Region with Pre-mRNA Branch Point and Promotion of Base Pairing with U2 snRNA. *Science*, 273(5282):1706–1709, sep 1996.

[49] Changwei Shao, Bo Yang, Tongbin Wu, Jie Huang, Peng Tang, Yu Zhou, Jie Zhou, Jinsong Qiu, Li Jiang, Hairi Li, Geng Chen, Hui Sun, Yi Zhang, Alain Denise, Dong-Er Zhang, and Xiang-Dong Fu. Mechanisms for U2AF to define 3′ splice sites and regulate alternative splicing in the human genome. *Nature Structural & Molecular Biology*, 21(11):997–1005, oct 2014.

[50] Kathi Zarnack, Julian König, Mojca Tajnik, Iñigo Martincorena, Sebastian Eustermann, Isabelle Stévant, Alejandro Reyes, Simon Anders, Nicholas M. Luscombe, and Jernej Ule. Direct Competition between hnRNP C and U2AF65 Protects the Transcriptome from the Exonization of Alu Elements. *Cell*, 152(3):453–466, jan 2013.

[51] C. Le Guiner, F. Lejeune, D. Galiana, L. Kister, R. Breathnach, J. Stevenin, and F. Del Gatto-Konczak. TIA-1 and TIAR Activate Splicing of Alternative Exons with Weak 5' Splice Sites followed by a U-rich Stretch on Their Own Pre-mRNAs. *Journal of Biological Chemistry*, 276(44):40638–40646, aug 2001.

[52] Wen-Juan Wei, Shi-Rong Mu, Monika Heiner, Xing Fu, Li-Juan Cao, Xiu-Feng Gong, Albrecht Bindereif, and Jingyi Hui. YB-1 binds to CAUC motifs

and stimulates exon inclusion by enhancing the recruitment of U2AF to weak polypyrimidine tracts. *Nucleic Acids Research*, 40(17):8622–8636, jun 2012.

[53] Yuanchao Xue, Yu Zhou, Tongbin Wu, Tuo Zhu, Xiong Ji, Young-Soo Kwon, Chao Zhang, Gene Yeo, Douglas L. Black, Hui Sun, Xiang-Dong Fu, and Yi Zhang. 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(6):996–1006, dec 2009.

[54] Markus C. Wahl, Cindy L. Will, and Reinhard Lhrmann. The Spliceosome: Design Principles of a Dynamic RNP Machine. *Cell*, 136(4):701–718, feb 2009.

[55] B. Kate Dredge, Alexandros D. Polydorides, and Robert B. Darnell. The splice of life: Alternative splicing and neurological disease. *Nature Reviews Neuroscience*, 2(1):43–50, jan 2001.

[56] Robert B. Darnell. RNA Logic in Time and Space. *Cell*, 110(5):545–550, sep 2002.

[57] Qun Pan, Ofer Shai, Leo J Lee, Brendan J Frey, and Benjamin J Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics*, 40(12):1413–1415, nov 2008.

[58] Jason M. Johnson, John Castle, Philip Garrett-Engele, Zhengyan Kan, Patrick M. Loerch, Christopher D. Armour, Ralph Santos, Eric E. Schadt, Roland Stoughton, and Daniel D. Shoemaker. Genome-Wide Survey of Human Alternative Pre-mRNA Splicing with Exon Junction Microarrays. *Science*, 302(5653):2141–2144, dec 2003.

[59] Marina M. Scotti and Maurice S. Swanson. RNA mis-splicing in disease. *Nature Reviews Genetics*, 17(1):19–32, nov 2015.

[60] Steffen Erkelenz, William F. Mueller, Melanie S. Evans, Anke Busch, Katrin Schöneweis, Klemens J. Herte, and Heiner Schaal. Position-dependent splicing activation and repression by SR and hnRNP proteins rely on common mechanisms. *RNA*, 19(1):96–102, nov 2012.

[61] Miriam Llorian, Schraga Schwartz, Tyson A Clark, Dror Hollander, Lit-Yeen Tan, Rachel Spellman, Adele Gordon, Anthony C Schweitzer, Pierre de la Grange, Gil Ast, and Christopher W J Smith. Position-dependent alternative splicing activity revealed by global profiling of alternative splicing events regulated by PTB. *Nature Structural & Molecular Biology*, 17(99):1114–1123, aug 2010.

[62] James R Tollervey, Tomaž Curk, Boris Rogelj, Michael Briese, Matteo Cereda, Melis Kayikci, Julian Knig, Tibor Hortobágyi, Agnes L Nishimura, Vera Župunski, Rickie Patani, Siddharthan Chandran, Gregor Rot, Blaž Zupan, Christopher E Shaw, and Jernej Ule. Characterizing the RNA targets and position-dependent splicing regulation by TDP-43. *Nature Neuroscience*, 14(4):452–458, feb 2011.

[63] Panagiota Kafasla, Ian Mickleburgh, Miriam Llorian, Miguel Coelho, Clare Gooding, Dmitry Cherny, Amar Joshi, Olga Kotik-Kogan, Stephen Curry, Ian C. Eperon, Richard J. Jackson, and Christopher W.J. Smith. Defining the roles and interactions of PTB. *Biochemical Society Transactions*, 40(4):815–820, aug 2012.

[64] Min-Yuan Chou, Jason G Underwood, Julia Nikolic, Martin H.T Luu, and Douglas L Black. Multisite RNA Binding and Release of Polypyrimidine Tract Binding Protein during the Regulation of c-src Neural-Specific Splicing. *Molecular Cell*, 5(6):949–957, jun 2000.

[65] Ravinder Singh and Juan Valcárcel. Building specificity with nonspecific RNA-binding proteins. *Nature Structural & Molecular Biology*, 12(8):645–653, aug 2005.

[66] F. C. Oberstrass. Structure of PTB Bound to RNA: Specific Binding and Implications for Splicing Regulation. *Science*, 309(5743):2054–2057, sep 2005.

[67] Caroline Clerte and Kathleen B. Hall. Characterization of multimeric complexes formed by the human PTB1 protein on RNA. *RNA*, 12(3):457–475, mar 2006.

[68] Dmitry Cherny, Clare Gooding, Giles E Eperon, Miguel B Coelho, Clive R Bagshaw, Christopher W J Smith, and Ian C Eperon. Stoichiometry of a regulatory splicing complex revealed by single-molecule analyses. *The EMBO Journal*, 29(13):2161–2172, may 2010.

[69] Andrea Ghetti, Serafin Pinol-Roma, W. Matthew Michael, Carlo Morandi, and Gideon Dreyfuss. hnRNP 1 the polyprimidine tract-binding protein: distinct nuclear localization and association with hnRNAs. *Nucl Acids Res*, 20(14):3671–3678, 1992.

[70] Niroshika Keppetipola, Shalini Sharma, Qin Li, and Douglas L. Black. Neuronal regulation of pre-mRNA splicing by polypyrimidine tract binding proteins PTBP1 and PTBP2. *Critical Reviews in Biochemistry and Molecular Biology*, 47(4):360–378, jun 2012.

[71] Chorng-Horng Lin and James G. Patton. Regulation of alternative 3' splice site selection by constitutive splicing factors. *RNA*, 1:234–45, May 1995.

[72] Eric J. Wagner and Mariano A. Garcia-Blanco. Polypyrimidine tract binding protein antagonizes exon definition. *Mol Cell Biol*, 21:3281–8, May 2001.

[73] J. Sauliere, A. Sureau, A. Expert-Bezancon, and J. Marie. The Polypyrimidine Tract Binding Protein (PTB) Represses Splicing of Exon 6B from the -Tropomyosin Pre-mRNA by Directly Interfering with the Binding of the U2AF65 Subunit. *Molecular and Cellular Biology*, 26(23):8755–8769, sep 2006.

[74] Kirsty Sawicka, Martin Bushell, Keith A. Spriggs, and Anne E. Willis. Polypyrimidine-tract-binding protein: a multifunctional RNA-binding protein. *Biochm. Soc. Trans.*, 36(4):641–647, aug 2008.

[75] David Brett, Heike Pospisil, Juan Valcárcel, Jens Reich, and Peer Bork. Alternative splicing and genome complexity. *Nature Genetics*, 30(1):29–30, dec 2001.

[76] Stefan Stamm, Shani Ben-Ari, Ilona Rafalska, Yesheng Tang, Zhaiyi Zhang, Debra Toiber, T.A. Thanaraj, and Hermona Soreq. Function of alternative splicing. *Gene*, 344:1–20, jan 2005.

[77] Mehmet Yabas, Hannah Elliott, and Gerard Hoyne. The Role of Alternative Splicing in the Control of Immune Homeostasis and Cellular Differentiation. *International Journal of Molecular Sciences*, 17(1):3, dec 2015.

[78] Jin Wang and James L Manley. Regulation of pre-mRNA splicing in metazoa. *Current Opinion in Genetics & Development*, 7(2):205–211, apr 1997.

[79] Stefan Stamm, Jian Zhu, Kenta Nakai, Peter Stoilov, Oliver Stoss, and Michael Q. Zhang. An Alternative-Exon Database and Its Statistical Analysis. *DNA and Cell Biology*, 19(12):739–756, dec 2000.

[80] Qiang Xu, Barmak Modrek, and Christopher Lee. Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic Acids Research*, 30(17):3754–3766, sep 2002.

[81] James P. Orengo and Thomas A. Cooper. Alternative Splicing in Disease. In *Advances in Experimental Medicine and Biology*, pages 212–223. Springer New York, 2007.

[82] Arianne J. Matlin, Francis Clark, and Christopher W. J. Smith. Understanding alternative splicing: towards a cellular code. *Nature Reviews Molecular Cell Biology*, 6(5):386–398, may 2005.

[83] L. O. F. Penalva and L. Sanchez. RNA Binding Protein Sex-Lethal (Sxl) and Control of Drosophila Sex Determination and Dosage Compensation. *Microbiology and Molecular Biology Reviews*, 67(3):343–359, sep 2003.

[84] Roland Tacke and James L. Manley. Functions of SR and Tra2 Proteins in Pre-mRNA Splicing Regulation. *Experimental Biology and Medicine*, 220(2):59–63, feb 1999.

[85] Qin Li, Sika Zheng, Areum Han, Chia-Ho Lin, Peter Stoilov, Xiang-Dong Fu, and Douglas L Black. The splicing regulator PTBP2 controls a program of embryonic splicing required for neuronal maturation. *eLife*, 3, jan 2014.

[86] Eric T. Wang, Rickard Sandberg, Shujun Luo, Irina Khrebtukova, Lu Zhang, Christine Mayr, Stephen F. Kingsmore, Gary P. Schroth, and Christopher B. Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, nov 2008.

[87] Francisco E. Baralle and Jimena Giudice. Alternative splicing as a regulator of development and tissue identity. *Nature Reviews Molecular Cell Biology*, may 2017.

[88] Mathieu Quesnel-Vallières, Manuel Irimia, Sabine P. Cordes, and Benjamin J. Blencowe. Essential roles for the splicing regulator nSR100/SRRM4 during nervous system development. *Genes & Development*, 29(7):746–759, apr 2015.

[89] Bushra Raj, Dave O'Hanlon, John P. Vessey, Qun Pan, Debashish Ray, Noel J. Buckley, Freda D. Miller, and Benjamin J. Blencowe. Cross-Regulation between an Alternative Splicing Activator and a Transcription Repressor Controls Neurogenesis. *Molecular Cell*, 43(5):843–850, sep 2011.

[90] Paul L. Boutz, Peter Stoilov, Qin Li, Chia-Ho Lin, Geetanjali Chawla, Kristin Ostrow, Lily Shiue, Manuel Ares Jr., and Douglas L. Black. A post-transcriptional regulatory switch in polypyrimidine tract-binding proteins re-

programs alternative splicing in developing neurons. *Genes & Development*, 21(13):1636–1652, jul 2007.

[91] Eugene V. Makeyev, Jiangwen Zhang, Monica A. Carrasco, and Tom Maniatis. The MicroRNA miR-124 Promotes Neuronal Differentiation by Triggering Brain-Specific Alternative Pre-mRNA Splicing. *Molecular Cell*, 27(3):435–448, aug 2007.

[92] Jernej Ule, Aljaž Ule, Joanna Spencer, Alan Williams, Jing-Shan Hu, Melissa Cline, Hui Wang, Tyson Clark, Claire Fraser, Matteo Ruggiu, Barry R Zeeberg, David Kane, John N Weinstein, John Blume, and Robert B Darnell. Nova regulates brain-specific splicing to shape the synapse. *Nature Genetics*, 37(8):844–852, jul 2005.

[93] Kirk B Jensen, B.Kate Dredge, Giovanni Stefani, Ru Zhong, Ronald J Buckanovich, Hirotaka J Okano, Yolanda Y.L Yang, and Robert B Darnell. Nova-1 Regulates Neuron-Specific Alternative Splicing and Is Essential for Neuronal Viability. *Neuron*, 25(2):359–371, feb 2000.

[94] Ronald J. Buckanovich, Jerome B. Posner, and Robert B. Darnell. Nova the paraneoplastic Ri antigen, is homologous to an RNA-binding protein and is specifically expressed in the developing motor system. *Neuron*, 11(4):657–672, oct 1993.

[95] Yolanda Y. L. Yang, Guang Lin Yin, and Robert B. Darnell. The neuronal RNA-binding protein Nova-2 is implicated as the autoantigen targeted in POMA patients with dementia. *Proceedings of the National Academy of Sciences*, 95(22):13254–13259, oct 1998.

[96] Robert B. Darnell and Jerome B. Posner. Paraneoplastic Syndromes Involving the Nervous System. *New England Journal of Medicine*, 349(16):1543–1554, oct 2003.

[97] Karl V. Voelkerding, Shale A. Dames, and Jacob D. Durtschi. Next-Generation Sequencing: From Basic Research to Diagnostics. *Clinical Chemistry*, 55(4):641–658, feb 2009.

[98] Tarek Hamed Attia and Maysaa Abdallah Saeed. Next Generation Sequencing Technologies: A Short Review. *Journal of Next Generation Sequencing & Applications*, 01(S1), 2015.

[99] Chandra Shekhar Pareek, Rafal Smoczynski, and Andrzej Tretyn. Sequencing technologies and genome sequencing. *Journal of Applied Genetics*, 52(4):413–435, jun 2011.

[100] James M. Heather and Benjamin Chain. The sequence of sequencers: The history of sequencing DNA. *Genomics*, 107(1):1–8, jan 2016.

[101] Michael L. Metzker. Sequencing technologies — the next generation. *Nature Reviews Genetics*, 11(1):31–46, dec 2009.

[102] Brendan P. Hodkinson and Elizabeth A. Grice. Next-Generation Sequencing: A Review of Technologies and Tools for Wound Microbiome Research. *Advances in Wound Care*, 4(1):50–58, jan 2015.

[103] M Margulies, M Egholm, WE Altman, S Attiya, JS Bader, LA Bemben, J Berka, MS Braverman, YJ Chen, Z Chen, SB Dewell, L Du, JM Fierro, XV Gomes, BC Godwin, W He, S Helgesen, CH Ho, GP Irzyk, SC Jando, ML Alenquer, TP Jarvie, KB Jirage, JB Kim, JR Knight, JR Lanza, JH Leamon, SM Lefkowitz, M Lei, J Li, KL Lohman, H Lu, VB Makhijani, KE McDade, MP McKenna, EW Myers, E Nickerson, JR Nobile, R Plant, BP Puc, MT Ronan, GT Roth, GJ Sarkis, JF Simons, JW Simpson, M Srinivasan, KR Tartaro, A Tomasz, KA Vogt, GA Volkmer, SH Wang, Y Wang, MP Weiner, P Yu, RF Begley, and JM Rothberg. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437:376–80, Sep 2005.

[104] José F. Siqueira, Ashraf F. Fouad, and Isabela N. Rôças. Pyrosequencing as a tool for better understanding of human microbiomes. *Journal of Oral Microbiology*, 4(1):10743, jan 2012.

[105] Sara Goodwin, John D. McPherson, and W. Richard McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333–351, may 2016.

[106] Jason A. Reuter, Damek V. Spacek, and Michael P. Snyder. High-Throughput Sequencing Technologies. *Molecular Cell*, 58(4):586–597, may 2015.

[107] Jia Guo, Ning Xu, Zengmin Li, Shenglong Zhang, Jian Wu, Dae Hyun Kim, Mong Sano Marma, Qinglin Meng, Huanyan Cao, Xiaoxu Li, Shundi Shi, Lin Yu, Sergey Kalachikov, James J. Russo, Nicholas J. Turro, and Jingyue Ju. Four-color DNA sequencing with 3'-O-modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides. *Proceedings of the National Academy of Sciences*, 105(27):9145–9150, jun 2008.

[108] Jay Shendure1, Gregory J. Porreca, Nikos B. Reppas, Xiaoxia Lin, John P. McCutcheon, Abraham M. Rosenbaum, Michael D. Wang, Kun Zhang, Robi D. Mitra, and George M. Church. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, 309:1728–32, Sep 2005.

[109] H.P.J. Buermans and J.T. den Dunnen. Next generation sequencing technology: Advances and applications. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 1842(10):1932–1941, oct 2014.

[110] Peter J. A. Cock, Christopher J. Fields, Naohisa Goto, Michael L. Heuer, and Peter M. Rice. The Sanger FASTQ file format for sequences with quality scores and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, 38(6):1767–1771, dec 2009.

[111] Edward I. Budowsky and Gulnara G. Abdurashidova. Polynucleotide—Protein Cross-Links Induced by Ultraviolet Light and Their Use

for Structural Investigation of Nucleoproteins. In *Progress in Nucleic Acid Research and Molecular Biology*, pages 1–65. Elsevier, 1989.

[112] Joel W. Hockensmith, William L. Kubasek, William R. Vorachek, Elisabeth M. Evertsz, and Peter H. von Hippel. [13] Laser cross-linking of protein-nucleic acid complexes. In *Protein \3- DNA Interactions*, pages 211–236. Elsevier, 1991.

[113] Martin D. Shetlar, John Carbone, Elaine Steady, and Kellie Hom. Photochemical addition of amino acids and peptides to polyuridylic acid. *Photochemistry and Photobiology*, 39(2):141–144, feb 1984.

[114] Martin D. Shetlar, John Christensen, and Kellie Hom. Photochemical addition of amino acids and peptides to dna. *Photochemistry and Photobiology*, 39(2):125–133, feb 1984.

[115] Martin D. Shetlar, Kellie Hom, John Carbone, David Moy, Elaine Steady, and Mark Watanabe. Photochemical addition of amino acids and peptides to homopolyribonucleotides of the major dna bases. *Photochemistry and Photobiology*, 39(2):135–140, feb 1984.

[116] T. Jellinek and R. B. Johns. The mechanism of photochemical addition of cysteine to uracil and formation of dihydrouracil. *Photochemistry and Photobiology*, 11(5):349–359, may 1970.

[117] Isao Saito, Hiroshi Sugiyama, and Teruo Matsuura. Photoinduced reactions. 151. Isolation and characterization of a thymine-lysine adduct in UV-irradiated nuclei. The role of thymine-lysine photoaddition in photo-cross-linking of proteins to DNA. *Journal of the American Chemical Society*, 105(23):6989–6991, nov 1983.

[118] Anthony A. Shaw and Martin D. Shetlar. Ring-opening photoreactions of cytosine and 5-methylcytosine with aliphatic alcohols. *Photochemistry and Photobiology*, 49(3):267–271, mar 1989.

[119] Anthony A. Shaw, Arnold M. Falick, and Martin D. Shetlar. Photoreactions of thymine and thymidine with N-acetyltyrosine. *Biochemistry*, 31(45):10976–10983, nov 1992.

[120] Yoichiro Sugimoto, Julian König, Shobbir Hussain, Blaž Zupan, Tomaž Curk, Michaela Frye, and Jernej Ule. Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions. *Genome Biology*, 13(8):R67, 2012.

[121] Serafn Piol-Roma, Stephen A.Adam, YangDo Choi, and Gideon Dreyfuss. Ultraviolet-induced cross-linking of RNA to proteins in vivo. *Methods Enzymol*, 180:410–8, 1989.

[122] Juri Rappsilber. The beginning of a beautiful friendship: Cross-linking/mass spectrometry and modelling of proteins and multi-protein complexes. *Journal of Structural Biology*, 173(3):530–540, mar 2011.

[123] Donny D. Licatalosi, Aldo Mele, John J. Fak, Jernej Ule, Melis Kayikci, Sung Wook Chi, Tyson A. Clark, Anthony C. Schweitzer, John E. Blume, Xuning Wang, Jennifer C. Darnell, and Robert B. Darnell. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, 456(7221):464–469, nov 2008.

