“A novel crosslinking and immunoprecipitation method reveals the function of CSTF2tau in alternative processing of snRNAs”

Dissertation

Zur Erlangung des Grades

Doktor der Naturwissenschaften

Am Fachbereich Biologie

Der Johannes Gutenberg-Universität Mainz

Yulia Kargapolova

Mainz, 2016

Dekan:

1. **Berichterstatter:**
2. **Berichterstatter:**

Tag der mündlichen Prüfung: 22.09.16

Declaration

I, Yulia Kargapolova, declare that the work presented in this thesis is my own. This work was carried out in the Center of Thrombosis and Hemostasis of the University Medical Center Mainz. Where any of the content presented is the result of input or data from a related collaborative research this is properly acknowledged in the text such that it is possible to ascertain how much of the work is my own. I have not already obtained a Ph.D. degree in Biology or elsewhere on the basis of this work. Furthermore, I took reasonable care to ensure that the work is original, and, to the best of my knowledge, does not breach copyright law, and has not been taken from other sources except where such work has been cited and acknowledged within the text.

Acknowledgements

I would like to thank my supervisor for giving me the opportunity to work in his lab, for the continuous support of my Ph.D. study and related research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis.

I would like to thank a postdoc in our lab, for her help with establishing and improving the protocols, for useful comments and discussions on analysis of the results. I would like to thank a computational scientist from HPC group of University Mainz, who supported me on the way of getting acquainted with a high-performance cluster Mogon, which was used for data analysis and provided help to install all necessary tools on Mogon. I would like to thank a group leader in Bioinformatics Department of Institute for Molecular Infection Biology in Würzburg, for his valuable advices, which were helpful for analysis of my data. I would like to thank other group members, who made me go further in my research work. Without these people my work would be incomplete.

Abstract

RNA-binding proteins (RBPs) play a crucial role in the regulation of gene expression on various levels. RNA exists in a form of RNA-protein (RNP) particle throughout the life span. The composition and the structure of RNP vary between cell types, upon various stimuli and in time, and effect the fate of RNA. Mutations in RNA-binding proteins therefore can tremendously change the transcriptome profile of the cell and may be detrimental for the cell fate. Numerous mutations in RBPs are linked to inherited human diseases. The study of the mechanisms of recognition of target genes by RBPs and their role on the fate of targets is crucial for understanding disease mechanisms, treating and diagnostics of the human diseases. It may also have important implications for the future diagnostic and potentially therapeutic strategies.

The current study aims to improve currently existing approaches for illuminating the specificity of RNA-binding proteins (such as HITS-CLIP and iCLIP). The existing protocols are capable of detecting the RNA-protein interactions in a living cell. Yet they are characterized by complex biochemistry, usage of radioactivity, which is banned in many laboratories, and a protocol for library synthesis, which is based on an inherently low linker ligation efficiency. These limitations restrict the broad scientific community, especially non-expert laboratories with bio-medical expertise from the usage of the protocols.

To overcome the above mentioned limitations, I modified the protocol. The modified protocol (conCLIP) is free of radioactivity, avoids using RNA-ligation on low input material, thus improving the complexity, robustness and reproducibility of the cDNA synthesis procedure. The protocol also omits the size selection of cDNA, a step necessary to remove adaptor-adaptor contaminants and requires as little as 10 cycles of PCR amplification (which is at least 15 cycles less than in other protocols), thereby reducing non-desired amplification artefacts.

I further designed a pipeline for the analysis of sequencing data, generated by conCLIP and confirmed the performance of both, the protocol and the pipeline by applying it to a known RNA-binding protein, CSTF2tau. The data, generated by conCLIP recapitulate

previously described RNA recognition properties of this protein, it also reveal yet undescribed binding capacities.

Particularly interesting properties of CSTF2tau, which have not been described before, are its binding to 5’ ends of replication-dependent histones as well as specific recognition of some small non-coding RNAs. My work also describes for the first time that small nuclear (sn)RNAs, involved in splicing, are significantly upregulated upon depletion of CSTF2tau. Interestingly, the same group of snRNAs is bound by the protein and the binding occurs at the 3’end of the molecules. I also reveal that a fraction of snRNAs is polyadenylated and this fraction decreases upon CSTF2tau depletion. Moreover, the depletion of the protein stabilizes snRNAs. I propose a model, which suggests a new mechanism by which the level of the snRNAs can be regulated via the activation of internal oligoadenylation sites. Upon depletion, internal polyadenylation sites become less efficient. In contrast, high level CSTF2tau increases the usage of cryptic cleavage sites, which triggers oligoadenylation of snRNAs (presumably with other processing components) and results in their fast degradation.

Table of contents

List of figures

List of tables

List of abbreviations

Abbreviation	Definition
3'-UTR	3' untranslated region
5'-UTR	5' untranslated region
ADP	Adenosine diphosphate
ATP	Adenosine triphosphate
APA	Alternative Polyadenylation
bp	Base pair
cDNA	Complementary DNA
CFIA/B	Cleavage factor IA/IB
CFIIm	Mammalian Cleavage factor II
CFIm	Mammalian cleavage factor I
ChrRNA-Seq	Chromatin-bound RNA sequencing
CLIP	Crosslinking and Immunoprecipitation
coIP	Co-immunoprecipitation
CPSF	Cleavage and polydenylation specificity factor
CS	Cleavage site
CSTF	Cleavage stimulatory factor
DNA	Deoxyribonucleic acid
DTT	Dithiothreitol
EDTA	Ethylenediaminetetraacetic acid
kDa	Kilodalton
NGS	New Generation Sequencing
nt	Nucleotide
RBP	RNA-binding protein
RNP	Ribonucleoprotein
ORF	Open reading frame
PAGE	Polyacrylamide gel electrophoresis
PCR	Polymerase chain reaction
PNK	Polynucleotide kinase
PolII	RNA-polymerase II
qPCR	Quantitative real-time PCR
RNA	Ribonucleic acid
RNA-Seq	RNA sequencing
RRM	RNA recognition motif
RT-qPCR	Reverse-transcription quantitative real-time PCR
scaRNA	Small Cajal body-specific RNA
SDS	Sodium dodecyl sulfate
snRNA	Small nuclear ribonucleic acid
snoRNA	Small nucleolar ribonucleic acid
UV	Ultra-violet
UTR	Untranslated Region

Introduction

1.1. Eukaryotic pre-mRNA processing

Human cells contain approximately 21,000 protein-coding and around 17,000 non-coding genes [1]. Normally not all genes are expressed at the same levels in a single cell. Gene expression varies between cell types and throughout the life span of the cell and is regulated at different levels: on the level of transcription, as well as post-transcriptionally, during RNA processing. Eukaryotic pre-mRNA processing is a complex mechanism of maturation of newly transcribed mRNA, which includes mutually interdependent processes, such as 5'end capping, splicing, 3'end cleavage and polyadenylation, mRNA transport and decay. The above mentioned processes are tightly regulated in the cell and provide additional steps of diversity and gene expression control [2, 3]. This regulation is achieved with help of RNA-binding proteins, which form RNA-protein complexes with changing composition and structure, guiding the RNAs and determining their fate during the life span (further detailed in the next sections).

1.1.1. Pre-mRNA 5' end capping

The first step of pre-mRNA processing, which occurs shortly after transcription initiation, is the 5' end capping (Figure 1). The cap consists of a guanine nucleotide methylated at the N^7 position [4-6]. It is linked to the 5' nucleotide of the RNA through a 5'-5' pyrophosphate linkage (m^7GpppN). The N^7 methyl group is essential for the recognition of cap binding proteins, CBP and eIF4E, as well as for efficient splicing, mRNA polyadenylation, export, translation and stability. Removal of the cap is catalyzed by decapping enzymes, DCP2 and Nudt16, releasing m^7 GDP and 5' monophosphate RNA [7]. The 5' monophosphorylated RNA is rapidly degraded by 5'-3' exoribonuclease XRN1 in cytoplasm.

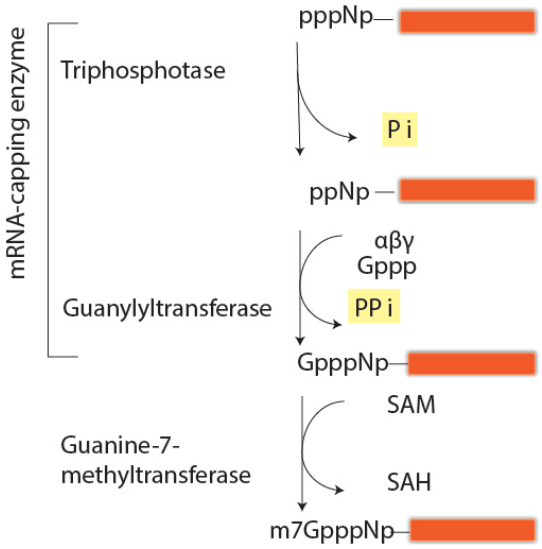


Figure 1 - Sequential steps of 5’ end capping

1.1.2. Pre-mRNA splicing

The pre-mRNA transcripts of eukaryotes often contain non-coding regions (introns), which are removed prior to translation by a process called splicing. Splicing is accomplished by a spliceosome, an RNA-protein complex, which recognizes *cis*-elements on the mRNA to carry out 2 essential transesterification steps of the splicing reaction (Figure 2) [8]. The spliceosome encompasses 5 RNP particles each of which contains a small nuclear (sn)RNAs (U1, U2, U4, U5 and U6) and are assembled on each intron. The major class of introns (U2 type introns) contains consensus sequences for the 5’ splice site, intron branch point and 3’ splice site (Figure 2). The spliceosome recognizes the 5’- (usually composed of AG/GURAGU) and the 3’ splice site (contains a polypyrimidine tract followed by an AG dinucleotide at the actual 3′ splice site) of the intron. Upstream of the 3’ splice site a so called branching point is located (Figure 2). The branching point is a sequence containing nucleophile for the first step of splicing. In the first step of splicing reaction, the 2’-hydroxyl group of a special A residue of the branch point attacks the phosphate at the 5’ splice site. This leads to a cleavage of 5’exon from the intron and a ligation of the intron 5’end to the branching point 2’-hydroxyl. As a result of this step, a 5’ exon and intron/3’-exon fragment in a lariat configuration are produced (Figure 2). The second transesterification reaction occurs when the phosphate at the 3’ end of the intron is attacked by the 3’-hydroxyl of the detached exon. This ligates the two exons and releases the intron in the form of a lariat.

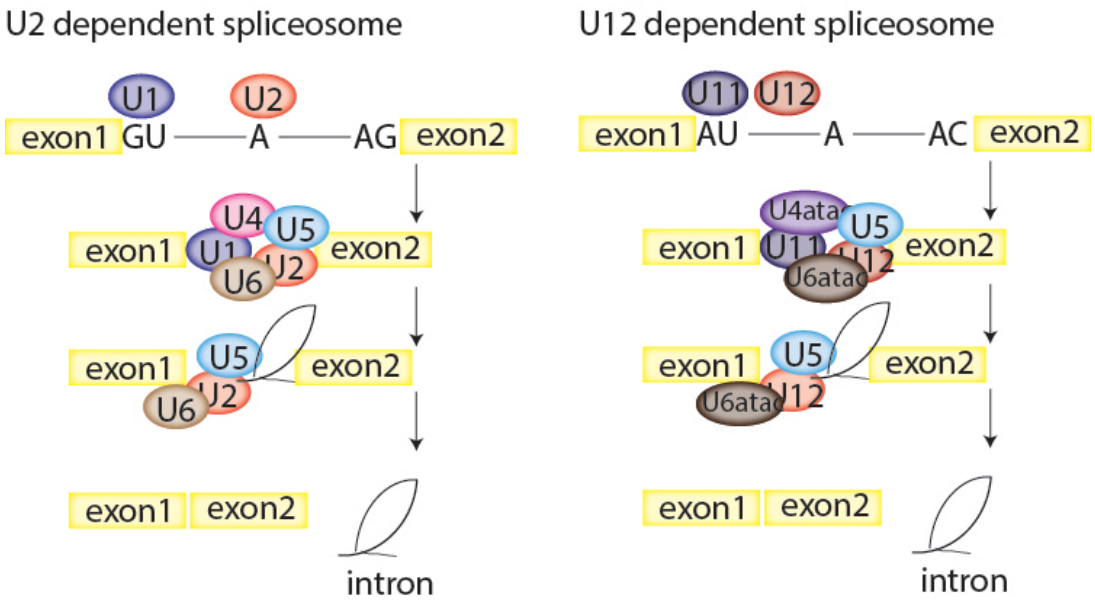


Figure 2 - Schematic overview of splicing carried out by major (U2 dependent) and minor (U12 dependent) spliceosome. The majority of introns in the human transcriptome are spliced by the major spliceosome (left), yet a small proportion (around 1%) is processed by the minor spliceosome complex (right). The introns differ in their *cis*-sequences and are accordingly recognized by different RNP-complexes.

In 1990's a novel class of introns, which includes less than 1% of total number of introns was found. These introns differ by the *cis*-elements and are spliced by so called minor spliceosome complex. The minor spliceosome also consists of 5 snRNAs (U11, U12, U4atac, U6atac and U5). The donor and acceptor sites of U12 type introns are higher conserved comparing to U2 type, whereas the polypyrimidine tract is missing or less pronounced [9].

Apart from the spliceosome, splicing is regulated by other activator and repressor proteins. Such RNA-binding proteins as for example SR proteins usually promote splicing, whereas hnRNPs inhibit it [10]. Splicing can be regulated such that different mRNAs may lack or contain exons. This process, called alternative splicing, is a mean of increasing protein diversity in cells [11]. In human around 95% of genes are alternatively spliced [12]. This allows synthesis of more than one protein isoform from one gene and therefore is an important regulatory step, determining the protein functionality of the expressed gene.

1.1.3. 3’ end processing of polyadenylated RNAs

The majority of transcribed RNAs with exception of replication-dependent histones and some non-coding RNAs, such as small non-coding RNAs, possess poly(A)-tails at their 3’ ends [13]. 3’end processing of polyadenylated pre-mRNAs occurs as a result of a two-step reaction. First the cleavage and polyadenylation site (CS) is recognized and the mRNA is endonucleolytically cleaved and subsequently polyadenylated (Figure 3). The cleavage and polyadenylation reaction is controlled by a complex machinery, which consist of more than 50 proteins [14]. The main components of the cleavage and polyadenylation machinery are described in the following section (1.1.5.). The length of the poly(A)-tail attached to the 3’end varies and is between 50 and 100 nt long [15]. In general polyadenylation of mRNAs is crucial for mRNA export, stability and efficient translation [16].

3’ end processing is tightly interconnected with splicing, transcription and translation [17-19]. An example of such interconnection has been recently reported; two components of cleavage and polyadenylation machinery have been shown to be involved in alternative splicing [20]. On the other hand, the usage of polyadenylation site can be affected by splicing machinery components as has been illustrated on example of inhibition of intronic alternative polyadenylation (APA, see below) sites by U1 snRNP [21, 22].

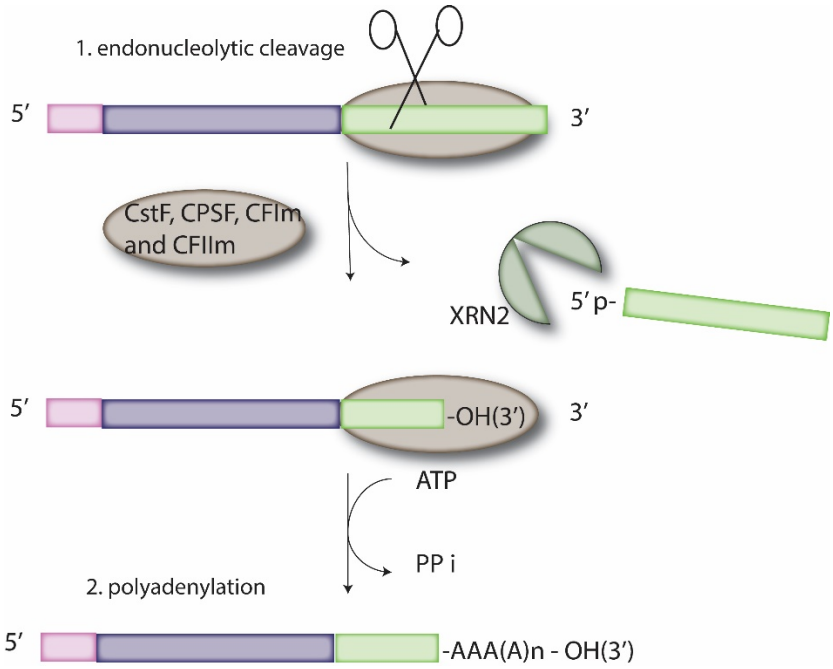


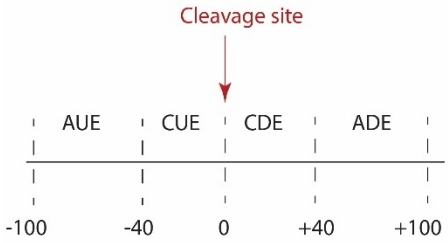
Figure 3 - Schematic overview of sequential steps of 3' end processing. The polyadenylation site is recognized by a cleavage and polyadenylation machinery, which consists of Cstf, CPSF, CFIm and CFIIm sub-complexes and carries out an endonucleolytic cleavage. The 5' end product of endonucleolytic cleavage is further polyadenylated, whereas the 3'-end product is degraded by exonuclease.

Over 70% of human transcripts possess multiple polyadenylation sites [23, 24]. Alternative polyadenylation (APA) is a result of differential selection of CS by the cleavage and polyadenylation machinery. The cleavage and polyadenylation sites are distinct in their nucleotide composition and relative location. APA occurs across different cell types, upon cell differentiation and proliferation, as a result of quantitative and qualitative changes of the components of the cleavage and polyadenylation machinery [25-28]. In some cases alternative polyadenylation occurs internally (internal APA), leading to shortening of the transcript and as a result to expression of a truncated protein, as reported for IgM heavy chain upon B-cells activation [29]. Alternatively, APA occurs within a 3' untranslated region (tandem APA), leading to processing of transcripts with different 3'-prime ends. Recent studies coupled tandem APA leading to global shortening of transcripts with cancer phenotype [30].

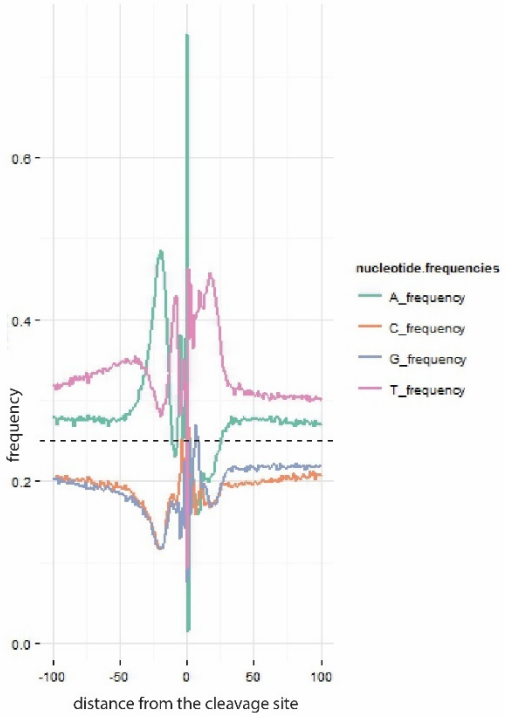
1.1.4. Sequence elements for mRNA 3'-end processing

3' end processing of polyadenylated transcripts relies on specific *cis*-elements, located in a conserved manner around the cleavage sites (see Figure 4A). Mammalian polyadenylation sites contain three main sequence types (polyadenylation signal, upstream sequence elements and downstream sequence elements) that define the polyadenylation site choice. The hexameric polyadenylation signal (PAS) located 10~30 nt upstream of cleavage site was the first discovered element [31]. The most frequently occurring hexamer is AAUAAA (58,2% of sites), next frequently used PAS contains AUUAAA hexamer (detected in 14,9% of sites) [32] (Figure 4C). Another type of sequence elements triggering effective site recognition are downstream sequence elements (core downstream elements, or "CDE" and auxiliary downstream elements, or "ADE"). The CDE elements are less conserved and usually are represented by U/GU rich sequences located 14-70 nucleotides downstream of the cleavage site [33, 34]. The ADE are located 40-100 nucleotides downstream of the cleavage site. They are less defined and are generally G-rich [35, 36] (Figure 4B and C). The elements located upstream of CS (upstream core elements, or "UCE" and auxiliary upstream elements "AUE") do not have a consensus sequence, but are usually U-rich sequences (UUUU) or contain similar motifs (UGUA, UAUA) [32] (Figure 4). Polyadenylation sites are divergent and some are lacking one or more consensus elements. For example more than 30% of human polyadenylation sites lack the AA(U)UAAA hexamer [34] and around 20% of polyadenylation sites possess no U/GU rich CDE [37, 38].

A.



B.



C.

Region	Name	Top 3 hexamers	Percent of Hits
-100/-41	AUE.1	GGGGAG, GUGGGG, GGGUGG	48%
	AUE.2	UUUGUA, GUAUUU, CUGUGU	93%
	AUE.3	UAUAUA, AUAUAU, UUUAUA	51%
	AUE.4	UGUAUA, AUGUAU, UGUAUU	82%
-40/-1	CUE.1	UAUUUU, UGUUUU, UUUUUU	82%
	CUE.2	AAUAAA, AUAAAG, AAAUAA	90%
+1/+40	CDE.1	GUGUCU, CUGCCU, UGUCUC	87%
	CDE.2	UUAUUU, UUUCUU, UGUUUU	90%
	CDE.3	UGUGUG, GUGUGU, CUGUGU	66%
	CDE.4	CUGGGG, UGUCUG, GUCUGU	68%
+41/+100	ADE.1	CCUCCC, CUCCCC, CACCCC	21%
	ADE.2	CCCGCC, CCCCGC, CCCGCG	29%
	ADE.3	GGUGGG, GGCUGG, GGGUGG	74%
	ADE.4	GGGCAG, GGCCAG, GGGGCC	27%
	ADE.5	GGGAGG, GGAGGG, GGGGAG	85%

Figure 4 - Sequence elements around the cleavage and polyadenylation sites. The region around the cleavage site can be split into sequence subsets (A). In each subset a particular cis-acting elements were identified [32]. The frequencies of each type of sequence elements are listed in the table (C). AUE – auxiliary upstream element, CUE – core upstream element, CDE – core downstream element, ADE – auxiliary downstream element. Nucleotide frequencies distribution around the cleavage site (CS; 0 position) calculated based on CS annotation in neuroblastoma cells (B) (own data). Sequences around the CS are in general A/U rich (dashed line marks expected frequency of 25%). Table by Hu et al. [32]

1.1.5. Protein factors for pre-mRNA processing

The mammalian 3' end processing complex contains several sub-complexes each of which recognizes *cis*-elements of the polyadenylation sites (Figure 5). These sub-complexes include the “cleavage and polyadenylation specificity factor” (CPSF), the “cleavage stimulation factor” (CstF), “cleavage factor I” (CFIm), and “cleavage factor II” (CFIIm).

Mammalian CPSF contains six subunits, CPSF-30, CPSF-73, CPSF-100, CPSF-160, WDR33 and hFIP1. The CPSF complex recognizes the AAUAAA hexamer of the polyadenylation site. Recent studies revealed that at least two subunits of the complex (CPSF-30 and WDR33) bind AAUAAA elements *in vivo* [39, 40]. Another component of CPSF

complex, FIP1, recognizes U-rich CUE elements *in vivo* [41, 42]. The cleavage occurs 15-30 nucleotides downstream of the AAUAAA hexamer and is performed by CPSF73 [43].

The mammalian CstF complex contains three proteins, CSTF2 or its paralog CSTF2tau, CSTF77 and CSTF50 [38]. This complex binds the CDE, with particular selectivity to CDE.3 type elements (see Figure 4), preferentially recognized by CSTF2 and its paralog CSTF2tau [44].

Other components of the cleavage and polyadenylation machinery are the CFIm and CFIIm sub-complexes. CFIm is composed of three proteins, namely CPSF5, CPSF7 and CPSF6. The CFIm complex has been shown to bind AUE, containing UGUA consensus [48]. All three proteins exhibit very specific positioning 40-50 nucleotides upstream of cleavage and polyadenylation site [41]. It has been speculated that the CFIm complex participates in recognition of polyadenylation sites at early stages as well as CstF complex and is needed to stabilize the CPSF and CstF sub-complexes on the RNA [41]. The CFIIm complex is composed of two proteins, PCF11 and Clp1 (Figure 5). The CFIIm complex however does not interact with RNA directly. PCF11 is a scaffolding protein, which interacts with PolII CTD, whereas the interacting partner Clp1, bridges CPSF with CFIm complex (Figure 5) [45]. Finally another newly identified 3' processing factor is RBBP6 (retinoblastoma binding protein 6) [14], it also functions by recognizing the RNA [46]. RBBP6 interacts with the Cstf complex and recognizes mRNAs with A/U-rich 3' untranslated regions (UTRs) [46].

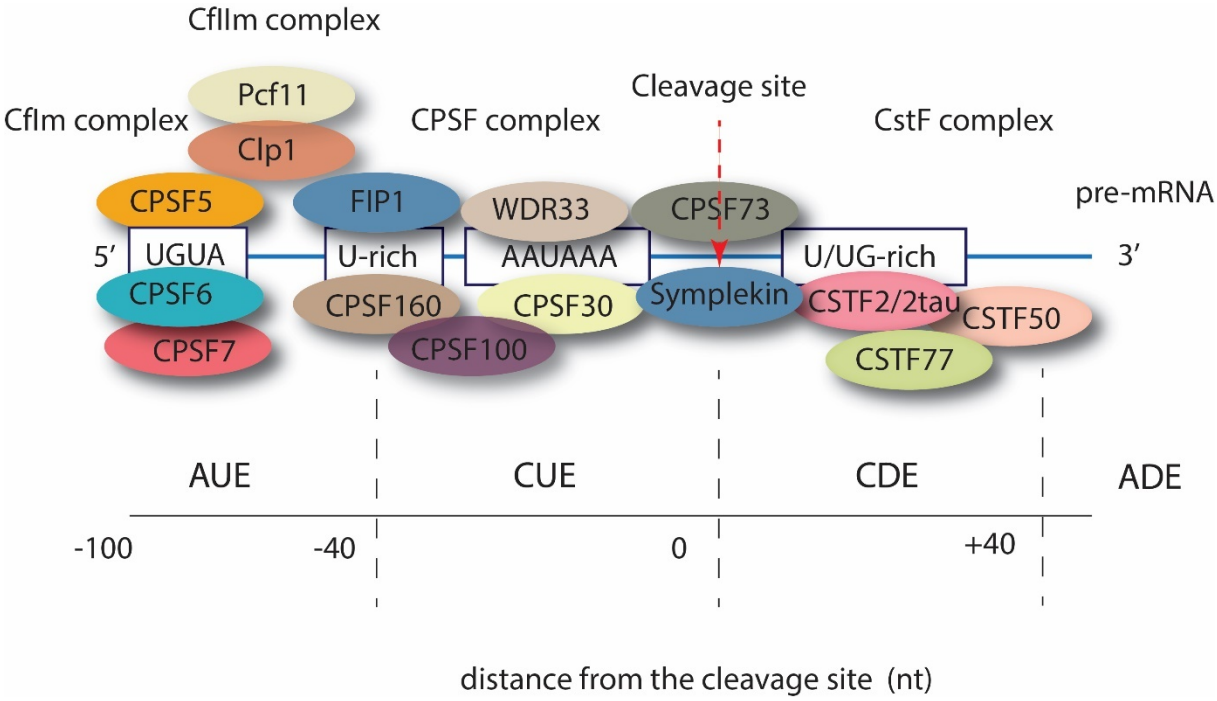


Figure 5 - Schematic overview of protein complexes participating in 3' end processing. The CstF complex is located downstream of the cleavage site and is composed of three proteins (CDE). The CPSF complex contains six subunits and is located -1-40 region upstream of the cleavage site (CUE). CPSF73 is the endonuclease that cleaves the pre-mRNA. The CfIm is located further upstream of cleavage site (-100 region) and recognizes auxiliary upstream sequence (AUE) with high preference for UGUA motifs. The CfIIm complex directly binds template and is associated with PolII and other protein complexes via protein-protein interactions.

1.2. 3' end processing of non-polyadenylated RNAs

Apart from mRNA molecules, whose processing and maturation occurs as described above, mammalian cells contain other types of RNAs processed via different pathways. The next two sections will serve to briefly describe the processing of replication-dependent histones and provide a brief overview of the maturation of non-coding RNAs, namely small nuclear (sn)RNAs and small nucleolar (sno)RNAs.

1.2.1. 3' end RNA processing of replication-dependent histones

Replication-dependent (RD) histone mRNAs are the only type of mRNAs transcribed by RNA polymerase II and believed to be not polyadenylated in the normal cell state [47]. Under certain conditions however polyadenylated ("misprocessed") histone RNA molecules can accumulate [48]. Constitutive 3'end processing of replication-dependent histones depends on two *cis*-elements. The first element is a stem-loop structure, which is bound by a so called

stem-loop binding protein “SLBP” (Figure 6). The second element is a histone downstream element (HDE), which is recognized by the U7 small ribonucleoprotein (U7snRNP), comprised of U7 snRNA, Lsm10, Lsm11 and FLASH proteins [49, 50]. The complex of proteins, containing the endonuclease CPSF73 and other polyadenylaton factors associate with U7snRNP [50]. After binding of these protein complexes, the cleavage of histone pre-mRNA occurs via endonucleolitic cleavage of CPSF73, five nucleotides downstream of the stem-loop (Figure 6). In addition to CPSF73 other proteins are involved in 3’end processing of replication-dependent histones such as CPSF100, Symplekin and CSTF2 [51]. U7 snRNA is possibly not the only snRNA involved in processing of replication-dependent histones. Recently snRNAs U2 and U12 have been shown to promote the processing of replication-dependent histones [52].