[124] Julian König, Kathi Zarnack, Gregor Rot, Tomaž Curk, Melis Kayikci, Blaž Zupan, Daniel J Turner, Nicholas M Luscombe, and Jernej Ule. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nature Structural & Molecular Biology*, 17(7):909–915, jul 2010.

[125] Henning Urlaub, Klaus Hartmuth, and Reinhard Lührmann. A two-tracked approach to analyze RNA–protein crosslinking sites in native nonlabeled small nuclear ribonucleoprotein particles. *Methods*, 26(2):170–181, feb 2002.

[126] Ina Huppertz, Jan Attig, Andrea D'Ambrogio, Laura E. Easton, Christopher R. Sibley, Yoichiro Sugimoto, Mojca Tajnik, Julian Knig, and Jernej

Ule. iCLIP: Protein–RNA interactions at nucleotide resolution. *Methods*, 65(3):274–287, feb 2014.

[127] Markus Hafner, Markus Landthaler, Lukas Burger, Mohsen Khorshid, Jean Hausser, Philipp Berninger, Andrea Rothballer, Manuel Ascano Jr., Anna-Carina Jungkamp, Mathias Munschauer, Alexander Ulrich, Greg S. Wardle, Scott Dewell, Mihaela Zavolan, and Thomas Tuschl. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, 141:129–41, Apr 2010.

[128] Anna-Carina Jungkamp, Marlon Stoeckius, Desirea Mecenas, Dominic Grn, Guido Mastrobuoni, Stefan Kempa, and Nikolaus Rajewsky. In Vivo and Transcriptome-wide Identification of RNA Binding Protein Target Sites. *Molecular Cell*, 44(5):828–840, dec 2011.

[129] Markus Hafner, Markus Landthaler, Lukas Burger, Mohsen Khorshid, Jean Hausser, Philipp Berninger, Andrea Rothballer, Manuel Ascano, Anna-Carina Jungkamp, Mathias Munschauer, Alexander Ulrich, Greg S. Wardle, Scott Dewell, Mihaela Zavolan, and Thomas Tuschl. Transcriptome-wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP. *Cell*, 141(1):129–141, apr 2010.

[130] Michael J Moore, Chaolin Zhang, Emily Conn Gantman, Aldo Mele, Jennifer C Darnell, and Robert B Darnell. Mapping Argonaute and conventional RNA-binding protein interactions with RNA at single-nucleotide resolution using HITS-CLIP and CIMS analysis. *Nat Protoc*, 9:263–93, Feb 2014.

[131] I Huppertz, J Attig, A D'Ambrogio, LE Easton, CR Sibley, Y Sugimoto, M Tajnik, J Knig, and J Ule. iCLIP: protein-RNA interactions at nucleotide resolution. *Methods*, 65:274–87, Feb 2014.

[132] BJ Zarnegar, RA Flynn, Y Shen, BT Do, HY Chang, and PA Khavari. irCLIP platform for efficient characterization of protein-RNA interactions. *Nat Methods*, 13:489–92, Jun 2016.

[133] Nostrand EL Van, GA Pratt, AA Shishkin, C Gelboin-Burkhart, MY Fang, B Sundararaman, SM Blue, TB Nguyen, C Surka, K Elkins, R Stanton, F Rigo, M Guttman, and GW Yeo. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat Methods*, 13:508–14, Jun 2016.

[134] Eric L Van Nostrand, Gabriel A Pratt, Alexander A Shishkin, Chelsea Gelboin-Burkhart, Mark Y Fang, Balaji Sundararaman, Steven M Blue, Thai B Nguyen, Christine Surka, Keri Elkins, Rebecca Stanton, Frank Rigo, Mitchell Guttman, and Gene W Yeo. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nature Methods*, 13(6):508–514, mar 2016.

[135] Keith Le, Katherine Mitsouras, Meenakshi Roy, Qi Wang, Qiang Xu, Stanley F. Nelson, and Christopher Lee. Detecting tissue-specific regulation of alternative splicing as a qualitative change in microarray data. *Nucleic Acids Research*, 32(22):e180–e180, dec 2004.

[136] DD Shoemaker, EE Schadt, CD Armour, YD He, P Garrett-Engele, PD McDonagh, PM Loerch, A Leonardson, PY Lum, G Cavet, LF Wu, SJ Altschuler, S Edwards, J King, JS Tsang, G Schimmack, JM Schelter, J Koch, M Ziman, MJ Marton, B Li, P Cundiff, T Ward, J Castle, M Krolewski, MR Meyer, M Mao, J Burchard, MJ Kidd, H Dai, JW Phillips, PS Linsley, R Stoughton, S Scherer, and MS Boguski. Experimental annotation of the human genome using microarray technology. *Nature*, 409:922–7, Feb 2001.

[137] Roger Bumgarner. Overview of DNA microarrays: types, applications, and their future. *Curr Protoc Mol Biol*, Chapter 22:Unit 22.1., Jan 2013.

[138] Francesca Finotello and Barbara Di Camillo. Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis. *Brief Funct Genomics*, 14:130–42, Mar 2015.

[139] mily A Chen, Tade Souaiaia, Jennifer S Herstein, Oleg V Evgrafov, Valeria N Spitsyna, Danea F Rebolini, and James A Knowles. Effect of RNA integrity on uniquely mapped reads in RNA-Seq. *BMC Res Notes*, 7:753, Oct 2014.

[140] Marcel C. Van Verk, Richard Hickman, Corne M.J.Pieterse, and Saskia C.M. Van Wees. RNA-Seq: revelation of the messengers. *Trends Plant Sci*, 18:175–9, Apr 2013.

[141] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10:57–63, Jan 2009.

[142] Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michal Wojciech Szczesniak, Daniel J. Gaffney, Laura L. Elo, Xuegong Zhang, and Ali Mortazavi. A survey of best practices for RNA-seq data analysis. *Genome Biol*, 17:13, Jan 2016.

[143] Paul L. Auer and R. W. Doerge. Statistical design and analysis of RNA sequencing data. *Genetics*, 185:405–16, Jun 2010.

[144] Nicholas J. Schurch, Pieta Schofield, Marek Gierlinski, Christian Cole, Alexander Sherstnev, Vijender Singh, Nicola Wrobel, Karim Gharbi, Gordon G. Simpson, Tom Owen-Hughes, Mark Blaxter, and Geoffrey J. Barton. Erratum: How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA*, 22:1641, Oct 2016.

[145] Supriyo De and Myriam Gorospe. Bioinformatic tools for analysis of CLIP ribonucleoprotein data. *Wiley Interdisciplinary Reviews: RNA*, page e1404, dec 2016.

[146] Table 2: Next-generation sequencing and bioinformatics details obtained from FastQC software (Andrews 2010) and QDD pipeline (Meglécz et al., 2010) for Lutjanus johnii, Protonibea diacanthus and Lethrinus laticaudis .

[147] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):10, may 2011.

[148] Gorges Martin, Andreas R. Gruber, and Mihaela Zavolan. Mapping Protein-RNA Interactions by CLIP. *Materials and Methods*, 1, sep 2011.

[149] Claudia Misale. Accelerating Bowtie2 with a lock-less concurrency approach and memory affinity. In *2014 22nd Euromicro International Conference on Parallel Distributed, and Network-Based Processing*. Institute of Electrical and Electronics Engineers (IEEE), feb 2014.

[150] Daehwan Kim, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L Salzberg. TopHat2: accurate alignment of transcriptomes in the presence of insertions deletions and gene fusions. *Genome Biology*, 14(4):R36, 2013.

[151] Alexander Dobin and Thomas R. Gingeras. Optimizing RNA-Seq Mapping with STAR. In *Methods in Molecular Biology*, pages 245–262. Springer Nature, 2016.

[152] Pär G Engström, Tamara Steijger, Botond Sipos, Gregory R Grant, Gunnar Rätsch Andre Kahles, The RGASP Consortium, Nick Goldman, Tim J Hubbard, Jennifer Harrow, and Roderic Guigo Paul Bertone. Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat Methods*, 10:1185–91, Dec 2013.

[153] Heng Li, Jue Ruan, and Richard Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, 18(11):1851–1858, nov 2008.

[154] Haibin Xu, Xiang Luo, Jun Qian, Xiaohui Pang, Jingyuan Song, Guangrui Qian, Jinhui Chen, and Shilin Chen. FastUniq: A Fast De Novo Duplicates Removal Tool for Paired Short Reads. *PLoS ONE*, 7(12):e52249, dec 2012.

[155] Erik Holmqvist, Patrick R Wright, Lei Li, Thorsten Bischler, Lars Barquist, Richard Reinhardt, Rolf Backofen, and Jrg Vogel. Global RNA recognition patterns of post-transcriptional regulators Hfq and CsrA revealed by UV crosslinkingin vivo. *The EMBO Journal*, 35(9):991–1011, apr 2016.

[156] Paula H Reyes-Herrera and Elisa Ficarra. Computational Methods for CLIP-seq Data Processing. *Bioinformatics and Biology Insights*, page 199, oct 2014.

[157] Philip J. Uren, Emad Bahrami-Samani, Suzanne C. Burns, Mei Qiao, Fedor V. Karginov, Emily Hodges, Gregory J. Hannon, Jeremy R. Sanford, Luiz O. F. Penalva, and Andrew D. Smith. Site identification in high-throughput RNA-protein interaction data. *Bioinformatics*, 28(23):3013–3020, sep 2012.

[158] Michael T Lovci, Dana Ghanem, Henry Marr, Justin Arnold, Sherry Gee, Marilyn Parra, Tiffany Y Liang, Thomas J Stark, Lauren T Gehman, Shawn Hoon, Katlin B Massirer, Gabriel A Pratt, Douglas L Black, Joe W Gray, John G Conboy, and Gene W Yeo. Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. *Nature Structural & Molecular Biology*, 20(12):1434–1442, nov 2013.

[159] YeoLab. CLIpper - clip peak enrichment recognition, 2013.

[160] Mahmoud M. Ibrahim, Scott A. Lacadie, and Uwe Ohler. JAMM: a peak finder for joint analysis of NGS replicates. *Bioinformatics*, 31(1):48–55, sep 2014.

[161] Hideaki Shimazaki and Shigeru Shinomoto. A Method for Selecting the Bin Size of a Time Histogram. *Neural Computation*, 19(6):1503–1527, jun 2007.

[162] Jeffrey D. Banfield and Adrian E. Raftery. Model-Based Gaussian and Non-Gaussian Clustering. *Biometrics*, 49(3):803, sep 1993.

[163] Gene W Yeo, Nicole G Coufal, Tiffany Y Liang, Grace E Peng, Xiang-Dong Fu, and Fred H Gage. An RNA code for the FOX2 splicing regulator revealed

by mapping RNA-protein interactions in stem cells. *Nature Structural & Molecular Biology*, 16(2):130–137, jan 2009.

[164] Eduardo Eyras. Pyicoteo - a suite of tools for the analysis of high-throughput sequencing data, 2013.

[165] A. Kucukural, H. Ozadam, G. Singh, M. J. Moore, and C. Cenik. ASPeak: an abundance sensitive peak detection algorithm for RIP-Seq. *Bioinformatics*, 29(19):2485–2486, aug 2013.

[166] Guramrit Singh, Alper Kucukural, Can Cenik, John D. Leszyk, Scott A. Shaffer, Zhiping Weng, and Melissa J. Moore. The Cellular EJC Interactome Reveals Higher-Order mRNP Structure and an EJC-SR Protein Nexus. *Cell*, 151(4):750–764, nov 2012.

[167] Beibei Chen, Jonghyun Yun, Min Kim, Joshua T Mendell, and Yang Xie. PIPE-CLIP: a comprehensive online tool for CLIP-seq data analysis. *Genome Biology*, 15(1):R18, 2014.

[168] Ankeeta Shah, Yingzhi Qian, Sebastien M. Weyn-Vanhentenryck, and Chaolin Zhang. CLIP Tool Kit (CTK): a flexible and robust pipeline to analyze CLIP sequencing data. *Bioinformatics*, page btw653, oct 2016.

[169] Tomaz Curk. iCount - protein-rna interaction analysis, 2016.

[170] Debashish Ray, Hilal Kazan, Kate B. Cook, Matthew T. Weirauch, Hamed S. Najafabadi, Xiao Li, Serge Gueroussov, Mihai Albu, Hong Zheng, Ally Yang, Hong Na, Manuel Irimia, Leah H. Matzat, Ryan K. Dale, Sarah A. Smith, Christopher A. Yarosh, Seth M. Kelly, Behnam Nabet, Desirea Mecenas, Weimin Li, Rakesh S. Laishram, Mei Qiao, Howard D. Lipshitz, Fabio Piano, Anita H. Corbett, Russ P. Carstens, Brendan J. Frey, Richard A. Anderson, Kristen W. Lynch, Luiz O. F. Penalva, Elissa P. Lei, Andrew G. Fraser, Benjamin J. Blencowe, Quaid D. Morris, and Timothy R. Hughes. A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, 499:172–7, Jul 2013.

[171] Timothy L. Bailey. DREME - motif discovery in transcription factor chip-seq data, 2011.

[172] Sophia S. Liu, Adam J. Hockenberry, Andrea Lancichinetti, Michael C. Jewett, and Luis A. N. Amaral. NullSeq: A Tool for Generating Random Coding Sequences with Desired Amino Acid and GC Contents. *PLoS Comput Biol*, 12:e1005184, Nov 2016.

[173] Glass Lab at UCSD. HOMER - software for motif discovery and next generation sequencing analysis, 2010.

[174] Xiao Li, Gerald Quon, Howard D. Lipshitz, and Quaid Morris. Predicting in vivo binding sites of RNA-binding proteins using mRNA secondary structure. *RNA*, 16(6):1096–1107, apr 2010.

[175] Hilal Kazan, Debashish Ray, Esther T. Chan, Timothy R. Hughes, and Quaid Morris. RNAcontext: A New Method for Learning the Sequence and Structure Binding Preferences of RNA-Binding Proteins. *PLoS Computational Biology*, 6(7):e1000832, jul 2010.

[176] Emad Bahrami-Samani, Luiz O.F. Penalva, Andrew D. Smith, and Philip J. Uren. Leveraging cross-link modification events in CLIP-seq for motif discovery. *Nucleic Acids Research*, 43(1):95–103, dec 2014.

[177] J. S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29(6-7):1105–1119, may 1990.

[178] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, sep 1995.

[179] Regression with Support Vector Machines. In *Knowledge Discovery with Support Vector Machines*, pages 193–208. Wiley-Blackwell, oct 2009.

[180] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106, 2010.

[181] Christian Hauer, Tomaz Curk, Simon Anders, Thomas Schwarzl, Anne-Marie Alleaume, Jana Sieber, Ina Hollerer, Madhuri Bhuvanagiri, Wolfgang Huber, Matthias W. Hentze, and Andreas E. Kulozik. Improved binding site assignment by high-resolution mapping of RNA–protein interactions using iCLIP. *Nature Communications*, 6:7921, aug 2015.

[182] Nicole Lambert, Alex Robertson, Mohini Jangi, Sean McGeary, Phillip A. Sharp, and Christopher B. Burge. RNA Bind-n-Seq: Quantitative Assessment of the Sequence and Structural Binding Specificity of RNA Binding Proteins. *Molecular Cell*, 54(5):887–900, jun 2014.

[183] Zhen Wang, Melis Kayikci, Michael Briese, Kathi Zarnack, Nicholas M. Luscombe, Gregor Rot, Blaž Zupan, Tomaž Curk, and Jernej Ule. iCLIP Predicts the Dual Splicing Effects of TIA-RNA Interactions. *PLoS Biology*, 8(10):e1000530, oct 2010.

[184] Ronny Lorenz, Stephan H Bernhart, Christian Hner zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. ViennaRNA Package 2.0. *Algorithms Mol Biol*, 6(1):26, 2011.

[185] Miguel B Coelho, Jan Attig, Nicolas Bellora, Julian König, Martina Hallegger, Melis Kayikci, Eduardo Eyras, Jernej Ule, and Christopher WJ Smith. Nuclear matrix protein Matrin3 regulates alternative splicing and forms overlapping regulatory networks with PTB. *The EMBO Journal*, 34(5):653–668, jan 2015.

[186] A Corvelo, M Hallegger, CW Smith, and E Eyras. Genome-wide association between branch point properties and alternative splicing. *PLoS Comput Biol*, 6:e1001016, Nov 2010.

[187] DA Bitton, C Rallis, DC Jeffares, GC Smith, YY Chen, S Codlin, S Marguerat, and J Bhler. LaSSO, a strategy for genome-wide mapping of intronic lariats and branch points using RNA-seq. *Genome Res*, 24:1169–79, Jul 2014.

[188] AJ Taggart, AM DeSimone, JS Shih, ME Filloux, and WG Fairbrother. Large-scale mapping of branchpoints in human pre-mRNA transcripts in vivo. *Nat Struct Mol Biol*, 19:719–21, Jun 2012.

[189] Burkhard Morgenstern, Nadine Werner, Sonja J. Prohaska, Rasmus Steinkamp, Isabelle Schneider, Amarendran R. Subramanian, Peter F. Stadler, and Jan Weyer-Menkhoff. Multiple sequence alignment with user-defined constraints at GOBICS. *Bioinformatics*, 21(7):1271–1273, nov 2004.

[190] Joel D Nelson, Oleg Denisenko, and Karol Bomsztyk. Protocol for the fast chromatin immunoprecipitation (ChIP) method. *Nature Protocols*, 1(1):179–185, jun 2006.

[191] Luiz O. F. Penalva, Scott A. Tenenbaum, and Jack D. Keene. Gene Expression Analysis of Messenger RNP Complexes. In *mRNA Processing and Metabolism*, pages 125–134. Springer Nature.

[192] Jernej Ule, Kirk Jensen, Aldo Mele, and Robert B. Darnell. CLIP: A method for identifying protein–RNA interaction sites in living cells. *Methods*, 37(4):376–386, dec 2005.

[193] Gregor Rot Tomaz Curk Melis Kayikci Blaz Zupan Daniel J Turner Nicholas M Luscombe Jernej Ule Julian König, Kathi Zarnack. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat Struct Mol Biol*, 17:909–15, Jul 2010.

[194] Julian König, Nicholas J. McGlincy, and Jernej Ule. Analysis of Protein-RNA Interactions with Single-Nucleotide Resolution Using iCLIP and Next-Generation Sequencing. In *Tag-Based Next Generation Sequencing*, pages 153–169. Wiley-Blackwell, jan 2012.

[195] MY Chou, JG Underwood, J Nikolic, MH Luu, and DL Black. Multisite RNA binding and release of polypyrimidine tract binding protein during the regulation of c-src neural-specific splicing. *Mol Cell*, 5:949–57, Jun 2000.

[196] Jernej Murn, Kathi Zarnack, Yawei J. Yang, Omer Durak, Elisabeth A. Murphy, Sihem Cheloufi, Dilenny M. Gonzalez, Marianna Teplova, Tomaž Curk, Johannes Zuber, Dinshaw J. Patel, Jernej Ule, Nicholas M. Luscombe, Li-Huei Tsai, Christopher A. Walsh, and Yang Shi. Control of a neuronal morphology program by an RNA-binding zinc finger protein Unkempt. *Genes & Development*, 29(5):501–512, mar 2015.

[197] Serena L. Chan, Ina Huppertz, Chengguo Yao, Lingjie Weng, James J. Moresco, John R. Yates, Jernej Ule, James L. Manley, and Yongsheng Shi. CPSF30 and Wdr33 directly bind to AAUAAA in mammalian mRNA 3′ processing. *Genes & Development*, 28(21):2370–2380, oct 2014.

[198] Sebastien M. Weyn-Vanhentenryck, Aldo Mele, Qinghong Yan, Shuying Sun, Natalie Farny, Zuo Zhang, Chenghai Xue, Margaret Herre, Pamela A. Silver, Michael Q. Zhang, Adrian R. Krainer, Robert B. Darnell, and Chaolin Zhang. HITS-CLIP and Integrative Modeling Define the Rbfox Splicing-Regulatory Network Linked to Brain Development and Autism. *Cell Reports*, 6(6):1139–1152, mar 2014.

[199] Brian J Zarnegar, Ryan A Flynn, Ying Shen, Brian T Do, Howard Y Chang, and Paul A Khavari. irCLIP platform for efficient characterization of protein–RNA interactions. *Nature Methods*, 13(6):489–492, apr 2016.

[200] Herve Le Hir, Elisa Izaurralde, Lynne E. Maquat, and Melissa J. Moore. The spliceosome deposits multiple proteins 20-24 nucleotides upstream of mRNA exon-exon junctions. *The EMBO Journal*, 19(24):6860–6869, dec 2000.

[201] Jérôme Saulière, Valentine Murigneux, Zhen Wang, Emélie Marquenet, Isabelle Barbosa, Olivier Le Tonquèze, Yann Audic, Luc Paillard, Hugues Roest Crollius, and Hervé Le Hir. CLIP-seq of eIF4AIII reveals transcriptome-wide mapping of the human exon junction complex. *Nature Structural & Molecular Biology*, 19(11):1124–1131, oct 2012.