The majority of replication-dependent histones contain a cryptic polyadenylation signal downstream of the U7-dependent site [53]. Upon depletion of 3’-end processing factors such as CSTF2 (but not CSTF2tau) the number of “misprocessed” (polyadenylated) replication-dependent histones increases [53]. These molecules are substrates for rapid degradation by the exosome complex [53]. Accordingly, the depletion of the nuclear catalytic subunit of the exosome, DIS3 and co-depletion of CSTF2 leads to an accumulation of unprocessed histone precursor H3C in HeLa cells [54].

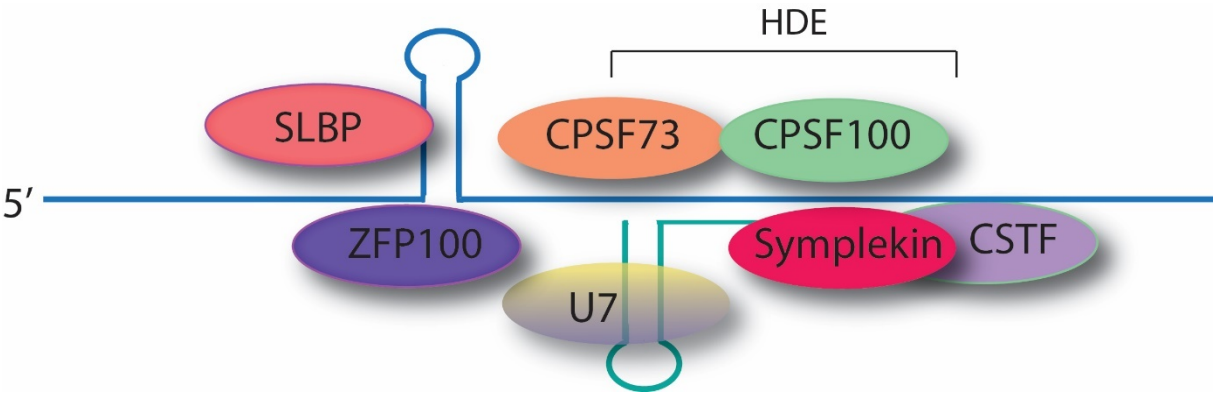


Figure 6 - Schematic view of RNA-protein complexes participating in the processing of replication-dependent histones.

1.2.2. Biogenesis, transcription and processing of non-coding RNAs

Non-coding RNAs (ncRNAs) commonly include RNAs such as microRNAs, snoRNAs as well as other small regulatory RNAs and longer transcripts many of which are poorly characterized [55]. In the context of this work, the biogenesis and processing of two groups of ncRNAs, namely sn- and snoRNAs, will be explained in more detail. Further information on the other ncRNAs can be found elsewhere [56, 57].

1.2.2.1. Processing and maturation of U-type snRNAs

U-type snRNAs are uridine-rich small ncRNAs. They form a moiety of the major (U1, U2, U4, U5 and U6) and minor (U11, U12, U4atac, U6atac and U5) spliceosome (see section 1.1.2). The exceptional U7 snRNA is not involved in splicing, but participates in 3'-end processing of replication-dependent histones (see section 1.2.1).

U-type snRNAs are transcribed by RNA-polymerase II (RNAPII) with exception of U6 and U6atac molecules, which are transcribed by RNA-polymerase III (RNAPIII) [58] and processed as typical RNAPIII transcribed transcripts [59-62]. The RNAPIII terminates at short T-stretches; as a result the nascent U6 snRNA particles are terminated with a hydroxyl group and can vary in length. Upon transcription termination, the U6-specific poly(U) polymerase TUTase, TUT1, adds up to 20 uridines to the 3'end of the molecule [63]. The mature U6 snRNA 3'-end is formed by the 3'-5'exonuclease Usb1, which removes the last uridine to form a 2'-3' cyclic phosphate [64]. This modification stabilizes the molecule. If lacking the cyclic phosphate, the U6 snRNAs are accessible for poly(A) polymerase and directed to degradation [65].

In contrast, the processing of RNAPII transcribed snRNAs (U1, U2, U4, U5, U11, U12 and U4atac) is carried out by a different mechanism. The first step of the processing occurs in the nucleus, where the 3' end endonucleolytic cleavage takes place (Figure 7). Recent evidences suggest that this cleavage is accomplished by the components of the integrator complex, IntS11 and IntS9 [66]. The closest homologs of IntS11/9 are CPSF73 and CPSF100. The exact composition of the integrator protein complexes responsible for maturation of human snRNAs remains to be discovered [67].

The correct processing of U-type snRNAs requires the promoter sequence, the 3' box located 9-19 nucleotides downstream of the mature 3'end of the molecules and the carboxy-terminal domain (CTD) of the largest RNAPII subunit, Rpb1 [59-62, 68] (Figure 7). The signals triggering the gathering of the Integrator complex on U-type snRNAs are still unknown. Most probably the complex is directed by the presence of several elements and factors mediating the interaction between Integrator complex and U-type pre-RNAs [59].

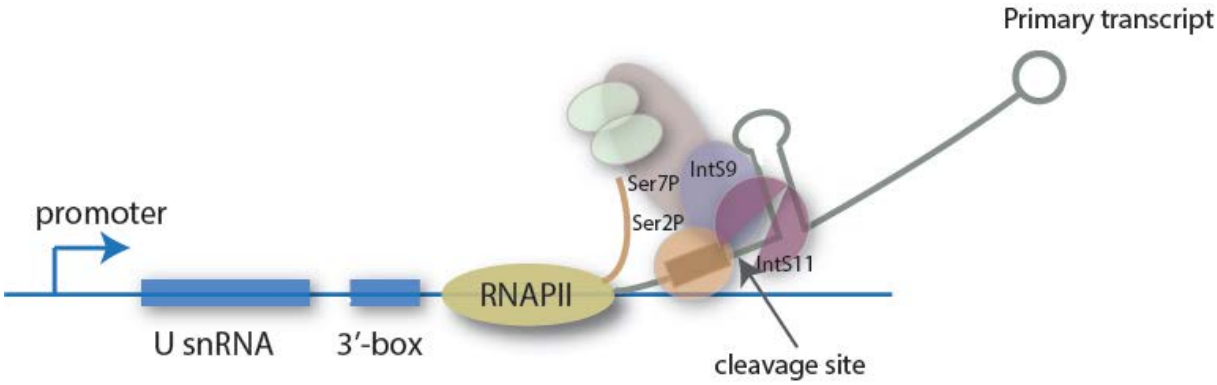


Figure 7 - Schematic view of U-type snRNA processing carried out by the Integrator complex. The cleavage of the primary transcript is carried out by IntS11, a homolog of endonuclease CPSF73, which cleaves the pre-mRNA.

The downstream steps of U-type snRNA maturation occur in the cytoplasm, where Sm-core particles are assembled (Figure 8). Following the assembly of the Sm core and association of SMN complex, the (m7G) cap of the snRNA is hypermethylated and a few nucleotides on the 3' termini are trimmed off. The snRNAs are then relocated back to the nucleus, where they further function in splicing and possibly histone processing [52].

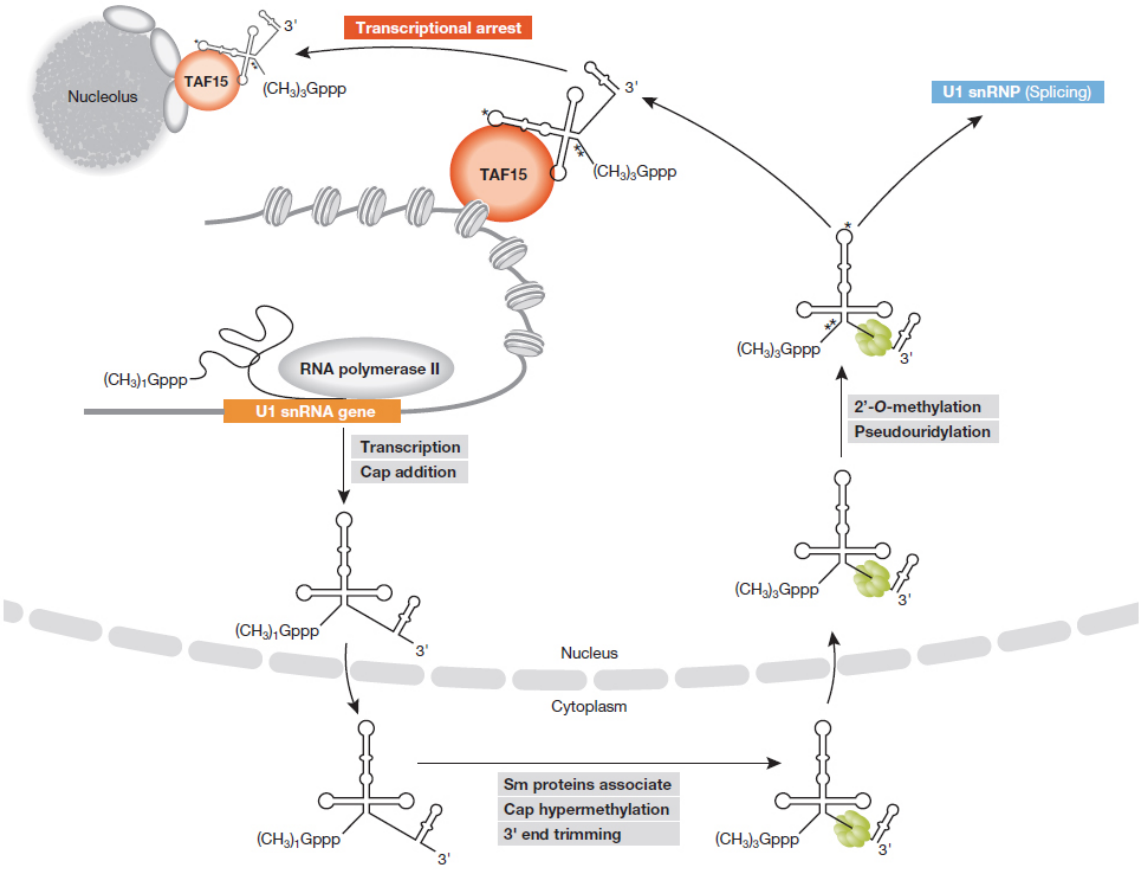


Figure 8 - Biogenesis of snRNAs exemplified for U1 snRNA. U1 is transcribed in the nucleus and transported to the cytoplasm, where it undergoes cap hypermethylation, 3' end trimming and associates with Sm proteins. The U1 snRNA is re-imported into nucleus for additional modification and assembly into the functional U1 snRNPs. PolII, RNA polymerase II; snRNA, small nuclear RNA; TAF15, TATA box binding protein-associated factor 15. Asterisks indicate snRNA modifications (Adapted from Kugel and Goodrich, 2009 [69]).

1.2.2.2. Processing and maturation of snoRNAs

snoRNAs are the components of ribonucleoprotein particles, which catalyze site specific RNA modifications of ribosomal (r)RNAs, snRNAs and transfer (t)RNAs. Recent reports suggest that levels of snoRNAs are changed in cancer cells [70]. In humans the majority of snoRNAs are intronically coded and transcribed together with the host gene and only few snoRNAs are transcribed independently (Table 1) [71]. snoRNAs are subdivided into 2 classes according to the modifications they perform. The box H/ACA snoRNAs guide pseudouridylation, whereas C/D snoRNAs guide 2'-O-methylation. C/D box snoRNAs contain conserved C (UGAUGA) and D (CUGA) boxes and form RNA-protein complex with

four proteins, NOP56, NOP58, 15.5 kDa and Fibrillarin, which catalyzes the 2'-O-methylation. H/ACA box snoRNAs carry H box (ANANNA) and ACA (ACA) box. Four proteins form ribonucleoprotein complexes together with H/ACA snoRNAs. These are NHP2, NOP10, Gar1 and Dyskerin, which catalyzes pseudouridilation.

SnoRNAs may be involved in other processes apart from RNA modification. For instance, a C/D box snoRNA, SNORD115 (HBII-52), which shows sequence complementarity to the serotonin receptor 2C pre-mRNA, influences alternative splicing of this pre-mRNA [72].

Table 1 - Organization of snoRNA genes in human genome.

Organism	snoRNAs	Genes	Independent		Intronic	
			Individual	Clustered	Individual	Clustered
H. sapiens	216	456 257 C/D; 181 H/ACA	42 15 C/D; 27 H/ACA	0	412 256 C/D; 34 H/ACA	0

Adapted from Dieci et al., 2009

The processing of snoRNAs at their 3' and 5' ends has been extensively studied in yeast. SnoRNAs are released from the precursor by a series of endonucleolytic cleavages by RNAse III homolog, Rnt1, followed by exonucleolytic processing by Xrn2/Rat1 [68, 73]. In *S. cerevisiae*, cells with mutations affecting the nuclear exosome contain polyadenylated precursors and intermediates of snoRNAs [74, 75].

In human cells, intronically encoded snoRNAs are processed co-transcriptionally. In case of C/D class snoRNAs, the processing is coupled with splicing [76]. Upon splicing and debranching, the snoRNAs are cleaved from both 5' and 3' ends by exosome compounds. The mature snoRNAs are protected by snoRNPs from further degradation by the exosome, snoRNPs binding determine the boundaries of the mature snoRNAs [55]. It has been observed

that some snoRNAs may be further processed to produce smaller RNAs, which are active in alternative splicing of a certain set of transcripts [77].

In mammalian cells snoRNAs can get polyadenylated via PAPD5, which adds oligo(A) tails to the last few nucleotides remaining after exonucleolytic degradation of the 3' flanking intron. These oligoadenylated processing intermediates are then trimmed by PARN [78].

1.3. RNA export

While being transcribed, pre-mRNAs form complexes with various RNA-binding proteins, which facilitate the 5'-end capping, splicing and 3'-end processing of a newly synthesized RNA. After transcription and processing are completed the mRNAs are exported to the cytoplasm. The nucleocytoplasmic export occurs via the nuclear pore complex (NPC) [79]. The ready-to-export mRNP diffuses through the interchromatin space direction NPC (Figure 9). The translocation of mRNP requires proteins in the mRNPs serve as adaptors for binding to export receptors. The export receptors in turn mediate the contact between mRNP and NPC. The vast majority of mRNPs use a specific heterodimer export receptor called NXF1:NXT1 [80]. In many cases the export occurs with help of the transcription-export complex (TREX), which associates with the elongating RNA polymerase II-nascent pre-mRNP. TREX consists of the multi-subunit THO complex and two export proteins, the RNA helicase UAP56 and Aly [81]. NXF1 functions in translocation of bulk mRNAs, yet the small RNAs such as for instance rRNAs and U type snRNAs are exported in a Crm1-dependent manner [82]. Crm1 is not an RNA-binding protein, therefore the contact between the RNAs and Crm1 occurs via adaptor proteins. There are several adaptor proteins found to function in the Crm1-dependent mRNA export pathway, for example RNA-binding protein human antigen (HuR), leucine-rich pentatricopeptide repeat protein (LRPPRC) and nuclear export factor 3 (Nxf3) [57]. The export of U type snRNAs is mediated by the adaptor protein PHAX, which binds to both the CBC and near the cap of U type snRNA [83]. The high affinity of PHAX for small RNAs of less than 200-300 nt distinguishes this RNA export pathway [84].

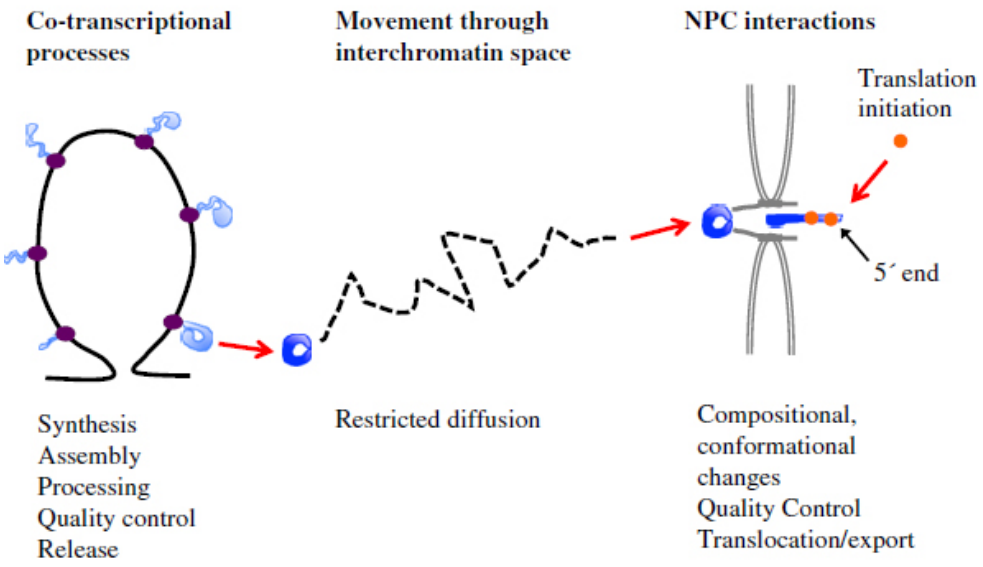


Figure 9 - Overview of the nuclear steps of gene expression and nucleocytoplasmic export.

(Adapted from Björk and Wieslander [85]).

1.4. RNA decay

The importance of mRNA decay has been assessed by large-scale studies that revealed the relative contribution of decay rate and transcription rate to differential mRNA abundance [86]. Interestingly, the decay rate predominates, as the median half-life of human transcripts is comparatively long. mRNAs differ in their stability rates, which is mainly determined via elements within non-coding 3' and 5' UTRs. The mRNA decay is conserved throughout evolution and is similar for different species. It can be subdivided into several steps, the first step is deadenylation, removal of the poly(A) tail. In mammalian deadenylation is carried out by CCR4-NOT complex (Figure 10). Upon deadenylation, two scenarios are possible. In the first scenario, the degradation continues via decapping by Dcp1-Dcp2 and is followed by the 5'-3' decay carried out by exonuclease Xrn1 (Figure 10). Alternatively, mRNAs are degraded by the exosome from the 3' end and decapping is later carried out by Dcp5 (Figure 10). The depletion of the main 5'-3' or 3'-5' degradation pathways does not result in accumulation of aberrant mRNA. This suggests that the two pathways can act redundantly [87].

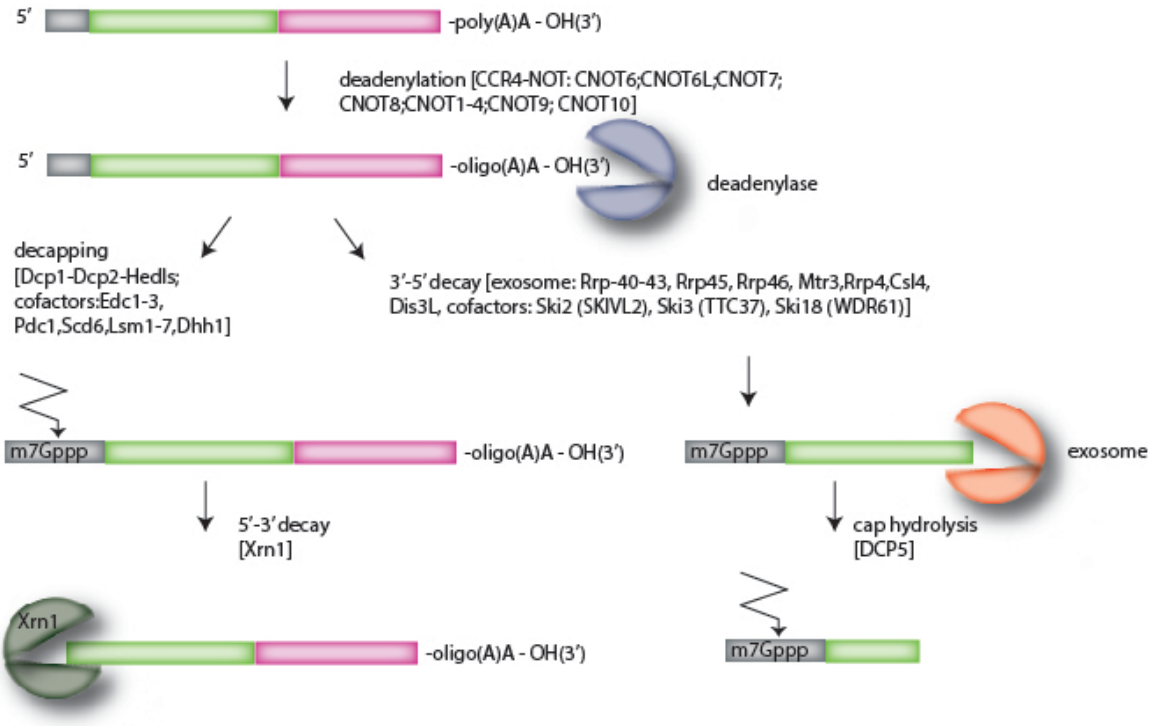


Figure 10 - Schematic overview of 3'-5' and 5'-3' mRNA decay pathways. The first step of mRNA degradation is deadenylation. In humans deadenylation is carried out by CCR4-NOT complex. Further execution of two degradation pathways is possible. 5'-3' degradation starts with decapping and continues with 5'-3' decay by Xrn1. The second pathway starts with 3'-5' degradation by exosome complex, following decapping.

Interestingly, polyadenylation of mammalian RNA can play to some extent opposite roles. On the one hand, polyadenylation of mRNA contributes to nuclear transport, translation and increased stability, on the other hand, oligoadenylation of molecules mark them as substrates for exonucleolytic degradation [88].

There are additional degradation pathways, which target specific RNAs. First pathway involves deadenylation-independent decapping, in which the decapping happens in presence of a long poly(A)-tail. For example, decay of RPS28B mRNA [89]. Another degradation pathway is specific for replication-dependent histones and happens via oligo-uridilation of the 3'ends, which serve a platform for exosome [90]. Alternatively, mRNAs harboring nonsense codons within their open reading frames (ORFs) are recognized by the nonsense-mediated RNA decay (NMD) pathway [91]. Additionally eukaryotic mRNAs can be degraded via

endonucleolytic cleavage, as exemplified by degradation of ß-globin mRNA, harboring a nonsense codon [92]. Furthermore, pathway specific for a particular set of RNAs can be mediated by miRNAs, which recognize specific sites on mRNAs and can cause their deadenylation and decay [93]. Else the decay of a particular set of transcripts, characterized by AU-rich (ARE-mediated decay) or GU-rich 3' UTRs may be regulated by RBPs [94].

1.5. RNA-binding proteins and their role in health and diseases

As highlighted above, RNA-binding proteins (RBPs) are key factors regulating the fate of different classes of RNA molecules throughout their life span. RBPs play a crucial role in normal cell growth and development by modulating the gene expression on a posttranscriptional level. Accumulating evidences suggest that many diseases are directly or indirectly caused by mutations and /or expression changes of RBPs [95]. The importance of RBPs and their role in disease-related changes becomes widely accepted. Thus understanding of the mechanisms of posttranscriptional gene regulation can have important diagnostic and even prognostic value [95]. The next sections will serve to describe the current attempts on discovering new RBPs, describe their function in health and disease and shed a light on current approaches to study RNA-protein interactions.

1.5.1. Discovering RNA-binding proteins and linking them to human diseases

Recently developed methods such as "interactome capture", which enable *in vivo* capturing and detection of RNA-binding proteins, revealed numerous novel RBPs [96, 97]. Of particular interest is the domain structure of RNA-binding proteins as well as their functionality. Many RBPs appear to be multifunctional; apart from binding to RNA several RBPs are involved in metabolic processes and possess enzymatic activity [98]. RBPs form complexes to regulate a fate of various RNAs (as detailed before). The structure and composition of these complexes can integrate external biological stimuli leading to rapid spatial remodeling of RBP-complexes.

Applying "interactome capture" Castello and coworkers identified 860 RNA-binding proteins in HeLa cells [99]. The comparison of the experimentally proven list of RNA-binding proteins with a database of diseases with Mendelian inheritance, OMIM database

(Online Mendelian Inheritance in Man; updated on 30 December 2014) revealed over 100 proteins linked to diseases. 707 proteins out of 860 were mentioned in OMIM, of them 131 protein were found to be linked with human diseases (18%). It was observed that the mutations in RNA-binding proteins are linked mostly to neurological, cancer, sensory, metabolic and muscular diseases (Figure 11) [99].

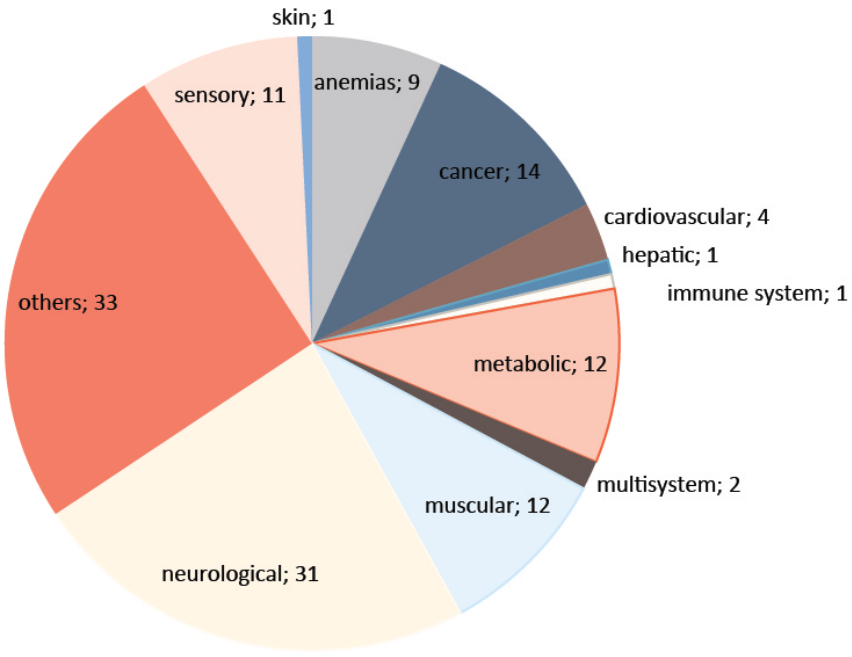


Figure 11 - Mutations in RNA-binding proteins are linked to various diseases. Numbers of proteins linked to each group of diseases are shown (own data).

1.5.2. Exemplifying diseases caused by mutations or miss-regulations of RBPs

Neurological disorders comprise the biggest group of diseases, caused by abnormal expression of RNA-binding proteins. One of the examples of disorders belonging to this group is an Amyotrophic lateral sclerosis (ALS), a fatal neurodegenerative disease, which involves progressive degeneration of motor neurons in the spinal cord, brainstem and motor cortex. ALS is a rare disease with a rapid progression, leading to death of 50 % of patients 2-3 years after diagnosis. ALS is inherited in 5-10 %; while other cases appear to occur randomly. Four RNA-binding proteins known so far were found to be linked to ALS, namely Ataxin2 (ATXN2), trans-active response (TAR) DNA-binding protein TDP43, matrin 3 (MATR3) and fused in sarcoma (FUS). All these RNA-binding proteins share a similar feature. FUS and

TDP43 are normally found in nucleus. By contrast, in pathological conditions they are present mainly in cytoplasm. Interestingly, Elden and coworkers have shown that Ataxin2 can associate with TDP43 in a complex, this interaction is RNA-dependent [100]. They also found that ATXN2 abnormally localizes in ALS patients' neurons and affects TDP43 toxicity. Further altered levels interfere with assembly of stress granules. All four proteins are involved in posttranscriptional processing of RNA molecules, transcription, alternative splicing, translation, RNA transport, stability and degradation [101-103]

Recent studies showed that TDP-43 and FUS proteins may regulate common RNA targets in neurons; 25 % of genes with altered gene expression levels and 10% of genes with alternatively spliced exons were common for FUS- and TDP-43-silenced primary cortical neurons [104, 105]. Further studies might help to illuminate the downstream mechanisms of action of these RNA-binding proteins.

Another large group of diseases caused by miss regulation of RNA-binding proteins is cancer. Surprisingly one of the RNA-binding proteins, HuR (ELAVL1), whose role in cancer as a tumor antigen was first discovered in early 1990s, is not present in OMIM database. The biological role of this protein in cancer tissues is well studied and broadly discussed in the literature. The expression level of HuR was shown to be increased in almost 300 cancerous tissues (oral, colorectal, gastric, lung, breast, ovarian, renal, skin carcinoma and mesothelioma [106-111]). Unlike previous examples, where expression changes or miss-processing happens because of mutations on DNA level, HuR protein overexpression seems to be regulated on a post-translational level and epigenetically. HuR stabilizes a number of RNA molecules harboring AU- and U-rich elements (ARE) on 5' and 3'UTRs. HuR is a shuttling protein, upon binding to its mRNA targets in the nucleus, it shuttles to the cytoplasm and protects the messenger RNA from digestion by exonucleases. The function of HuR in cancer development and progression is accomplished by stabilization and enhanced of translation of oncogenes such as *c-myc*, *c-fos*, EGF and transcripts which promotes enhanced cell division; Protα, which helps to overcome apoptosis and many other oncogenes harboring ARE-elements and recognized by HuR [110, 112].

In spite of numerous attempts to characterize the mechanisms the underlying functionality of RBPs in health and disease, only a small proportion of them has been addressed. The diversity and variety of action of RNA-binding proteins is enormous and, in many cases, remains to be studied. Nevertheless, the attempts to study mechanisms of action of RNA-binding proteins in health and diseases already provide intriguing insights.

1.5.3. CLIP as a method to study RNA-protein interactions

Defining the mRNA targets is crucial for understanding the role of RBPs and their contribution to human pathologies and possibly identifying new ways to treat and diagnose disorders in the future. A method, which can be successfully applied to study RNA-protein interactions, should meet certain requirements. Firstly, the method should enable to investigate rapid changes emerging in a living cell *in vivo*. Secondly, it should allow estimating quantitative changes of RNA molecules bound by a particular protein.

Ultraviolet (UV) crosslinking and immunoprecipitation (CLIP) and its variants have been successfully applied to identify specific RNA-protein interactions both in the cell culture as well as in a living organism or tissue [113, 114]. CLIP was first applied for the Nova 1 and 2 proteins [115]. Nova is a family of RNA-binding proteins specifically expressed in neurons; they regulate spicing of RNAs encoding synaptic proteins. Nova1 and Nova 2 are targeted in a paraneoplastic syndrome in which inhibitory control of motor systems is affected [116-118].

1.5.4. CLIP and its variants

The CLIP technique is capable of studying RNA-protein interactions *in vivo*. In this method the RNA-protein interactions are being preserved in living cells by means of ultraviolet irradiation (Figure 10, step 1). UV-C (245 nm) light triggers formation of covalent bounds between amino acid residues localized in close proximity or bound directly to the RNA molecule at the moment of irradiation. The mechanisms involved in formation of covalent bounds between amino acid residues and nucleic acids is more explicitly described by Meisenheimer, 1997 [119]. Some amino acids are most reactive toward the nucleic acids, these are Cys, Lys, Phe, Trp, and Tyr, whereas His, Glu, and Asp are moderately reactive. Following UV-irradiation cells are solubilized and the RNA is partially digested. Only those

parts of RNA molecules, which are spatially covered and thus protected by protein remain intact. The RNA-protein complexes are visualized by radioactive labeling of RNA. Next the RNA-protein complexes are purified under stringent conditions by immunoprecipitation using antibody directed against the protein of interest (Figure 12, step 2). The second dimension of purification is carried out next with help of protein electrophoresis and subsequent transfer onto a nitrocellulose membrane. This step separates the RNA-protein complexes by size and eliminates free RNA, which migrates through the nitrocellulose membrane. Thereafter a band corresponding to the RNA-protein complex is isolated and treated with proteinase K, which digests the protein and releases the intact RNA (Figure 12, step 3). Typically these are short fragments of 20 to 60 nucleotides. Ultimately, the RNA is used to generate a cDNA library (Figure 12, step 4). In the last step the cDNA library is sequenced and, upon mapping on the transcriptome, the whole repertoire of RBP RNA targets can be determined (Figure 12, step 5). Current CLIP techniques are able to detect thousands of RNA targets and target sites.

The CLIP technique had recently undergone several modifications and improvements. They were mainly targeting the crosslinking of the RNA to the protein as well as library preparation and computational analysis. These improvements enhanced the crosslinking efficiency and increased the cDNA library complexity. With the time, the method gained better resolution and efficiency of detection of RNA-binding sites. For example, so called “PAR-CLIP” (Photoactivatable Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation) (Figure 12 compare PAR-CLIP vs. HITS-CLIP and iCLIP) [120] introduces photoactive ribonucleoside analogs (4-thiouridine or 6-thioguanosine) in the RNA of cultured cells. A photoreactive group incorporated into the nucleic acid facilitates the crosslinking and separation of a signal from noise. When 4-thiourudune PAR-CLIP is being used, a T to C change is typically observed at the binding site in 50-70 % of cases [121]. However, PAR-CLIP is only applicable for cultured and highly proliferating cells. iCLIP is another approach claiming to be capable of distinguishing the protein binding site with an individual nucleotide resolution (Figure 12). The method is based on the observation that the reverse transcriptase frequently stops at the site of crosslinking. A modified protocol for cDNA synthesis enables to capture such pre-terminated events. Analysis of reverse transcription stop sites thus gives precise information on exact position of crosslinking. Yet,

the efficiency of crosslinking varies from protein to protein and depends on the amino acid content of the RNA-recognition motif of the protein as well as on spatial organization of protein-RNA complexes [119]. Some proteins were shown to be less efficiently crosslinked [96, 122].

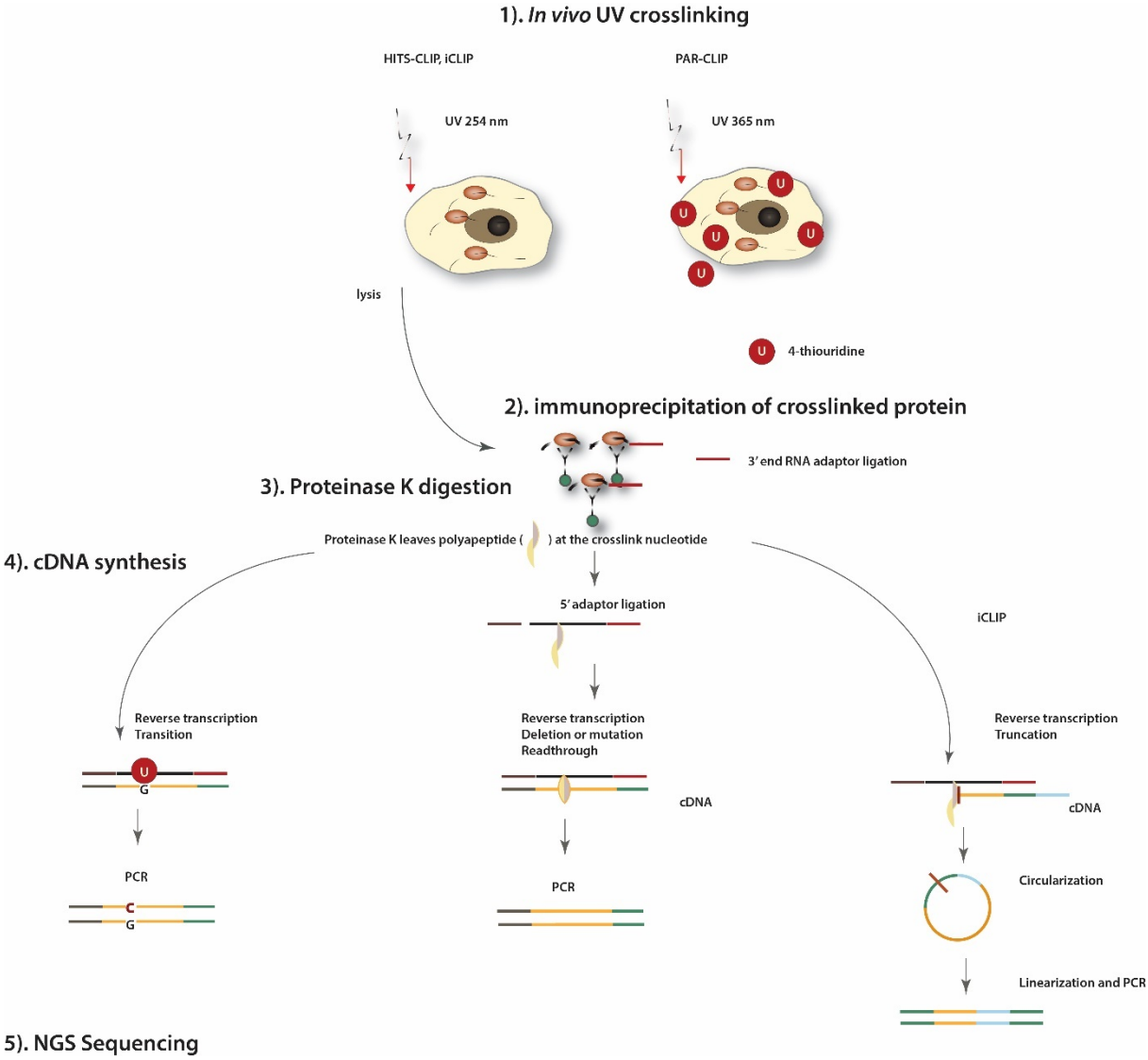


Figure 12 - Schematic representation of three variants of CLIP protocol. HITS-CLIP and iCLIP rely on UV 254 nm crosslinking (1). In case of PAR-CLIP cells are feed with 4-thiouridine and are later crosslinked with UV 365 nm (1). Next in all three protocols the desired protein is immunoprecipitated (2). The 3' end adaptor is then ligated to the co-immunoprecipitated RNA. Next the protein bound to the RNA is digested with help of Proteinase K (3). Depending on the CLIP variant different steps follow up. In case of iCLIP the cDNA synthesis

and circularization takes place (4). In case of HITS-CLIP and PAR-CLIP the 5' adaptors are ligated first, after that the cDNA is synthesized. The last steps of all variants are PCR and next generation sequencing (5) [123].

Recently it has been shown that the reverse transcriptase does not always stall at the crosslinked sites [124]. Therefore considering the overlapping start sites as the crosslinking position with a nucleotide resolution is not always optimal [124].

CLIP techniques include several steps with complicated biochemistry. One of the steps is immunoprecipitation of the protein of interest. The efficiency of immunoprecipitation critically depends on the quality of antibodies used. In many cases, CLIP was thus applied not on endogenously expressed proteins but on exogenously introduced peptides harboring affine tags useful for standardized immunoprecipitation, for example FLAG-tag (CLIP, HITS-CLIP, iCLIP), streptavidine and polyhistidine epitopes (iCLAP [125]), 6x-histidine and Protein A tags (CRAC [126]). Those modifications, on one hand overcome an issue of inefficient antibodies, but on the other hand deprive the method of its benefits to study endogenously expressed proteins in native conditions.

In spite of numerous attempts of last years to improve the CLIP variants, there are still common limitations, which characterize current methods. For example, all techniques require radioactive labeling of the RNA, high levels of input materials and rely on complex protocol for library synthesis. These limitations restrict the broad usage of the technique and make it inapplicable to non-expert laboratories.

1.5.5. Challenges with interpretation of CLIP results

CLIP and its variants are very powerful approaches to study RNA-protein interactions. Yet along with the challenging protocol the interpretation of data is demanding as well. The method shows the whole picture as a snapshot, and therefore it is important to make attempts to separate the seen in space and time. Thus, CLIP is usually complemented by other transcriptome-wide methods, such as RNA-sequencing upon depletion of the studied factor, poly(A)-sequencing, chromatin immunoprecipitation sequencing (ChIP-Seq), overexpression and mutation of the protein of interest.

As the crosslinking captures both transient as well as stable RNA-RBP interactions, it is necessary to rank the target sites in order to reveal the most physiologically important targets. The power of the method is at the same time its limit, as the CLIP techniques are characterized by a low reproducibility, which possibly can be explained by the capacity of tracing transient/changing interactions. It has not been intensively studied if the method is quantitative. The future perspectives of CLIP most probably lies in a field of studying dynamic RNA-protein interaction [127].

1.5.6. CLIP a method to study dynamic RNA-protein interactions

One of the biggest potential of the CLIP method not broadly and efficiently implemented so far is its applicability to study dynamic rather than steady state composition of RNA-protein complexes. Tollervey and coauthors applied iCLIP to reveal differential binding of TDP-43 protein in cortical tissues from a post-mortem brain from normal (healthy) subjects and subjects with sporadic frontotemporal lobar degeneration (FTLD-TDP), where TDP-43 was mostly primarily localized in cytoplasm [128]. They observed that the TDP-43 protein interacts with a number of RNAs, which play important roles in the brain, such as transcription factor AP-2 alpha (TFAP2A), ciliary neurotrophic factor receptor (CNTFR), transducin-like enhancer of split 1 (TLE1) and many others [128]. Moreover the mislocalization of TDP-43 protein in the cytoplasm of FTLD patients leads to a tenfold increment of affinity of TDP-43 to 3'UTR of target RNAs. At the same time, decrease of TDP-43 protein in the nucleus affected alternative splicing of transcripts important for normal neuronal development and function. Another example of dynamic RNA-protein interactions study is a paper by Beckmann and coauthors [129]. The authors explore the binding specificities of a metabolic enzyme (the dehydrogenase HSD17B10), which binds to mitochondrial RNAs with high preference to 5' ends of tRNAs. In contrast, disease associated variants of HSD17B10 (R130C), which causes HSD10 disease [130] exhibits decreased binding signals to several pre-tRNAs [129].

1.5.7. Aims of the thesis

This work aims to further improve the existing CLIP techniques for studying RNA-protein interactions. In particular it aims to improve the cDNA library synthesis protocol in

order to achieve robustly reproducible results, to apply a radioactive-free labeling of the RNA and to decrease the amount of input material needed for library synthesis. These improvements may improve the applicability and accuracy of the technique. Additionally, this work aims to apply the established technique for transcriptome-wide study of RNA-RBP interactions and to reveal novel binding capacities of studied RBPs.

Materials and Methods

2.1. Molecular biology techniques

2.1.1. Western Blot

Total protein lysate was separated in 7% polyacrylamide Bis-Tris gel system, as described by Updyke and Engelhorn (Life Technologies) using MOPS-SDS running buffer. Proteins were transferred onto a nitrocellulose membrane (Hybond ECL, GE Healthcare). Upon transfer the membrane was stained with Ponceau red for 5 minutes, fixed with 0,5% acetic acid and rinsed with water. Ponceau staining was used as a measurement for equal loading. The nitrocellulose membrane was blocked with 5% nonfat-dried milk in 1x TBS-T solution. Upon blocking the membrane was incubated with primary antibodies in 5% milk-TBS-T solution for 1 hour, rinsed with TBS-T and incubated with HRP-coupled secondary antibodies for another 1 hour. Signal was detected by applying ECL Select reagent (GE Healthcare).

Buffer	Composition	Purpose
TBS-T (x10)	Tris-Cl (100mM, pH 7,4); NaCl (9%); Tween (1%)	Western Blot blocking, washing, antibody incubation
Ponceau	Ponceau S (0,1%); TCA (Trichloroacetic Acid) (7%) in water	Nitrocellulose membrane staining
Blotting Buffer	Tris-Cl (50 mM, pH 7,4); Glycin (40 mM); MeOH (10%); SDS (0,04%)	Protein transfer from the gel onto the membrane

2.1.2. RNA isolation

RNA from human cell culture was isolated using peqGold TriFast solution (PEQLAB) according to manufacturer instructions. For total RNA and small RNA isolation a miRNeasy kit (Qiagen) was used. RNA quantity was measured with a NanoDrop Spectrophotometer (Thermo Fisher Scientific). RNA quality and integrity was assessed by agarose gel electrophoresis or using the Agilent RNA Nano kit and Bioanalyzer (Agilent). The size of small RNA molecules was determined by Urea-PAGE gel separation.

Gel/Buffer	Composition	Purpose
Urea-PAGE gel (8%)	Gel 40 (19:1) 10ml; Urea 21g; 10x TBE 5ml; APS(10%) 210 µl; TEMED 50 µl; water up to 50ml	Low range RNA separation; determining size and integrity
TBE x10	Tris Base (890mM); Boric Acid (890mM); EDTA (pH 8,0) (20mM)	Running buffer

2.1.3. Protein isolation

Total proteins from mammalian cells were isolated using lysis buffer containing Empigen BB detergent. The amount of the lysis buffer varied depending on the cell number. To disrupt DNA an ultrasonic homogenizer (Bandelin Sonopuls) was used. Lysates were sonicated in average 5 times using 20 pulses at 50-60% power with 15 sec pauses. The extraction was performed on ice.

Buffer	Composition	Purpose
Empigen BB lysis buffer	Tris-HCl (pH 7,6) (50mM); NaCl (500mM); EDTA (1 mM); Empigen BB (0,5%)	Protein lysis buffer

2.1.4. Reverse transcription and quantitative PCR

Reverse transcription was performed with oligo d(T) (16mer) primers, according to a standard protocol (see below). For small RNAs and other non-polyadenylated RNA classes miScript II RT kit (Qiagen) was used. The reaction was carried over according to manufacturer instructions.

Reagent	Volume/Amount
Total RNA	0,5 – 2 µg
Oligod(T) primer (0,5 µg/ µl)	1 µl
Mix primer and total RNA; incubate 5 min at 70 ^{o}C; cool on ice or at 4 ^{o}C	
Add the following reaction mixture and incubate in PCR cycle at 42 ^{o}C 60 min; at 80 ^{o}C 5min	
5x RT-buffer	4 µl
dNTPs (5mM)	2 µl
DTT (0,1 mM)	2 µl
RNAsin	0,5 µl
Reverse transcriptase	0,5 µl

For quantitative PCR OneTaq x2 master mixture with standard buffer was used. For DNA detection a fluorescent DNA binding dye SYBR green was applied. The measurements were carried out using BioRad Real-Time detection system.

Reagent	Volume/Amount
cDNA (diluted 1:10)	5 µl
One-Taq 2x mix	7,5 µl
SYBR green	0,3 µl
10 µM prim mix	0,3 µl
1,9 ul H_2O	1,9 µl

2.1.5. 3' end RACE

To sequence 3'ends of polyadenylated transcripts the anchored oligo dT primers carrying adapter sequence on 5' end were used for reverse transcription. The reverse transcription was followed by a PCR reaction where a specific forward primer and universal reverse primer complementary to adaptor sequence were used. Amplified fragments were cloned into CloneJET PCR cloning vector (Thermo Fisher Scientific) as recommended by the manufacturer. An aliquot of ligation mixture was used to transform chemically competent *E.coli* cells. Analysis of recombinant clones was performed by a colony PCR. Recombinant plasmids were purified using NucleoSpin Plasmid kit (Macherey-Nagel). Sequencing of plasmids was performed at Eurofins Genomics.

Primer	Sequence (5'-3')
OligodT anchored primer	TTTCCCTACACGACGCTCTTCCGATCTTTTTTTTTTTTTTTV
Universal primer	TTTCCCTACACGACGCTCTT
GPI_d_1	ATGAGGCTCAAGCAAGTGCCCTG
GPI_d_2	GGTCTGGAGACAAAGGAGCATTCAC
GPI_p_1	CTACATGGGATGTGAACAACGTGAACATG
GPI_p_2	CACTGCATGTTCCTGGACACCAC
OAZ1_p_1	AGGGTCTCCCTCCACTGCTGTAG
OAZ1_p_2	ATCCTCGTCTTGTCGTTGGACGTTAG
OAZ1_d_1	TGCATCGGGTGTAATCACTTTATTGGC
OAZ1_d_2	ATTGCTGGTTTAAGATTCAGATTATCCTTGTAC
CANX_p_1	GTAATCAATACATTTGGGAAAGTCTGCTATGTAG
CANX_p_2	CACTCCACAGTGTTTATATTGGGAAGATATTG
CANX_d_1b	TCCAGCATCCTGATTAAATGTCTGACAC
CANX_d_2b	CATTTGACAACAATGGAACAGGTAACCAGC
PCMT1_d_1	AGAACTAGGACACAGCTTACACAGCATG
PCMT1_d_2	GTAAACACTCAGCTGTTCAGATTGGACATAAC
PCMT1_p_1	TAAAGTAATATTACACAAATTTATACTTTGCATCCTTCCTCTA
PCMT1_p_2	ACAGTGGATTGCTCATCTCAGTCCTCAAAG

DDX5_p_1	GAGAGAATGGGGAGAAAAATCACATTTATT
DDX5_p_2	GTGGGCCTTTTAATTTGTAAACACTGAAATG
DDX5_d_1	CAGTCTTCAAATATTTTTTATTGGAAGGCCATGC
DDX5_d_2	CATGGGAATTGCAGAAATGACTGCAGTG
HSBP1_p_1	GATTTCAGGCACCTTTATTCATGGCAG
HSBP1_p_2	CTCGGAAGTGGCAAATGGAAATGATATG
HSBP1_d_1	GCAAGAACCGCCAAAGTTTTAGTGTTTAATAATG
HSBP1_d_2	CTGGCATTTTTCCAAGCCAAGAGAAGATC

2.1.6. Poly(A)-tail length assay (e-PAT)

To tag the polyadenylated RNAs and measure the length of poly(A) tail an ePAT method was applied as described by Janicke and coauthors [131]. e-PAT products were sequenced by the procedure as described for 3'-end RACE (see section 2.1.5).

Primer	Sequence
PAT-anchor-1	GCGAGCTCCGCGGCCGCGTTTTTTTTTTTT
PAT-Uni-R1	GCGAGCTCCGCGGCCGCG
U1 snRNA_1	GATACCATGATCACGAAGGTGGTT
U1 snRNA_2	AAATTATGCAGTCGAGTTTCCCAC
U2 snRNA_1	AAATTATGCAGTCGAGTTTCCCAC
U2 snRNA_2	AATCCATTTAATATATTGTCCTCGGATAGA
U4 snRNA_1	GCGCGATTATTGCTAATTGAAA
U4 snRNA_2	AATTGCCAATGCCGACTATAT
U5 snRNA_1	GGTTTCTCTTCAGATCGCATAAATC
U5 snRNA_2	AAAAAATTGGGTTAAGACTCAGAGTT
U6 snRNA_1	GCTTCGGCAGCACATATACTAAAAT
U6 snRNA_2	AATTTGCGTGTCATCCTTGCG
U11 snRNA_1	GTGCGGAATCGACATCAAGAG
U11 snRNA_2	CGGGACCAACGATCAC
U12 snRNA_1	AACTTATGAGTAAGGAAAATAACGATTCG
U12 snRNA_2	CCTTTACCCGCTCAAAAATT
U4atac snRNA_1	GCGCATAGTGAGGGCAGTACT
U4atac snRNA_2	CCAAAATAAAGCAAAAGCTCTAGTT
U6atac snRNA_1	CCAAAATAAAGCAAAAGCTCTAGTT
U6atac snRNA_2	CAATGCCTTAACCGTATGACG

Reagent	Volume/Amount
Total RNA (miRNeasy kit)	0,5 – 2 µg
100mM PAT-anchor-1	1 µl
Mix primer and total RNA; incubate 5 min at 80 ^{o}C; 2 min 25 ^{o}C	
Spin down and add the following reaction mixture and incubate in PCR cycle at 25 ^{o}C 60 min; at 80 ^{o}C 10 min, 55 ^{o}C 2 min	
H_2O	4 µl
5x SSIII buffer	4 µl
DTT (0,1 mM)	1 µl
10 mM dNTPs	1 µl

RNAsin	1 µl
Klenow (5 U)	1 µl
While maintaining the tubes at that temperature in the block, add: 1 µL (200 U) of Superscript III and mix; incubate in a PCR cycler 55 ^{o}C 60 min; 80 ^{o}C 10 min; 4 ^{o}C hold; add 180 µl of H_2O; continue with e-PAT PCR	
H_2O	8,2 µl
5x Phire Buffer	4 µl
10 mM dNTPs	0,4 µl
10 µM Primer A (PAT-Uni-R1)	1 µl
10 µM Primer B (Specific primer (forward))	1 µl
cDNA	5 µl
Phire Hot Start II	0,4 µl
Incubate in a PCR cycler with the following settings: 98 ^{o}C 30s; 35 cycles of 98 ^{o}C 5s; 60 ^{o}C 5s; 72 30s; and 1 min incubation at 72 ^{o}C; visualize products on PAGE gel	

2.1.7. DNA analysis and visualization

To analyze the length of PCR products agarose or polyacrylamide gels were used. Polyacrylamide gels were applied to resolve fragments shorter than 400 base pairs.

Gel/Buffer	Composition	Purpose
DNA PAGE gel (8%)	Gel 40 (19:1) 10ml; 10x TBE 5ml; APS(10%) 210 µl; TEMED 50 µl; water up to 50ml	Low range DNA separation
TBE x10	Tris Base (890mM); Boric Acid (890mM); EDTA (pH 8,0) (20mM)	Running buffer

2.2. Mammalian Cell Culture techniques

2.2.1. General cell culturing procedure

Cell lines used in this work were HEK-293 (embryonic kidney) and BE(2)-C (neuroblastoma). The medium used was Dulbecco's Modified Eagle's Medium (high glucose) with the following components added: heat inactivated fetal bovine serum at a final concentration 10% and Penicillin-Streptomycin at a final concentration 1%. Both cell lines are adhesive; for splitting the medium was discarded and the cells were rinsed with sterile Phosphate Buffered Saline (PBS). Next 4 ml of Trypsin-EDTA were added and left on cells for 5 min. After that 6-10 ml of complete growth medium was added; cells were aspirated by pipetting. Cells were subcultured 1:5 every 2-3 days.

2.2.2. Plasmid and siRNA transfections

For siRNA transfection 100,000 cells were seeded in 12 well plates 12-16 hours before transfection. Roti-Fect siRNA transfection kit (Carl Roth) was used.

Reagent	Volume/Amount
OPTI-MEM	98 µl
Roti-Fect (Roti-Fect siRNA kit)	2 µl
Mix components	
OPTI-MEM	97,5 µl
siRNA(20µM)	2,5 µl
Mix components; prepare transfection mixture by mixing 100 µl of Roti-Fect mix and 100 µl of siRNA mix together; incubate 20 min at RT	
Wash cells with PBS; add 800 µl of antibiotic-free DMEM-12,5% serum media to the cells; add 200 µl of the transfection mixture to the cells; change medium after 48 hours; use complete medium as replacement	

Transfection of cells with plasmid DNA was carried out by using Roti-Fect Plus (Carl Roth) plasmid transfection reagent. For plasmid DNA transfection 1 million cells were seeded in 10 cm dishes 12-16 hours before transfection. For transfection 10 µg of plasmid DNA in 5 ml of medium (without serum and antibiotic) were mixed with 8 µl (in 5 ml medium); upon 20 minutes incubation the nucleic acid/lipid complex was added to the cells. After 2-6 hours at 37 °C in a CO_2 incubator transfection medium was exchanged by “complete” medium.

2.3. conCLIP method

2.3.1. Crosslinking and immunoprecipitation

Crosslinking of cultured cells was performed on ice with two consequently applied pulses of energy of 1500 mJ/cm^2 using CL-1000 ultraviolet crosslinker .

Immunoprecipitation was carried out with help of Protein G Dynabeads (Life Technology) according to manufacturer instructions. 30 µl of beads were used per experiment; beads were coupled with 2-5 µl of primary antibody or with equivalent amount of IgG (control). For one experiment a minimum of 700 µg of total cell protein in 500 µl cell lysis buffer was used. Protein lysate was treated with 5 µl of Turbo DNase (Life Technologies) and RNaseT1 (1:1000 diluted) for 10 minutes at 37 ^oC . The reaction was then stopped by adding 5 µl of 10% SDS. RNase and DNase treated lysate was added to the

antibody-coupled beads and incubated for 1 hour with gentle rotation. Next, the beads were washed 3 times with pre-cooled high-salt buffer and 1 time with pre-cooled PNK-buffer.

Gel/Buffer	Composition	Purpose
High-salt Buffer	Tris-HCl (pH 7,6) (20mM); NaCl (1M); EDTA 1mM; Empigen BB (0,5%)	High salt beads wash buffer
PNK-buffer	Tris-HCl (pH 7,6) (20 mM); $MgCl_2$ (10mM); Igepal (0,1%)	Beads wash buffer

Next, a second RNase digestion was performed on the beads in 250 µl mild lysis buffer and 5 µl RNaseI (Ambion). The reaction was incubated exactly 5 min and put immediately on ice. Upon second RNase digestion 2 sequential washings were carried out using high-salt buffer followed by washing with PNK-buffer. RNA bound to proteins was treated with PNK 30 min at 37 ^{o}C according to the following reaction:

Reagent	Volume/Amount
PNK-buffer	2 µl
H_2O	14,2 µl
PNK	0,8 µl
ATP 10 mM	2 µl
RNaseIn	1 µl

2.3.2. RNA labeling

Beads were washed once more with PNK-buffer and split into 2 tubes; 1/3 of reaction was labeled and 2/3 used for library synthesis. To label 1/3 of the reaction we used RNA-ligase (NEB) and biotinylated ADPs (Jena Bioscience) according to the following reaction:

Reagent	Volume/Amount
T4 RNA-ligase	4 µl
ATP 10 mM	3 µl
PEG 8000	6 µl
Biotinylated-ADP	0,2 µl
RNaseIn	1 µl
H_2O	16 µl

Labeling was performed overnight. Next the beads were washed with PNK-buffer once. The RNA-protein complex was eluted at 65 ^{o}C using 18 µl of elution buffer and 18 µl of 2x SDS sample buffer; incubate 5 min at 95 ^{o}C before loading onto the gel.

Gel/Buffer	Composition	Purpose
Elution buffer	Tris-HCl (pH 7,6) (20mM); NaCl (1M); EDTA 1mM; Empigen BB (0,5%); SDS (1%)	Elution of RNA-protein complexes from the beads
2x SDS sample buffer	Tris-HCl (pH 6,8) (100 mM); SDS (4%); Bromphenol blue (0,2%); glycerol (20% v/v); DTT (200 mM)	Loading dye for protein electrophoresis

The protein gel of suitable percentage was prepared and run as described in the Western blot procedure (2.1.1). Proteins were transferred onto nitrocellulose membrane. Next the RNA-protein complex was visualized.

2.3.3. Visualization of RNA-protein complexes

The membrane was incubated with membrane wash buffer for 10 min; followed by membrane blocking buffer 15 min; Streptavidin-HRP Conjugation buffer 30 min; again with membrane blocking buffer 10 min; 3 times with membrane wash buffer 10 min each time and finally rinsed with 1x PBS. After that the ECL Select reagent was applied; membrane was exposed to CCD camera (Bio-Rad) for 15 sec -10 min depending on the signal intensity.

Buffer	Composition	Purpose
Membrane Wash Buffer	1x PBS; SDS (0,5%)	Washing membrane before visualization
Membrane Blocking Buffer	1x PBS; SDS (0,5%); Aurora (0,1%)	Blocking membrane before visualization
Streptavidin-HRP Conjugation Buffer	1x PBS; SDS (0,5%);Aurora (0,1%); Straptavidin-HRP conjugate (Sigma-Aldrich) (1:10000)	Conjugation buffer for visualization
PBS x10	Na_2HPO_4 (400 mM); NaH_2PO_4 (100 mM); NaCl (1M); adjust pH to 7,4	Main component of other buffers of visualization protocol

2.3.4. Elution of RNA from RNP-complexes

The RNA was next eluted from the RNA-protein complexes transferred to the nitrocellulose membrane with help of proteinase K as described before by [132]. Eluted RNA was precipitated by Ethanol:Isopropanol (1:1) overnight at -80 ^{o}C.