[202] Corinna Giorgi, Gene W. Yeo, Martha E. Stone, Donald B. Katz, Christopher Burge, Gina Turrigiano, and Melissa J. Moore. The EJC Factor eIF4AIII Modulates Synaptic Strength and Neuronal Protein Expression. *Cell*, 130(1):179–191, jul 2007.

[203] Toshiharu Shibuya, Thomas Ø Tange, Nahum Sonenberg, and Melissa J Moore. eIF4AIII binds spliced mRNA in the exon junction complex and is essential for nonsense-mediated decay. *Nat Struct Mol Biol*, 11(4):346–351, mar 2004.

[204] Isabel M. Palacios, David Gatfield, Daniel St Johnston, and Elisa Izaurralde. An eIF4AIII-containing complex required for mRNA localization and nonsense-mediated mRNA decay. *Nature*, 427(6976):753–757, feb 2004.

[205] Chia C. Chan, Josee Dostie, Michael D. Diem, Wenqin Feng, Matthias Mann, Juri Rappsilber, , and Gideon Dreyfuss. eIF4A3 is a novel component of the exon junction complex. *RNA*, 10(2):200–209, feb 2004.

[206] Dennis M. Mishler, Alexander B. Christ, and Joan A. Steitz. Flexibility in the site of exon junction complex deposition revealed by functional group and RNA secondary structure alterations in the splicing substrate. *RNA*, 14(12):2657–2670, oct 2008.

[207] Yuanchao Xue, Kunfu Ouyang, Jie Huang, Yu Zhou, Hong Ouyang, Hairi Li, Gang Wang, Qijia Wu, Chaoliang Wei, Yanzhen Bi, Li Jiang, Zhiqiang Cai, Hui Sun, Kang Zhang, Yi Zhang, Ju Chen, and Xiang-Dong Fu. Direct Conversion of Fibroblasts to Neurons by Reprogramming PTB-Regulated MicroRNA Circuits. *Cell*, 152(1-2):82–96, jan 2013.

[208] Chaolin Zhang and Robert B Darnell. Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. *Nature Biotechnology*, 29(7):607–614, jun 2011.

[209] Olivier Cordin, Josette Banroques, N. Kyle Tanner, and Patrick Linder. The DEAD-box protein family of RNA helicases. *Gene*, 367:17–37, feb 2006.

[210] Markus Hafner, Markus Landthaler, Lukas Burger, Mohsen Khorshid, Jean Hausser, Philipp Berninger, Andrea Rothballer, Manuel Ascano, Anna-Carina Jungkamp, Mathias Munschauer, Alexander Ulrich, Greg S. Wardle, Scott Dewell, Mihaela Zavolan, and Thomas Tuschl. PAR-CliP - A Method to Identify Transcriptome-wide the Binding Sites of RNA Binding Proteins. *Journal of Visualized Experiments*, (41), jul 2010.

[211] Rachel Spellman and Christopher W.J. Smith. Novel modes of splicing repression by PTB. *Trends in Biochemical Sciences*, 31(2):73–76, feb 2006.

[212] A.-L. Steckelberg, J. Altmueller, C. Dieterich, and N. H. Gehring. CWC22-dependent pre-mRNA splicing and eIF4A3 binding enables global deposition of exon junction complexes. *Nucleic Acids Research*, 43(9):4687–4700, apr 2015.

[213] A. Alexandrov, D. Colognori, M.-D. Shu, and J. A. Steitz. Human spliceosomal protein CWC22 plays a role in coupling splicing to exon junction complex deposition and nonsense-mediated decay. *Proceedings of the National Academy of Sciences*, 109(52):21313–21318, dec 2012.

[214] Isabelle Barbosa, Nazmul Haque, Francesca Fiorini, Charlotte Barrandon, Catherine Tomasetto, Marco Blanchette, and Hervé Le Hir. Human CWC22 escorts the helicase eIF4AIII to spliceosomes and promotes exon junction complex assembly. *Nature Structural & Molecular Biology*, 19(10):983–990, sep 2012.

[215] Iren Wang, Janosch Hennig, Pravin Kumar, Ankush Jagtap, Miriam Sonntag, Juan Valcrcel, and Michael Sattler. Structure dynamics and RNA binding of the multi-domain splicing factor TIA-1. *Nucleic Acids Research*, 42(9):5949–5966, mar 2014.

[216] Joshua T. Witten and Jernej Ule. Understanding splicing regulation through RNA splicing maps. *Trends in Genetics*, 27(3):89–97, mar 2011.

[217] S. A. Amero, G. Raychaudhuri, C. L. Cass, W. J. van Venrooij, W. J. Habets, A. R. Krainer, and A. L. Beyer. Independent deposition of heterogeneous nuclear ribonucleoproteins and small nuclear ribonucleoprotein particles at sites of transcription. *Proceedings of the National Academy of Sciences*, 89(18):8409–8413, sep 1992.

[218] Robin Reed. Mechanisms of fidelity in pre-mRNA splicing. *Current Opinion in Cell Biology*, 12(3):340–345, jun 2000.

[219] Jan Attig, Igor Ruiz de los Mozos, Nejc Haberman, Zhen Wang, Warren Emmett, Kathi Zarnack, Julian König, and Jernej Ule. Splicing repression allows the gradual emergence of new Alu-exons in primate evolution. *eLife*, 5, nov 2016.

[220] Qi Liu, Xue Zhong, Blair B. Madison, Anil K. Rustgi, and Yu Shyr. Assessing Computational Steps for CLIP-Seq Data Analysis. *BioMed Research International*, 2015:1–10, 2015.

[221] Michael Uhl, Torsten Houwaart, Gianluca Corrado, Patrick R. Wright, and Rolf Backofen. Computational analysis of CLIP-seq data. *Methods*, 118-119:60–72, apr 2017.

[222] Peter J Skene and Steven Henikoff. A simple method for generating high-resolution maps of genome-wide protein binding. *eLife*, 4, jun 2015.

[223] Philip J. Uren, Emad Bahrami-Samani, Suzanne C. Burns, Mei Qiao, Fedor V. Karginov, Emily Hodges, Gregory J. Hannon, Jeremy R. Sanford, Luiz O. F. Penalva, and Andrew D. Smith. Site identification in high-throughput RNA–protein interaction data. *Bioinformatics*, 28(23):3013–3020, sep 2012.

[224] Tao Wang, Guanghua Xiao, Yongjun Chu, Michael Q. Zhang, David R. Corey, and Yang Xie. Design and bioinformatics analysis of genome-wide CLIP experiments. *Nucleic Acids Research*, 43(11):5263–5274, may 2015.

[225] Cheryl Stork and Sika Zheng. Genome-Wide Profiling of RNA–Protein Interactions Using CLIP-Seq. In *Methods in Molecular Biology*, pages 137–151. Springer New York, 2016.

[226] M Bennett, S Michaud, J Kingston, and R Reed. Protein components specifically associated with prespliceosome and spliceosome complexes. *Genes & Development*, 6(10):1986–2000, oct 1992.

[227] Maria Dolores Chiara, Leon Palandjian, Rebecca Feld Kramer, and Robin Reed. Evidence that U5 snRNP recognizes the 3' splice site for catalytic step II in mammals. *The EMBO Journal*, 16(15):4746–4759, aug 1997.

[228] Anne Tisserant and Harald König. Signal-regulated Pre-mRNA occupancy by the general splicing factor U2AF. *PLoS One*, 3:e1418, Jan 2008.

[229] H. Hornig, M. Aebi, and C. Weissmann. Effect of mutations at the lariat branch acceptor site on -globin pre-mRNA splicing in vitro. *Nature*, 324(6097):589–591, dec 1986.

[230] L. Stirling Churchman and Jonathan S. Weissman. Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature*, 469(7330):368–373, jan 2011.

[231] Nicholas T. Ingolia, Sina Ghaemmaghami, John R. S. Newman, and Jonathan S. Weissman. Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. *Science*, 324(5924):218–223, apr 2009.

[232] Christian Kambach, Stefan Walket, and Kiyoshi Nagai. Structure and assembly of the spliceosomal small nuclear ribonucleoprotein particles. *Current Opinion in Structural Biology*, 9(2):222–230, apr 1999.

[233] Henning Urlaub, Veronica A. Raker, Susanne Kostka, and Reinhard Lührmann. Sm protein-Sm site RNA interactions within the inner ring of

the spliceosomal snRNP core structure. *The EMBO Journal*, 20(1):187–196, jan 2001.

[234] Ali R. Awan, Amanda Manfredo, and Jeffrey A. Pleiss. Lariat sequencing in a unicellular yeast identifies regulated alternative splicing of exons that are evolutionarily conserved with humans. *Proc Natl Acad Sci U S A*, 110:12762–7, Jul 2013.

[235] André Corvelo, Martina Hallegger, Christopher W. J. Smith, and Eduardo Eyras. Genome-Wide Association between Branch Point Properties and Alternative Splicing. *PLoS Computational Biology*, 6(11):e1001016, nov 2010.

[236] Guy Kol, Galit Lev-Maor, and Gil Ast. Human-mouse comparative analysis reveals that branch-site plasticity contributes to splicing regulation. *Human Molecular Genetics*, 14(11):1559–1568, apr 2005.

[237] Christian Kambach, Stefan Walke, Robert Young, Johanna M. Avis, Eric de la Fortelle, Veronica A. Raker, Reinhard Lhrmann, Jade Li, and Kiyoshi Nagai. Crystal Structures of Two Sm Protein Complexes and Their Implications for the Assembly of the Spliceosomal snRNPs. *Cell*, 96(3):375–387, feb 1999.

[238] Danny A. Bitton, Charalampos Rallis, Daniel C. Jeffares, Graeme C. Smith, Yuan Y.C. Chen, Sandra Codlin, Samuel Marguerat, and Jürg Bähler. LaSSO a strategy for genome-wide mapping of intronic lariats and branch points using RNA-seq. *Genome Research*, 24(7):1169–1179, apr 2014.

[239] Allison J Taggart, Alec M DeSimone, Janice S Shih, Madeleine E Filloux, and William G Fairbrother. Large-scale mapping of branchpoints in human pre-mRNA transcripts in vivo. *Nature Structural & Molecular Biology*, 19(7):719–721, jun 2012.

[240] Cindy L. Will, Henning Urlaub, Tilmann Achsel, Marc Gentzel, Matthias Wilm, and Reinhard Lührmann. Characterization of novel SF3b and 17S U2

snRNP proteins including a human Prp5p homologue and an SF3b DEAD-box protein. *The EMBO Journal*, 21(18):4978–4988, sep 2002.

[241] O Gozani, R Feld, and R Reed. Evidence that sequence-independent binding of highly conserved U2 snRNP proteins upstream of the branch site is required for assembly of spliceosomal complex A. *Genes & Development*, 10(2):233–243, jan 1996.

[242] Or Gozani, Judith Potashkin, and Robin Reed. A Potential Role for U2AF-SAP 155 Interactions in Recruiting U2 snRNP to the Branch Site. *Molecular and Cellular Biology*, 18(8):4752–4760, aug 1998.

[243] Charles C. Query, Patrick S. Mccaw, and Phillip A. Sharp. A minimal spliceosomal complex A recognizes the branch site and polypyrimidine tract. *Mol Cell Biol*, 17:2944–53, May 1997.

[244] Cindy L. Will, Claudia Schneider, Andrew M. MacMillan, Nikos F. Katopodis, Gitte Neubauer, Matthias Wilm, Reinhard Lührmann, and Charles C. Query. A novel U2 and U11/U12 snRNP protein that associates with the pre-mRNA branch site. *EMBO J*, 20:4536–46, Aug 2001.

[245] S. Alsafadi, A. Houy, A. Battistella, T. Popova, M. Wassef, E. Henry, F. Tirode, A. Constantinou, S. Piperno-Neumann, S. Roman-Roman, M. Dutertre, and M.H. Stern. Cancer-associated SF3B1 mutations affect alternative splicing by promoting alternative branchpoint usage. *European Journal of Cancer*, 61:S94–S95, jul 2016.

[246] Anna Corrionero, Belen Minana, and Juan Valcarcel. Reduced fidelity of branch point recognition and alternative splicing induced by the anti-tumor drug spliceostatin A. *Genes & Development*, 25(5):445–459, mar 2011.

[247] Rachel B. Darman, Michael Seiler, Anant A. Agrawal, Kian H. Lim, Shouyong Peng, Daniel Aird, Suzanna L. Bailey, Erica B. Bhavsar, Betty Chan, Simona Colla, Laura Corson, Jacob Feala, Peter Fekkes, Kana Ichikawa,

Gregg F. Keaney, Linda Lee, Pavan Kumar, Kaiko Kunii, Crystal MacKenzie, Mark Matijevic, Yoshiharu Mizui, Khin Myint, Eun Sun Park, Xiaoling Puyang, Anand Selvaraj, Michael P. Thomas, Jennifer Tsai, John Y. Wang, Markus Warmuth, Hui Yang, Ping Zhu, Guillermo Garcia-Manero, Richard R. Furman, Lihua Yu, Peter G. Smith, and Silvia Buonamici. Cancer-Associated SF3B1 Hotspot Mutations Induce Cryptic 3′ Splice Site Selection through Use of a Different Branch Point. *Cell Reports*, 13(5):1033–1045, nov 2015.

[248] Kevin Talbot, Irene Miguel-Aliaga, Payam Mohaghegh, Chris P. Ponting, and Kay E. Davies. Characterization of a gene encoding survival motor neuron (SMN)-related protein a constituent of the spliceosome complex. *Human Molecular Genetics*, 7(13):2149–2156, dec 1998.

[249] Gitte Neubauer, Angus King, Juri Rappsilber, Cinzia Calvio, Mark Watson, Paul Ajuh, Judith Sleeman Angus Lamond, and Matthias Mann. Mass spectrometry and EST-database searching allows characterization of the multi-protein spliceosome complex. *Nat Genet*, 20:46–50, Sep 1998.

[250] Rita Das, Zhaolan Zhou, and Robin Reed. Functional Association of U2 snRNP with the ATP-Independent Spliceosomal Complex E. *Molecular Cell*, 5(5):779–787, may 2000.

[251] Patrik Förch, Oscar Puig, Nancy Kedersha, Concepción Martínez, Sander Granneman, Bertrand Séraphin, Paul Anderson, and Juan Valcárcel. The Apoptosis-Promoting Factor TIA-1 Is a Regulator of Alternative Pre-mRNA Splicing. *Molecular Cell*, 6(5):1089–1098, nov 2000.

[252] Allison J. Taggart, Chien-Ling Lin, Barsha Shrestha, Claire Heintzelman, Seongwon Kim, and William G. Fairbrother. Large-scale analysis of branchpoint usage across species and cell lines. *Genome Res*, 27:639–649, Apr 2017.

[253] Nuno Andre Faustino and Thomas A. Cooper. Pre-mRNA splicing and human disease. *Genes & Development*, 17(4):419–437, feb 2003.

[254] Mariano A Garcia-Blanco, Andrew P Baraniak, and Erika L Lasda. Alternative splicing in disease and therapy. *Nature Biotechnology*, 22(5):535–546, may 2004.

[255] Alexander G. Baltz, Mathias Munschauer, Bjrn Schwanhusser, Alexandra Vasile, Yasuhiro Murakawa, Markus Schueler, Noah Youngs, Duncan Penfold-Brown, Kevin Drew, Miha Milek, Emanuel Wyler, Richard Bonneau, Matthias Selbach, Christoph Dieterich, and Markus Landthaler. The mRNA-Bound Proteome and Its Global Occupancy Profile on Protein-Coding Transcripts. *Molecular Cell*, 46(5):674–690, jun 2012.

[256] Markus Schueler, Mathias Munschauer, Lea Haarup Gregersen, Ana Finzel, Alexander Loewer, Wei Chen, Markus Landthaler, and Christoph Dieterich. Differential protein occupancy profiling of the mRNA transcriptome. *Genome Biology*, 15(1):R15, 2014.

[257] S Chul Kwon, Hyerim Yi, Katrin Eichelbaum, Sophia Fhr, Bernd Fischer, Kwon Tae You, Alfredo Castello, Jeroen Krijgsveld, Matthias W Hentze, and V Narry Kim. The RNA-binding protein repertoire of embryonic stem cells. *Nature Structural & Molecular Biology*, 20(9):1122–1130, aug 2013.

[258] Vihandha O. Wickramasinghe, Mar Gonzàlez-Porta, David Perera, Arthur R. Bartolozzi, Christopher R. Sibley, Martina Hallegger, Jernej Ule, John C. Marioni, and Ashok R. Venkitaraman. Regulation of constitutive and alternative mRNA splicing across the human transcriptome by PRPF8 is determined by 5′ splice site strength. *Genome Biology*, 16(1), sep 2015.

[259] Yu-Cheng T Yang, Chao Di, Boqin Hu, Meifeng Zhou, Yifang Liu, Nanxi Song, Yang Li, Jumpei Umetsu, and Zhi Lu. CLIPdb: a CLIP-seq database for protein-RNA interactions. *BMC Genomics*, 16(1):51, 2015.

[260] Tao Wang, Yang Xie, and Guanghua Xiao. dCLIP: a computational approach for comparative CLIP-seq analyses. *Genome Biology*, 15(1):R11, 2014.

[261] Davide Cirillo, Federico Agostini, and Gian Gaetano Tartaglia. Predictions of protein-RNA interactions. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 3(2):161–175, sep 2012.

[262] Petr Klus, Benedetta Bolognesi, Federico Agostini, Domenica Marchese, Andreas Zanzoni, and Gian Gaetano Tartaglia. The cleverSuite approach for protein characterization: predictions of structural properties solubility, chaperone requirements and RNA-binding abilities. *Bioinformatics*, 30(11):1601–1608, feb 2014.

[263] Xiang-Dong Fu and Manuel Ares Jr. Context-dependent control of alternative splicing by RNA-binding proteins. *Nat Rev Genet*, 15:689–701, Oct 2014.

[264] Christoph Dieterich and Peter F. Stadler. Computational biology of RNA interactions. *Wiley Interdiscip Rev RNA*, 4:107–20, Jan 2013.

[265] Eckhard Jankowsky and Michael E. Harris. Specificity and nonspecificity in RNA–protein interactions. *Nature Reviews Molecular Cell Biology*, 16(9):533–544, aug 2015.

[266] Yang Li, Mahlon Collins, Rachel Geiser, Nadine Bakkar, David Riascos, and Robert Bowser. RBM45 homo-oligomerization mediates association with ALS-linked proteins and stress granules. *Scientific Reports*, 5(1), sep 2015.

Appendix

Publication

Nejc Haberman*, Ina Huppertz*, Jan Attig, Julian König, Zhen Wang, Christian Hauer, Matthias W. Hentze, Andreas E. Kulozik, Herve Le Hir, Tomaz Curk, Christopher R. Sibley, Kathi Zarnack and Jernej Ule (2017), Insights into the design and interpretation of iCLIP experiments. *Genome Biology*,

Genome Biology

RESEARCH Open Access

Insights into the design and interpretation of iCLIP experiments

Nejc Haberman[1,2†], Ina Huppertz[1,3,4†], Jan Attig[1,2], Julian König[1,5], Zhen Wang[6,7], Christian Hauer[4,8,9], Matthias W. Hentze[4,8], Andreas E. Kulozik[8,9], Hervé Le Hir[6,7], Tomaž Curk[10], Christopher R. Sibley[1,11], Kathi Zarnack[12*] and Jernej Ule[1,2*]

Abstract

Background: Ultraviolet (UV) crosslinking and immunoprecipitation (CLIP) identifies the sites on RNAs that are in direct contact with RNA-binding proteins (RBPs). Several variants of CLIP exist, which require different computational approaches for analysis. This variety of approaches can create challenges for a novice user and can hamper insights from multi-study comparisons. Here, we produce data with multiple variants of CLIP and evaluate the data with various computational methods to better understand their suitability.

Results: We perform experiments for PTBP1 and eIF4A3 using individual-nucleotide resolution CLIP (iCLIP), employing either UV-C or photoactivatable 4-thiouridine (4SU) combined with UV-A crosslinking and compare the results with published data. As previously noted, the positions of complementary DNA (cDNA)-starts depend on cDNA length in several iCLIP experiments and we now find that this is caused by constrained cDNA-ends, which can result from the sequence and structure constraints of RNA fragmentation. These constraints are overcome when fragmentation by RNase I is efficient and when a broad cDNA size range is obtained. Our study also shows that if RNase does not efficiently cut within the binding sites, the original CLIP method is less capable of identifying the longer binding sites of RBPs. In contrast, we show that a broad size range of cDNAs in iCLIP allows the cDNA-starts to efficiently delineate the complete RNA-binding sites.