2.3.5. Library preparation and deep sequencing

Eluted RNA was first polyadenylated for 30 min at 37 ^{o}C according to the following protocol:

Reagent	Volume/Amount
ATP 10mM	0,5 µl
Poly(A) polymerase	0,12 µl
Poly(A) polymerase buffer	0,25 µl
RNaseIn	1 µl

Following polyadenylation the RNA was again precipitated overnight.

Next the cDNA was synthesized following the procedure and using primers as exemplified:

Primer for cDNA synthesis (example)	CGATTGAGGCCGGTAATACGACTCACTATAGGGGTTCAGAGTTCTACAGTCCGACGATCNNNNN ACGTTGTTTTTTTTTTTTTTTTTTTTTTV

Reagent	Volume/Amount
Primer (25 ng/ul)	0,5 µl
RNase free H_2O	0,6 µl
Polyadenylated RNA (pellet)	Pellet
Primer annealing 70 ^{o}C 10 min; move on ice; add the following reaction mixture	
First strand buffer (Ambion kit)	0,2 µl
dNTPs	0,4 µl
RNase Inhibitor	0,1 µl
Incubate 2 hours at 42 ^{o}C; move to ice	

Second strand synthesis was accomplished according to the following protocol:

Reagent	Volume/Amount
RNase free H_2O	6,3 µl
Second strand buffer (Ambion kit)	1,0 µl
dNTPs	0,4 µl
DNA Pol	0,2 µl
RNase H	0,1 µl
Incubate 2 hours at 16 ^{o}C; move to ice	

After the second strand synthesis the cDNA was purified using Ambion kit purification columns following the manufacturer instructions. cDNA was eluted with 16 µl of RNase free water and next in vitro transcription was carried out overnight.

Reagent	Volume/Amount
cDNA	16 µl
ATP	4 µl
GTP	4 µl
CTP	4 µl
UTP	0,1 µl
10x T7 buffer (Ambion)	4 µl
T7 enzyme (Ambion)	4 µl
Incubate overnight at 37 ^{o}C; move to ice	

RNA was purified using Ambion kit columns and consequently dephosphorylated by phosphatase and phosphorylated by polynucleotide kinase according to the following protocol:

Reagent	Volume/Amount
Purified RNA	16 µl
10x Phosphatase buffer	2 µl
Antarctic phosphatase	1 µl
RNase Inhibitor	1 µl
Incubate 30 min at 37 ^{o}C and 5 min at 65 ^{o}C move to ice	

Reagent	Volume/Amount
Water	17 µl
Phosphatase treated RNA	20 µl
Phosphatase buffer	5 µl
ATP (10 mM)	5 µl
RNase Inhibitor	1 µl
PNK	2 µl
Incubate 60 min at 37 ^{o}C; move to ice	

Phosphorylated RNA was purified using miRNeasy kit (Qiagen) according to the manufacturer instructions. The sample was concentrated to 10 µl; half of the sample was used for 3' adapter ligation.

Reagent	Volume/Amount
RNA 3'-end adapter (RA3; Illumina)	1 µl
Phosphatase and PNK treated RNA	5 µl
Mix RNA with adapter; incubate in thermal cycler 2 min at 70 ^{o}C; immediately place on ice	
5x HM Ligation Buffer (HML; Illumina)	2 µl
RNase Inhibitor	1 µl
T4 RNA-ligase 2, truncated (NEB)	1 µl
Incubate 1 hour at 28 ^{o}C; with the reaction tube remaining on the thermal cycler add 1 µl of Stop Solution (STP, Illumina), mix thoroughly by pipetting, continue to incubate at 28 ^{o}C another 15 min, and then place the tube on ice; add 3 µl of water	

For the Reverse transcription reaction half of Adapter-ligated RNA reaction was used, another half stored at -80 ^{o}C.

Reagent	Volume/Amount
Adapter-ligated RNA	6 µl
RNA RT Primer (RTP; Illumina)	1 µl
Mix RNA with adapter; incubate in thermal cycler 2 min at 70 ^{o}C; immediately place on ice; add 5,5 µl of the following mixture	
5x First Strand Buffer	2 µl
12,5 mM dNTPs	0,5 µl
100 mM DTT	1 µl
RNase Inhibitor	1 µl
SuperScript II RT	1 µl
Incubate 1 hour at 50 ^{o}C; place the tube on ice	

Next the total reaction was used for final PCR reaction according to the following protocol:

Reagent	Volume/Amount
RPI primer (Illumina)	2 µl
RPIX primer (Illumina)	2 µl
cDNA	12,5 µl
PML master mix	25 µl
Ultra-pure water	8,5 µl
Incubate in thermal cycler with the following settings: 30s 98 ^{o}C; 10s 98 ^{o}C, 30 s 60 ^{o}C, 30 s 72 ^{o}C (9-11 cycles); 10 min 72 ^{o}C; hold on 4 ^{o}C	

The PCR product was purified with AMPure XP beads according to manufacturer instructions. The product was eluted in a final volume of 10 µl. Subsequently purified product was analyzed by Qubit and Bioanalyzer; concentration of above 1 ng/ µl and average size of 230 bp was usually observed. The libraries were sequenced on MiSeq Platform (Illumina) if less than 5 samples were multiplexed and with a NextSeq 500 (Illumina) if 5 or more samples were multiplexed. In this case the sequencing was carried out by GeneCore EMBL core facility.

2.4. Bioinformatics

2.4.1. Bioinformatical pipeline for conCLIP analysis

Next Generation sequencing data were analyzed with help of High Performance Computer (Mogon) at Mainz University. The analysis has been carried out on the basis of currently available NGS sequencing tools (see table) as well as with in house pipelines and algorithms.

Tool	Year	Application	Reference/source	Programming language
Trimmomatic	2014	Trimming of adapter sequences	[133]	Java
FastQC		Read quality assessment	[134]	Java

Samtools	2009	Reading/writing/editing/indexing/viewing SAM/BAM/CRAM format	[135]	C
BEDOPS	2012	A toolkit which performs statistical calculations, archiving, conversion and other management of genomic data of arbitrary scale	[136]	NA
HTSeq-count	2015	a tool that preprocesses RNA-Seq data for differential expression analysis by counting the overlap of reads with genes	[137]	Python
BEDtools	2010	tools to compare large datasets (e.g. next-generation sequencing data) with both public and custom genome annotation tracks	[138]	C++
CLIPper	2013	Peak finder	[139]	Python
HOMER	2010	A software for NGS sequencing data analysis	[140]	Perl, C++
IGVTools	2011	A tool for interactive exploration of large, integrated genomic datasets	[141, 142]	Java
edgeR	2010	Differential expression analysis of RNA-seq expression profiles with biological replication	[143, 144]	R
DEXSeq	2012	The package is focused on finding differential exon usage using RNA-seq exon counts between samples with different experimental designs	[145]	R

Results

3.1 Chapter 1

3.1.1. Establishment of conCLIP method

Crosslinking and Immunoprecipitation is a powerful technique to decipher RNA sequence motifs to which RBPs bind to (discussed in introduction, section 1.5.5.). Figure 12 illustrates the principle and key steps of CLIP variants. As an initial step of my research I aimed to optimize the CLIP technology and make it suitable for the main goal of our project – studying of dynamic RNA-protein interactions and gaining new insights into protein binding capacities with high resolution. Specifically it was intended to omit radioactivity, decrease the input material and improve the cDNA library synthesis protocol.

3.1.2 TIA-1 immunoprecipitation as a starting point of conCLIP establishment

In the first place TIA-1, a protein whose binding profile was studied using CLIP method before, was picked in order to have a reference point for setting up the CLIP protocol in the context of the present work [125]. This protein has three RNA-recognition motifs. In the nucleus it regulates alternative splicing by binding to U-rich sequences adjacent to the 5′ splice site and recruiting U1-C to promote exon inclusion [146]. It may also regulate the splicing of its own mRNA [147]. In the cytoplasm TIA-1 functions as translational silencer by binding to the 3′ untranslated region (3′ UTR) of mRNAs [148].

As described in section 1.5.5, the first step of the CLIP protocol is UV crosslinking. The protocol was applied to the HEK293 cell line transiently transfected with a plasmid carrying the TIA-1 open reading frame. The cells were washed twice with cold PBS prior to crosslinking. The crosslinking conditions were optimized such that a maximum amount of RNA signal could be detected, while irradiating the minimum time.

3.1.3 Establishing a non-radioactive labeling of RNA to visualize RNA-protein complexes

The conventional CLIP method as well as all variants of it are based on radioactive visualization of the co-immunoprecipitated RNA (see section 1.5.4). To avoid radioactivity

while keeping the same sensitivity, labeling of RNA was conducted using a biotin-coupled nucleotide as a substrate and a T4 RNA-ligase as an enzyme. To visualize the RNA molecules caring a biotin label, a biotin detection procedure described in Materials and Methods (section 2.3.3.) was performed.

The purity of the IP was confirmed by silver staining (data not shown). The specificity of the IP was confirmed by spike-in experiments (Figure 13A). Before starting the work, it was known that TIA-1 binds to a very specific sequence motif, which can be found in a group of transcripts with low efficiency 3`end processing signals. This sequence motif is situated in the 3'-UTR upstream of polyadenylation signals (Upstream Sequence Element, USE) and has been shown to promote 3'end processing. Amongst many proteins binding to the USE, TIA-1 has been demonstrated to recognize this sequence element [149]. The synthetic RNA oligonucleotides harboring the USE motif or containing another unrelated motif as shown in the table below were applied to control the specificity of immunoprecipitation. Both oligonucleotides were labeled with a biotin at the 5'end. These oligonucleotides were added to the cell lysate and crosslinked by UV irradiation. Subsequently TIA-1 was immunoprecipitated from this lysate and the biotin label was then detected.

Table 2 - RNA oligonucleotides used for spike-in experiments.

Molecule	Sequence 5'-3'	Modification 5' end
USE	AGAUUAUUUUUGUGUUUCUA	BIOTIN
USE unrelated	AGAACGAGACGAGCGCGCCUA	BIOTIN

As shown in Figure 13A it was possible to confirm the specificity of the immunoprecipitation protocol. Binding of TIA-1 to the USE is conserved and stabilized upon UV-irradiation (compare lanes 3 and 4). At the same time this does not result in a non-specific binding of an unrelated sequence motif (lanes 1 and 2). Thus I confirmed that the setup of my crosslinking and IP protocols is highly specific; it preserves previously identified RNA-protein interactions [149] without affecting the specificity.

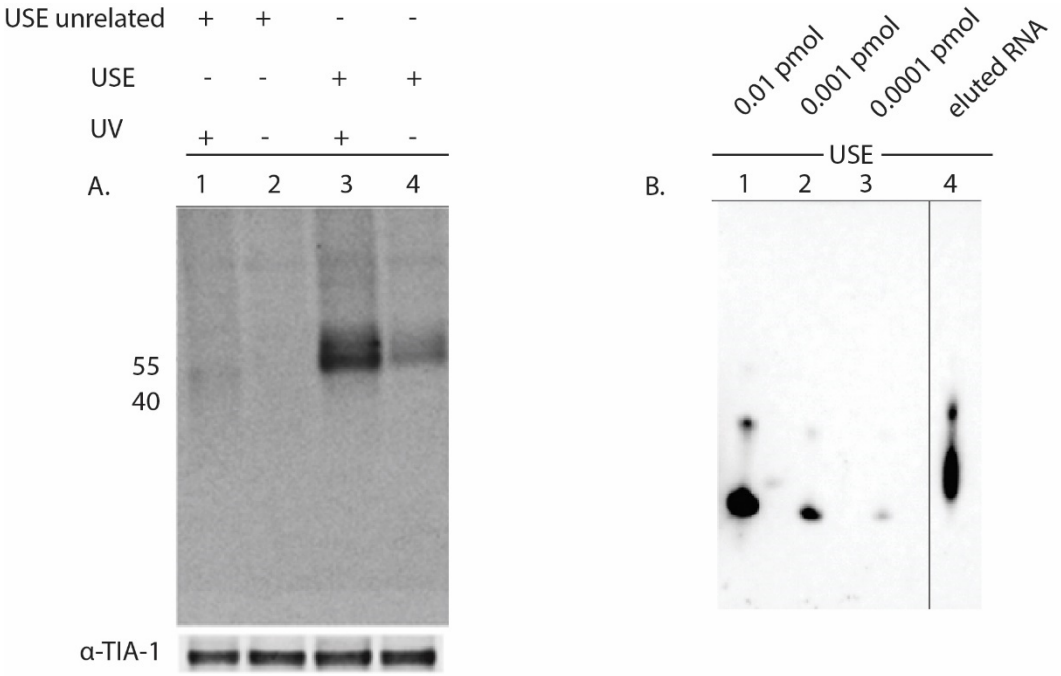


Figure 13 - Visualization of spiked-in biotinylated RNA oligonucleotides confirms specificity of immunoprecipitation and crosslinking (A). The spike-in molecule can be successfully eluted from the nitrocellulose membrane by a proteinase K digestion and detected upon elution (B).

To test whether the next step of the protocol (elution and release of RNA from RNP-complex) is working, the bands representing the RNP-complex were cut from the nitrocellulose membrane. The protein was digested by proteinase K according to the protocol. After elution, the RNA was separated on an 8% UREA PAGE gel, transferred to a nylon membrane and visualized by biotin detection procedure (Figure 13B). Serial dilutions of USE oligonucleotide (0,01-0,0001 pmol) were loaded to estimate the approximate amount of eluted RNA (Figure 13B lanes 1-3) by comparing the signal observed from known amounts of RNA with eluted RNA. As Figure 13B illustrates the amount of USE oligo eluted from RNP-complex is in a range of 0,02-0,01 pmol (approximately 100 pg). Figure 13B also shows that the migration pattern of the RNA eluted from the RNP-complex had slightly changed (compare lane 4 vs. lane 1). It is expected that during proteinase K digestion single amino acid residues covalently bound to RNA may remain attached, which has been shown earlier [132].

Next the same technique was applied to visualize TIA-1 complex with endogenous RNA, partially digested with RNases (Figure 14A). The CLIP protocol is quite sensitive to the amount of RNase used for partial RNA digestion and the type of enzymes. It was shown, that digestion with an endonucleases RNase T1 or RNase A, which degrades ssRNA at G or C and U residues accordingly, may introduce bias. Accordingly, the use of endoribonuclease RNase I is strongly recommended (no sequence specificity) [150]. Both RNases I and T1 were tested in context of this work. Ultimately, a combination of both RNase T1 and RNase I is used in the conCLIP protocol. Depending on the digestion conditions RNA molecules as short as 60 mer can be eluted from the RNA-protein complexes (Figure 14 B, lane 3). When comparing lanes 3, 4 and 5, it was detected, that lower RNase concentrations (Figure 14 B, lane 5) preserve longer RNA fragments attached to the protein, while higher RNase concentrations, preserve shorter RNA fragments of the similar length (Figure 14 B, lane 4). In contrast, when no UV light was used or no antibody added during the immunoprecipitation, expectedly no RNA signal can be detected (Figure 14B, lane 1 and 2). Thus, a protocol was successfully established, which allows to immunoprecipitate RNA-protein complexes and visualize them without using radioactive labeling.

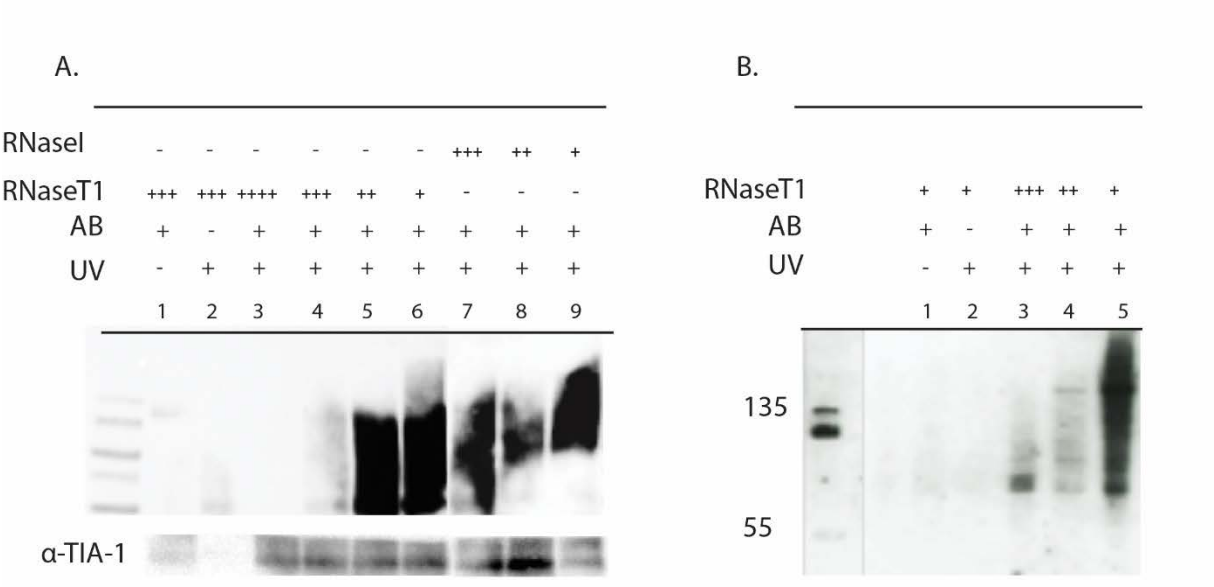


Figure 14 - Visualization of endogenous RNA co-purified with TIA-1 protein (A). Elution of endogenous RNA from TIA-1 - RNA complexes (B). Amount of signal and size of eluted RNA clearly correlates with the amount of RNase T1.

3.1.4 Improving cDNA synthesis protocol

The CLIP technique is known to be a tedious procedure due to many steps it contains (see section 1.5.4.). The main challenge of the method is the cDNA library preparation. In currently existing and published variants of the CLIP protocol the first step of cDNA library preparation is RNA ligation. It is though well known that RNA-ligase is an enzyme with poor efficiency (around 10%). Therefore the ligation step is the most crucial and a limiting step of the protocol. Inefficient RNA-ligation lowers the library complexity. Since our aim is to use the CLIP technique quantitatively it is important for us to preserve the complexity of the library on the highest level possible. Therefore the use of RNA-ligase was eliminated on the low input material. Instead of RNA ligation in the first step of the protocol, the RNA was polyadenylated (Figure 15).

As a polyadenylation reaction is the basis of the conCLIP protocol library synthesis I carefully estimated whether the poly(A) polymerase may have substrate preferences. To address this question the USE element containing oligonucleotide was used. Four types of molecules that differ only by the 3'end terminal nucleotide were synthesized. Next two types of experiments were performed. The first experiment was designed to assess the behavior of the poly(A) polymerase in a time course experiment (30 sec, 3 min and 10 min incubation). Figure 15 A illustrates that there is a slight substrate preference of the poly(A) polymerase. For instance oligonucleotides terminated with A and G are most efficiently tailed, whereas C and G terminated oligonucleotides were less efficiently tailed. The second type of experiment assessed the behavior of the poly(A) polymerase upon usage of different amounts of ATP. Figure 15 B illustrates the same tendency towards a slight preference over A and G terminated substrate. Although there is a slight difference in substrate polyadenylation efficiency, this difference is not striking and can be eliminated by prolonging the reaction time and increasing the substrate amount as illustrated in Figure 15 C. Figure 15 C shows that even 5 µg of RNA oligonucleotide was fully polyadenylated upon overnight incubation with 2mM ATP (lane 3). It is important to note that the amount of RNA tested in this experiment is much higher than that processed in the conCLIP protocol. As estimated before in the context of current work, the expected amount of RNA, eluted from RNA-protein complex may equal to 100 pg. Therefore although the substrate preference cannot be completely excluded, there are no

obvious evidences confirming that the poly(A) polymerase may introduce a significant bias under the conditions used here.

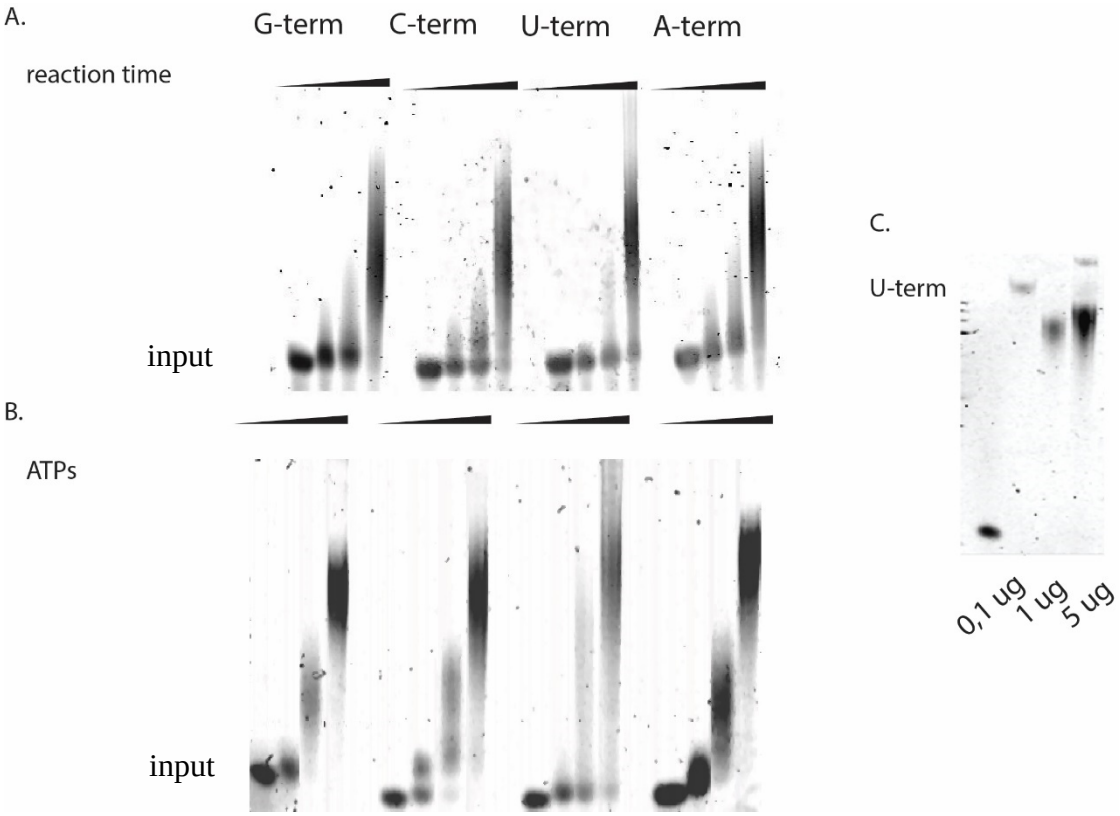


Figure 15 - Testing the polyadenylation efficiency of poly(A)-polymerase on four substrates varying by the terminal nucleotide. U- and C-terminated molecules are less efficiently polyadenylated in time (A) and need access of ATP to get fully polyadenylated (B). U-terminated substrate may be fully polyadenylated after prolonging the reaction time and increasing the amount of ATPs in the reaction mixture (C).

Next the polyadenylated RNA was reversely transcribed using anchored oligo (dT) primers. Three different approaches were further applied to proceed with the library synthesis (Figure 16 Step3). Approach A is based on G-tailing of cDNA using terminal deoxynucleotidyl transferase and subsequent PCR amplification of the product. Approach B employs circular ligase activity, an enzyme which was successfully applied in a recently published iCLIP protocol [151]. Whereas approach C utilizes *in vitro* transcription reaction, serving to boost the amount of RNA in a first step and to work with higher amounts of input

material in the later stages. This approach was first applied for single cell RNA sequencing experiments and is suitable for low input material [152].

The best performance was observed upon using approach C for the library synthesis (details not shown). This approach does not require an additional step of cDNA size selection and therefore avoids loosing valuable material. Importantly approach C contributes to improving the library complexity. To amplify the library for sequencing only as few as 9-11 cycles of PCR are needed. In contrast, currently existing protocols, such as iCLIP and HITS-CLIP use a minimum of 25 cycles of PCR amplification [151, 153]. A high number of PCR cycles typically results in a disproportionate amplification of sequences with different nucleotide composition. Thus the protocol established here is conceptually superior. Despite the fact that the number of cycles used in the conCLIP method is low (below 11 cycles), an amplification bias cannot be excluded. In order to correct for the potential amplification bias unique molecular identifiers were used to label each single molecule as described by Islam and co-workers [154].

It is important to mention that the capacities of modern platforms for NGS sequencing are quite high. For instance a NextSeq system may produce 400 million reads per flow cell (NextSeq 500). As a consequence there is a possibility to pool different samples preliminary labeled by distinct barcodes together. This diminishes the sequencing costs per sample without having a drastic effect on sequencing depth and gained information. This option is being exploited in the conCLIP protocol as well. Thus the oligo (dT) primer contains a 6 digit barcode sequence, which helps to label and distinguish different experimental samples pooled together.

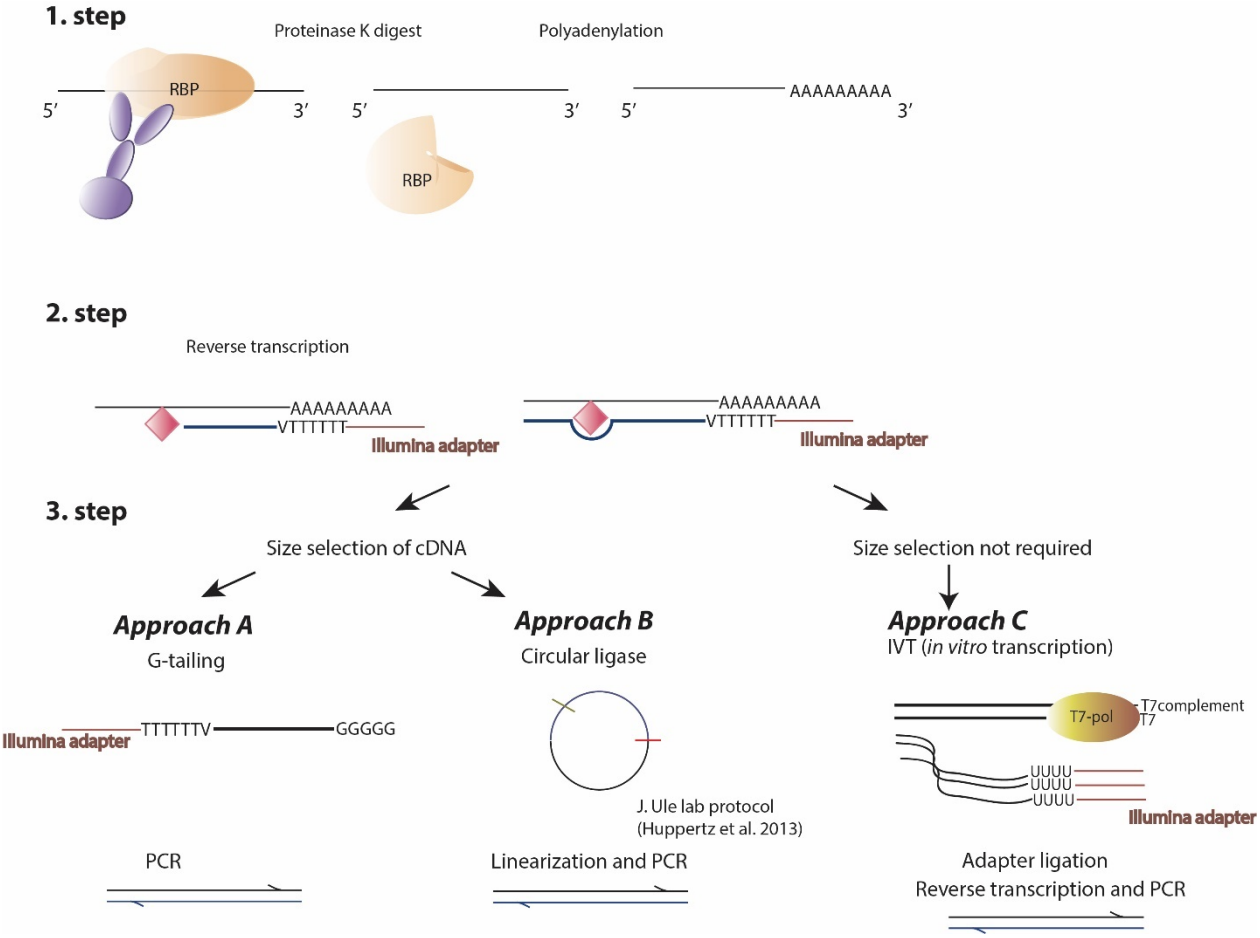


Figure 16 - Schematic representation of three approaches applied for library synthesis. Approach A is based on G-tailing of cDNA; approach B uses circular ligase to circulate the cDNA and approach C applies *in vitro* transcription to amplify input material. Final conCLIP protocol is based on approach C.

3.1.5 Designing a pipeline for conCLIP tags analysis (conCLIP-pip)

To analyze the conCLIP sequencing data, I developed a pipeline. The pipeline approach automatized the process of analysis and enabled to standardize this procedure for all samples. The pipeline is written in shell language and integrates a number of programs listed in Materials and Methods (section 2.4.1.). Few scripts written using R language were integrated into the pipeline or used for data analysis as standalone scripts. Figure 17 illustrates the main steps of the conCLIP pipeline. A more detailed logic of conCLIP pipeline is illustrated in Figure 18.

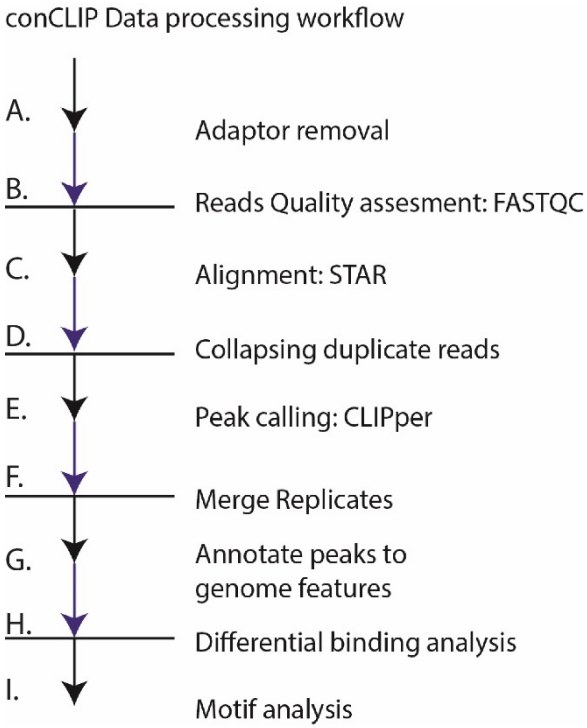


Figure 17 - ConCLIP pipeline workflow.

Briefly, the analysis starts with a demultiplexing algorithm, which separates the pooled samples based on experimental barcode sequence. For this a standalone script is used. After demultiplexing, the reads are trimmed in order to remove adaptor sequences and the reads shorter than 30 nucleotides are discarded (Figure 17 A). The following step of the pipeline identifies quality of libraries by evaluating such parameters as complexity, nucleotide composition and length distribution of sequences (Figure 17B). Next the reads are aligned onto the human genome (Figure 17 C).

The next step of the pipeline serves to remove the reads produced as a result of amplification bias. The reads are removed based on their exact position in the genome, strand and the unique molecular identifier (Figure 17 D). Removal of duplicates has been applied in many variants of the CLIP protocols, particularly in iCLIP and FAST-iCLIP [151, 155]. The principle of deduplication and its importance are explained in Figure 19. The reads aligned to the genome, tend to form dense regions. The positions of these regions can be computationally calculated by a peak calling algorithm. If the amplification duplicates are not removed, the peak calling algorithm may falsely detect the dense regions, which do not reflect

the real binding of the protein, but stem from an amplification introduced bias. If the duplicates were removed before, the dense regions can be correctly assigned (Figure 19 B and C). After the peaks are defined they are classified by the p-value, and peaks with a significant p-value (<0.05) are kept (Figure 18, step 6). To count for a potential background contamination, the peaks are also determined for the negative control experiment. In case of the negative control, the experiment is carried out exactly the same way, except from the immunoglobulin G is used as the antibody. Thus the peaks detected in the negative control are composed of sequences with high affinity to the beads or antibodies, but not the protein of interest. The peaks defined in the negative control are subtracted from peaks determined in the experimental (positive) sample (Figure 18, step 7).

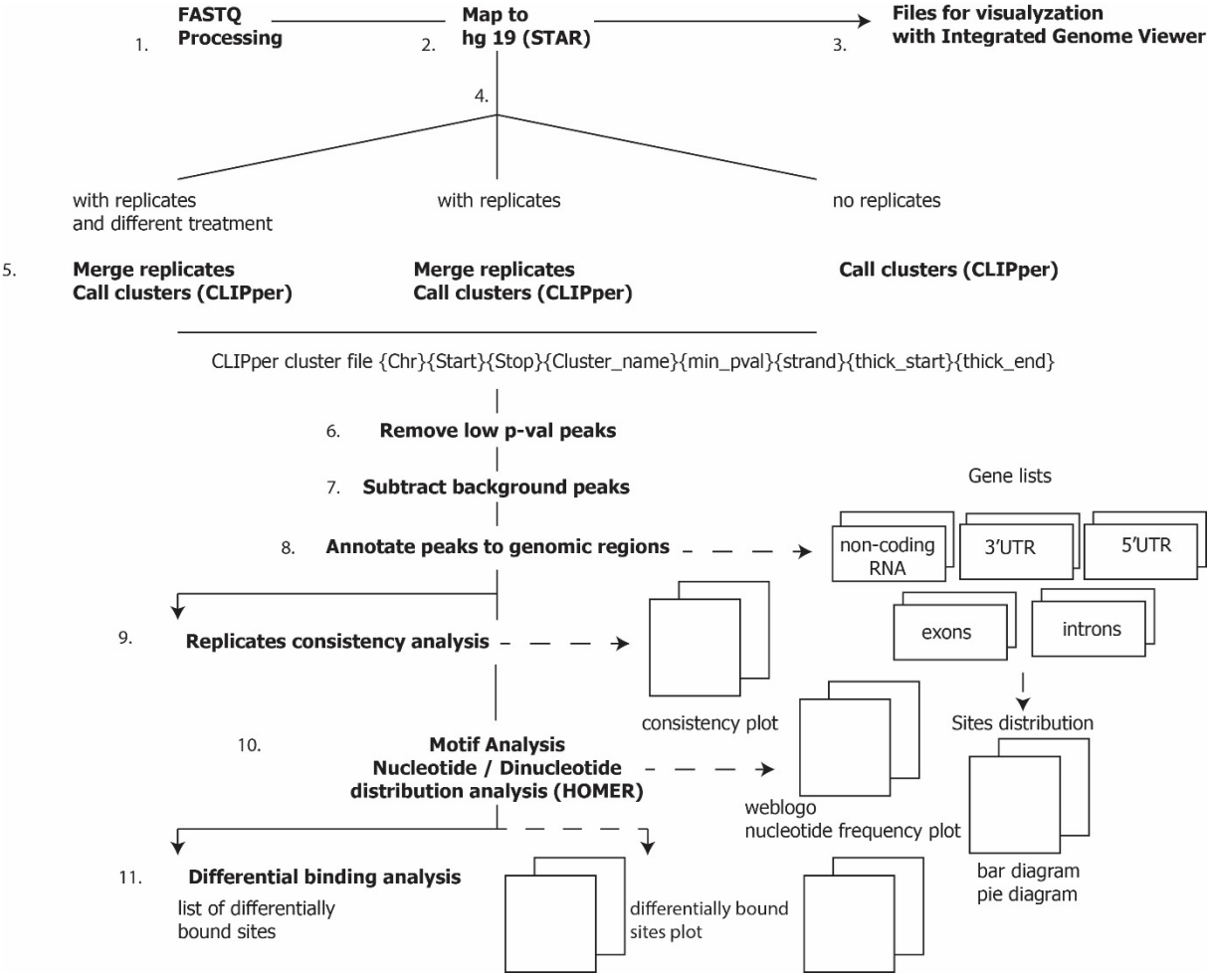


Figure 18 - Schematic overview of conCLIP pipeline

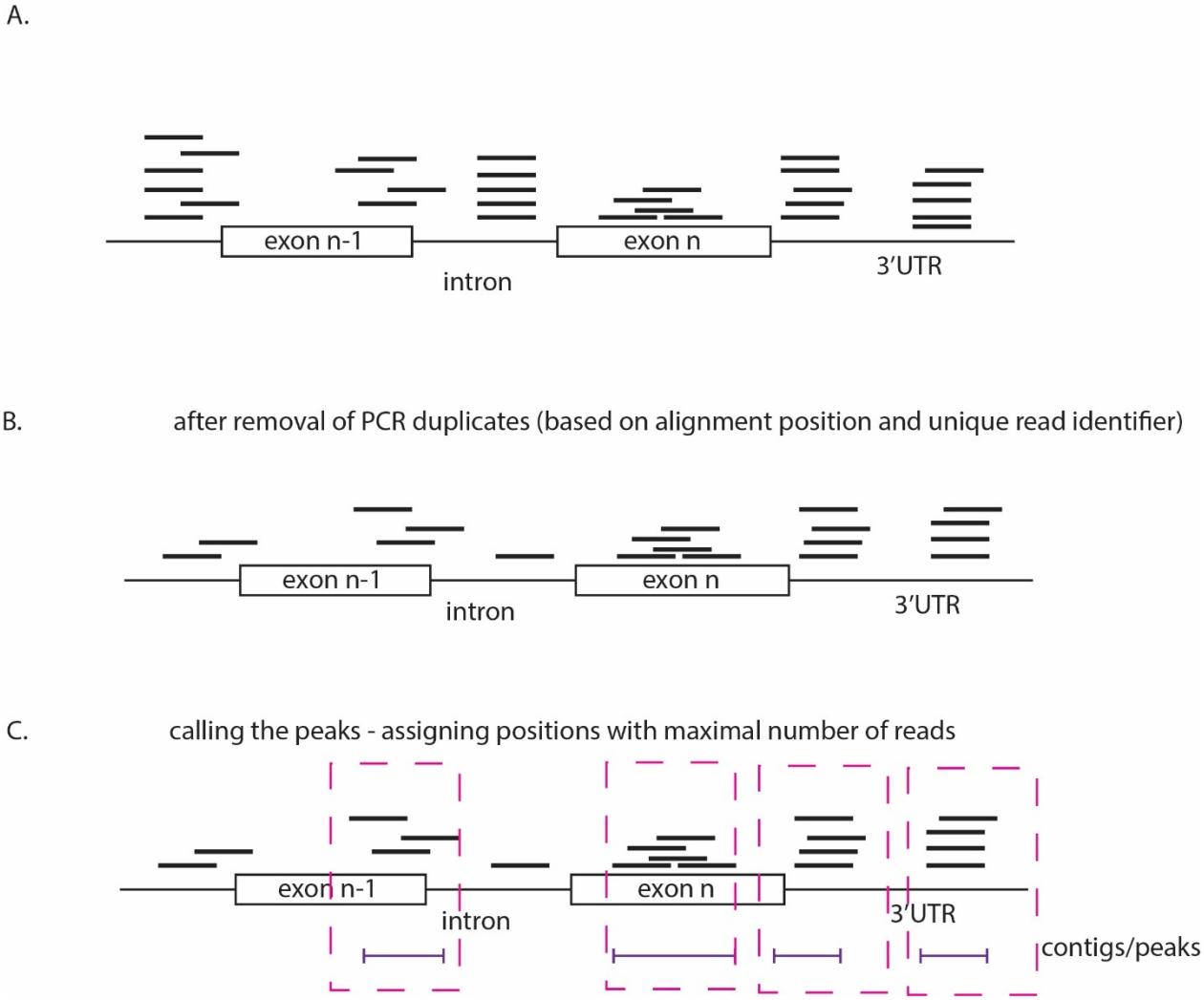


Figure 19 - Schematic overview of duplicate removing principle and its importance for analysis. Upon alignment the genomic coordinates of each read are determined (A). As an example, the duplicated reads form a dense region in the intron (A). After the removal of duplicates less reads are present in the intron (B). The peaks are detected upon removal of duplicates, no false peaks are defined (C).

Next the peaks are annotated onto genomic regions (Figure 18, step 8). The conCLIP pipeline distinguishes between non-coding and coding types of RNAs (Figure 18, step 8). Peaks are classified as being aligned to 3' UTRs, 5' UTRs, exons, introns or intergenic regions (Figure 18, step 8). An additional script is utilized to accomplish the visualization of the quantitative distribution of the peaks on the genomic region (Figure 18, step 8). The peaks belonging to the same type of feature are combined into lists (Figure 18, step 8). Significant peaks detected in more than one replicate are further used to search for enriched motifs (Figure 18, step 10). As an additional assessment of the sequence composition of the sites, a

nucleotide and dinucleotide distribution analysis is performed (Figure 18, step 10). In case of motif and the sequence composition analysis, the peaks are newly assigned by defining a -50 to +50 nt regions relative to the center of the original peak.

The conCLIP pipeline is designed such that the user can analyze results from experiments with different setup. For instance, the setup may include experiments with replicates and different treatments or else experiments without replicates can be analyzed.

In addition to the summary files, the pipeline outputs several plots and diagrams, which make visual analysis of data easy and fast. For example, if several replicates are used in the experiment, a replicate consistency plot is produced (Figure 18 step 10). If the pipeline is used to compare protein binding capacity in different conditions an edgeR Bioconductor package is applied to detect differentially bound sites (Figure 18 step 11).

The conCLIP pipeline designed and written in context of the current work thus provides the investigator with a tool for a comprehensive analysis of conCLIP sequencing data. It uses different existing tools for NGS data analysis and provides a fast, standardized and automatized analysis. The fast speed of the alignment is achieved by the usage of STAR program. It enables a fast and accurate alignment and provides flexible control over alignment process. A peak defining algorithm efficiently defines the peaks and assigns significance thresholds on a gene-by-gene basis and not by a genome-wide cutoff as the peak searching algorithms used for chromatin immunoprecipitation sequencing data analysis [139]. This tool has shown the best performance in the type of data analyzed in the context of current work. The conCLIP pipeline permits some flexibility of analysis and is suitable for all three types of experimental setups. The main results of analysis are visualized in form of diagrams and graphs (see next chapters).

3.2 Chapter 2

3.2.1 Transcriptome-wide occupancy of the core polyadenylation machinery factor CSTF2tau in BE(2)-C cells

CSTF2tau is a RNA-binding protein directly involved in RNA processing by guiding the core cleavage and polyadenylation machinery to the particular cleavage and

polyadenylation sites on mRNAs [44, 156]. CSTF2tau RNA binding capacity and specificity has been studied with help of HITS-CLIP as well as iCLIP approaches independently by two research groups [41, 44, 157]. The protein is an ideal candidate for evaluating the conCLIP method as the binding preferences of it have been well described before. Moreover, the binding of CSTF2tau in proximity of the cleavage sites predicts the cleavage site usage with high probability [41]. CLIPs of this protein can be thus used as a marker to trace the APA sites usage in a dynamic fashion, for instance upon applying various conditions, initiating different cellular programs or depleting factors involved in APA to follow the APA changes.

The conCLIP method was applied to study binding preferences of CSTF2tau protein in the BE(2)-C cell line. The immunoprecipitation and library synthesis were carried out according to the protocol established in this study. The computational analysis was performed with help of the conCLIP pipeline designed and written in the context of this work.

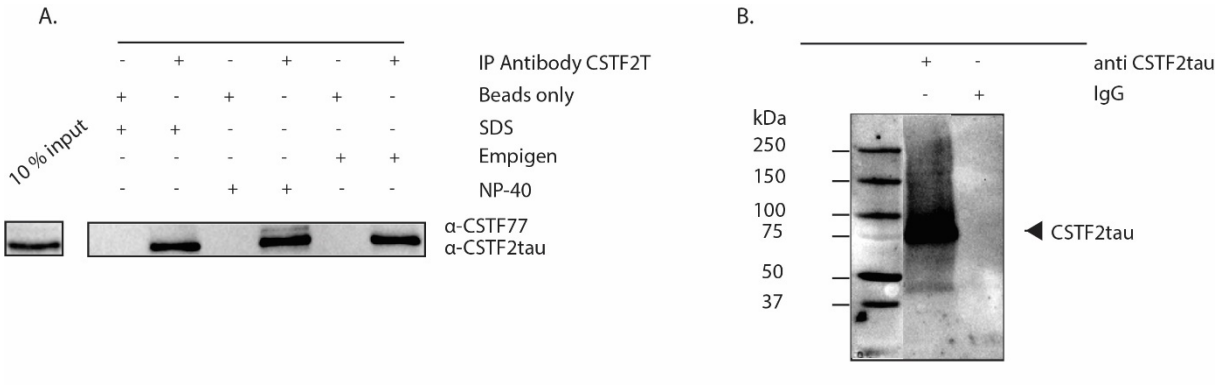


Figure 20 - Immunoprecipitation the CSTF2tau protein. Efficiency and specificity of immunoprecipitation of the protein were assessed by western blotting (A). Various detergents were tested to specifically immunoprecipitate CSTF2tau (A). Usage of washing buffer containing EmpigenBB allows a specific immunoprecipitation (A). CSTF2tau RNA-protein complexes were visualized by biotin detection procedure (B). Lane with IgG used for immunoprecipitation serves as negative control and is empty (B).

Endogenously CSTF2tau was specifically immunoprecipitated from BE(2)-C cells. Three different conditions were tested to improve the specificity and efficiency of immunoprecipitation. The specificity was controlled by silver staining (not shown) and Western blot analysis (Figure 20A).

CLIP type approaches are very sensitive to immunoprecipitation efficiency and purity. A co-immunoprecipitation of other interacting partners together with the protein of interest

may lead to a false discovery of RNA species. Therefore it is crucial to assess the purity of the IP. In this study proteins, which may be potentially co-immunoprecipitated were assessed by Western blot. CSTF77 has been shown to interact with CSTF2tau [158]. Therefore the membrane with immunoprecipitate was probed with antibodies against CSTF77 (Figure 20A). As illustrated, the usage of NP-40 containing buffer leads to co-immunoprecipitation of CSTF77 together with CSTF2tau (Figure 20A). As the co-immunoprecipitation of other proteins is not desired, the NP-40 containing buffer was no longer used in the current protocol.

Upon immunoprecipitation the RNA-protein complexes co-pulled together with the protein were visualized. As shown in Figure 20B, CSTF2tau immunoprecipitates high quantities of RNA (Figure 20B).

The RNA digestion conditions were specifically optimized for CSTF2tau protein with the aim to get the RNA digested as close to the binding site as possible. I observed that intensive digestion helps to improve reproducibility. Therefore, the conditions and the time of digestions were thoroughly controlled.

CLIP techniques are usually characterized by rather low reproducibility [150]. In order to evaluate the success of the experiment on the first hand, I was assessing the reproducibility of biological replicates. Figure 21 illustrates the consistency between 2 replicates of the conCLIP method applied to CSTF2tau protein. Figure 21 shows the correlation between the normalized read counts of the most conserved binding sites defined within the two replicates. The correlation between replicates was calculated using Pearson’s correlation method and equaled to 0.9739 (Figure 21). As reported previously, correlations between replicas produced by CLIP variants are different and usually vary between 0.6 and 0.9 [150]. A correlation of above 0.97 observed in this study by applying conCLIP technique suggests that this technique is highly reproducible.

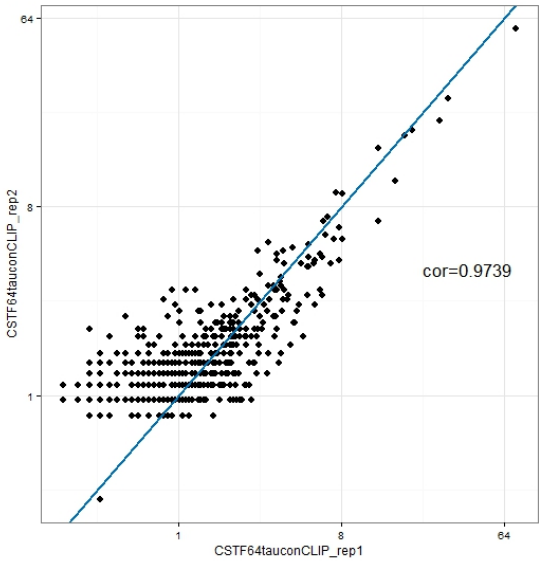


Figure 21 - Assessing consistency between two conCLIP replicates. Consistency plot shows the ratio between reads coverage on conserved CSTF2tau binding sites of 2 replicates. Pearson correlation between replicates was estimated as to be 0.9739.

Further analysis of the CSTF2tau conCLIP using the conCLIP pipeline revealed the binding preferences of this protein. Figure 22 illustrates the genome-wide distribution of CSTF2tau sites. As shown in Figure 22, 38% of the sites are localized in introns; 33% are aligned onto 3'UTRs, 7% belong to 5' UTRs, 4% of sites were detected on non-coding genes and 2% on exons (Figure 22, pie diagram). After correction for the length of each feature, it was observed that 3' UTRs and 5' UTRs are covered by the protein at most (Figure 22, bar diagram). The observed specificities of the protein, which particularly tends to bind 3' UTRs, recapitulates previously described properties of CSTF2tau.

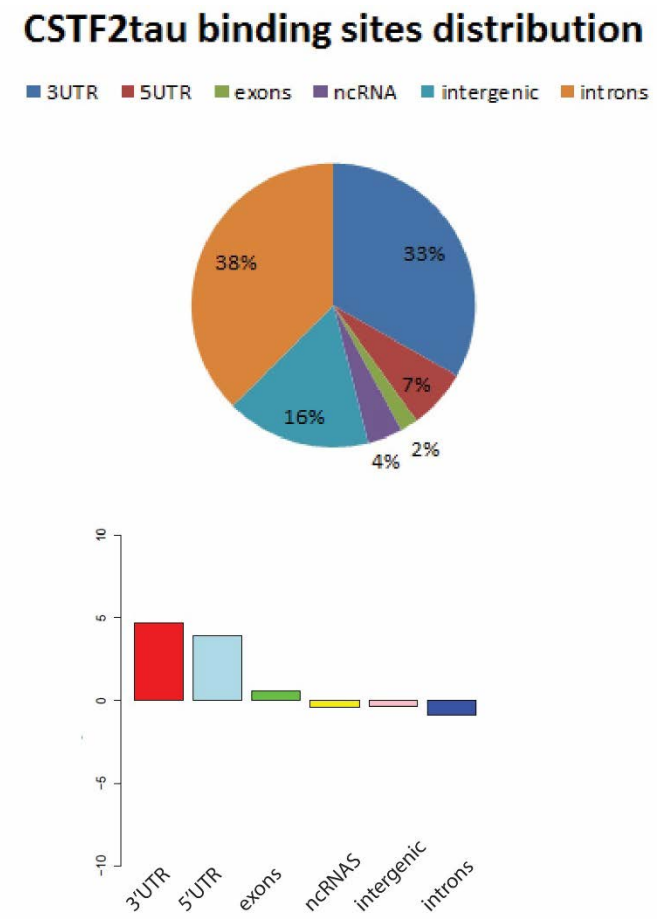


Figure 22 - Distribution of CSTF2tau binding sites on genomic features. The pie diagram shows the distribution of binding sites of the CSTF2tau on the genomic features in percent. The bar diagram shows enrichment in coverage of sites over the total length of the feature.

Further analysis of the exact location of the binding sites within the 3' UTRs revealed that the peaks are focused and co-localized with cleavage and polyadenylation sites (Figure 23, marked in red color).

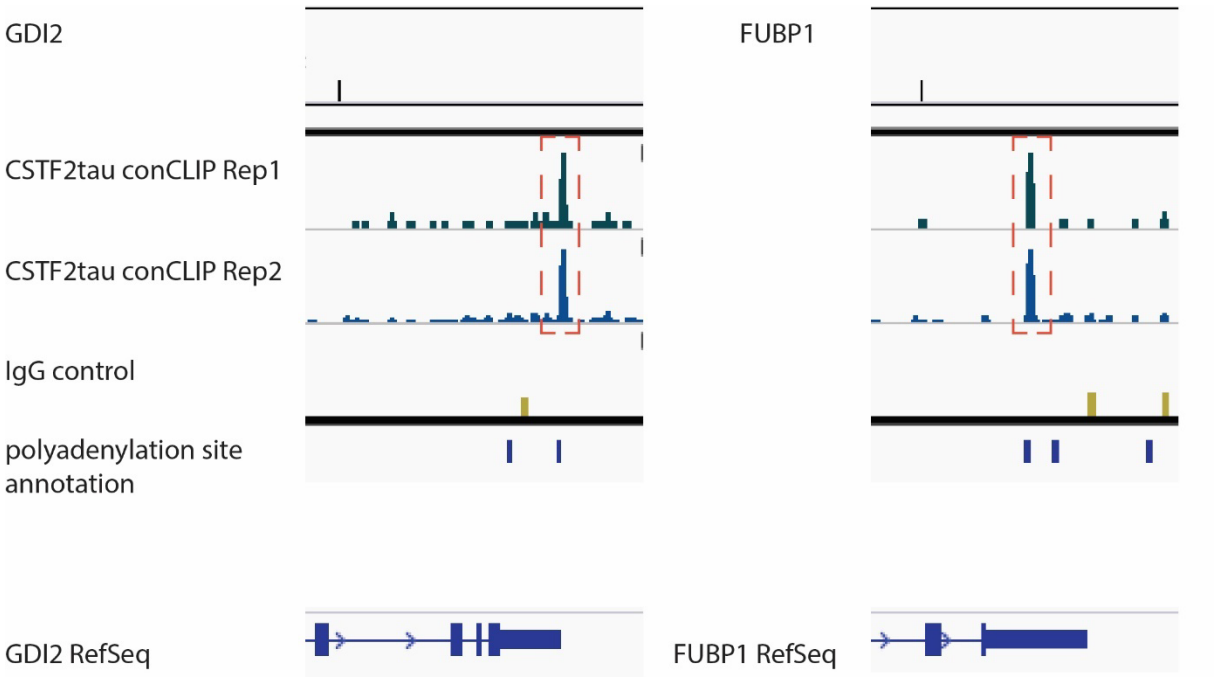


Figure 23 – Exemplified binding of CSTF2tau protein on GDI2 and FUBP1 coding transcripts visualized by Integrative Genomics Viewer. The majority of peaks formed by CSTF2tau co-localize with transcript cleavage sites. Peaks formed by CSTF2tau are focused. Samples, which were generated by immunoprecipitation with an IgG antibody, serve as a control.

Next, I performed the analysis, which allowed me to assess the exact position of the binding sites' centers in comparison to cleavage sites transcriptome-wide. The analysis revealed that the binding sites centers are at most frequent within 30-80 nucleotides downstream of the cleavage and polyadenylation sites (Figure 24). Approximately 11% (1654 sites out of 13110) of all detected binding sites are localized within 100 nucleotides of polyadenylation sites shown to be used in the BE(2)-C cell line (Danckwardt lab, unpublished data). These results recapitulate previous finding about the location of the CstF complex on the 3'UTR and relative to the cleavage site. As discussed in the introduction (section 1.1.4 and 1.1.5), this part of the cleavage and polyadenylation machinery recognizes the *cis*-elements located downstream of the cleavage site.

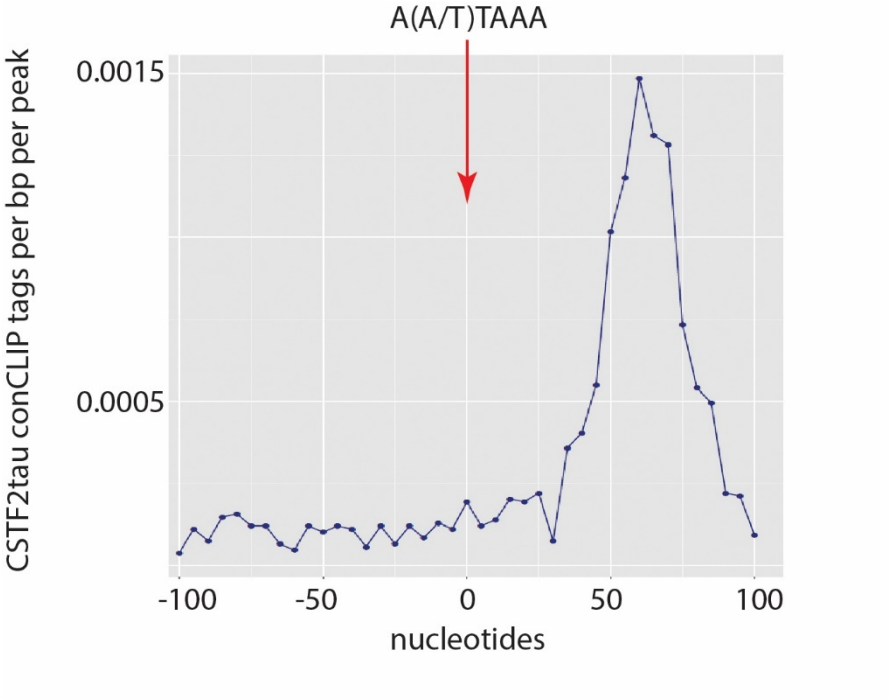


Figure 24 - Relative position of centers of CSTF2tau binding sites around the cleavage sites (0 position represents the position of A(A/U)UAAA hexamer).

In the following analysis I aimed to determine the sequence preferences of the CSTF2tau protein. In order to achieve that, a nucleotide composition of conserved sites was analyzed. Figure 25 A provides an overview on the distribution of nucleotide frequencies along the length of cumulative CSTF2tau binding sites. As illustrated, frequencies of A and T (U) nucleotides are higher than expected (black line). High frequencies of As and Ts are observed along the whole analyzed region, in particularly they are high at the beginning of the motif (Figure 25 A). Figure 25 B presents the frequency of dinucleotides along the binding regions. Frequencies of dinucleotides AA, TG (UG) and TT (UU) are higher than expected (marked by black line) (Figure 25 B).