Conclusions: We demonstrate the advantage of iCLIP and related methods that can amplify cDNAs that truncate at crosslink sites and we show that computational analyses based on cDNAs-starts are appropriate for such methods.

Keywords: Protein–RNA interactions, iCLIP, eCLIP, irCLIP, Binding site assignment, High-throughput sequencing, Polypyrimidine tract binding protein 1 (PTBP1), Eukaryotic initiation factor 4A-III (eIF4A3), Exon-junction complex

Background

RNA-binding proteins (RBPs) play crucial roles in all aspects of post-transcriptional gene regulation. To understand the mechanisms of their action, it is essential to identify the endogenous sites of protein–RNA interactions, which has been aided by the development of ultraviolet (UV) crosslinking and immunoprecipitation (CLIP) [1, 2]. During the CLIP protocol, crosslinked protein–RNA complexes are purified and the RNA fragments are released by digesting the protein, resulting in RNAs with a covalently bound peptide at the crosslink site. This is followed by reverse transcription, during which the bound peptide can lead to truncation of complementary DNAs (cDNA) at the crosslink site. The CLIP protocol prepares the cDNA library in a way that requires the reverse transcriptase to read through this peptide, thereby generating only ‘readthrough cDNAs’. Therefore, individual-nucleotide resolution CLIP (iCLIP) was also developed to exploit the ‘truncated cDNAs’ [3]. The cDNA-starts of these truncated cDNAs identify the nucleotide just downstream of the crosslinked peptide. Even though iCLIP amplifies both truncated and readthrough cDNAs, computational comparisons of CLIP and iCLIP cDNAs estimated that over 80% of iCLIP cDNAs truncate at the crosslink sites of most RBPs [4]. Recently, further variants were developed that also amplify truncated cDNAs, including BrdU-CLIP [5], eCLIP [6] and irCLIP [7]. Therefore, understanding the proportion

* Correspondence: kathi.zarnack@bmls.de; j.ule@ucl.ac.uk
†Equal contributors
[12]Buchmann Institute for Molecular Life Sciences (BMLS), Goethe University Frankfurt, Max-von-Laue-Str. 15, 60438 Frankfurt am Main, Germany
[1]Department of Molecular Neuroscience, UCL Institute of Neurology, Queen Square, London WC1N 3BG, UK
Full list of author information is available at the end of the article

and characteristics of truncated cDNAs in these protocols is essential.

The computational methods that use cDNA-starts to assign RNA-binding sites have been developed along with iCLIP. However, a recent study observed that the starts of long and short iCLIP cDNAs often map to different genomic positions for several RBPs, which leads to non-coinciding cDNA-starts [8]. Here, we focused on experiments produced for polypyrimidine tract binding protein 1 (PTBP1), eukaryotic initiation factor 4A-III (eIF4A3) and the splicing factor U2 auxiliary factor 65 kDa subunit (U2AF2), which represent examples of non-coinciding or coinciding cDNA-starts in introns or exons. eIF4A3 is a component of the exon junction complex (EJC). In vitro biochemical experiments with several splicing substrates demonstrated that the site of EJC deposition is normally expected at nucleotides −20 to −24 upstream of the exon-exon junction (−24..−20 nt) [9]. However, further studies showed that the sequence and structure of a nascent messenger RNA (mRNA) can shift EJC deposition as far as 10 nt away from this expected site [10]. The non-coinciding cDNA-starts in eIF4A3 iCLIP data produced by the previous study were shifted upstream of this expected region and it was proposed that the presence of non-coinciding cDNA-starts might be related to this shift [8]. The study concluded that the use of cDNA-starts may not be appropriate in iCLIP whenever non-coinciding cDNA-starts are prevalent.

To understand if cDNA-starts can be used to assign RNA-binding sites, we further analysed the iCLIP data with high frequency of non-coinciding cDNA-starts. We first examined the position and prevalence of crosslink-induced mutations to confirm previous findings, showing that such mutations are generally >5-fold less common within iCLIP than CLIP cDNAs, regardless of the presence of non-coinciding cDNA-starts [4]. Moreover, we identified RNA motifs that are commonly associated with crosslink sites and found them most highly enriched at cDNA deletions in CLIP, and cDNA-starts in iCLIP, eCLIP and irCLIP, even if non-coinciding cDNA-starts are prevalent. Interestingly, when using the photoactivatable 4-thiouridine (4SU)-based crosslinking in combination with iCLIP, the motifs were more highly enriched at cDNA-starts than at T-to-C transitions. These results demonstrate that the cDNA-starts can reliably be used to determine crosslink sites in iCLIP, regardless of the crosslinking method.

Further analyses demonstrated that presence of sequence and structural constraints at cDNA-ends is the cause of the non-coinciding cDNA-starts. To experimentally validate this finding, we produced additional PTBP1 and eIF4A3 iCLIP experiments, which demonstrate that the prevalence of the non-coinciding cDNA-starts is directly correlated with the extent of cDNA-end constraints. We show that the broad size range of iCLIP cDNAs in these new experiments allows the cDNA-starts to assign binding sites that align with the expected binding motifs (PTBP1) or binding regions (eIF4A3). We conclude that the use of the iCLIP cDNA-starts is appropriate to assign the protein–RNA crosslink sites in iCLIP and related methods.

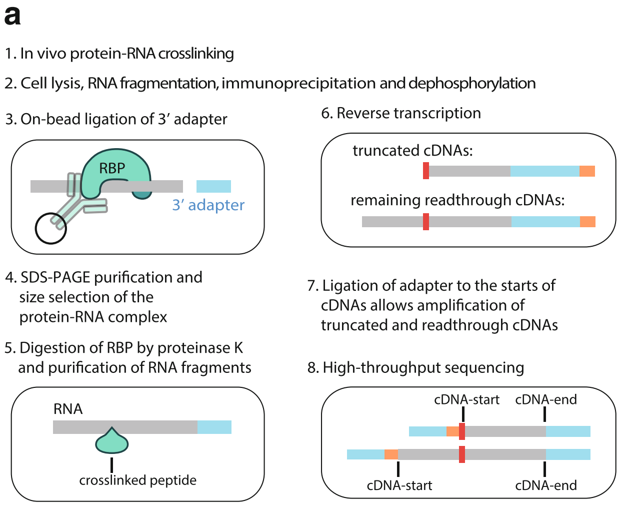
Results

Crosslink sites are identified by cDNA-starts in iCLIP

The iCLIP protocol is composed of eight principal experimental steps (Fig. 1a). First, cells or tissues are irradiated with UV light, which can create covalent bonds between an RBP and RNA. Cell lysates are then treated with RNase and the crosslinked RNA fragments are co-immunoprecipitated with the RBP. In the third step, an oligonucleotide adapter is ligated to the 3′ end of RNA fragments. The immunoprecipitated complexes are then separated and visualised by SDS-PAGE and the protein–RNA complex is isolated in a size-specific manner (Additional file 1: Figure S1). The RBP is removed from the RNA through proteinase K digestion, leaving a small peptide at the crosslink site. This impairs the reverse transcription and commonly leads to the truncation of cDNAs at the crosslinked peptide. Hence, iCLIP cDNAs start at the nucleotide just downstream of the crosslinked peptide and they end at the site of RNase cleavage.

To assess how variations in experimental conditions affect the assigned binding sites, we compared published and newly produced experiments for eIF4A3, PTBP1 and U2AF2. For the ease of comparisons, we numerically label the different experiments produced by the same method (Fig. 1b). eIF4A3-iCLIP1 refers to data generated in the previous study [8], while eIF4A3-iCLIP2 and eIF4A3-iCLIP3 were newly produced by the Le Hir and Ule labs, respectively. These are compared to the published eIF4A3 CLIP [11]. The PTBP1-iCLIP1 also refers to data generated in the previous study [12], while PTBP1-iCLIP2 and PTBP1-iCLIP3 were newly produced with deliberate protocol differences. Specifically, 4SU was used to induce crosslinking and RNase I conditions were adjusted in PTBP1-iCLIP2, and the 3′ dephosphorylation step was omitted in PTBP1-iCLIP3. These are compared to the published PTBP1 CLIP [13], eCLIP [6] and irCLIP data [7]. Finally, we also compare the PTBP1 data to U2AF2 CLIP [14] and iCLIP [15].

It was proposed that presence of non-coinciding cDNA-starts might indicate that some of these cDNAs have read through the crosslink site during reverse transcription [8]. It has been shown previously that such readthrough cDNAs often contain deletions, which are introduced into cDNAs at the crosslink site during reverse transcription


a
1. In vivo protein-RNA crosslinking
2. Cell lysis, RNA fragmentation, immunoprecipitation and dephosphorylation
3. On-bead ligation of 3′ adapter
RBP
3′ adapter
4. SDS-PAGE purification and size selection of the protein-RNA complex
5. Digestion of RBP by proteinase K and purification of RNA fragments
RNA
crosslinked peptide
6. Reverse transcription
truncated cDNAs:
remaining readthrough cDNAs:
7. Ligation of adapter to the starts of cDNAs allows amplification of truncated and readthrough cDNAs
8. High-throughput sequencing
cDNA-start
cDNA-end
cDNA-start
cDNA-end

b

protein	method & experiment number	Pubmed ID	accession no.	cell line	total number of unique cDNAs
eIF4A3	CLIP	23085716	GSM1001330	HeLa	11,690,349
eIF4A3	iCLIP1	26260686	E-MTAB-2599	HeLa	7,148,538
eIF4A3	iCLIP2	new	E-MTAB-3618	HEK293	14,454,772
eIF4A3	iCLIP3	new	E-MTAB-4000	HeLa	11,935,788
PTBP1	CLIP	23313552	GSE19323	HeLa	1,779,318
PTBP1	iCLIP1	25599992	E-MTAB-3108	HeLa	8,447,229
PTBP1	iCLIP2	new: 4SU, RNase	E-MTAB-5027	HEK293	9,211,541
PTBP1	iCLIP3	new: dephospo	E-MTAB-5026	HeLa	3,275,592
PTBP1	eCLIP	27018577	ENCSR981WKN	K562	6,060,266
mock	eCLIP	27018577	ENCSR445FZX	K562	5,669,907
PTBP1	irCLIP	27111506	CSR981WKN	HeLa	65,593,070
U2AF2	CLIP	25326705	GSM1509288	HeLa	4,702,278
U2AF2	iCLIP	23374342	E-MTAB-1371	HeLa	116,771,612

Fig. 1 An overview of methods and experiments. **a** A simplified *schematic* of the iCLIP protocol [17]. Before, cells or tissues are irradiated with UV light, which creates covalent bonds between proteins and RNAs that are in direct contact (step 1). After lysis, the crosslinked RNA is fragmented by limited concentration of RNase I and RNA fragments are then co-immunoprecipitated with the RBP (step 2), followed by ligation of a 3′ adapter (step 3). After SDS-PAGE purification (step 4), the crosslinked RBP is removed through proteinase K digestion and purification of RNA fragments (step 5). Reverse transcription is performed with a primer that includes a barcode (orange) containing both an experimental identifier and a unique molecular identifier (UMI) (step 6). The peptide that is on the crosslink site impairs reverse transcription and commonly leads to truncation of cDNAs at the crosslink site. Therefore, two types of cDNAs are generated: truncated cDNAs and readthrough cDNAs. In iCLIP, the cDNA library is prepared in such a way that both truncated and readthrough cDNAs are amplified (step 7). After PCR amplification and sequencing (step 8), both truncated and readthrough cDNAs are present. **b** *Table* summarising the experiments used in this study. *4SU* using 4SU combined with UV-A crosslinking, *RNase* optimised RNase digest conditions including antiRNase inhibitor and increased RNase I concentration, *dephospho* omitting 3′ dephosphorylation

[4, 16]. We compared the proportion of cDNAs with deletions in the different eIF4A3 datasets. Since the rate of sequencing errors rises with increasing cDNA length, we only examined cDNAs shorter than 40 nt for this purpose. Strikingly, a bimodal distribution of deletions is apparent in all datasets, with one peak of deletions close to the cDNA-starts (5..8th nt) and the second close to the cDNA-centres (22..27th nt, Fig. 2a). Thus, the deletions present in iCLIP show the same features as in CLIP and likely inform on the presence of readthrough cDNAs. Importantly, the proportion of deletions is lower by a factor of 5 or more in all eIF4A3 iCLIP experiments compared to CLIP, indicating that readthrough cDNAs represent a minor proportion of iCLIP data.

We used sequence motifs as a second feature that can serve as an identifier of crosslink sites. We defined these

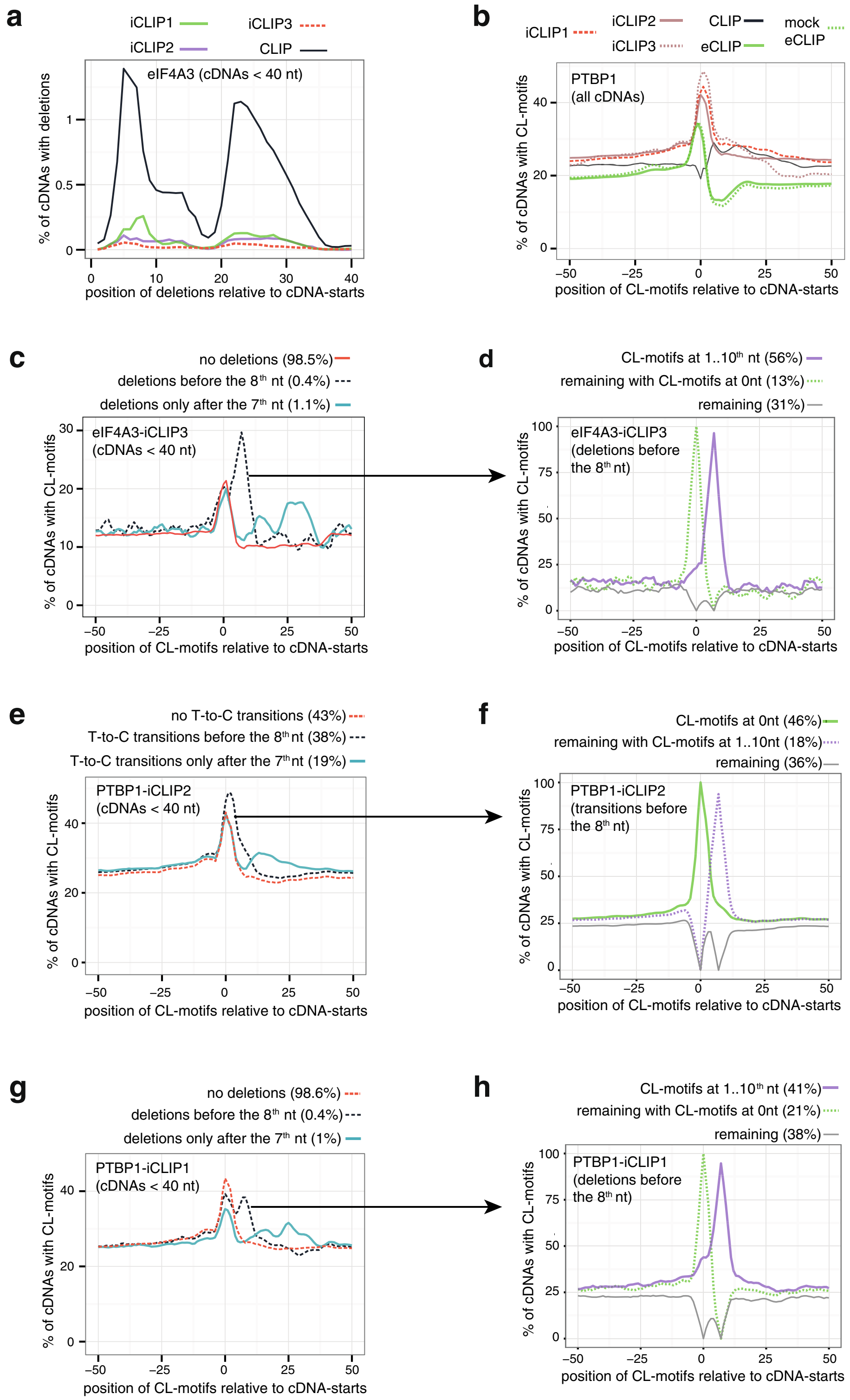


Fig. 2 (See legend on next page.)

(See figure on previous page.)

Fig. 2 Crosslink-associated (CL)-motifs are enriched at cDNA deletions and cDNA-starts in iCLIP. **a** Proportion of eIF4A3 cDNAs with deletion at each position relative to the cDNA-start. Only cDNAs shorter than 40 nt are examined. **b** Analysis of all PTBP1 experiments examined in this study shows the proportion of cDNAs from each experiment that overlap with a CL-motif at each position relative to the cDNA-start. **c** Proportion of eIF4A3-CLIP3 cDNAs that overlap with a CL-motif at each position relative to the cDNA-start. Only cDNAs shorter than 40 nt are examined; they are divided into those lacking deletions or containing a deletion within the first 7 nt or anywhere in the remaining portion of the cDNA. **d** The cDNAs of eIF4A3-CLIP3 containing a deletion within the first 7 nt are further sub-divided into three categories. First, cDNAs with CL-motifs between the 1st and 10th nucleotide of the cDNA. Second, the remaining cDNAs that contain CL-motifs at the position 0. And third, all remaining cDNAs. The proportion of cDNAs that overlap with a CL-motif at each position relative to the cDNA-start is then plotted for each sub-category. **e** Proportion of PTBP1-iCLIP2 cDNAs that overlap with a CL-motif at each position relative to the cDNA-start. Only cDNAs shorter than 40 nt are examined and are divided into those lacking T-to-C transitions or containing a transition within the first 7 nt or anywhere in the remaining portion of the cDNA. **f** The cDNAs of PTBP1-iCLIP2 containing a T-to-C transition within the first 7 nt are further sub-divided into three categories. First, cDNAs with CL-motifs overlapping the position 0. Second, the remaining cDNAs that contain CL-motifs between the 1st and 10th nucleotide of the cDNA. And third, all remaining cDNAs. Visualisation as in (**d**). **g** Same as (**c**), but for PTBP1-iCLIP1. **h** Same as (**d**), but for PTBP1-iCLIP1

sequence motifs based on analysis of eCLIP mock input data that were produced along with the PTBP1 eCLIP [6]. Even though no immunoprecipitation is done, the eCLIP mock data represent RNA fragments crosslinked to RBPs, because the lysate is loaded onto the gel and transferred to a nitrocellulose membrane and the non-crosslinked RNA migrates out of the gel or through the membrane. Thus, eCLIP mock data represent RNAs crosslinked to many different RBPs and should reflect the sequence preferences at crosslink sites that are common to a mixture of RBPs. We identified 10 tetramers that are enriched at cDNA-starts by a factor of 1.5 or more compared to the 10 nt region upstream of the cDNA-starts. Since they serve as a signature of crosslink sites, we refer to them as ‘CL-motifs’ (for UV crosslink-associated motifs). On one hand, these CL-motifs could represent sequence preferences of one or few unknown RBPs that dominate the eCLIP mock input data. On the other hand, all CL-motifs are rich in uridines (see ‘Methods’), which would agree with the hypothesis of preferred UV-C crosslinking to uridines [4]. The CL-motifs are rich in polypyrimidine sequences that are preferentially bound by PTBP1 and U2AF2 [18] and thus it is expected that their enrichment should be especially high at crosslink sites of these proteins. While no further increase in CL-motif enrichment is seen at cDNA-starts of PTBP1-eCLIP, it is reassuring to find their increased enrichment at cDNA-starts of all PTBP1 and U2AF2 iCLIP experiments (Fig. 2b, Additional file 1: Figure S2A).

We also found significant enrichment of CL-motifs at cDNA-starts of all eIF4A3 iCLIP experiments (Fig. 2c, Additional file 1: Figure S2B). eIF4A3 is not thought to bind RNA with sequence specificity based on biochemical and transcriptomic studies [9, 11, 19] and its sequence-independent interaction with RNA is consistent with the properties of DEAD-box proteins [20]. Moreover, we did not find any generic enrichment of CL-motifs at nucleotides −20 to −24 upstream of the exon-exon junctions, where EJC normally binds (data not shown). Thus, it is most likely that CL-motifs only reflect crosslinking preferences in the case of eIF4A3 iCLIP. In contrast to their enrichment at cDNA-starts of all iCLIP experiments, CL-motifs are depleted from the cDNA-starts of all CLIP experiments and instead they are enriched within the sequence of CLIP cDNAs (Fig. 2b, Additional file 1: Figure S2A, B). This agrees with the expected prevalence of truncated cDNAs in iCLIP and readthrough cDNAs in CLIP.