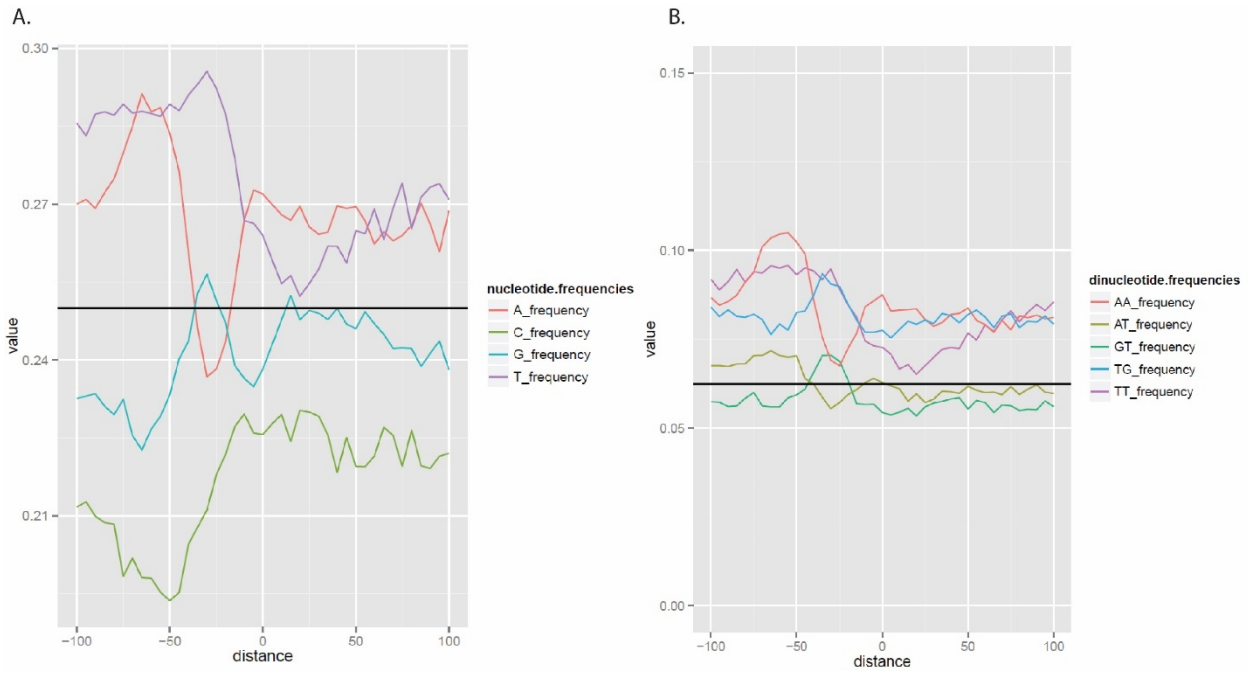


Figure 25 – Sequence composition around the CSTF2tau binding sites. Distribution of nucleotide (A) and dinucleotide (B) frequencies along the binding region of CSTF2tau protein.

The over-representation of A and T rich regions within the binding sites located near the cleavage and polyadenylation sites' annotation can be also observed when applying *de novo* motif search algorithm to all determined sites. *De novo* motif analysis reveals enrichment of several motifs (Figure 26). The first motif contains sequence resembling an AAUAAA hexamer sequence, the other two motifs are U/GU rich and most probably represent the downstream U/GU/UG rich sequence elements [32] (Figure 26), required for efficient processing and typically recognized by the CSTF2/2tau proteins [44]. This finding adds additional piece of evidence suggesting that the conCLIP method precisely and reliably identifies positions of binding of endogenously expressed CSTF2tau protein, in accordance to previously published data [44].

Figure 26 - Motifs, recognized by CSTF2tau protein. First most significantly enriched motifs are shown.

To reveal the functional categories of CSTF2tau bound genes, the gene list was submitted to the DAVID functional annotation tool. The gene Ontology analysis shows that the genes bound by CSTF2tau protein are enriched in such categories as mRNA processing, transcription and translation (Figure 27). To recapitulate, the CSTF2tau protein preferably binds the transcripts involved in RNA processing the same category the CSTF2tau itself belongs to. This suggests that there might be a feedback regulatory mechanism, controlling the RNA processing for genes involved in this process.

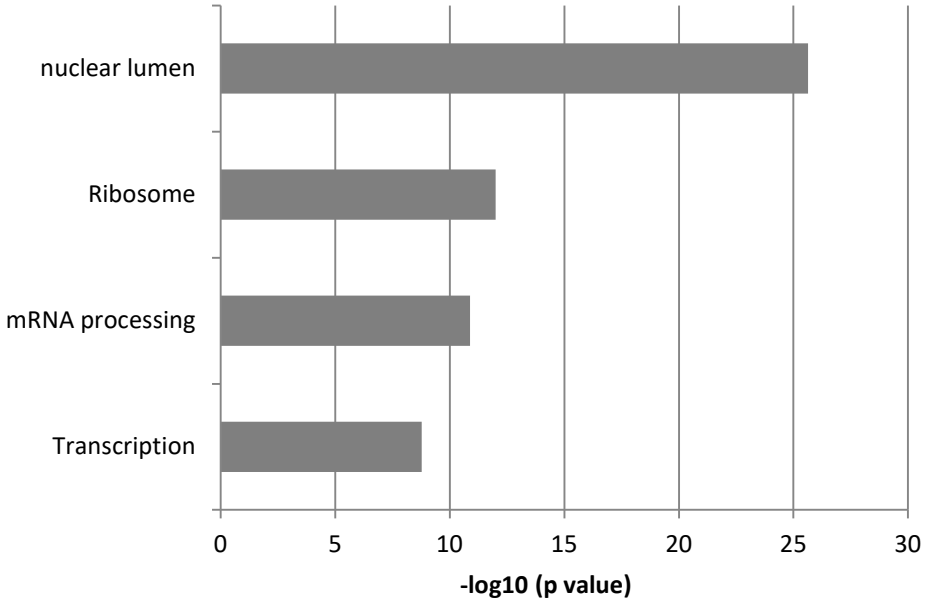


Figure 27 - Enriched Gene Ontology categories of genes bound by CSTF2tau.

To address the question whether the binding profile of CSTF2tau protein in BE(2)-C cells predicts the cleavage site usage, the probability of a predominant binding around the major cleavage site was calculated. Based on polyA sequencing data generated in our lab (A. Ogorodnikov, unpublished data) genes with multiple polyadenylation sites were selected. These genes were ranked by their expression level and only top 3000 genes were further analyzed. Out of top 3000 genes with major cleavage sites, which accumulate at least 90 percent of reads (or more), were selected (Figure 28). The window of 120 nucleotides around the major cleavage site was set as a 0 window. The annotation of selected genes was expanded by 2000 nucleotides towards the 3'end in order to include the most distal cleavage sites, which are located further downstream of the 3'end of the gene annotation. Next the selected genes' loci were split into windows with a reference 0 window being located around the strongest cleavage site (Figure 28). Each window except from the very last 3' and 5' end windows were set to be 120 nt long. Figure 28 schematically illustrates the approach described above. An example of this calculation using an arbitrary gene A is given in Figure 28. This gene contains 3 annotated cleavage sites (CS1, CS2, CS3). Cleavage site 3 (CS3) is the majorly used cleavage site and is covered by 90 % of the reads (90 reads out of 100). Let the reference window be set around the CS3. The 0 window coordinates are now determined and the gene A is split into windows of 120 nucleotides each. The windows are numbered: those, which are located downstream of the reference window are numbered positively. Windows located upstream of the reference window are numbered negatively. Next, the number of reads generated by the conCLIP experiment and aligned to each window is calculated. If the maximal number of reads was calculated within the windows -1, 0 or 1, the binding of CSTF2tau protein was considered to be predictive for a cleavage site usage. The same logic is later applied for less strongly used cleavage sites with the varying thresholds (> 70% of reads; > 60% of reads; >50% of reads, see Table 3). Table 3 illustrates that the binding of CSTF2tau in proximity to cleavage site promotes usage of this site in 25 % of transcripts possessing multiple APA sites in neuroblastoma cells. This result suggests that CSTF2tau protein binding pattern has a predictive power over the dominant site usage, yet the percentage of prediction of site usage in BE(2)-C cell line is lower than previously shown in HeLa cells [41].

To conclude, CSTF2tau protein RNA-binding preferences were addressed in the context of the current study. It has been shown that the protein binds with higher preference to 3' UTRs and with less preference to 5' UTRs. This study confirms that there is a noticeable number of peaks centralized 50 to 60 nucleotides downstream of the A(A/U)UAAA consensus sequence. This finding highlights an importance of CSTF2tau protein in recognition of the polyadenylation sites. The current study also reveals that usage of up to 25 % of major cleavage sites in a cohort of genes with multiple cleavage sites are predicted by binding of CSTF2tau protein in BE(2)-C cells.

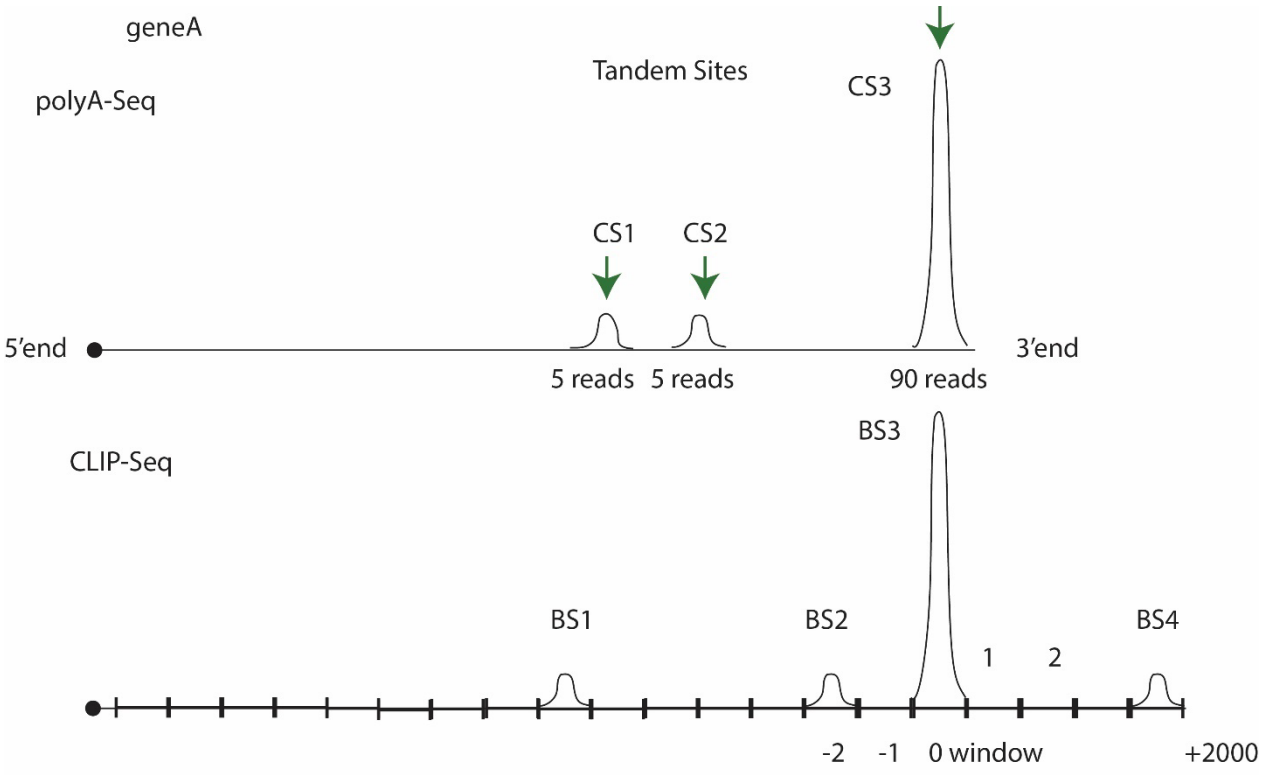


Figure 28 - Schematic overview of logic underlying the calculations of probability of cleavage site prediction. The upper part of the figure represents an arbitrary gene A with three cleavage sites (CS) detected by a poly(A)-sequencing approach [159]. The lower part of the figure shows the same gene with regions bound by the CSTF2tau protein and labeled as binding sites (BS) 1 to 4.

Table 3 - Proportion of genes with tandem cleavage sites in which CSTF2tau has the maximum binding at the dominant as opposed to alternative cleavage sites.

Reads at the dominant CS	>50%	>60%	>70%	>90%
Total number of genes	2296 (100%)	2008 (100%)	1760 (100%)	662 (100%)
conCLIP (CSTF2tau) Rep1	473 (20,6%)	426 (21,2%)	379 (21,21%)	153 (23,11%)
conCLIP(CSTF2tau) Rep2	563 (24,5%)	495 (24,6%)	432 (24,65%)	175 (26,43%)

3.2.3 Dynamic conCLIP: studying the dynamic changes in binding of CSTF2tau protein upon knockdown of CFIIm complex component PCF11

To ultimately address the question whether CSTF2tau binding reflects APA, the conCLIP protocol was applied upon the depletion of PCF11, a protein regulating poly(A) site usage (Danckwardt lab, unpublished data; [160]. To this end, PCF11 was depleted by siRNA treatment. Cells treated with siRNA against C. elegans (siC. el) gene were used as control. Efficiency of knockdown was assessed by western blotting; the protein was depleted down to 25% (Figure 29 compare lanes siRNA1 and siRNA2 vs siRNA C.el).

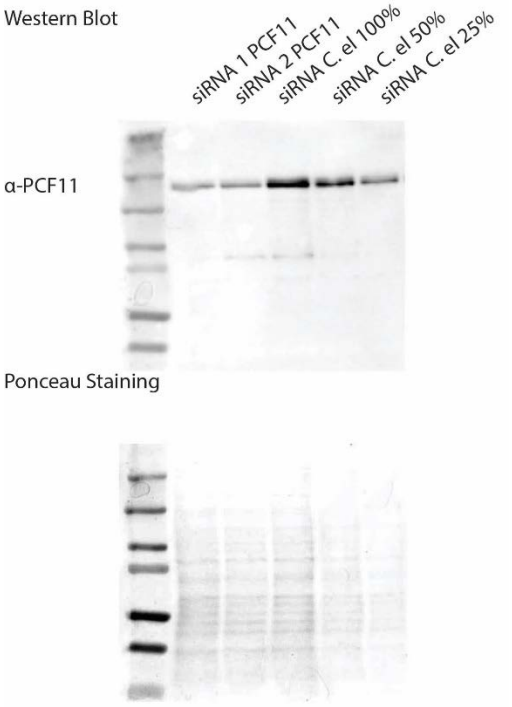


Figure 29 - The knockdown efficiency of PCF11 was assessed by western blotting. The knockdown efficiency of PCF11 obtained by both siRNAs 1 and 2 was down to 25% of residual protein levels. Equal loading was assessed by Ponceau Red staining.

I next carried out dynamic conCLIP experiments aiming to assess changing in binding preferences of the CSTF2tau protein upon PCF11 knockdown. The consistency between replicates was determined by assessing Pearson's correlation coefficient. Figure 30 illustrates a heat map of correlation coefficients between 4 samples where conCLIP was carried out after BE(2)-C cells were depleted from PCF11 protein (Figure PCF11_1.1,PCF11_1.2, PCF11_2.1 and PCF11_2.2) and a control sample (Figure 30 contr_1.1). As illustrated, the correlation

coefficient between siRNA 1 and siRNA treated samples are 0.93 and 0.99, accordingly. Interestingly, conCLIP experiments performed on cells treated with siRNA 2 against PCF11 protein correlated at least with the control.

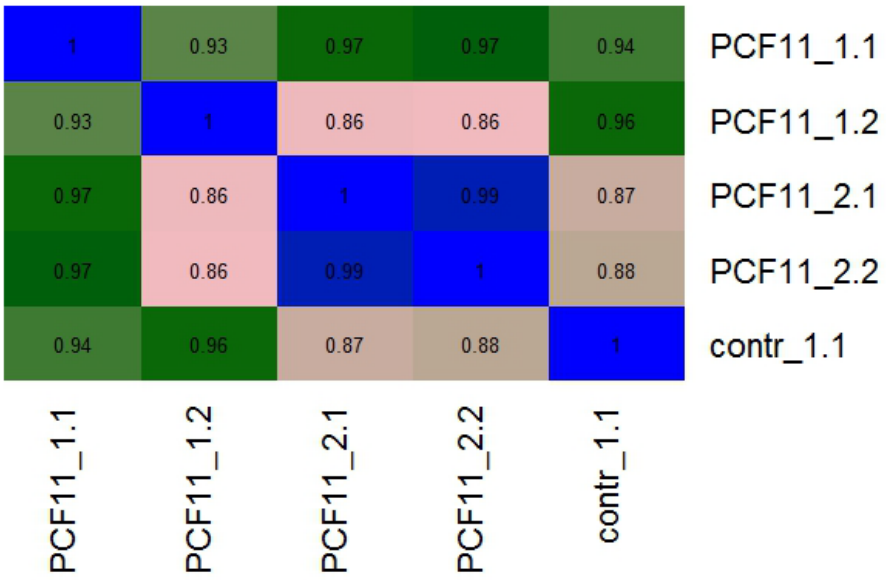


Figure 30 - Heat map of correlations between conCLIP protocol replicates. Correlation was calculated by counting the reads aligned to 3'untranslated region of the gene, namely -40 +80 region around the cleavage sites.

To analyze the quantitative changes of CSTF2tau binding downstream of polyadenylation sites, the conCLIP coverage on these regions was compared between control and knockdown samples. Based on the results of the location of CSTF2tau peaks generated in the context of this work (see Figure 24), the regions of -40 +80 around the cleavage sites were selected to serve as an annotation. The number of reads mapped strictly to the selected regions was count for untreated and treated samples. The EdgeR Bioconductor package was used to estimate quantitative changes of CSTF2tau binding in 2 conditions. As binding of CSTF2tau upon PCF11 depletion using siRNA2 was the most distinct from control, these 2 replicates were further analyzed (PCF11_2.1, PCF11_2.2). In total 59 sites were detected to be significantly regulated (threshold of false discovery rate <0.05) (see Figure 31). Similarly the differential expression of APA isoforms was determined (data not shown).

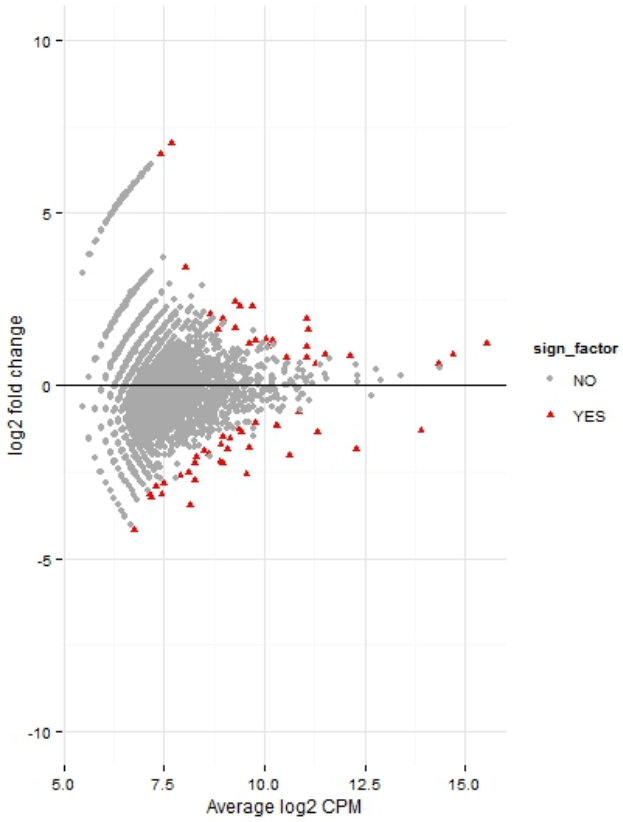


Figure 31 - Differential binding of CSTF2tau upon PCF11 depletion. EdgeR Bioconductor package reveals differentially bound sites (FDR<0.05), colored in red.

Interestingly, a small overlap between the differentially used APA isoforms and differentially bound CSTF2tau sites was observed. However, when plotting the fold change of the most significantly changed poly(A)-seq retrieved APA sites versus the fold change of the most significant differentially bound CSTF2tau sites, a striking positive correlation was observed (Figure 32, further discussed in the discussion section).

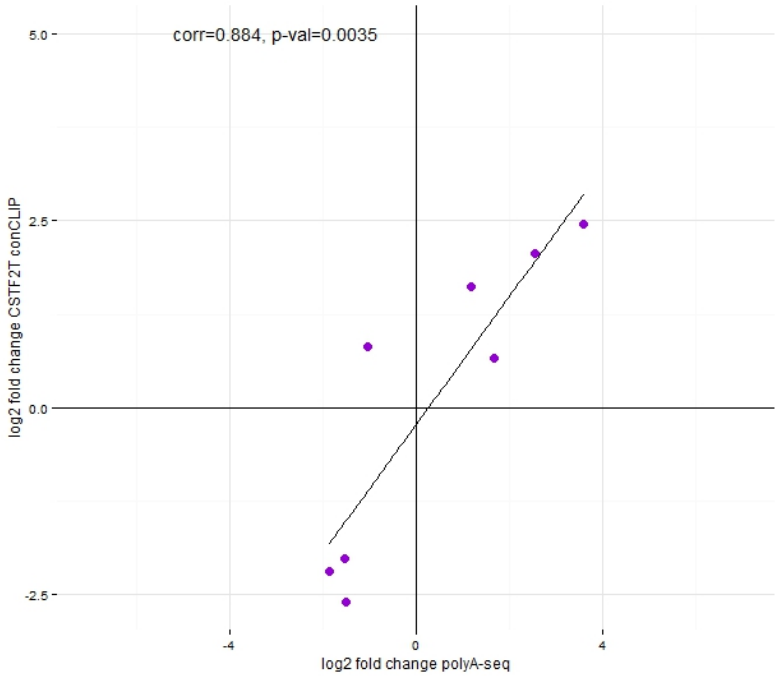


Figure 32 - Significant positive correlation of fold changes of APA and CSTF2tau binding sites upon PCF11 depletion. Correlation between most significant dynamically bound CSTF2tau sites and alternative polyadenylation assessed by EdgeR Bioconductor package (FDR<0.05) was calculated by Person's correlation method. Each dot represents an APA site, undergoing significant regulation and significantly differentially bound by CSTF2tau protein.

Figure 33 illustrates two example genes, in which the binding of CSTF2tau reflects APA. Both genes, LONRF2 and MXD4, undergo APA upon PCF11 knockdown (Figure 33 compare poly(A)-seq PCF11 KD versus mock control). As shown in Figure 33, the longer transcript isoforms of LONRF2 and MXD4 are more prevalent upon PCF11 knockdown. At the same time, CSTF2tau protein (which apparently recognizes only distal cleavage site of these transcripts) binds more efficiently downstream of the distal poly(A) site upon PCF11 knockdown and thus reflects the cleavage site usage (Figure 33 compare CSTF2tau conCLIP, PCF11 KD_Rep1 and Rep2 versus mock control), further discussed in the Discussion section.

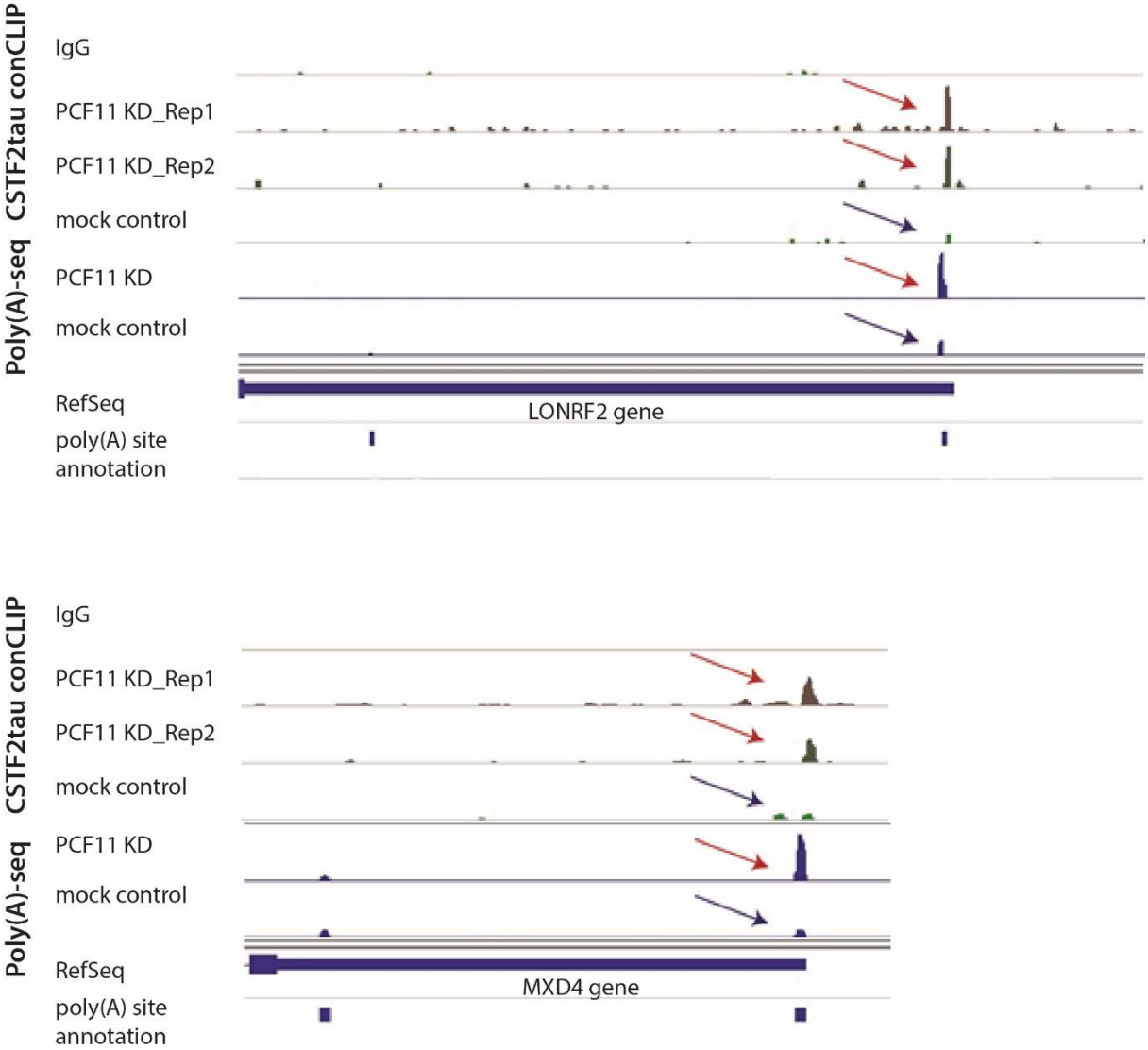


Figure 33 - IGV snapshots illustrate predictive manner of CSTF2tau binding for poly(A) site choice. Increased binding of CSTF2tau protein upon PCF11 knockdown reflects alternated usage of distal cleavage site (on example of 2 genes LONRF2 and MXD4).

In conclusion, the conCLIP method, established in the context of this thesis, reproduces previously published data [44]. The protocol is carried out in the absence of radioactivity. Furthermore, the protocol omits a size selection of cDNA, the step necessary to remove adaptor-adaptor contaminants and requires as little as 10 PCR-cycles to generate sufficiently complex libraries. Finally, the conCLIP method appears to be suited to study dynamic RNA-protein interactions. Accordingly, differential CSTF2tau conCLIP quantitatively reflects (some of the known) APA changes.

3.3 Chapter 3

3.3.1 Binding of CSTF2tau on histones and non-coding RNAs

ConCLIP is capable of recognizing more than 16,000 binding sites of CSTF2tau protein (see Chapter 2). As described in the previous chapter, CSTF2tau recognizes 3’ UTRs of coding RNAs at the highest degree. It has been shown before that the protein is involved in 3’ end processing [41, 44]. In this context binding of CSTF2tau to the 3’ ends is easily explained and even expected. Yet the deeper analysis of other CLIP-tags identified here may shed the light on novel functional aspect of CSTF2tau.

Previously it has been reported that CSTF2 (paralog of CSTF2tau) was found in a complex with U7snRNP and functions in the 3’ end processing of replication-dependent histones [53]. CSTF2tau in turn can be recruited by U7snRNP complex in mouse embryonic stem cells upon depletion of CSTF2 [53]. In the context of the current study, which was carried out in human neuroblastoma cell line, I observe the binding of CSTF2tau on replication-dependent histones as well as on U7snRNA. This suggests that CSTF2tau protein forms complexes with U7snRNP in the presence of CSTF2 in neuroblastoma cells.

The position of the CSTF2tau binding sites on replication-dependent histones deserves special attention. Unlike previously described binding preferences of the protein with a tendency to recognize 3’ end regions of mRNAs, binding of CSTF2tau on replication-dependent histones occurs at the 5’ end of the reading frame. This observation may be explained by different composition and spatial organization of the 3’ end processing machinery. Figure 34 illustrates the difference of the spatial location of the binding sites on replication-independent histones, which are processed by conventional cleavage and polyadenylation machinery (for example H3F3A and H1FX, Figure 34A) and replication-dependent histones, which are processed by a unique 3’ end histone-processing mechanism (HIST1H4E and HIST1H3D). Further examples showing histones with binding sites of CSTF2tau protein in the same logic are listed in tables 4 and 5.

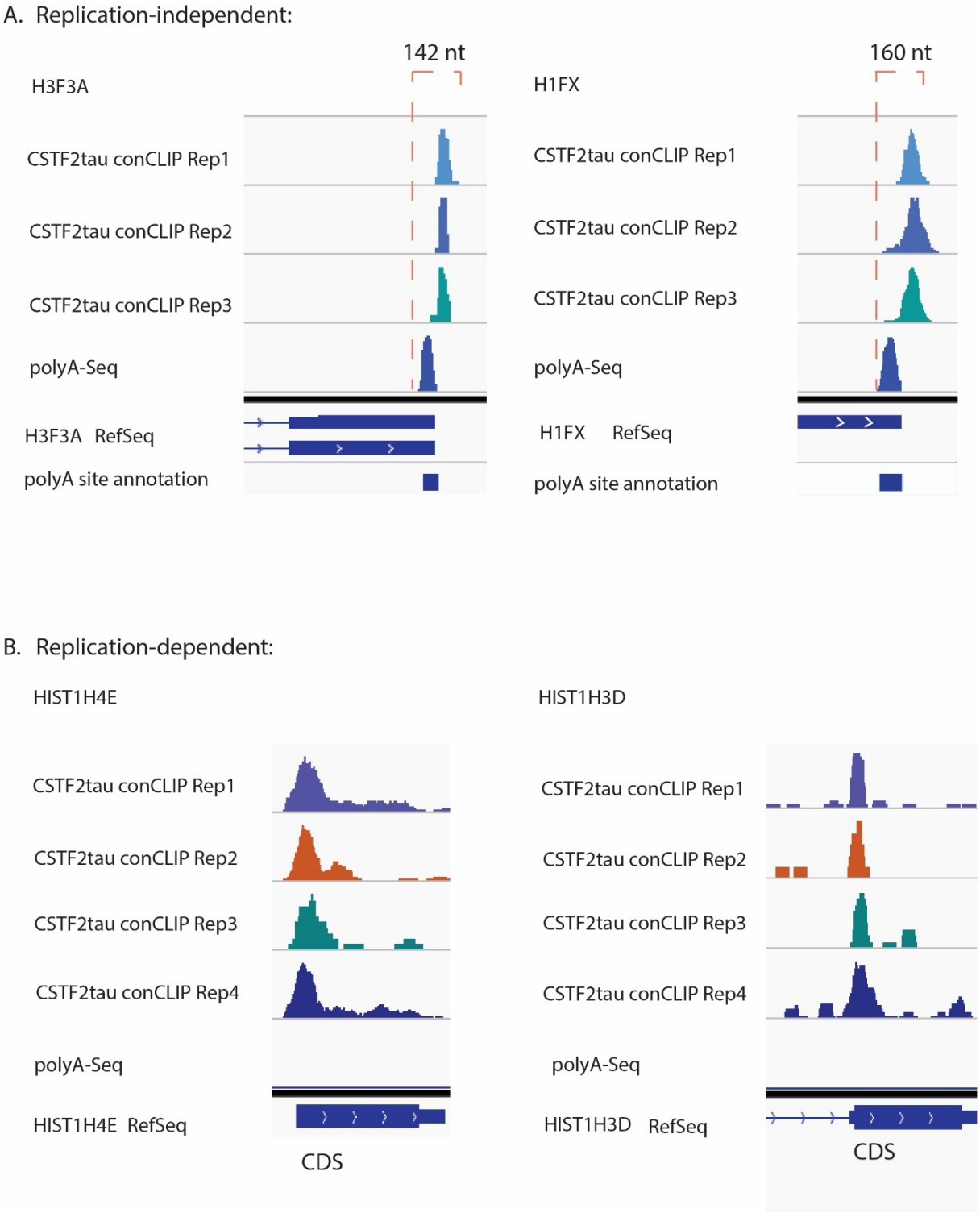


Figure 34 - CSTF2tau binds replication-independent and replication-dependent histones. The binding occurs at the 3' end of replication-independent histones (A) and at the 5' end of CDS of replication-dependent histones (B).

Table 4 - Binding of CSTF2tau on replication-independent histones (3’ end binding)

SYMBOL	NAME	CHROMOSOME
H1FX	H1 histone family, member X	3q21.3
H2AFY2	H2A histone family, member Y2	10q22.1
H3F3A	H3 histone, family 3A	1q42.12
H2AFX	H2 histone, family 3B (H3.3B)	11q23.3

Table 5 - Binding of CSTF2tau on replication-dependent histones (5’-end binding)

SYMBOL	NAME	CHROMOSOME
HIST1H1AE	H1 Histone Family, Member 1	6p22.2
HIST2H2BE	H2B Histone Family, Member Q	1q21.2
HIST1H4C	H4 Histone Family, Member G	6p22.2
HIST2H2BC	Histone 2, H2bc	1q21.2
HIST1H2AM	2A Histone Family, Member N	6p22.1
HIST1H4B	H4 Histone Family, Member I	6p22.2
HIST1H2BD	H2B Histone Family, Member B	6p22.2
HIST1H4E	H4 Histone Family, Member J	6p22.2
HIST1H3D	H3 Histone Family, Member B	6p22.2

Apart from previously reported interactions of CSTF2tau with transcripts of protein-coding genes, we also observed numerous binding sites on non-coding transcripts. About 10% of genes bound by the protein belong to the category “non-coding genes” (Figure 35 A). The majority of non-coding transcripts bound by CSTF2tau protein belong to antisense (63), lincRNA (67) and pseudogene (36) type of transcripts (Figure 35 B).

Hypergeometric analysis was next carried out to discover RNA types (Figure 36), which are over- and under- represented in the cohort of CSTF2tau targets. Strikingly, transcripts of protein-coding genes were under-represented in the targets list, whereas some categories of non-coding genes, for example sense intronic, linc RNAs and antisense RNAs were over-represented (Figure 36). This observation suggests that the CSTF2tau protein binds protein-coding genes selectively. Over-represented binding of the CSTF2tau protein on non-coding RNA types was not reported before and suggests that the protein possibly possesses novel, yet not identified functions.

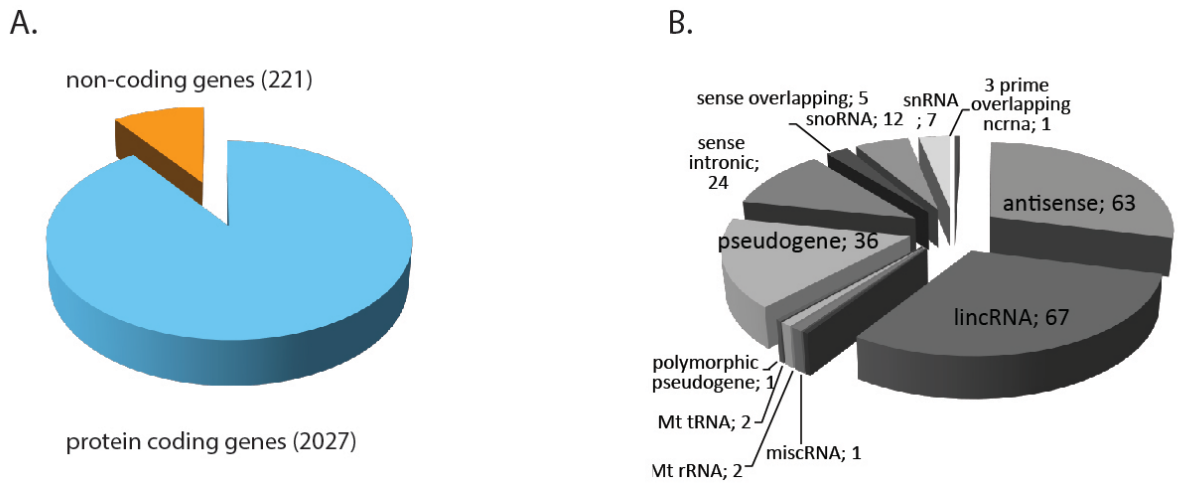


Figure 35 - Genome-wide distribution of CSTF2tau binding sites. CSTF2tau protein binds coding (90%) and non-coding (10%) genes (A). Distribution of binding sites over the non-coding type of genes (B).

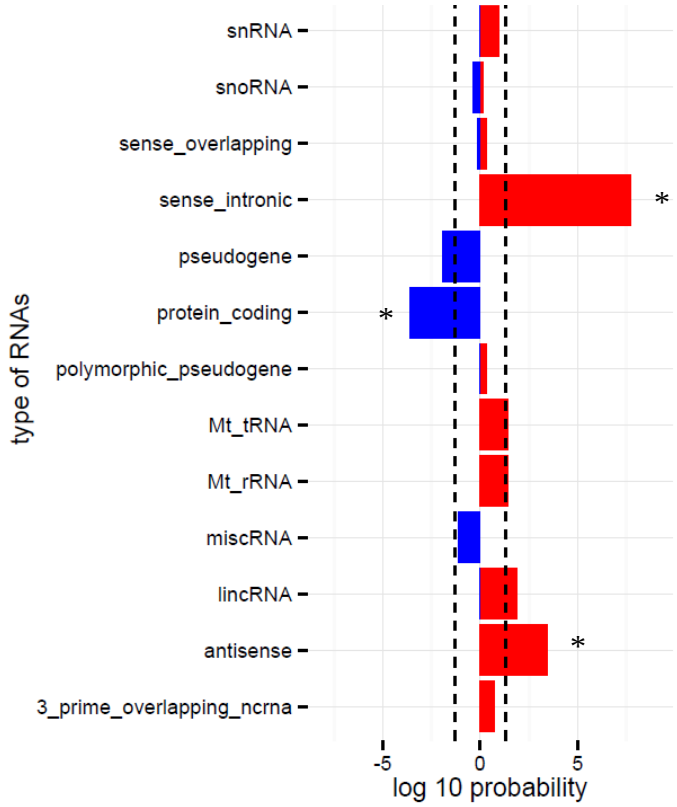


Figure 36 - Hypergeometric test reveals significant over-representation of antisense and sense intronic non-coding and linc RNAs within the cohort of CSTF2tau bound genes. Dashed line labels probability below 5%. Red bars indicate over representation; blue bars show under representation.

When looking at individual examples of non-coding genes bound by the protein, I observed that the CSTF2tau binds Small Cajal body-specific RNAs (SCARNAs). SCARNAs

are a class of small nucleolar RNAs (snoRNAs) that are localized specifically in the Cajal body, an organelle involved in biogenesis of snRNPs. Among SCARNAs SCARNA7 and SCARNA9 were found. These molecules guide the modification of RNA polymerase II transcribed spliceosomal RNAs U1 and U2 [161].

Interestingly, among other non-coding RNA molecules U11, and U12 as well as U4, U4ATAC and U5 were found to be bound by CSTF2tau. The protein recognizes the 3' region of these molecules, as exemplified in the Figure 37 for U11, U5, U12 and U4 snRNAs.

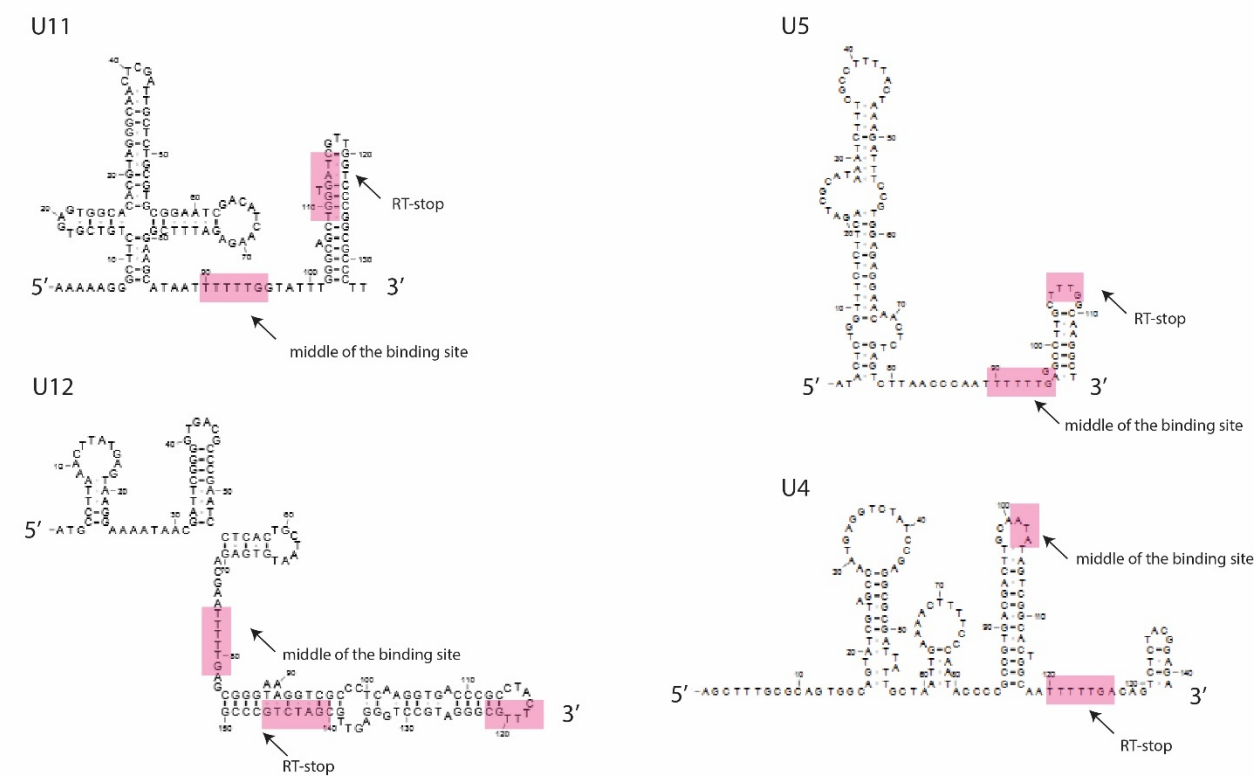


Figure 37 - CSTF2tau binding sites are located at the 3' end of snRNAs U11, U5, U12 and U4. The red color marks the high density regions of the protein binding, arrow bars point to start and stop of high density regions as well as to the probable transcription stop (RT-stop, where majority of reads are terminated).

The strength of crosslinking and immunoprecipitation protocols lies in deep and multi-layer target identification as shown here. Yet, it does not prove the functionality of the protein for the identified targets. The first priority would be therefore to subdivide the target list into direct interacting partners, whose abundance changes upon depletion of this protein.

3.3.2 Functional analysis of the effect of CSTF2tau depletion on gene regulation

To further address the functional importance of CSTF2tau protein for gene expression, we performed RNA-sequencing experiment in the presence and absence of the protein. BE(2)-C cells were transiently transfected with a pool of 4 siRNAs targeting CSTF2tau. The depletion efficiency was confirmed by western blot analysis (Figure 38). We also analyzed the protein abundance change of CSTF2, the paralog of CSTF2tau. As Figure 38 illustrates, CSTF2 was not noticeably regulated upon CSTF2tau knockdown in neuroblastoma cells, whereas depletion of CSTF2 led to tremendously increased levels of CSTF2tau (Figure 38, compare 2 and 3 or 4). It has been speculated before that the CSTF2tau and its paralog possess similar functions, and can functionally substitute each other [44]. As there is no change in the level of CSTF2tau upon the depletion of its paralog, we do not expect compensatory effects on the 3'end processing and downstream fate of RNAs bound by CSTF2tau.

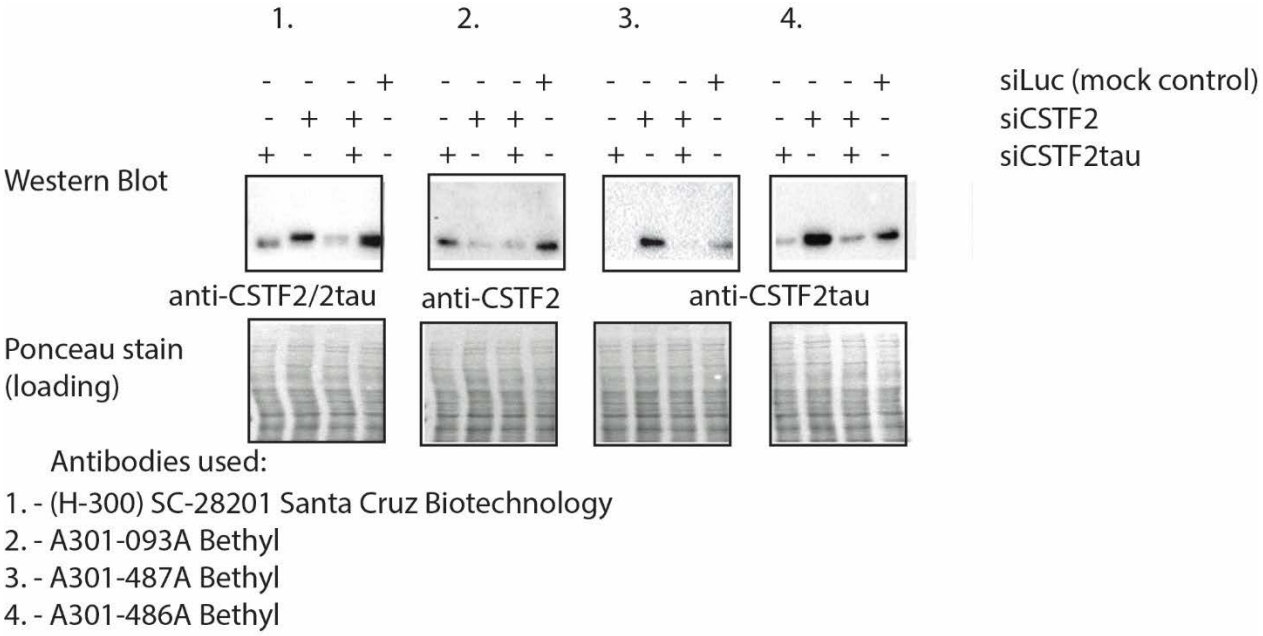


Figure 38 - CSTF2tau knockdown efficiency assessed by western blot. Three different antibodies were used to visualize isoform-specific knockdowns.

On average 80 million reads per sample were retrieved by RNA sequencing of the respective sample. The relative similarity between replicates was assessed by multidimensional scaling (MDS) plot (Figure 39). The MDS plot shows that the biological replicates are highly consistent between each other (gene expression from replicate to

replicate differs by 10%). Of note, different treatments are well separated from each other, suggesting that one can expect to find many differentially expressed genes (Figure 39).

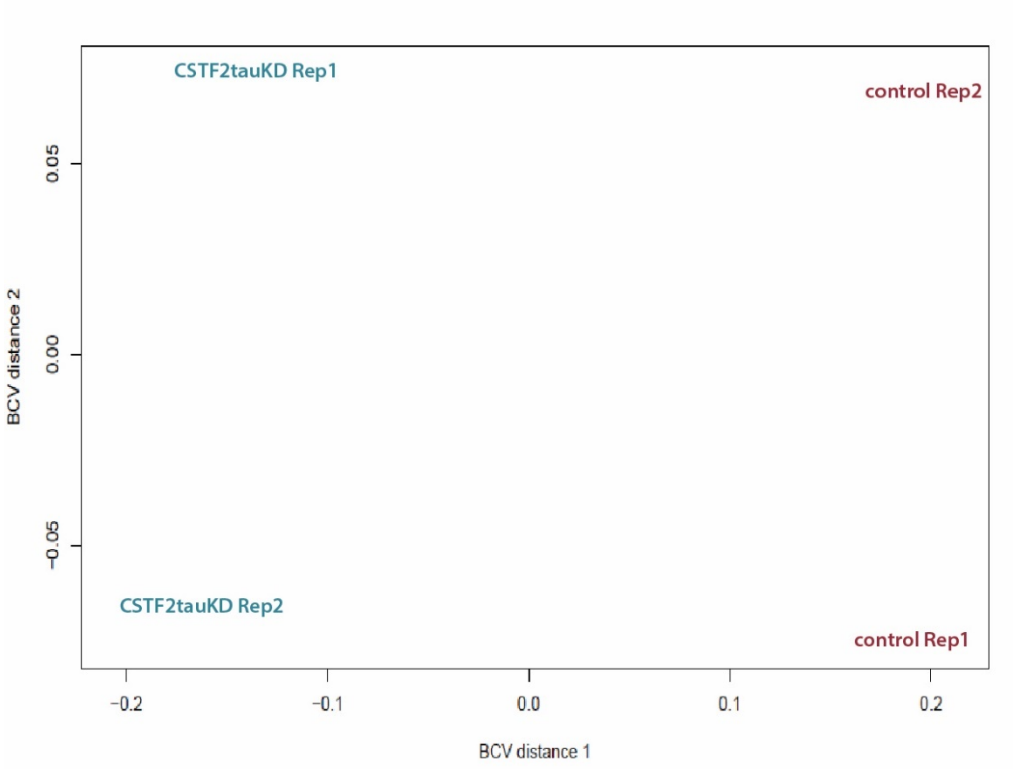


Figure 39 - MDS plot represents relative similarities between control and CSTF2tau knockdown samples. Distances between samples correspond to the biological coefficient of variation between each pair of samples.

Indeed, further analysis of differentially expressed genes revealed over 2000 genes to be differentially regulated. Figure 40 shows a volcano plot (A) and a MA plot (B), in which differentially regulated genes are depicted as red dots.

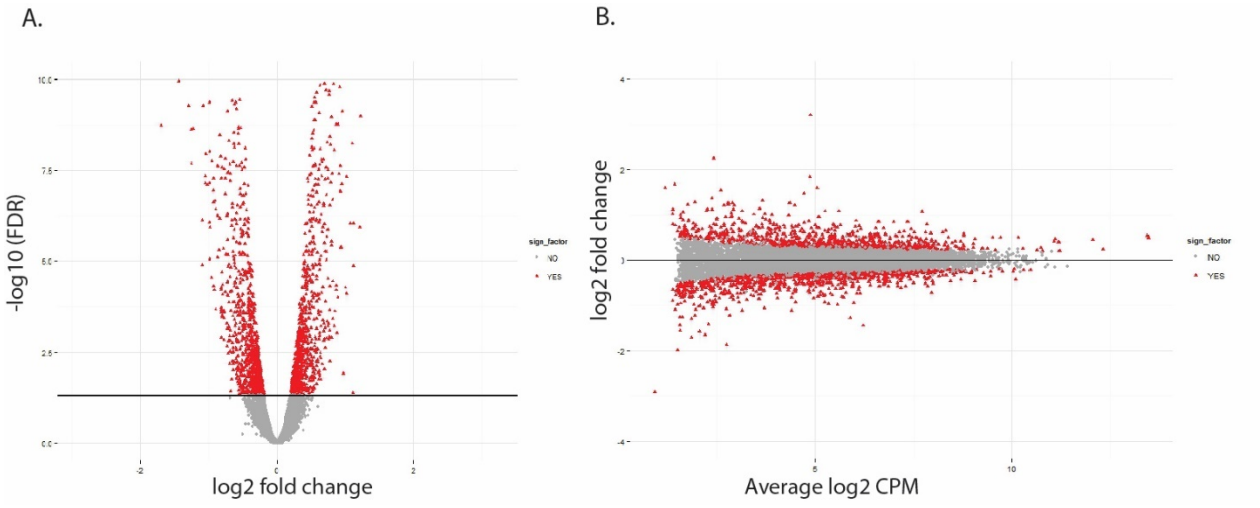


Figure 40 - Genes differentially expressed upon CSTF2tau depletion. The volcano plot represents differentially expressed genes log2 fold change is plotted vs. –log10 FDR (A). The MA-plot represents differentially expressed genes, average log 2 counts per million reads plotted against log2 fold change (B).

Interestingly, there was no obvious general directionality of change in a steady state RNA level; RNAs were both decreased and increased in abundance (p-val 0.05). Only 215 genes were highly regulated (>1.5 fold). Interestingly the steady-state mRNA expression of the majority of genes (88%) with high degree of regulation were upregulated in CSTF2tau depleted cells (Table 6).

Table 6 - Statistics of differential expression analysis upon CSTF2tau knockdown in BE(2)-C cells

Total # of analyzed genes	12039
Number of significantly regulated genes (FDR<0,05)	2099
Number of genes significantly upregulated upon depletion of CSTF2tau	1010
Number of genes significantly downregulated upon depletion of CSTF2tau	1089
Number of not regulated genes	9940
Number of genes downregulated upon depletion by >1.5 fold	26
Number of genes upregulated upon depletion by > 1.5 fold	189

Next, the list of genes, whose expression was changed by at least 25% upon CSTF2tau depletion, was analyzed using Protein ANalysis THrough Evolutionary Relationships (PANTHER) classification system, which allows to classify genes and reveal relationships between them [162]. Interestingly, the analysis of over represented signaling pathways, present in the cohort of genes up- or downregulated upon CSTF2tau knockdown, revealed a

significant enrichment of up-regulated genes belonging to Heterotrimeric G-protein signaling pathway and a significant enrichment of down-regulated genes belonging to Wnt signaling pathway. Thus, a high proportion of Wnt signaling pathway components are downregulated and a high proportion of Heterotrimeric G-protein signaling pathway components are up-regulated upon CSTF2tau depletion (Table 7).

Table 7 - PANTHER pathway analysis

PANTHER Pathways	# expressed genes	# observed genes	expected	Fold enrichment	P-value (Bonferroni corrected)	Directionality of gene regulation
Heterotrimeric G-protein signaling pathway-Gq alpha and Go alpha mediated pathway	77	14	7.66	2.8	1.46E-02	Up
Wnt signaling pathway	29	12	2.88	4.16	5.78E-03	Down

The hypergeometric analysis revealed that only a minority of protein-coding genes were highly regulated upon CSTF2tau depletion. In contrast, the non-coding RNAs, especially sn- and snoRNAs were highly regulated (Figure 41).

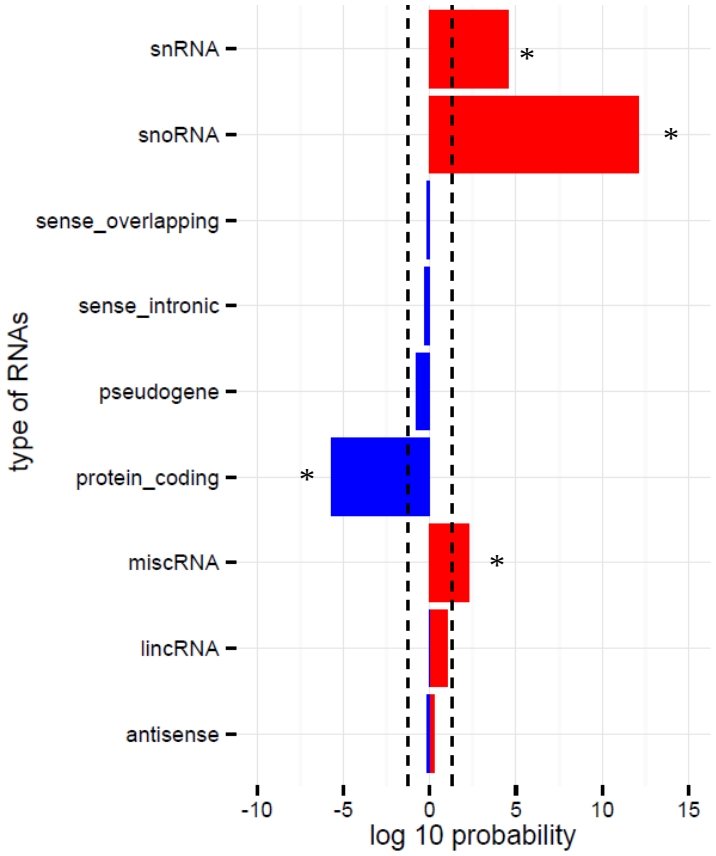


Figure 41 - Hypergeometric test reveals significant over-representation of snRNA and snoRNAs within the cohort of genes differentially expressed upon CSTF2tau depletion. Dashed line labels probability below 5%. Red bars indicate over representation; blue bars show under representation.

Strikingly, the genes belonging to sn- and snoRNA type were exclusively up-regulated even when a less stringent threshold of regulation was applied (Table 8). If the regulation were unbiased, the expected number of downregulated sn/snoRNA genes should be around 24, instead no gene is found. The probability of bias observed here is very low (1.398881e-14), suggesting that the directional regulation of snRNA and snoRNA gene type upon CSTF2tau knockdown is caused by the depletion of the protein and does not occur by chance.

Table 8 - Analysis of directionality of regulation of sn/snoRNA genes in comparison with other gene types.

	Positively regulated	Negatively Regulated
Sn/snoRNAs	56	0
Other Genes	575	463

In order to identify transcripts, which are directly bound by CSTF2tau and regulated, I related the target list of binding with the list of differentially expressed genes. Strikingly, the protein-coding genes are under-represented in the overlapping cohort (Figure 42). This finding suggests that there is no direct effect of CSTF2tau depletion on the abundance of its mRNA targets. This finding is in line with an analysis of CSTF2tau protein knockout in mouse testis cells, which did not support the direct effect of CSTF2tau on expression of its targets [156]. However, our analysis reveals that the snRNA genes are over-represented in the list of regulated CSTF2tau targets (Figure 42). This finding allows speculating that CSTF2tau binding might have a direct effect on the abundance of the snRNAs.

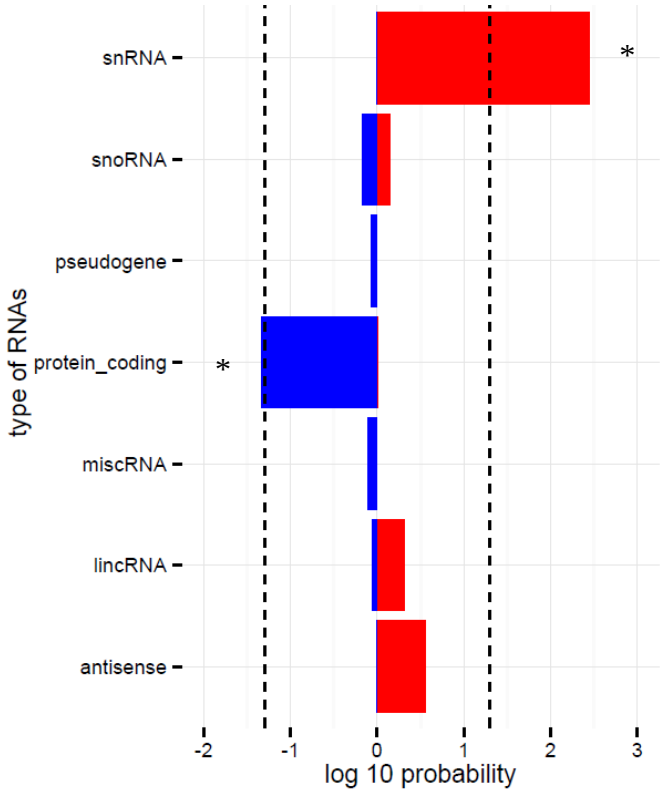


Figure 42 - Hypergeometric test reveals significant over-representation of snRNA and under-representation of protein-coding transcripts within the cohort of genes differentially expressed and bound by CSTF2tau protein. Dashed line labels the probability below 5%. Red bars indicate over representation, blue bars show under representation.