To further assess the validity of CL-motifs, we exploited the bimodal distribution of deletions in the cDNAs shorter than 40 nt, where one peak of deletions is seen in the first 7 nt and a second peak around the centre of cDNAs (Fig. 2a). We separated the cDNAs into three classes: those with deletions in the first 7 nt, those with deletions elsewhere and those with no deletions. If cDNAs contain a deletion in PTBP1 and U2AF2 iCLIP, CL-motifs are most highly enriched at the position of deletion, but not at cDNA-starts, which confirms that they represent readthrough cDNAs (Additional file 1: Figure S2C, D). However, in iCLIP of all three proteins, >90% of cDNAs lack deletions; in these cDNAs, CL-motifs are enriched exclusively at the cDNA-starts, confirming that these largely correspond to truncated cDNAs (Fig. 2c, Additional file 1: Figure S2C, D). In conclusion, analysis of cDNA deletions and CL-motifs indicates that the position of crosslink sites can generally be defined by cDNA-starts in iCLIP.

cDNA-starts assign crosslink sites in iCLIP regardless of the crosslinking method

We noticed that the cDNAs with deletions in eIF4A3-iCLIP3 contain some CL-motif enrichment at cDNA-starts in addition to the position of deletions (Fig. 2c). To better understand this dual enrichment, we separated those cDNAs with deletion before the 8th nt into three classes (Fig. 2d). Fifty-six percent of cDNAs had the CL-motifs overlapping with the deletion and these had no additional motif enrichment at cDNA-starts. Thirteen percent had CL-motifs at their start, but not at the position of the deletion, and 31% had CL-motifs at neither position. A possible explanation for the dual enrichment is that about 80% of deletions correspond to crosslink sites and about 20% are a result of sequencing errors within truncated rather than readthrough cDNAs. This further indicates that readthrough cDNAs correspond to a minor proportion of iCLIP reads.

Since cDNAs with deletions are rare in iCLIP, we performed a new PTBP1-iCLIP experiment (PTBP1-iCLIP2) in which we incubated cells with 4SU and induced crosslinking with UV-A. We additionally optimised the RNase conditions (see below). In PAR-CLIP, which originally introduced 4SU-mediated crosslinking, the presence of T-to-C transitions indicates the position of crosslink sites [21] and therefore we wished to examine if the same applies to iCLIP when 4SU is used to induce crosslinking. We used CL-motifs to examine the alignment of cDNA-truncations and transitions to crosslink sites. The CL-motifs are CU-rich (see 'Methods') and correspond well to the known binding motifs of PTBP1 [18]. Thus, even if 4SU-mediated crosslinking has different sequence preferences, we expect that the CL-motifs should align well to the crosslink sites of PTBP1 due to its binding preferences. The further benefit of using the same CL-motifs for all analyses is that it allows us to directly compare the extent of their enrichment across all different experiments. Notably, we obtained an unexpected misalignment between CL-motifs and transitions in PTBP1-iCLIP2: 57% of cDNAs contained deletions, but CL-motifs were enriched mainly at cDNA-starts, just like in PTBP1-iCLIP1 (compare Fig. 2e and Additional file 1: Figure S2C). In 67% of cDNAs with transitions, the position of the transition mapped to the first few nucleotides close to the cDNA-start. We therefore examined these cDNAs in more detail by dividing them into three classes (Fig. 2f): 46% of these cDNAs contain CL-motifs at the cDNA-start. Eighteen percent contain CL-motifs at the site of transition rather than the cDNA-start. Thirty-six percent contain no CL-motif at either position. A possible explanation for this pattern of enrichment is that about 20% of transitions correspond to crosslink sites and about 75% are a result of other causes. In conclusion, presence of transitions does not separate readthrough and truncated cDNAs in iCLIP, since CL-motifs are equally enriched at cDNA-starts of cDNAs containing or lacking transitions.

While transitions do not overlap well with CL-motifs in PTBP1-iCLIP2, the overlap is better with deletions in PTBP1-iCLIP1. Even though only 1.4% of PTBP1-iCLIP1 cDNAs contain deletions (Fig. 2g), a greater proportion contain CL-motifs at the position of the deletion than at cDNA-starts (Fig. 2h). This indicates that deletions are more reliable than transition to identify crosslink sites in readthrough cDNAs, even when 4SU is used for crosslinking in iCLIP. Taken together, analysis of deletions, transitions and CL-motifs indicates that the incidence of readthrough cDNAs is generally low and that the majority of cDNAs truncate at crosslink sites in iCLIP regardless of the crosslinking method.

Non-coinciding cDNA-starts result from constrained cDNA-ends

In addition to the model of readthrough cDNAs, the previous study also discussed an alternative model, in which the non-coinciding cDNA-starts could originate from constraints on RNase cleavage, particularly when these are combined with the presence of long binding sites [8]. We examined this alternative model in more detail, since the prevalence of readthrough cDNAs in iCLIP did not appear sufficient to account for the non-coinciding cDNA-starts. We used the tool developed by the previous study (iCLIPro) to examine the prevalence of non-coinciding cDNA-starts in the PTBP1-iCLIP1 dataset. This tool compares the cDNA-start positions of shorter and longer cDNAs and displays overlapping starts by enrichment at position 0, while non-coinciding starts are enriched at other positions. As seen previously [8], we find that the PTBP1-iCLIP1 library contains non-coinciding cDNA-starts (Fig. 3a). To understand if the prevalence of non-coinciding cDNA-starts depends on the length of binding sites, we first identified regions on RNAs where cDNA-starts are significantly clustered, and we refer to these as 'crosslink clusters' (see 'Methods' for more detail). Notably, the proportion of non-coinciding cDNA-starts increases within crosslink clusters that are longer than 5 nt (Fig. 3b) and this increase is particularly dramatic in clusters longer than 30 nt (Fig. 3c). This analysis reveals that the non-coinciding cDNA-starts originate mainly from long binding sites.

To understand the possible causes and effects of the constrained RNase cleavage, we examined the new PTBP1-iCLIP2 experiment, in which we had also modified the conditions of RNase treatment: in this experiment, we included an inhibitor of endogenous RNases into the lysis buffer (antiRNase which does not inhibit RNase I) and slightly increased the concentration of RNase I compared to PTBP1-iCLIP1. In this way, we hoped to ensure that RNase I, which is not thought to have any sequence specificity, was responsible for fragmenting the RNAs in PTBP1-iCLIP2. Interestingly, the proportion of non-coinciding cDNA-starts is decreased in PTBP1-iCLIP2 (Fig. 3d), and this decrease is particularly apparent when analysing long crosslink clusters (Fig. 3e, f). In addition to the overlapping cDNA-starts, the cDNA-ends in PTBP1-iCLIP2 also often overlap with cDNA-starts (diagonal enrichment in Fig. 3d-f). Both RNase I and UV crosslinking require single-stranded RNA and thus their similar RNA structure

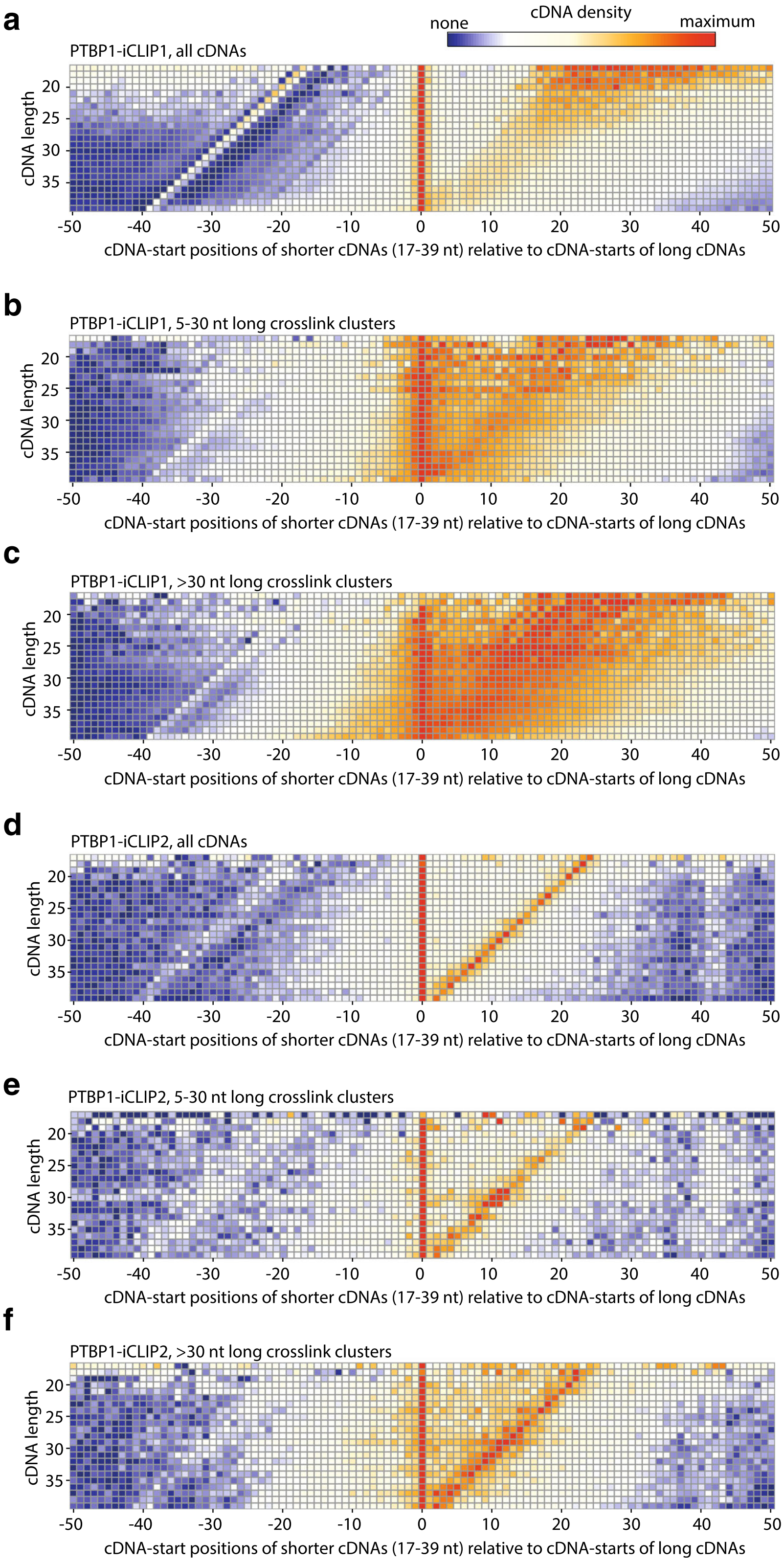


Fig. 3 (See legend on next page.)

(See figure on previous page.)
Fig. 3 Proportion of non-coinciding cDNA-starts differs between PTBP1 iCLIP experiments. **a** *Heatmap* for PTBP1-iCLIP1 generated using the previously developed software iCLIPro [8] to show the relative positioning of cDNA-starts of shorter iCLIP cDNAs (17–39 nt) compared to cDNA-starts of long cDNAs (longer than 39 nt). **b** As in (**a**), but for cDNAs of PTBP1-iCLIP1 that overlap with 5–30 nt long crosslink clusters. **c** As in (**a**), but for cDNAs of PTBP1-iCLIP1 that overlap with >30 nt long crosslink clusters. **d** As in (**a**), but for PTBP1-iCLIP2. **e** As in (**a**), but for cDNAs of PTBP1-iCLIP2 that overlap with 5–30 nt long crosslink clusters. **f** As in (**a**), but for cDNAs of PTBP1-iCLIP2 that overlap with >30 nt long crosslink clusters

preferences are the likely cause for their increased chance of overlap. Taken together, our results show that the prevalence and position of non-coinciding cDNA-starts can vary greatly between different iCLIP experiments performed for the same RBP and thus they are most likely a result of technical differences between these experiments.

To understand the technical features that lead to the non-coinciding cDNA-starts in PTBP1-iCLIP1, we examined in more detail the long crosslink clusters in which such cDNAs are most prominent. We restricted all following analyses to the 1000 crosslink clusters with the highest cDNA count to ensure that they have high coverage of diverse cDNA lengths. We identified the position within each crosslink cluster with the highest count of cDNA-starts (cDNA-start peak) and the position downstream of each crosslink cluster with the highest count of cDNA-ends (cDNA-end peak). Next, we categorised cDNAs based on their length and plotted their starts and ends around cDNA-start peaks (Fig. 4a) or cDNA-end peaks (Fig. 4b). As expected for long crosslink clusters, cDNA-starts are broadly distributed around the cDNA-start peaks. We measured the empirical cumulative distribution of cDNA-starts around cDNA-start peaks (Fig. 4a, c – inset), which demonstrates that the distribution of cDNA lengths has a much stronger influence on the position of cDNA-starts in PTBP1-iCLIP1 (Fig. 4a) than PTBP1-iCLIP2 (Fig. 4c). Strikingly, the cDNA-ends of all length categories precisely overlap at the position of cDNA-end peaks in PTBP1-iCLIP1 (Fig. 4b), while they are enriched over a broader region downstream of the cDNA-end peaks in PTBP1-iCLIP2 (Fig. 4d). Indeed, this tight constraint of cDNA-ends in PTBP1-iCLIP1 reveals three distinct peaks of cDNA-starts for each category of cDNA lengths (Fig. 4b), while these peaks are less prominent in the PTBP1-iCLIP2 library, in which the fold change for cDNA-end constraint is decreased by half (Fig. 4b, d – inner box plot). We

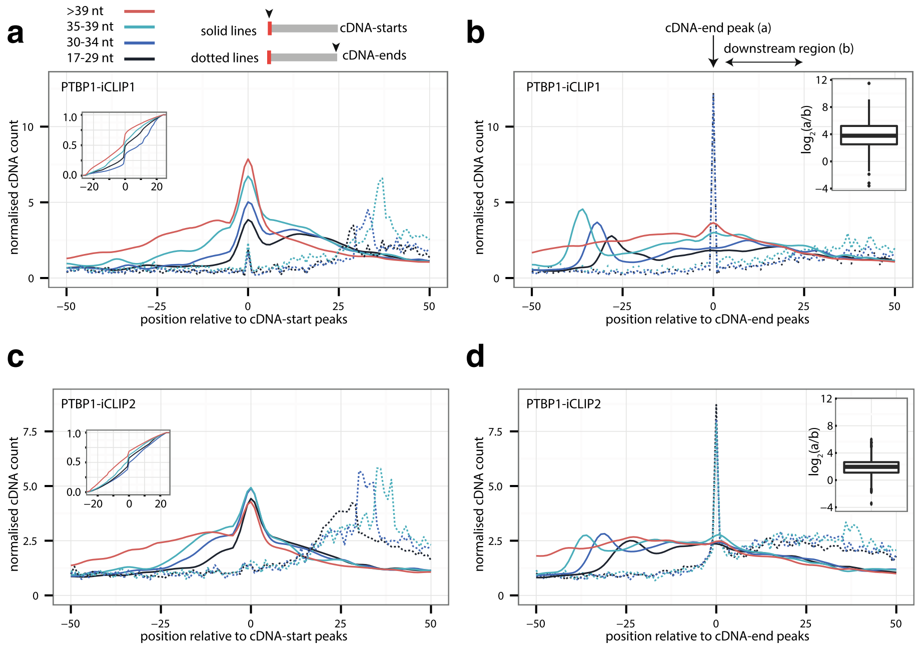


Fig. 4 Non-coinciding cDNA-starts are a result of constrained cDNA-ends. **a** The cDNA-starts (*solid lines*) and cDNA-ends (*dotted lines*) of PTBP1-iCLIP1 are plotted around the cDNA-start peak that was identified within each of the 1000 > 30 nt long crosslink clusters that have the highest total cDNA count. cDNAs are divided into four length categories: 17–29 nt, 30–34 nt, 35–39 nt and >39 nt. The *inner plot* shows the empirical cumulative distribution from all four length categories in the region between −25 nt and 25 nt around cDNA-start peaks. **b** As in (**a**), but plotted around the cDNA-end peak that was identified within the 30 nt downstream of each of the 1000 > 30 nt long crosslink clusters that have the highest total cDNA count. The *inner box plot* shows the ratio of cDNA counts ($\log_2$) at the position 0 (overlapping with cDNA-end peak) compared to the average count of cDNAs in the region from 5 nt to 25 nt downstream of the cDNA-end peak (marked by *horizontal arrow*). **c** As in (**a**), but for PTBP1-iCLIP2. **d** As in (**b**), but for PTBP1-iCLIP2

conclude that the presence of non-coinciding cDNA-starts is reduced in PTBP1-iCLIP2 due to the decreased constraints at cDNA-ends.

PTBP1 binding sites can be assigned correctly despite non-coinciding cDNA-starts

Analysis of the PTBP1-iCLIP1 demonstrated that the non-coinciding cDNA-starts are most enriched within long crosslink clusters. Thus, we speculated that analysis of long binding sites might detect non-coinciding cDNA-starts also in those iCLIP datasets previously analysed by the iCLIPro tool, even if they had not been detected by iCLIPro. For example, U2AF2-iCLIP appears to lack non-coinciding cDNA-starts when analysed by iCLIPro [8] and so we looked at iCLIP data for this protein in more detail.

Both PTBP1 and U2AF2 preferentially bind to pyrimidine tracts (Y-tracts) [14, 15, 18] and therefore we defined the coordinates of potential PTBP1 and U2AF2 binding sites independently of iCLIP data. Specifically, we compared the crosslinking of these two proteins across the longest computationally identified Y-tracts that are annotated in the human genome as T-rich or TC-rich ‘low complexity sequences’ and are located mainly at deep intronic positions. Interestingly, PTBP1-iCLIP1 and U2AF2-iCLIP have a similar presence of non-coinciding cDNA-starts within these long Y-tracts. The short iCLIP cDNAs identify the crosslink sites close to the 3′ region of the Y-tracts, while longer cDNAs identify crosslink sites that are located further towards the 5′ region (Fig. 5a, b). These non-coinciding cDNA-starts of all cDNA length categories correctly identify crosslink sites, because they are enriched almost exclusively within the Y-tracts, which these proteins are known to bind (Additional file 1: Figure S3).

Notably, the cDNA-ends are constrained to positions downstream of the Y-tracts both in PTBP1-iCLIP1 and U2AF2-iCLIP (Fig. 5c, d). Since cDNA-ends represent the sites of RNase cleavage, this most likely indicates inefficient RNase cleavage within the Y-tracts. Thus, the presence of non-coinciding cDNA-starts reflects constrained positions of cDNA-ends. In this context, the broad size range of iCLIP cDNAs is crucial to overcome the constraints at cDNA-ends, thereby enabling the non-coinciding cDNA-starts to detect crosslink events across the long Y-tracts. Moreover, the long cDNAs are particularly important, since they can truncate at crosslink sites that are located far from the site of RNase cleavage. In doing so, they identify crosslink sites across the entire length of the long binding sites.

Our analysis of Y-tracts indicates that non-coinciding cDNA-starts do not necessarily have a negative effect on the assignment of binding sites. To examine this notion more comprehensively, we compared the sequence features of crosslink clusters defined by PTBP1-iCLIP1 and PTBP1-iCLIP2. First, we identified PTBP1-binding motifs by searching for pentamers that are most highly enriched around the cDNA-starts in PTBP1-iCLIP2 (see ‘Methods’). Then, we visualised the position of these PTBP1-binding motifs around the crosslink clusters that were identified by cDNA-starts in the different iCLIP, eCLIP or irCLIP experiments (Fig. 5e). This confirmed that enrichment of the PTBP1-binding motifs correctly overlaps with the starts and ends of crosslink clusters regardless of which crosslinking type or which variant of library preparation protocol was used. Moreover, the high prevalence of non-coinciding cDNA-starts in PTBP1-iCLIP1 does not affect the high resolution of the method. Taken together, we conclude that the use of cDNA-starts is appropriate for the computational analysis of data produced by iCLIP or any related method that is capable of efficiently amplifying truncated cDNAs.

Efficient RNase I-mediated RNA fragmentation minimises the cDNA-end constraints

The cDNA-end corresponds to the position where the RNA fragment was cleaved by the RNase (Fig. 1a, step 2). As described earlier, we optimised the conditions of RNase treatment in the PTBP1-iCLIP2 experiment to ensure that RNase I is the primary cause of RNA fragmentation. This indicates that RNA fragmentation by other factors might have caused the high cDNA-end constraints in PTBP1-iCLIP1. To investigate this possibility, we first assessed cDNA positions at the 3′ splice sites, since these are subject to endogenous RNA cleavage by the spliceosome. While PTBP1 binds to Y-tracts at specific 3′ splice sites to repress alternative splicing, U2AF2 binds to most 3′ splice sites [14, 15, 18]. Interestingly, a peak of cDNA-ends is present at the last intronic nucleotide, even though most cDNA-ends are in the exonic sequence (Fig. 6a). The U2AF2-iCLIP cDNAs of all length categories that end in terminal part of introns have non-coinciding cDNA-starts (Fig. 6b), while the cDNAs that end in the exon have fully coinciding cDNA-starts (Fig. 6c).

The intronic U2AF2-iCLIP cDNA-ends are constrained to the position where the 3′ splice site is cleaved by the endogenous spliceosome. However, the cDNA-ends in exons are not constrained, most likely because they result from cleavage of pre-mRNA by RNase I. To test this hypothesis, we exploited the fact that intron lariats lack a phosphate at their 3′ end and thus no 3′ dephosphorylation is needed at step 2 of the protocol to detect them in iCLIP (Fig. 1a). We therefore produced another PTBP1 iCLIP experiment (PTBP1-iCLIP3), in which we omitted dephosphorylation from step 2 and continued directly to ligation of the 3′ adapter in step 3 (Fig. 1a). Since RNA fragments cleaved

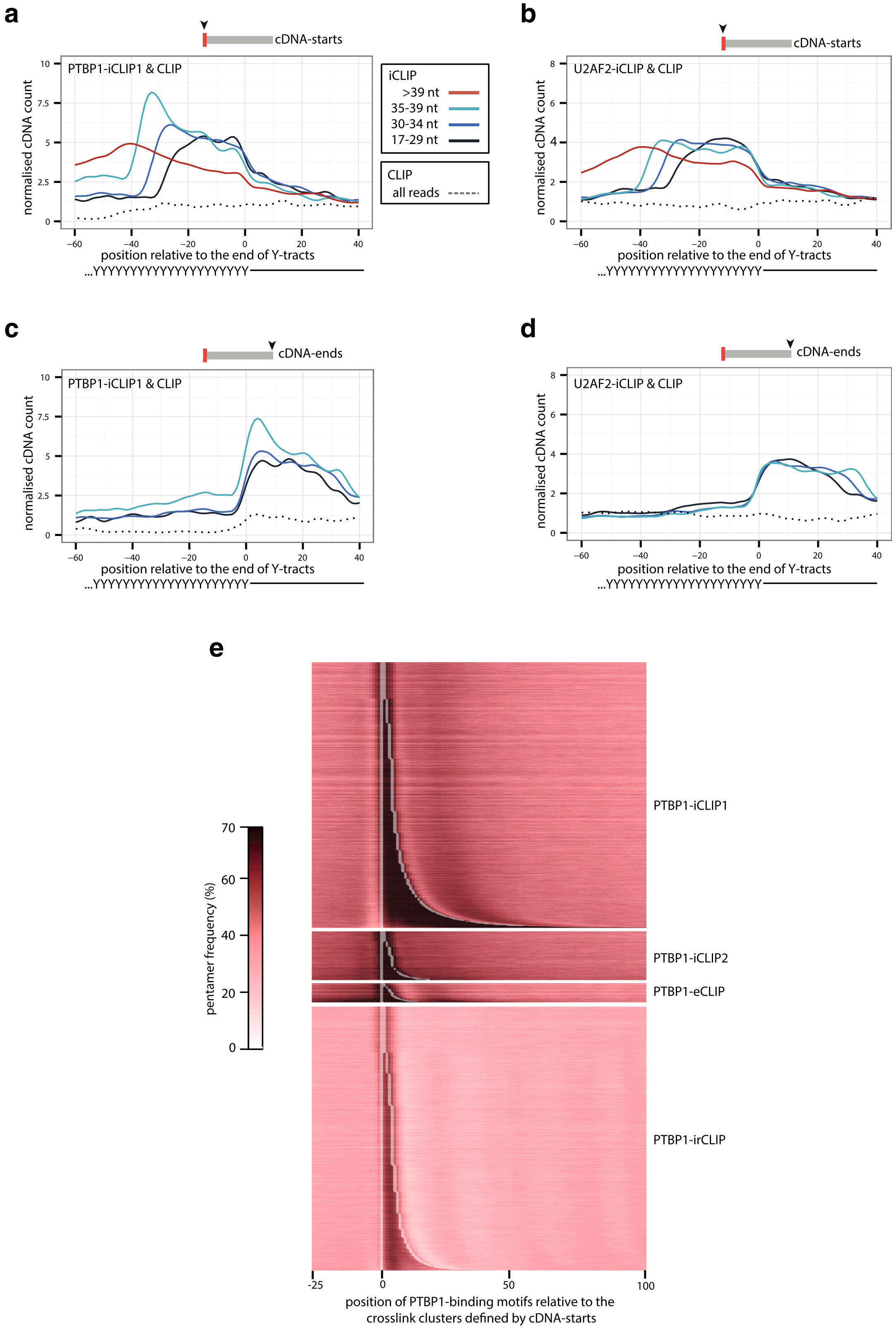


Fig. 5 (See legend on next page.)

(See figure on previous page.)
Fig. 5 Non-coinciding cDNA-starts are required to map the crosslink sites within Y-tracts. **a** The cDNA-starts of PTBP1-iCLIP1 and CLIP experiments are plotted around the ends of >35 nt Y-tracts that are annotated as T-rich or TC-rich low-complexity sequence in the human genome (hg19). cDNAs of PTBP1-iCLIP1 are divided into four length categories: 17–29 nt, 30–34 nt, 35–39 nt and >39 nt**. b** Same as (**a**), but using U2AF2-iCLIP and CLIP cDNAs. **c** Same as (**a**), but showing the positions of cDNA-ends. **d** Same as (**b**), but showing the positions of cDNA-ends. **e** *Heatmap* showing the coverage of PTBP1-binding motifs at the PTBP1-iCLIP1, PTBP1-iCLIP2, PTBP1-eCLIP or PTBP1-irCLIP crosslink clusters that were defined with a 3-nt clustering window. Each *row* shows the average coverage for 300 clusters of similar length, sorted from shortest to longest clusters. The *white line* marks the nucleotide preceding the start and the nucleotide following the median end of all clusters that were combined in each row. A *colour key* for the coverage per nucleotide of the PTBP1-binding motifs is shown on the *right*

at their 3′ end by RNase I contain a 3′ phosphate, they were not ligated to the 3′ adapter (Fig. 1a, step 3) and therefore only those RNA fragments cleaved by other means were amplified in PTBP1-iCLIP3. Notably, both in PTBP1-iCLIP1 and PTBP1-iCLIP3, the cDNA-ends at 3′ splice sites are strongly constrained to the end of introns, while this constraint is minor in PTBP1-iCLIP2 (Fig. 6d–f). Thus, non-coinciding cDNA-starts predominate at 3′ splice sites in PTBP1-iCLIP1 and PTBP1-iCLIP3, while in PTBP1-iCLIP2 most cDNA-starts coincide in the region of 20 nt to 5 nt upstream of the intron-exon junction. This suggests that the RNAs overlapping with the 3′ splice sites were fragmented by spliceosome-mediated cleavage in PTBP1-iCLIP1 and PTBP1-iCLIP3 and by RNase I in PTBP1-iCLIP2 and in U2AF2-iCLIP. It is this difference that appears to explain the variation in the prevalence of non-coinciding cDNA-starts at 3′ splice sites.

To further compare the characteristics at cDNA-ends, we visualised the nucleotide composition of cDNA-ends for the three PTBP1 iCLIP experiments. In PTBP1-iCLIP2, for which we used the increased RNase I concentration, we observed almost no sequence biases at cDNA-ends, in agreement with the lack of sequence specificity of RNase I (Fig. 6h). In contrast, a preference for adenosines was observed at the cDNA-ends in PTBP1-iCLIP1 and PTBP1-iCLIP3, suggesting that this preference results from an RNase I-independent fragmentation of RNAs (Fig. 6g–i). Spliceosome-mediated RNA cleavage contributes to only about 0.1% of these fragments (Fig. 6j) and therefore the primary cause of RNase I-independent fragmentation remains to be identified. Nevertheless, we can clearly conclude that the efficient use of RNase I avoids the constraints at cDNA-ends in iCLIP and this decreases the incidence of non-coinciding cDNA-starts.

Sequence or structure preferences of RNA fragmentation can constrain the cDNA-ends

Both U2AF2 and PTBP1 primarily bind to pre-mRNAs and therefore we also wished to assess the impact that constraints at cDNA-ends may have on RBPs binding mature mRNAs. For this purpose, we examined the iCLIP and CLIP cDNA libraries produced for eIF4A3. The position of cDNA-starts varies greatly between different eIF4A3 experiments (Fig. 7a). As observed by the previous study, the cDNA-starts in eIF4A3-iCLIP1 are shifted to positions upstream of the expected EJC-binding region (yellow rectangle in Fig. 7a) [8]. In contrast, the cDNA-starts of eIF4A3-iCLIP2 and eIF4A3-iCLIP3 overlap with the expected binding region. The cDNA-starts of eIF4A3-CLIP are shifted upstream of eIF4A3-iCLIP2 and eIF4A3-iCLIP3, which agrees with the likely prevalence of truncated cDNAs in iCLIP and readthrough cDNAs in CLIP.

Next, we asked how the position of cDNA-ends may influence the position of cDNA-starts. For this purpose, we first examined the positions of cDNA-ends by summarising them across all exon-exon junctions. The cDNA-ends in eIF4A3-iCLIP1 are highly enriched in a region immediately downstream of the expected EJC-binding region (mainly positions −18..0 nt relative to the junctions), but they are also more broadly distributed further downstream of the expected EJC-binding region, including in the downstream exon (Fig. 7b). To understand why the positions of cDNA-ends are so different in eIF4A3-iCLIP1 compared to the remaining experiments, we first identified the cDNA-end peak, corresponding to the position with the highest count of cDNA-ends at each junction. We then grouped all exon-exon junctions that had the same position of the cDNA-end peak relative to the junction. Both in eIF4A3-iCLIP1 and eIF4A3-iCLIP2, each junction has a preferred position of cDNA-ends, indicating that both experiments show a strong cDNA-end constraint at individual junctions (marked by blue rectangle in Additional file 1: Figure S4A, B).

To further understand the causes for different cDNA-end positions in eIF4A3-iCLIP1, we assessed the RNA sequence and structure preference at cDNA-ends. The cDNA-end peak in eIF4A3-iCLIP2, but not eIF4A3-iCLIP1, coincides with a strong decrease in pairing probability (Additional file 1: Figure S4C, D). Since most endonucleases cut at single-stranded RNA, this indicates that the RNA fragments were cut at their 3′ end by an endonuclease in eIF4A3-iCLIP2, while additional factors may contribute to the RNA fragmentation in eIF4A3-iCLIP1. In eIF2A3-iCLIP2, but not eIF4A3-iCLIP1, we also observe a second peak of cDNA-ends precisely at

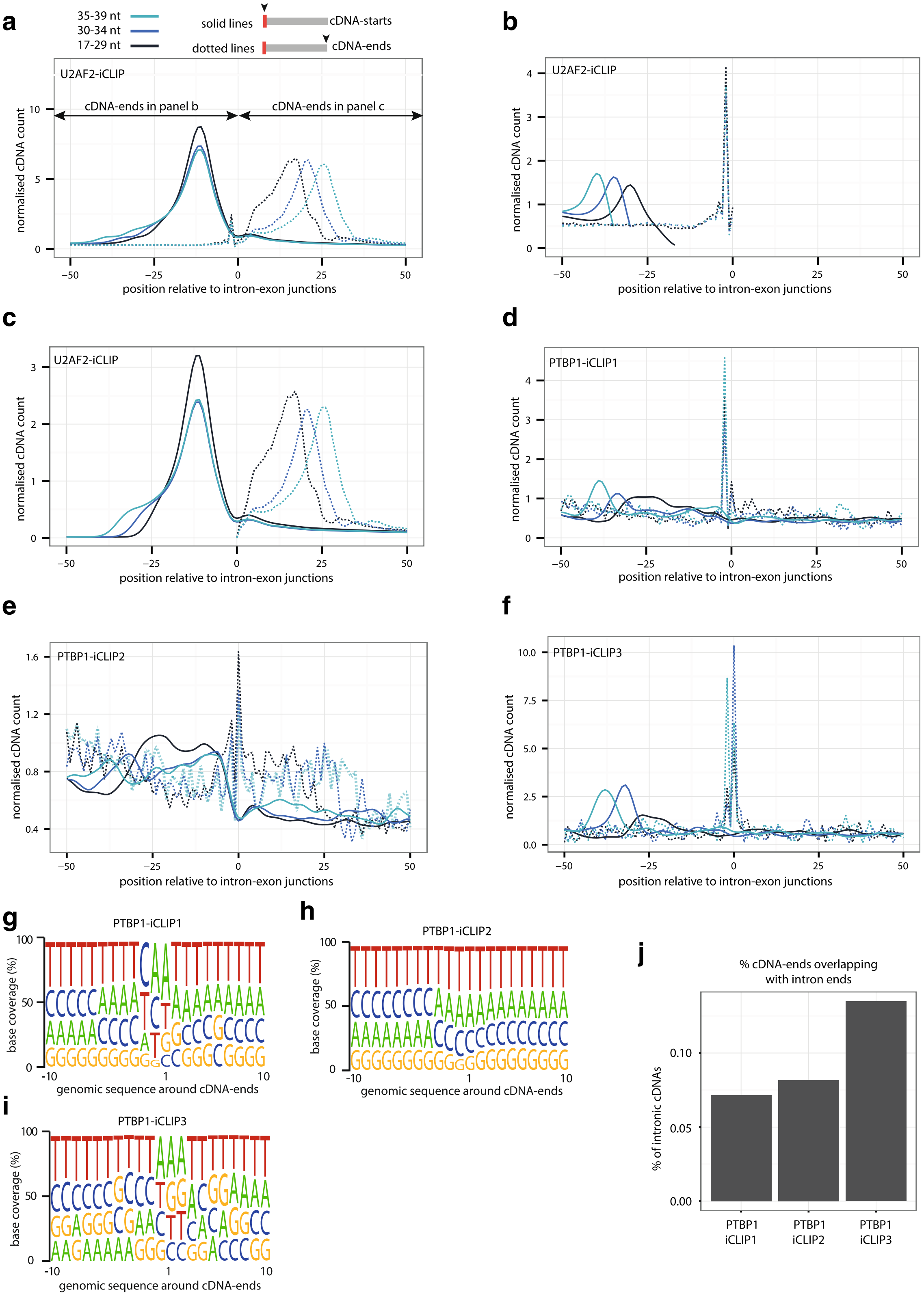


Fig. 6 (See legend on next page.)

(See figure on previous page.)

Fig. 6 Constrained cDNA-ends affect the cDNA-starts at 3′ splice sites. **a** The cDNA-starts (*solid lines*) and cDNA-ends (*dotted lines*) of U2AF2-iCLIP are plotted around intron-exon junctions (position 0 being the first nucleotide of the exon). cDNAs are divided into three length categories: 17–29 nt, 30–34 nt and 35-39 nt; the distribution of all cDNAs together is shown in *grey*. **b** Same as (**a**), but using only cDNAs that end in the intron. **c** Same as (**a**), but using only cDNAs that end in the exon. **d** Same as (**a**), but showing PTBP1-iCLIP1 cDNA-starts (*full lines*) and cDNA-ends (*dotted lines*). **e** Same as (**a**), but showing PTBP1-iCLIP2 (using 4SU and optimised RNase conditions) cDNA-starts (*full lines*) and cDNA-ends (*dotted lines*). **f** Same as (**a**), but showing PTBP1-iCLIP3 (omitting 3′ dephosphorylation) cDNA-starts (*full lines*) and cDNA-ends (*dotted lines*). **g** The composition of genomic nucleotides around iCLIP cDNA-ends from PTBP1-iCLIP1. **h** Same as (**g**), but showing PTBP1-iCLIP2. **i** Same as (**g**), but showing PTBP1-iCLIP3. **j** Proportions of cDNAs that map to introns which contain cDNA-ends at positions overlapping the last two nucleotides of introns. PTBP1-iCLIP1 and PTBP1-iCLIP2 are compared to PTBP1-iCLIP3 iCLIP, which was performed without using PNK to dephosphorylate RNAs in step 2. This enriches for RNAs that contain a 3′ OH, which can occur when they are cleaved at their 3′ end independently of RNase I, such as the 3′ ends of intron lariats

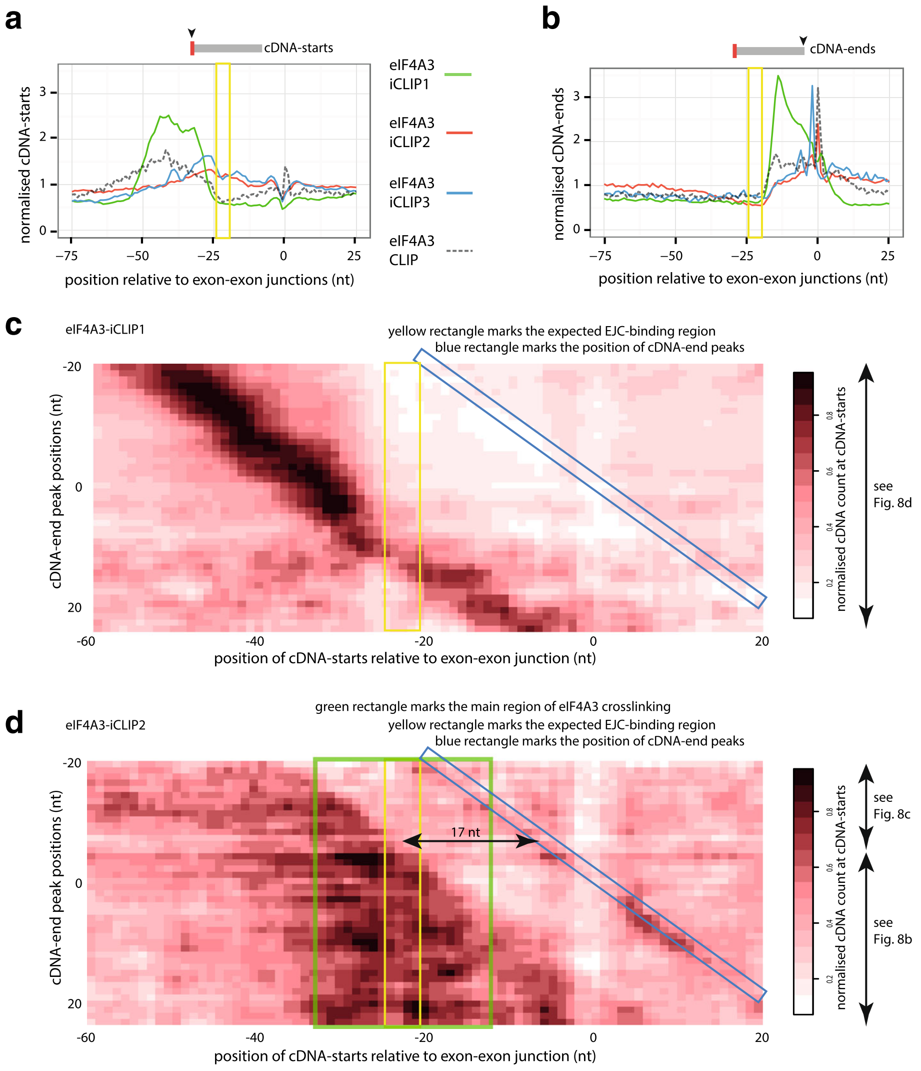


Fig. 7 A broad cDNA length range ameliorates the effects of constrained cDNA-ends. **a** The cDNA-starts of eIF4A3 iCLIP and CLIP experiments are plotted around the 1000 exon-exon junctions with the highest number of cDNAs. **b** Same as (**a**), but showing cDNA-ends. **c** *Heatmap* showing the position of cDNA-starts in eIF4A3-iCLIP1 around the 1000 exon-exon junctions with the highest number of cDNAs. Junctions are sorted according to their cDNA-end peak position. Each *row* shows the average of cDNA counts at all junctions with a cDNA-end peak at the indicated position. The values are normalised against the maximum value across all rows. On the *right*, the *arrows* mark parts of the figure in which binding site assignment corresponds to the *schematic* shown in Fig. 8d. *Coloured rectangles* mark the main region of eIF4A3 crosslinking (*green*), the expected EJC-binding region (*yellow*) and the position of the cDNA-end peak (*blue*). **d** Same as (**c**), but for eIF4A3-iCLIP2. The *arrow* in the figure marks the 17 nt minimal distance between cDNA-starts and the expected EJC-binding region that is required for cDNA-starts to be able to identify crosslink sites within the binding site. On the *right*, the *arrows* mark sections that correspond to the schematics shown in Fig. 8c, b

the end of the exon (position −1) (Fig. 7b, Additional file 1: Figure S4B). This probably reflects the deposition of eIF4A3 on the spliced exon intermediate, as would be expected based on previous biochemical studies [22–24]. The nucleotide composition at cDNA-ends also differs between eIF4A3-iCLIP1 and eIF4A3-iCLIP2 (Additional file 1: Figure S4E, F). These differences suggest that RNA was fragmented by different mechanisms in eIF4A3-iCLIP1 and eIF4A3-iCLIP2, but this remains to be fully understood. Both eIF4A3-iCLIP2 and eIF4A3-iCLIP3 have a strong enrichment of adenosine at the position following the cDNA-end peak, indicating a potential for RNase I-independent fragmentation (Additional file 1: Figure S4G). However, the published eIF4A3-CLIP data used RNase T1 to fragment the RNA [11], which prefers to cut after the guanosine, in agreement with a guanosine enrichment at the position preceding the cDNA-end peaks (Additional file 1: Figure S4H). Taken together, these findings indicate that differences in RNA fragmentation conditions can dramatically affect the structural and sequence features at cDNA-ends in CLIP and iCLIP experiments and thus they can constrain the positions of cDNA-ends in multiple different ways. The impact of these constraints becomes clear when aligning the junctions to the position of cDNA-end peak, which demonstrates that the position of each length category of cDNAs is defined by the position of cDNA-ends (Additional file 1: Figure S4I–K).

To understand the constraints at cDNA-ends at the level of individual exon-exon junctions, we examined the exon-exon junctions with highest coverage of cDNAs in greater detail. For this purpose, we focused on the 1000 junctions with the highest cDNA count. This demonstrates that the cDNA-ends are largely restricted to a single position in the eIF4A3-iCLIP3 experiment, while they are more variable in eIF4A3-iCLIP1. As a result, the cDNA-starts often coincide in eIF4A3-iCLIP1, but are fully non-coinciding in eIF4A3-iCLIP3 (Additional file 1: Figure S5). This again demonstrates that the cDNA-end constraints are the primary cause of non-coinciding cDNA-starts in iCLIP. These constraints therefore need to be considered when interpreting the position of binding sites assigned by iCLIP and related methods.

A broad range of cDNA lengths compensates for the constrained cDNA-ends

To understand how the constraints at cDNA-ends influence the positions of cDNA-starts, we grouped all exon-exon junctions that had the same position of the cDNA-end peak and visualised the position of cDNA-starts (Fig. 7c, d). This confirms that the position of cDNA-ends (marked with blue rectangle) has a strong effect on the position of the identified crosslink sites. Notably, this effect is a lot more pronounced for eIF4A3-iCLIP1, because cDNA-starts are enriched within a narrowly defined distance from the cDNA-ends (Fig. 7c). Analysis of the cDNA length distribution of the examined experiments shows that eIF4A3-iCLIP1 has the largest proportion of cDNAs that are shorter than 39 nt (58%) and most of these cDNAs are in the range of 27–38 nt long (Additional file 1: Figure S6). This indicates that a narrow range of cDNA lengths dominates the eIF4A3-iCLIP1 library and this range of cDNA lengths defines the distance at which cDNA-starts are positioned relative to cDNA-ends. For comparison, only 36% of cDNAs are shorter than 39 nt in eIF4A3-iCLIP3 and the size distribution is more even in eIF4A3-iCLIP2 (Additional file 1: Figure S6A). As a result of the narrow range of cDNA-starts, the cDNA-starts in eIF4A3-iCLIP3 rarely identify crosslink sites within the expected EJC-binding region (marked by the yellow rectangle).

In eIF4A3-iCLIP2, cDNAs have a broad range of lengths, which allows to identify crosslink positions over a broad area upstream of the cDNA-end peak, including the expected EJC-binding region (Fig. 7d). Nevertheless, eIF4A3-iCLIP2 does not identify crosslinking within the expected EJC-binding region at a subset of exon-exon junctions; at these junctions, the cDNA-ends are positioned closer than 17 nt to this region (top portion of the heatmap in Fig. 7d). Since only cDNAs longer than 16 nt are normally isolated by the iCLIP procedure and short cDNAs rarely map to a unique genomic position, it would be very challenging to identify crosslink sites closer than 17 nt to cDNA-ends. This demonstrates that to comprehensively identify crosslink sites within the binding region, the cDNA-ends should ideally be at least 17 nt away from the binding region.

In eIF4A3-iCLIP2, the large majority of cDNA-ends are present more than 17 nt downstream of the expected EJC-binding region (Fig. 7b). This decreased constraint on cDNA-ends and the broad range of cDNA lengths are the most likely reasons for the capacity of eIF4A3-iCLIP2 to identify crosslink sites over the EJC-binding region at most exon-exon junctions. Indeed, most crosslinking in eIF4A3-iCLIP2 is seen within the expected EJC-binding region, as well as approximately 10 nt on each side of this region (marked with green rectangle in Fig. 7d). In conclusion, the broad range of cDNA lengths can overcome the cDNA-end constraints by producing the non-coinciding cDNA-starts that can more comprehensively identify crosslink sites.

Discussion

Our study demonstrates that use of cDNA-starts is appropriate to assign protein–RNA crosslink sites with iCLIP and related methods, regardless of the crosslinking method. Our findings also underscore the importance of fully optimising the iCLIP conditions with the goal to produce a broad range of cDNA lengths with a minimal cDNA-end constraint. We find that crosslink sites are assigned by cDNA-starts even if non-coinciding cDNA-starts are present, since these are a result of cDNA-end constraints, which can have diverse causes (Fig. 8). The constrained cDNA-ends become problematic when they are present close to the binding site (Fig. 8c) or when a narrow range of cDNA lengths dominates the library (Fig. 8d). In these cases, only a portion of the binding site might be assigned and this portion is likely to locate at the upstream region of binding sites (Fig. 8c, d). In contrast, a broad range of cDNA lengths can compensate for

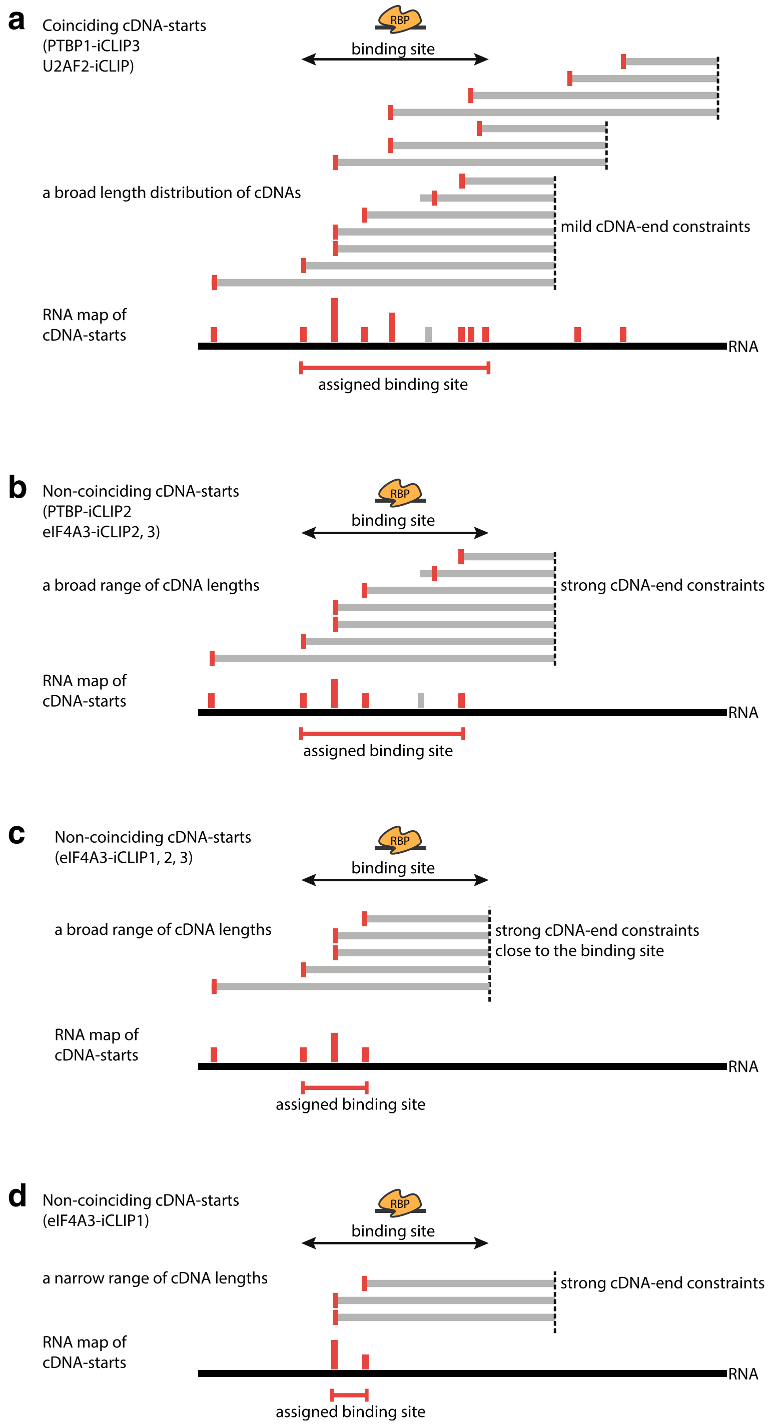


Fig. 8 (See legend on next page.)

(See figure on previous page.)
Fig. 8 A *schematic* explaining how different extents of cDNA-end constraints affect binding site assignment. **a** If the iCLIP library contains a broad range of cDNA lengths and unconstrained positions of cDNA-ends, then crosslink sites are identified in an unbiased manner, allowing assignment of the full binding site (RNA map at the bottom). The crosslink sites assigned by cDNA-starts are marked in *red bars* and a *grey bar* marks a crosslink site that is incorrectly assigned by a readthrough cDNA. **b** If cDNA-ends are constrained, most likely as a result of biased RNase cleavage, then the resulting cDNA-starts do not coincide. Nevertheless, if a broad distribution of cDNA lengths is available and the cDNA-ends are placed far enough from the binding site, then crosslink sites can still be identified across the full binding site, allowing correct assignment, as was seen in the case of eIF4A3-iCLIP2 (Fig. 7d). **c** If cDNA-ends are constrained to a position very close to the binding site, then those cDNAs that truncate at crosslink sites in the 3′ region of the binding site are too short to be isolated and mapped to the genome. Therefore, crosslink sites are identified only in the 5′ region of the binding site, leading to an overly narrow assignment of binding sites, as was seen in some of the sites identified by eIF4A3-iCLIP1 and eIF4A3-iCLIP2 (Fig. 7c, d). **d** If cDNA-ends are constrained and an iCLIP library contains a narrow distribution of cDNA sizes, then cDNA-end constraints lead to an overly narrow assignment of binding regions, as was seen in the case of eIF4A3-iCLIP1 (Fig. 7c)

cDNA-end constraints and in this case the presence of non-coinciding cDNA-starts does not detrimentally influence binding site assignment (Fig. 8b).

A previous study suggested that the non-coinciding cDNA-starts might reflect a prevalence of readthrough cDNAs [8]. Here, we show four independent approaches to examine non-coinciding cDNA-starts in PTBP1 and eIF4A3 iCLIP, all of which lead us to conclude that non-coinciding cDNA-starts are unrelated to readthrough cDNAs. First, we show that CL-motifs are enriched mainly at cDNA-starts in iCLIP, but not in CLIP. This also applies to the PTBP1-iCLIP2 experiment in which 4SU was used for crosslinking. Second, we find a much lower proportion of crosslink-induced deletions in eIF4A3 iCLIP compared to CLIP data, as was also observed previously for other RBPs [4]. Moreover, even though the proportion of transitions is high in the PTBP1-iCLIP2, analysis of CL-motifs indicates that a minority of transitions correspond to crosslink sites, while most crosslink sites overlap with cDNA truncations. This agrees with the enrichment of binding motifs at cDNA-starts in CPSF30 iCLIP, where 4SU was also used for crosslinking [25]. This conclusion is specific for iCLIP, since transitions can precisely assign the crosslink site in PAR-CLIP [21], because only readthrough cDNAs are amplified and usually only one transition is present in the sequenced read. Third, we show that the non-coinciding cDNA-starts in iCLIP result from the constrained cDNA-ends and that their prevalence is greatly diminished when RNase I is the primary source of RNA fragmentation. Fourth, while cDNA-starts of readthrough cDNAs could lead to spurious assignment of crosslink sites upstream of the expected binding regions, we find that the binding sites are correctly assigned by PTBP1-iCLIP1 as well as by eIF4A3-iCLIP2 and eIF4A3-iCLIP3. Thus, we find that prevalence of non-coinciding cDNA-starts is unrelated to the presence of readthrough cDNAs.

The presence of non-coinciding cDNA-starts previously served as an argument for using cDNA-centres instead of cDNA-starts because the cDNA-starts are shifted to the region upstream of the expected binding sites in eIF4A3-iCLIP1 [8]. However, we now find that other eIF4A3 iCLIP experiments also contain non-coinciding cDNA-starts, but their cDNA-starts correctly identify crosslink sites; this indicates that the non-coinciding cDNA-starts are not the cause of shifted binding site assignment in eIF4A3-iCLIP1. We now show that this shift is caused by the presence of cDNA-ends just downstream of binding sites, which is unique to eIF4A3-iCLIP1. We also show that the non-coinciding cDNA-starts are an indirect signature of sequence and structure biases at cDNA-ends, which reflect RNase preferences. It has recently been shown that the sequence bias of RNases can be incorporated into models that predict protein-RNA binding [26]. It remains unclear what causes the unusually high constraints at cDNA-ends in some of the experiments. Multiple sources of RNA fragmentation could lead to such preferences, including the cleavage of intron-exon boundaries upon splicing, specificity of RNA cleavage by exogenous or endogenous RNases, RNA fragmentation during sonication or spontaneous hydrolysis. Non-specific RNases, such as RNase I, should be used instead of the sequence-specific RNases, such as the RNase A, T1 or micrococcal nuclease. Moreover, as we demonstrate with the PTBP1-iCLIP2 experiment, it is important that cleavage by RNase I is the dominant source of RNA fragments. The optimal RNase conditions can be confirmed by visualisation of protein-RNA complexes after their separation with SDS-PAGE, as in the published guidelines [3, 17, 27].

We also show that additional aspects of the iCLIP protocol need careful optimisation to avoid cDNA-end constraints. The 3′ dephosphorylation of RNA fragments needs to be efficient (Fig. 1, step 2), since this is necessary for efficient 3′ adapter ligation to the RNA 3′ ends produced by RNase I (Fig. 1, step 3). While previous studies found sequence and structural biases in the RNA ligase-mediated 3′ adapter ligation [28, 29], we show that PTBP1-iCLIP2 cDNA-ends do not have much sequence bias, indicating that RNA ligation is not the reason for the constraints in the other experiments. However, it is important that the ligation is efficient, so that ideally most RNA fragments become ligated to the 3′ adapter, which minimises potential biases. Finally, purification of cDNAs should be performed in a way that maintains a broad range of cDNA lengths in the final amplified library. This should ideally include isolation of both short and long cDNAs to maximise mapping of crosslink sites that are located either close or far from the site of RNase cleavage, respectively. Moreover, it indicates that special procedures for genomic mapping of short cDNAs may be beneficial; for example, due to the problem that short reads often map at multiple genomic positions, mapping of short cDNAs could be narrowed down to the genomic regions where longer cDNAs map. In sum, it is important to ensure that RNase I is the primary source of RNA fragmentation, that 3′ dephosphorylation of RNA fragments is efficient and that the cDNA library has a broad range of cDNA sizes.

We show that use of cDNA-starts is appropriate to assign protein–RNA crosslink sites with iCLIP. Interestingly, we find that the number of assigned crosslink clusters can vary greatly between the different datasets in a manner that does not necessarily correlate with the number of unique cDNAs that are present in the library (Figs. 1b, 5e). These differences might reflect variable amounts of co-purified non-specific RNAs in the different experiments, which could result from the use of different antibodies and purification conditions. To draw more solid conclusions, direct comparisons between multiple methods will need to be done for a larger number of diverse RBPs.

It is clear, however, that identification of long binding sites can be particularly challenging. Presence of long cDNAs is required to identify crosslink sites across the complete length of long PTBP1 binding sites. Moreover, RNase cleavage sites need to be far enough from the EJC-binding site in order to identify contacts within 10 nt on either side of the expected EJC-binding region by the eIF4A3 iCLIP. This is compatible with the previous findings that the precise position of EJC binding can vary between different junctions, which can be influenced by RNA sequence and structure, or by other RBPs that bind in the vicinity [10, 11, 19]. Moreover, crosslink sites positioned further from the expected binding site might reflect eIF4A3 interactions that are independent of its DEAD-box domain. Thus, analysis of long binding sites with iCLIP experiments can provide valuable insights into mechanisms of protein-RNA complexes.

Conclusions

We find that the presence of non-coinciding cDNA-starts in iCLIP is not a signature of readthrough cDNAs, but instead reflects cDNA-end constraints. These can particularly affect the assignment of long binding sites of RBPs. To overcome these constraints, multiple technical aspects of iCLIP need to be optimised, including the conditions of RNase fragmentation, RNA ligation and cDNA purification. This produces cDNA libraries with a broad cDNA length distribution and low cDNA-end constraints, which are well suited for assigning the complete RNA binding sites of RBPs. These considerations apply to all protocols that amplify truncated cDNAs, including iCLIP, eCLIP and irCLIP, and they ensure that cDNA-starts comprehensively identify protein-RNA crosslink sites across the transcriptome.

Methods

iCLIP experiments

iCLIP experiments are based on the previously described protocol [17] with minor modifications. In PTBP1-iCLIP1 (which was already used for a previous publication [12]), no antiRNase was used and the concentration of RNase I was 0.5 U/mL. In PTBP1-iCLIP2, 4SU was used for crosslinking as previously described [17] and the RNase conditions were optimised to ensure efficient RNase I-dependent fragmentation. In detail, HEK293T cells were grown on 10 cm^2 dishes, incubated for 8 h with 100 μM 4SU and crosslinked with 2× 400 mJ/cm^2 365 nm UV light. Protein A Dynabeads were used for immunoprecipitations (IP). Eighty microlitres of beads were washed in iCLIP lysis buffer (50 mM Tris-HCl pH 7.4, 100 mM NaCl, 1% NP-40, 0.1% SDS, 0.5% sodium deoxycholate). For the preparation of the cell lysate, 2 million cells were lysed in 1 mL of iCLIP lysis buffer and the remaining cell pellet was dissolved in 50 μL urea lysis buffer (50 mM Tris-HCl, pH 7.4, 100 mM NaH_2PO_4, 7 M urea, 1 mM DTT). After the pellet had dissolved, the mixture was diluted with CLIP lysis buffer to 1000 μL, an additional centrifugation was performed and the two lysates were pooled before proceeding (2 mL total volume). As control for purity of protein–RNA complexes, we used a high-RNase condition for analysis by SDS-PAGE gel, but not for further preparation of cDNA library (Additional file 1: Figure S1A). For the full experiment, we incubated 2 mL of lysate with 4 U/mL of RNase I and 2 μL antiRNase (1/1000, AM2690, Thermo Fisher) at 37 °C for 3 min and centrifuged (Additional file 1: Figure S1B). We took care to prepare the initial dilution of RNase in water, since we found that RNase I gradually loses its activity when diluted in the lysis buffer. In total, 1.5 mL of the supernatant was then added to the beads, incubated at 4 °C for 4 h and cDNA library was prepared based on the standard protocol. In PTBP1-iCLIP3, the dephosphorylation step was omitted from step 2 (Fig. 1a) and the rest was same as the standard protocol.

eIF4A3-iCLIP2 was performed from HEK293 cells crosslinked with 0.15 mJ/cm^2 254 nm UV light. To prepare the cell lysate, the cells were lysed with 1 mL iCLIP lysis buffer (50 mM Tris-HCl pH 7.4, 100 mM NaCl, 1% NP-40, 0.1% SDS, 0.5% sodium deoxycholate, final concentration 2 mg/mL) and sonicated (Bioruptor, 5 × 5 s on/off). Two replicates were produced, one with 1 U/mL and the other with 2 U/mL of RNase I in 1 mL of lysate. The SDS-PAGE analysis showed the size of the resulting protein-RNA complexes to be similar (Additional file 1: Figure S1C) and therefore we grouped these two replicates for all analyses of eIF4A3-iCLIP2. After RNase treatment, the samples were centrifuged. For each IP, 100 μL of Protein G Dynabeads were washed in iCLIP lysis buffer and incubated with the anti-eIF4A3 antibody produced in the Le Hir laboratory [11]. The samples were rotated at 4 °C for 2 h. The beads were then washed with high-salt washing buffer (50 mM Tris-HCl pH 7.4, 1 M NaCl, 1% NP-40, 0.1% SDS, 0.5% sodium deoxycholate). After the first round of washes, the samples proceed through 3′ adapter addition, an additional phosphorylation (0.2 μL PNK, 0.4 μL cold ATP [1 mM], 0.4 μL 10× PNK buffer, 3 μL water). After SDS-PAGE separation, the guidelines recommend isolating radioactive RBP-RNA complexes that migrate 20–100 kDa higher than the RBP alone, which leads to isolation of RNA molecules of 50–300 nt. Since we expect that most cDNAs truncate at crosslink sites within these RNA molecules, we prepared the iCLIP library with a heterogeneous population of cDNAs that were 30–140 nt long (Additional file 1: Figure S1D). We then produced sequence reads of 150 nt using the Illumina MiSeq platform for PTBP1-iCLIP2 and 120 nt using the Illumina HiSeq platform for eIF4A3-iCLIP2, thereby obtaining sequences for cDNAs up to a length of 139 or 109 nt, respectively (after removal of adapters).

eIF4A3-iCLIP3 was performed from HeLa cells based on the previously described protocol [17] with minor modifications. Briefly, HeLa cells were grown on 10 cm^2 dishes and crosslinked with 0.15 mJ/cm^2 254 nm UV light. Protein A Dynabeads were used for IPs. For each IP, 40 μL of beads were washed in iCLIP lysis buffer and incubated with 5 μL of anti-eIF4A3 antibody produced in the Le Hir laboratory [11]. For the preparation of the cell lysate, the cells were scraped from a 10 cm^2 dish and lysed in 1 mL of iCLIP lysis buffer, incubated with of 1 U/mL of RNase I at 37 °C for 3 min and centrifuged. The supernatant was then added to the beads and incubated at 4 °C for 2 h. Afterwards, the beads were washed with IP buffer (10 mM Tris, 150 mM NaCl, 2.5 mM $MgCl_2$, 1% NP-40), RIPA-S buffer (50 mM Tris, 1 M NaCl, 5 mM EDTA, 2 M urea, 0.5% NP-40, 0.1% SDS, 1% sodium deoxycolate) and PNK buffer before proceeding to the iCLIP protocol for cDNA library preparation and Illumina HiSeq sequencing produced 50 nt sequence reads (Additional file 1: Figure S1E, F).

Computational analyses

All the source codes used for the analyses in this paper are released under an open source license compliant with OSI (http://opensource.org/licenses) and are available at the GitHub (https://github.com/jernejule/non-coinciding_cDNA_starts) and Zenodo repository (https://zenodo.org/badge/latestdoi/57377213).

Trimming of the adapter sequences

Before mapping the cDNAs, we removed unique molecular identifiers (UMIs) and trimmed the 3′ Solexa adapter sequence. Adapter sequences were trimmed with the FASTX-Toolkit 0.0.13 adapter removal software, using the following parameters: -Q 33 -a AGATCGGAAG -c -n -l 26.

Mapping of iCLIP sequence data

To map iCLIP sequence data for PTBP1 and all RBPs other than eIF4A3, we used the UCSC hg19/GRCh37 genome assembly and the Bowtie2 version 2.1 alignment software with default settings. More than 81% (1,642,850 of 2,007,824) and 85% (8,585,142 out of 9,634,025) of all cDNAs from the published and newly generated iCLIP data, respectively, mapped uniquely to a single genomic position. To map the eIF4A3 iCLIP data, we compiled a set of the longest mRNA sequence available for each multi-exon gene from BioMart Ensembl Genes 79. We mapped to these mRNAs with the Bowtie2 version 2.1 alignment software, allowing two mismatches. More than 50% (11,935,475 of 23,040,243) of cDNAs from all eIF4A3 iCLIP datasets mapped to a unique mRNA position.

The first 9 nt of the sequenced iCLIP read correspond to the barcode, which contains the experimental identifier that allows to separate experimental replicates, and the UMIs, which allow to avoid artefacts caused by variable PCR amplification of different cDNAs (Fig. 1a, step 8, orange) [3]. We used these UMIs to quantify the number of unique cDNAs that mapped to each position in the genome (for PTBP1) or transcriptome (for eIF4A3) by collapsing cDNAs with the same UMI that mapped to the same starting position to a single cDNA.

Definition of crosslink-associated (CL) motifs

We reasoned that sequence motifs enriched directly at the starts of the mock eCLIP cDNAs might uncover preferences of UV crosslinking, since they are thought to represent a mixture of crosslink sites for many different RBPs and thus they should not reflect sequence specificity of any specific RBP [6]. We therefore examined occurrence of tetramers that overlapped with the nucleotide preceding the cDNA-starts (position –1 nt) in comparison with the ones overlapping with the 10th nucleotide preceding the cDNA-starts (position –10 nt) in PTBP1 mock input eCLIP. We excluded the TTTT tetramer from further analyses, since it is often part of longer tracts of Ts, and therefore its inclusion decreases the resolution of analysis. Thus, the tetramers that are enriched over 1.5-fold at position –1 compared to –10 include TTTG, TTTC, TTGG, TTTA, ATTG, ATTT, TCGT, TTGA, TTCT and CTTT, and these were considered for all analyses of 'CL-motifs'.

Definition of crosslink clusters

The crosslink clusters were identified by False Discovery Rate peak finding algorithm from iCount (https://github.com/tomazc/iCount), by assessing the enrichment of cDNA-starts at specific sites compared to shuffled data as described previously [30], with the following additional details. At each cDNA-start, the counts of all cDNAs containing their cDNA-start at a maximum spacing of 15 nt were summed up (or at 3 nt spacing for Fig. 5e). Then crosslink clusters were defined by using the cDNA-starts with counts that passed the false discovery rate < 0.05 significance threshold (determined by comparing the count distribution to shuffled data). Neighbouring clusters that were less than 21 nt apart (or 3 nt apart for Fig. 5e) were then merged into single clusters.

Definition of cDNA-start peak and cDNA-end peak

The peak position of cDNA-starts was identified by comparing the counts at each cDNA-start and choosing the position with the maximum count within each defined region from the top 1000 exon-exon junctions that contain the highest number of cDNAs. Peak positions with a low number of cDNAs (less then median number of all top cDNA-start peaks) were ignored. If more than one position of cDNA-starts had equal count, then the position with maximum count that was located closest to the start of the defined region was chosen. The same approach was used to define cDNA-end peaks.

Definition of PTBP1-binding motifs

To identify the motifs bound by PTBP1, we searched for pentamers enriched in the region [−10..10] around the cDNA-start peaks identified in each crosslink cluster defined by PTBP1-iCLIP2. Sixty-nine pentamers had enrichment z-score > 299 and were used as PTBP1-binding pentamers for further analyses. Their sequences are: TCTTT, CTTTC, TCTTC, CTTCT, TCTCT, CTCTC, TTTCT, TTCTC, TTCTT, TTTTC, TCCTT, CTCTT, ATTTC, TTCCT, CTTCC, TTTCC, CCTTT, CTTTT, CCTTC, TCTGT, TTCTG, TCCTC, CTTCA, ATCTT, TGTCT, TCTGC, CTCCT, CCTCT, GTCTT, TCTAT, TCTCC, ATTCC, TTCTA, CTTTG, TATCT, ACTTC, TTATC, CTTAT, CTATT, TTCAT, TTCCA, TCTTG, TTGTC, TTGCT, CTCTA, CTCTG, TATTT, TCCCT, TCATT, TTCCC, CATTT, ATTCT, TTTAC, GTTCT, CTATC, TCATC, CTTTA, TGTTC, TATTC, CATCT, TACTT, CTGTT, CTTGC, ACCTT, TTTCA, TTTGT, TGTTT, CTTGT, ACTTT. All of these pentamers are enriched in pyrimidines, in agreement with the known preference of PTBP1 for UC-rich binding motifs [31].

Visualisation of cDNA distributions with the density graphs (used in Figs. 4a–d, 5a–d, 6a–f, 7a and b, Additional file 1: Figure 3A–E, Additional file 1: Figure 4I–K, Additional file 1: Figure 5A and B)

All normalisations were performed in R (version 3.1.0) together with the 'ggplot2' and the 'smoother' package for the final graphical output. For the analysis of eIF4A3 iCLIP, each density graph (RNA map) shows a distribution of cDNA-starts and cDNA-ends relative to positions of exon-exon junctions or end peaks in mRNAs. To avoid any border effects, we examined only exon-exon junctions within coding regions, excluding the first or the last junction. The number of cDNAs starting or ending at each position on the graph was normalised by the number of all cDNAs mapped to representative mRNAs, the mRNA length and the number of examined exon-exon junction positions, as described below:

$$\mathrm{RNAmap[n]} = \Big((\mathrm{cDNAs[n]/sum(cDNAs)})\Big) * \mathrm{length(mRNAs)/count(exon_{exon}junctions)}$$

where [n] stands for a specific position on the density graph.

To draw the graph, we then used the Gaussian method with a 5-nt smoothing window.

For the analysis of PTBP1 iCLIP and CLIP, each density graph (RNA map) shows a distribution of cDNA-starts and cDNA-ends relative to positions of its binding sites, which were defined using the position of Y-tracts. We obtained genomic positions of all TC-rich and T-rich low complexity sequences that are present in introns in protein-coding genes in the human genome by using the UCSC table browser.

To avoid the effects of variable abundance of intronic RNAs (and occasional presence of highly abundant non-coding transcripts, such as snoRNAs), we then normalised counts at each binding site by the density of cDNAs in the same region. For this purpose, we examined the region of the binding site, as well as 120 nt 5′ and 3′ of the binding site, to find the nucleotide with the largest count of cDNA-starts or ends (depending on whether starts or ends were plotted on the graph), which is referred to as ‘MaxCount’. We thus obtained ‘MaxCount-normalised cDNA counts’ at each position (which were between 0 and 1). For drawing RNA maps, we wished to examine the enrichment of cDNA counts within binding sites compared to nearby regions outside of binding sites. We therefore calculated the average ‘MaxCount-normalised cDNA counts’ at each position across the evaluated binding sites and divided this count at each position by the average ‘MaxCount-normalised cDNA counts’ in the region 50–100 nt downstream of the binding site, as described in the formula below:

$$\begin{aligned}\mathrm{RNAmap}[\mathrm{n}] \;=\; & \text{average normalised cDNAs}[\mathrm{n}] \\ & / \text{ average normalised cDNAs} \\ & \times [50 - 100 \text{ nt downstream of the binding site}]\end{aligned}$$

where [n] stands for a specific position on the density graph.

To draw the graph, we then used the Gaussian method with a 10-nt smoothing window. The empirical cumulative distribution (Fig. 4a, c) were generated in R with stat_ecdf function from ggplot2 package by using frequency of raw cDNA-start counts for each length category in region 25 nt upstream and downstream from cDNA start peak.

Assignment of the cDNA-end peak in eIF4A3 iCLIP

For cDNA-end peak assignment in eIFA3 iCLIP data, we used exons longer than 100 nt that were in the top 50% of the distribution of exons based on cDNA coverage. This ensured that sufficient cDNAs were available for assignment of the putative binding sites. We then summarised all cDNA-end positions in the region −20 to +25 around exon-exon junctions and selected the position with the maximum cDNA count as the ‘cDNA-end peak’.

Analysis of pairing probability

Computational prediction of the secondary structure around the cDNA-end peaks was performed using the RNAfold program with the default parameters [32].

Analysis of cDNA transitions

Density of C-to-T transitions across cDNAs was performed by using the samtools software with the following parameters: samtools calmd –u –u genomic.fasta input_BAM > BAM_with_transitions. This pipeline replaces BAM format mapped cDNA sequences with transitions relative to genomic reference. In the next step of the following pipeline we used a custom python script (available on github repository) that returns a density array of C-to-T transitions for cDNAs that are shorter than 40 nt. For the final visualisation of density graphs, we used the same approach as for all other density figures without additional normalisations.

Additional file

Additional file 1: Figure S1. Quality control of the PTBP1 and eIF4A3 iCLIP. **Figure S2.** CL-motifs are enriched at cDNA deletions and cDNA-starts in U2AF2-iCLIP. **Figure S3.** Analysis of cDNA-starts and cDNA-ends at the start of Y-tracts. **Figure S4.** Constrained cDNA-ends in eIF4A3 iCLIP. **Figure S5.** The impact of cDNA-end constraints on cDNA-starts in eIF4A3 iCLIP. **Figure S6.** Distribution of cDNA sizes in the studied experiments. (PDF 5.42 mb)

Acknowledgements
The authors thank all members of the Ule laboratory for their assistance and discussion; and Prof. Chris Smith for sharing the PTBP1 antiserum.

Funding
This work was supported by the European Research Council (617837-Translate to JU), the Wellcome Trust (103760/Z/14/Z to JU), the German Research Foundation (FOR2333 to KZ) and by the Francis Crick Institute, which receives its core funding from Cancer Research UK (FC001002), the UK Medical Research Council (FC001002) and the Wellcome Trust (FC001002).

Authors’ contributions
NH, IH and JU designed the study and experiments. CH performed the eIF4A3-iCLIP1, ZW the eIF4A3-iCLIP2, IH the eIF4A3-iCLIP3, JK the PTBP1-iCLIP1, PTBP1-iCLIP3 and U2AF iCLIP, and JA the PTBP1-iCLIP2. NH performed all computational analyses. JU and KZ wrote the manuscript with contributions from all co-authors. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Ethics approval and consent to participate
Not applicable.

Author details
[1]Department of Molecular Neuroscience, UCL Institute of Neurology, Queen Square, London WC1N 3BG, UK. [2]The Crick Institute, 1 Midland Road, London NW1 1AT, UK. [3]MRC Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge CB2 0QH, UK. [4]European Molecular Biology Laboratory (EMBL), Meyerhofstrasse 1, 69117 Heidelberg, Germany. [5]Institute of Molecular Biology (IMB), Ackermannweg 4, 55128 Mainz, Germany. [6]Institut de Biologie de l’ENS (IBENS), 46 rue d’Ulm, Paris F-75005, France. [7]CNRS UMR 8197, Paris Cedex 05 75230, France. [8]Molecular Medicine Partnership Unit (MMPU), Im Neuenheimer Feld 350, 69120 Heidelberg, Germany. [9]Department of Pediatric Oncology, Hematology and Immunology, University of Heidelberg, Im Neuenheimer Feld 430, 69120 Heidelberg, Germany. [10]Faculty of Computer and Information Science, University of Ljubljana, Tržaška cesta 25, 1000 Ljubljana, Slovenia. [11]Division of Brain Sciences, Department of Medicine, Imperial College London, London, UK. [12]Buchmann Institute for Molecular Life Sciences (BMLS), Goethe University Frankfurt, Max-von-Laue-Str. 15, 60438 Frankfurt am Main, Germany.

Received: 21 April 2016 Accepted: 8 December 2016
Published online: 16 January 2017

References
1. Ule J, Jensen KB, Ruggiu M, Mele A, Ule A, Darnell RB. CLIP identifies Nova-regulated RNA networks in the brain. Science. 2003;302:1212–5.
2. König J, Zarnack K, Luscombe NM, Ule J. Protein-RNA interactions: new genomic technologies and perspectives. Nat Rev Genet. 2012;13:77–83.
3. König J, Zarnack K, Rot G, Curk T, Kayikci M, Zupan B, et al. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. Nat Struct Mol Biol. 2010;17:909–15.
4. Sugimoto Y, König J, Hussain S, Zupan B, Curk T, Frye M, et al. Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions. Genome Biol. 2012;13:R67.
5. Weyn-Vanhentenryck SM, Mele A, Yan Q, Sun S, Farny N, Zhang Z, et al. HITS-CLIP and integrative modeling define the Rbfox splicing-regulatory network linked to brain development and autism. Cell Rep. 2014;6:1139–52.

6. Van Nostrand EL, Pratt GA, Shishkin AA, Gelboin-Burkhart C, Fang MY, Sundararaman B, et al. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). Nat Methods. 2016;13:508–14.
7. Zarnegar BJ, Flynn RA, Shen Y, Do BT, Chang HY, Khavari PA. irCLIP platform for efficient characterization of protein-RNA interactions. Nat Methods. 2016;13:489–92.
8. Hauer C, Curk T, Anders S, Schwarzl T, Alleaume AM, Sieber J, et al. Improved binding site assignment by high-resolution mapping of RNA-protein interactions using iCLIP. Nat Commun. 2015;6:7921.
9. Le Hir H, Izaurralde E, Maquat LE, Moore MJ. The spliceosome deposits multiple proteins 20-24 nucleotides upstream of mRNA exon-exon junctions. EMBO J. 2000;19:6860–9.
10. Mishler DM, Christ AB, Steitz JA. Flexibility in the site of exon junction complex deposition revealed by functional group and RNA secondary structure alterations in the splicing substrate. RNA. 2008;14:2657–70.
11. Sauliere J, Murigneux V, Wang Z, Marquenet E, Barbosa I, Le Tonqueze O, et al. CLIP-seq of eIF4AIII reveals transcriptome-wide mapping of the human exon junction complex. Nat Struct Mol Biol. 2012;19:1124–31.
12. Coelho MB, Attig J, Bellora N, König J, Hallegger M, Kayikci M, et al. Nuclear matrix protein Matrin3 regulates alternative splicing and forms overlapping regulatory networks with PTB. EMBO J. 2015;34:653–68.
13. Xue Y, Ouyang K, Huang J, Zhou Y, Ouyang H, Li H, et al. Direct conversion of fibroblasts to neurons by reprogramming PTB-regulated microRNA circuits. Cell. 2013;152:82–96.
14. Shao C, Yang B, Wu T, Huang J, Tang P, Zhou Y, et al. Mechanisms for U2AF to define 3′ splice sites and regulate alternative splicing in the human genome. Nat Struct Mol Biol. 2014;21:997–1005.
15. Zarnack K, König J, Tajnik M, Martincorena I, Eustermann S, Stevant I, et al. Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements. Cell. 2013;152:453–66.
16. Zhang C, Darnell RB. Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. Nat Biotech. 2011;29:607–14.
17. Huppertz I, Attig J, D'Ambrogio A, Easton LE, Sibley CR, Sugimoto Y, et al. iCLIP: protein-RNA interactions at nucleotide resolution. Methods. 2014;65:274–87.
18. Spellman R, Smith CW. Novel modes of splicing repression by PTB. Trends Biochem Sci. 2006;31:73–6.
19. Singh G, Kucukural A, Cenik C, Leszyk JD, Shaffer SA, Weng Z, et al. The cellular EJC interactome reveals higher-order mRNP structure and an EJC-SR protein nexus. Cell. 2012;151:750–64.
20. Cordin O, Banroques J, Tanner NK, Linder P. The DEAD-box protein family of RNA helicases. Gene. 2006;367:17–37.
21. Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, et al. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell. 2010;141:129–41.
22. Alexandrov A, Colognori D, Shu MD, Steitz JA. Human spliceosomal protein CWC22 plays a role in coupling splicing to exon junction complex deposition and nonsense-mediated decay. Proc Natl Acad Sci U S A. 2012;109:21313–8.
23. Barbosa I, Haque N, Fiorini F, Barrandon C, Tomasetto C, Blanchette M, et al. Human CWC22 escorts the helicase eIF4AIII to spliceosomes and promotes exon junction complex assembly. Nat Struct Mol Biol. 2012;19:983–90.
24. Steckelberg AL, Altmueller J, Dieterich C, Gehring NH. CWC22-dependent pre-mRNA splicing and eIF4A3 binding enables global deposition of exon junction complexes. Nucleic Acids Res. 2015;43:4687–700.
25. Chan SL, Huppertz I, Yao C, Weng L, Moresco JJ, Yates 3rd JR, et al. CPSF30 and Wdr33 directly bind to AAUAAA in mammalian mRNA 3′ processing. Genes Dev. 2014;28:2370–80.
26. Orenstein Y, Hosur R, Simmons S, Bienkoswka J, Berger B. Sequence biases in CLIP experimental data are incorporated in protein RNA-binding models. bioRxiv. 2016. doi:10.1101/075259
27. Ule J, Jensen K, Mele A, Darnell RB. CLIP: A method for identifying protein-RNA interaction sites in living cells. Methods. 2005;37:376–86.
28. Hafner M, Renwick N, Brown M, Mihailovic A, Holoch D, Lin C, et al. RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries. RNA. 2011;17:1697–712.
29. Zhuang F, Fuchs RT, Sun Z, Zheng Y, Robb GB. Structural bias in T4 RNA ligase-mediated 3′-adapter ligation. Nucleic Acids Res. 2012;40, e54.
30. Wang Z, Kayikci M, Briese M, Zarnack K, Luscombe NM, Rot G, et al. iCLIP predicts the dual splicing effects of TIA-RNA interactions. PLoS Biol. 2010;8, e1000530.
31. Oberstrass FC, Auweter SD, Erat M, Hargous Y, Henning A, Wenter P, et al. Structure of PTB bound to RNA: specific binding and implications for splicing regulation. Science. 2005;309:2054–7.
32. Lorenz R, Bernhart SH, Honer Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, et al. ViennaRNA Package 2.0. Algorithms Mol Biol. 2011;6:26.