3.3.3 Human snRNAs are polyadenylated

CSTF2tau is a protein known to play an important role in mRNA 3' end cleavage and polyadenylation [41]. Based on this functionality we wondered whether the binding of

CSTF2tau to the snRNA observed here might explain the RNA abundance change via differential 3’ end processing / polyadenylation. Currently numerous methods based on NGS sequencing are used to address the question of APA of mature mRNA [41, 163, 164]. Whether small RNAs are polyadenylated and to which extent this occurs, remains uncovered. Due to the peculiarities of modern sequencing techniques, namely size exclusion, small molecules such as snRNAs are normally lost. For example, the TAIL-seq protocol [15], which aims to discover the length and nucleotide composition of polyA tails also loses snRNAs because of the size selection step. Yet there is evidently a small fraction of snRNAs which can be polyadenylated. For example, U2 snRNA was shown to be polyadenylated in yeast upon introducing point mutations into CFIA complex protein RNA15 [62].

To answer the question whether the human U-type snRNAs are polyadenylated an ePAT assay was carried out. The ePAT method is designed to measure the length of poly(A) tail and was successfully applied on polyadenylated mRNA molecules [131]. In the context of this thesis, ePAT was applied to reveal the degree of polyadenylation of snRNA molecules. Upon ePAT PCR, the products were blunt-end cloned into a pJET vector and sequenced using a Sanger sequencing approach. ePAT confirmed that U1, U2 (not shown), U11, U4atac, U4, U5 and U12 (Figure 43) snRNAs population contain polyadenylated fraction. For U4atac, U4 and U12 the length of the poly(A) tail was rather short (13 nts), but the occurrence of the tail could not be explained by internal priming. In contrast, U5 (Figure 43), U1 and U2 (not shown) snRNAs contain longer poly(A) tails.

U4ATAC

```
U4ATAC   ACCATCCTTTTCTTGGGGTTGCGCTACTGTCCAATGAGCGCATAGTGAGGGCAGTACTGCTAACGCCTGAACAACACACCCGCATCAACTAGAGCTTTTGCTTTATTTTGGTGCAATTTTTGGAAAA---------
Clone1   ACCATCCTTTTCTTGGGGTTGCGCTACTGTCCAATGAGCGCATAGTGAGGGCAGTACTGCTAACGCCTGAACAACACACCCGCATCAACTAGAGCTTTTGCTTTATTTTGGTGCAATTTTTGGAAAAAAAAAAAAA
Clone2   ACCATCCTTTTCTTGGGGTTGCGCTACTGTCCAATGAGCGCATAGTGAGGGCAGTACTGCTAACGCCTGAACAACACACCCGCATCAACTAGAGCTTTTGCTTTATTTTGGTGCAATTTTTGGAAAAAAAAAAAAA
clone3   ACCATCCTTTTCTTGGGGTTGCGCTACTGTCCAATGAGCGCATAGTGAGGGCAGTACTGCTAACGCCTGAACAACACACCCGCATCAACTAGAGCTTTTGCTTTATTTTGGTGCAATTTTTGGAAAAAAAAAAAA
Clone4   ACCATCCTTTTCTTGGGGTTGCGCTACTGTCCAATGAGCGCATAGTGAGGGCAGTACTGCTAACGCCTGAACAACACACCCGCATCAACTAGAGCTTTTGCTTTATTTTGGTGCAATTTTTGGAAAAAAAAAAAA
```

U11

```
U11      AAAAAGGGCTTTGTCGTGAGTGGCACACGTAGGGCAACTCGATTGCTCTGCGTGCGGAATCGACATCAAGAGATTTCGGAAGCATAATTTTTTGGTATTTGGGCAGCTGGTGATCGTTGGTCCCGGCGCC
Clone1   AAAAAGGGCTTTGTCGTGAGTGGCACACGTAGGGCAACTCGATTGCTCTGCGTGCGGAATCGACATCAAGAGATTTCGGAAGCATAATTTTTTGGTATTTGGGCAGCTGGTGATCGTTAAAAAAAAAAAAA
Clone2   AAAAAGGGCTTTGTCGTGAGTGGCACACGTAGGGCAACTCGATTGCTCTGCGTGCGGAATCGACATCAAGAGATTTCGGAAGCATAATTTTTTGGTATTTGGGCAGCTGGTGATCGTTGAAAAAAAAAAAAA
Clone3   AAAAAGGGCTTTGTCGTGAGTGGCACACGTAGGGCAACTCGATTGCTCTGCGTGCGGAATCGACATCAAGAGATTTCGGAAGCATAATTTTTTGGTATTTGGGCAGCTGGTGATCGTTGAAAAAAAAAAAA
Clone4   AAAAAGGGCTTTGTCGTGAGTGGCACACGTAGGGCAACTCGATTGCTCTGCGTGCGGAATCGACATCAAGAGATTTCGGAAGCATAATTTTTTGGTATTTGGGCAGCTGGTGATCGTTGAAAAAAAAAAA(A)137
```

U12

```
U12      ATGCCTTAAACTTATGAGTAAGGAAAATAACGATTCGGGGTGACGCCCGAATCCTCACTGCTAATGTGAGACGAATTTTTGAGCGGGTAAAGGTCGCCCTCAAGGTGACCCGCCTACTTTGCGGGATGCCTGGGAGTTGCGATCTGCCG-----------
Clone    ATGCCTTAAACTTATGAGTAAGGAAAATAACGA-TCGGGGTGACGCCCGAATCCTCACTGCTAATGTGAGACGAATTTTTGAGCGGGTAAAGGTCGCCCTCAAGGTGACCCGCCTACTTTGCGGGATGCCTGGGAGTTGCGATCTGTAAAAAAAAAAAA
```

U4-2

```
U4-2     AGCTTTGCGCAGTGGCAGTATCGTAGCCAATGAGGTTTATCCGAGGCGCGATTATTGCTAATTGAAAACTTTTCCCAATACCCCGCCATGACGACTTGAAATATAGTCGGCATTGGCAATTTTTGACAGTCTCTACGGAGA
Clone1   AGCTTTGCGCAGTGGCAGTATCGTAGCCAATGAGGTCTATCCGAGGCGCGATTATTGCTAATTGAAAACAAAAAAAAAAAAA------------------------------------------------------------
Clone2   AGCTTTGCGCAGTGGCAGTATCGTAGCCAATGAGGTTTATCCGATGCGCGATTATTGCTAATTGAAAAAAAAAAAAAAA--------------------------------------------------------------
Clone3   AGCTTTGCGCAGTGGCAGTATCGTAGCCAATGAGGTTTATCCGAGGCGCGATTATTGCTAATTGAAAAAAAAAAAAAAA--------------------------------------------------------------
Clone4   AGCTTTGCGCAGTGGCAGTATCGTAGCCAATGAGGTTTATCCGAGGCGCGATTATTGCTAATTGAAAAAAAAAAAAAAA--------------------------------------------------------------
Clone5   AGCTTTGCGCAGTGGCAGTATCGTAGCCAATGAGGTTTATCCGAGGCGCGATTATTGCTAATTGAAAAAAAAAAAAAAA--------------------------------------------------------------
Clone6   AGCTTTGCGCAGTGGCAGTATCGTAGCCAATGAGGTTTATCCGAGGCGCGATTATTGCTAATTGAAAAAAAAAAAAAAAAAAAA---------------------------------------------------------
Clone7   AGCTTTGCGCAGTGGCAGTATCGTAGCCAATGAGGTTTATCCGAGGCGCGATTATTGCTAATTGAAAAAAAAAAAAA----------------------------------------------------------------
Clone8   AGCTTTGCGCAGTGGCAGTATCGTAGCCAATGAGGTTTATCCGAGGCGCGATTATTGCTAATTGAAAAAAAAAAAA-----------------------------------------------------------------
Clone9   AGCTTTGCGCAGTGGCAGTATCGTAGCCAATGAGGTTTATCCGAGGCGCGATTATTGCTAATTGAAAAAAAAAAAAAAAAAAAA---------------------------------------------------------
```

U5

```
U5A-1    ATACTCTGGTTTCTCTTCAGATCGCATAAATCTTTCGCCTTTTACTAAAGATTTCCGTGGAGAGGAACAACTCTGAGTCTTAACCCAATTTTTTGAGGCCTTGCTTTGGCAAGGCTGGTT
Clone1   ATACTCTGGTTTCTCTTCAGATCGCATAAATCTTTCGCCTTTTACTAAAGATTTCCGTGGAGAGGAACAAAAAAAAAAAAA----------------------------------------
Clone2   ATACTCTGGTTTCTCTTCAGATCGCATAAATCTTTCGCCTTTTACTAAAGATTTCCGTGGAGAGGAACAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
Clone3   ATACTCTGGTTTCTCTTCAGATCGCATAAATCTTTCGCCTTTTACTAAAGATTTCCGTGGAGAGGAACAACTCTGAAAAAAAAAAAAAA--------------------------------
Clone4   ATACTCTGGTTTCTCTTCAGATCGCATAAATCTTTCGCCTTTAAAAAAAAAAAAA-----------------------------------------------------------------
Clone5   ATACTCTGGTTTCTCTTCAGATCGCATAAATCTTTCGCCTTTTACTAAAGAAAAAAAAAAAAA--------------------------------------------------------
Clone6   ATACTCTGGTTTCTCTTCAGATCGCATAAATCTTTCGCCTTTTACTAAAGATTTCCATGGAGAGGAACAAAAAAAAAAAAAAAAAAAA--------------------------------
```

Figure 43 - Human snRNAs contain polyadenylated fraction. U4atac, U11, U4-2 and U5A-1 cDNA sequence retrieved from ENSEMBL database is aligned to sequences generated by ePAT approach. Poly(A) tails of different length were observed (in analogy also observed for U1 and U2).

As detected by ePAT and sequencing, the U5, U11, U1, U4 and U2 polyadenylated molecules are shorter than the annotated cDNA and thus these molecules might be trimmed before being polyadenylated. The binding of the protein thus occurs downstream of the polyadenylated region and apparently takes place before trimming (Figure 44).

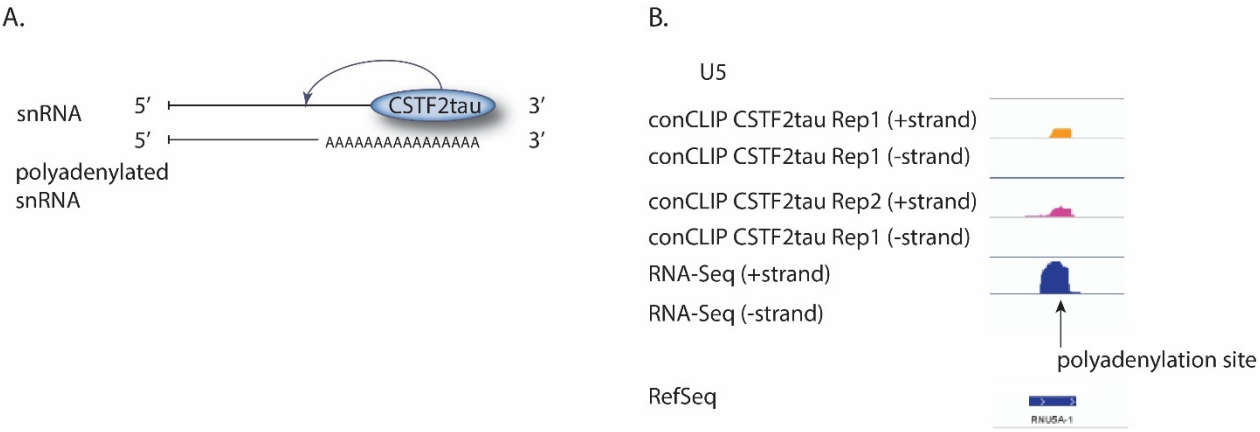


Figure 44 – Illustration of the binding of CSTF2tau occuring at the 3' end of the snRNAs or further downstream (A). Binding of CSTF2tau occurs downstream of the internal cleavage site of U5 snRNA (B).

As described above, the RNA-abundance of snRNAs such as U4 and U5 changed more than 1.5 fold upon CSTF2tau depletion. Of note, this gene expression measurement is based on random priming and thus both polyadenylated as well as non-polyadenylated RNA fractions are determined. In the context of this work, it was found that a fraction of snRNAs is polyadenylated. The next question of interest was whether the amount of the polyadenylated fraction of snRNAs changes upon CSTF2tau depletion. To address this question, CSTF2tau was depleted by single siRNA. The successful depletion was confirmed by western blot, in which less than 25% of protein was observed after 48 h of siRNA treatment (data not shown). Next, the RNA was extracted and the cDNA was synthesized. To catch both fractions, non-polyadenylated (majority) and polyadenylated (minority), the cDNA was synthesized either with the help of a miRNA reverse transcription kit (Qiagen) or after oligo (dT) priming and reverse transcription. Next, a pair of primers was used to assess the quantities of snRNAs present in both cases. The quantitative PCR revealed that the depletion of CSTF2tau protein leads to significant downregulation of polyadenylated snRNAs U11, U1 and U5 (Figure 45).

Previously it has been observed that oligoadenylation of snoRNAs initiates their degradation [78]. We therefore analyzed next whether the inhibition of polyadenylation upon CSTF2tau depletion affects the RNA decay of the U-type snRNAs studied here.

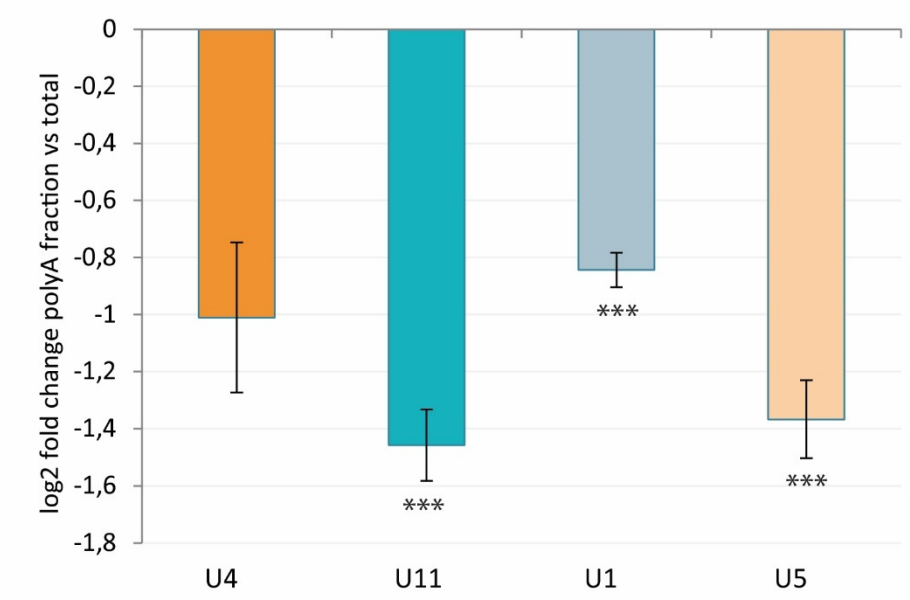


Figure 45 - Abundance change of polyadenylated fraction of snRNAs U4, U11, U1 and U5 observed upon CSTF2tau depletion (*pval<0.05).

3.3.4. The depletion of CSTF2tau protein effects the stability of snRNAs

To address whether the depletion of CSTF2tau effects the expression of snRNAs or their stability, actinomycin D (ActD) - RNA decay experiments were performed. Actinomycin D blocks activity of RNAPII, which transcribes majority of snRNAs. The cells were treated with siRNA against CSTF2tau for 48 hours and by ActD for 0, 2, 4 and 6 hours respectively. For each time point RNA was harvested and transcribed using Qiagen total RNA kit. Next the levels of U4, U5, U11 and U1 were assesed by quantitative PCR. The levels of RNAs were normalized by a housekeeping gene (GAPDH). Comparison of stability of 2 housekeeping genes against each other (ACTB and GAPDH) did not reveal signifficant difference between their levels upon depletion of the CSTF2tau (data not shown). Figure 46 illustrates that the relative stability of a proto-oncogene Jun does not differ between mock control and CSTF2tau knockdown samples, whereas both variants of U4, U1 as well as U5 snRNAs are faster degraded in the control sample in comparison to the CSTF2tau depleted sample (Figure 46).

This observation confirms our hypothesis that the decrease of the polyadenylated snRNAs fraction increases the stability of the studied snRNAs.

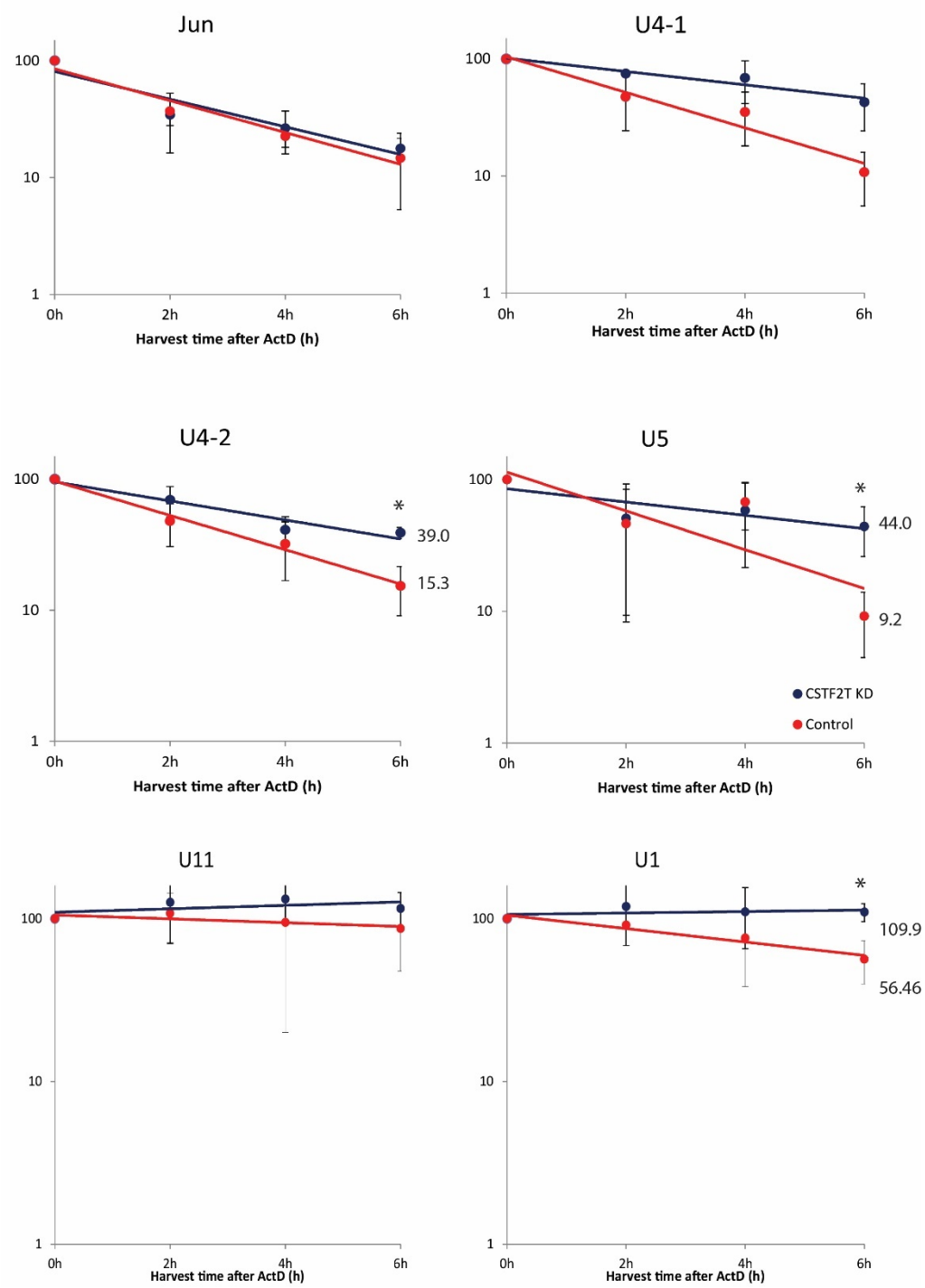


Figure 46 - Elevated levels of snRNAs are observed due to increased stability of the molecules upon CSTF2tau knockdown (* pval<0.05). The stability of proto-oncogene c-Jun is not changed.

To conclude, the analysis of gene expression upon CSTF2tau knockdown revealed a high proportion of significantly regulated genes, at the same time only 215 genes were regulated more than by 1.5 fold (Table 6). Interestingly, the protein-coding genes are not the majorly regulated group. Moreover, the protein-coding genes being both bound by the CSTF2tau and regulated upon depletion are under-represented (Figure 42). This finding suggests that the observed regulation of protein-coding genes is caused by a cascade of secondary (indirect) effects.

On the other hand, non-coding RNAs, in particular snoRNAs and snRNAs, are over-represented within the group of highly regulated genes (Figure 42). Further analysis revealed that the snRNAs bound by the protein are also regulated upon depletion. This finding points toward direct effects of the CSTF2tau depletion on snRNAs abundance.

Interestingly, the sn- and snoRNAs are exclusively upregulated upon CSTF2tau depletion. As CSTF2tau is involved in 3'end processing of coding RNAs, I propose that the mechanism, by which the amount of snRNAs is regulated, might be due to alternative processing. Recently it has been reported that snRNA U2 can be alternatively processed and be a source of small (19-22 nt) fragments [165]. In the context of this work, I show that human snRNAs contain a polyadenylated fraction. Importantly, the abundance of polyadenylated fraction declines uponCSTF2tau knockdown. It has been previously reported that in yeast the processing of snRNAs is mediated by the general cleavage and polyadenylation machinery. Although the cleavage of snRNAs is uncoupled from polyadenylation, still a small proportion of snRNAs are polyadenylated. In humans, snRNAs are processed by the Integrator complex and have not been detected to carry poly(A) tails. Yet the current study provides evidence that a fraction of human snRNAs contain short oligo(A) tails. Previously it has been reported that mammalian snoRNAs intermediates undergo several cycles of oligoadenylation/deadenylation, favouring the deadenylated state [78]. The oligoadenylation triggers more rapid decay of the molecules. Strikingly, it has been observed in the context of this study that, upon the depletion of CSTF2tau, the amount of polyadenylated fraction of snRNAs decreases, whereas the stability of snRNAs U4, U1 and U5 increases.

Discussion

The aim of my thesis was to improve currently available methods to study RNA-protein interactions via crosslinking and immunoprecipitation (CLIP). The current work describes a modified CLIP approach, named as “conCLIP”. This approach addresses and overcomes several limiting steps of previously described CLIP versions, such as HITS-CLIP, PAR-CLIP and iCLIP. The features of the methods are summarized in a table below. Firstly, current CLIP protocols rely on radioactive labeling of RNA in order to prove the success of the covalent crosslinking and specific co-immunoprecipitation of the RNA along with the protein [113, 121, 151, 166]. The conCLIP method, established in the context of this work uses labeling of RNA with biotin and thus avoids radioactivity. This acknowledges the increasingly widespread banning of radioactivity due to safety measures in laboratories. Secondly, conCLIP utilizes a cDNA synthesis protocol optimized for small input material and previously described for single cell sequencing technique [152]. The optimized technique provides a possibility to work with lesser material (as little as 700 µg of protein) and retrieve highly reproducible results (consistency between replicates above 0.95). Additionally, conCLIP exploits the benefits of an experimental barcoding approach and thus allows multiplexing of several samples. In the context of dynamic CLIP (comparing two or more conditions), this has the advantage that all subsequent steps of library preparation are carried out in one batch, allowing to reduce technical variabilities between the experiments. Further it reduces material and sequencing costs. On the other hand, conCLIP exploits a second type of barcodes (random barcodes), which allows distinguishing and discarding amplification artefacts. Most importantly, the libraries are prepared with only 10-11 cycles of PCR amplification, which are 15 cycles less than in the iCLIP protocol [132]. The conCLIP library protocol does not require size selection of the cDNA, as adaptor contamination of conCLIP libraries is virtually impossible. Therefore loss of material on the step of size selection cannot happen, thereby preventing a low complexity of the library due to artificial multiplication of single reads. The computational pipeline applied for conCLIP relies on previous knowledge on CLIP-Seq analysis [167]. It implements the high-speed mapping algorithm (STAR), which significantly shortens the time needed for the completion of the analysis [168].

Table 9 - Comparison of current CLIP variants with the conCLIP method

Method	CLIP [114, 167, 169]	HITS-CLIP [153]	PAR-CLIP [121]	iCLIP [132]	FAST-iCLIP [155]	conCLIP (present protocol)
Number of PCR cycles	25-35	25-35	16-25	20-30	20-30	**<=10**
Number of RNA-ligations	2	2	2	1	1	1
Model (cell culture/organs/tissues	All	All	Cell culture/only proliferating cells	All	All	All
Protein origin (endogenous/exogenous)	Both	Both	Both	Both?	Both?	Endogenous w/o problem and on low input material
Reproducibility	NA	NA	80-96%	NA	NA	**>95%**
Individual molecules recognition	No	No	No	Yes	Yes	Yes
Capturing of premature terminated reverse transcription events	No	No	No	Yes	Yes	Yes
Amount of input material	1-2 mg	NA	NA	NA	NA	**700 µg of protein**
Size selection of cDNA to avoid primer template products	crucial	crucial	crucial	crucial	crucial	**Not required**
Radioactivity	Yes	Yes	Yes	Yes	Yes	**No**

In my thesis, I prove that conCLIP is a reliable and quantitative approach (Figure 21, 23, 30, 34). At first, I applied the established conCLIP method on CSTF2tau protein. The method successfully recapitulates previously described binding preferences of the protein. ConCLIP confirms that the CSTF2tau protein binds 3' UTRs most prevalently. The results of the conCLIP confirm that CSTF2tau binds downstream of the cleavage and polyadenylation sites of mRNAs (Figure 24). In line with previously published data, I show that the binding of CSTF2tau downstream of polyadenylation site predicts the usage of the site by cleavage and polyadenylation machinery (Table 3). Yet, in contrast to previously published results, the

percentage of predicted sites is lower here (20-25% vs 60-64% described before [41]). This discrepancy can possibly be explained by (1) the use of a different cell line (here BE(2)-C cells vs HeLa). These two cell lines are of different origin and thus have different expression profiles. (2) It is worth considering, that the studied protein has a paralog CSTF2, expressed in both cell lines. Earlier it has been shown that both proteins recognize a similar set of transcripts, possess same motif preferences and are possibly redundant in their functionality [44]. Discrepancies in (3) read depth and (4) analysis may account for different percentage of predicted sites.

The future potential of CLIP type techniques is the application for studies of dynamic RNA-RBP interactions. In this context, I tested if dynamic CSTF2tau conCLIP is predictive for differential poly(A) site choice, observed by polyA-sequencing upon knockdown of PCF11 (a key modulator controlling PAS choice, Danckwardt lab unpublished, [160]). The analysis of dynamic conCLIPs of CSTF2tau protein revealed distinct patterns of protein binding between two conditions. I detected that the overlap between dynamically recognized CSTF2tau sites and alternative poly(A) isoforms upon PCF11 depletion is rather low. Yet the fold change of regulation of dynamically recognized CSTF2tau sites and dynamically changed APA sites showed high degree of positive correlation. This discrepancy might be explained by the fact, that not all APA-regulated sites are recognized by a CSTF2tau-containing 3' end processing complex. Alternatively, this observation reflects that the shift in polyadenylation site usage is not solely caused by changes in processing, but also by a difference in stability rates of the transcript isoforms [170]. Finally, reading depth may explain this discrepancy.

In the final part of this thesis complex and multi-layer data generated by conCLIP were used to study yet poorly described binding partners of CSTF2tau. In the context of this work, the binding of CSTF2tau on histones is being described (Figure 34). Applying conCLIP technique, I observed that the protein recognizes not only replication-independent histones, which are polyadenylated by the canonical cleavage and polyadenylation machinery, but also replication-dependent (RD) histones. In contrast to replication-independent histones, the processing of RD histones is uncoupled from polyadenylation and is carried out by a unique processing complex [158]. This complex recruits the components of the canonical

polyadenylation machinery, for example CPSF73, which performs the nucleolytic cleavage [50]. It has been reported that such a complex also contains the CSTF2 protein, the paralog of CSTF2tau [47]. In the context of this study, I reveal that in neuroblastoma cells the CSTF2tau protein is bound to RD histones and can also be a part of the cleavage complex. Of particular interest is the location of peaks on RD histones; they are located at the 5'end of the cDNA of the RD histone. Possibly such a localization can be explained by the spatial organization of the histone mRNAs, which brings the 3' and 5' end of the molecule together. Additionally, the conCLIP reveals non-coding RNAs being targets of CSTF2tau (Figure 35). The protein binding sites are over-represented on sense intronic, long intervening and antisense noncoding RNAs.

CLIP techniques are capable to identify thousands of binding sites, yet binding does not necessarily reflect function. To reveal which binding events reflect functionality, I compared the conCLIP data with total RNA-seq data after depletion of CSTF2tau. Although the CSTF2tau protein is believed to contribute to cleavage and polyadenylation site choice, the depletion of CSTF2tau has a relatively small effect on APA [44, 157]. As APA isoforms may differ by their stability rates [30], it is expected that the APA switch is reflected on the steady-state levels of RNAs. The RNA sequencing data reveals more than 2000 genes, whose steady-state levels are significantly changed upon the depletion of CSTF2tau (Table 6). Importantly, a few genes show high level of regulation, whereas the majority of changes are small. When sub-selecting genes regulated over a 50% threshold, the regulation becomes directional with the majority of transcripts showing increased levels of steady-state RNA upon depletion of CSTF2tau.

In line with previous reports revealing small effects on APA after CSTF2tau depletion [44, 157], the proportion of protein-coding genes recognized by the protein and exhibiting different steady-state RNA levels is low. In contrast, the non-coding transcripts bound by CSTF2tau, in particular transcripts belonging to the snRNA category, are regulated at high degree (Figure 42). To summarize, the effect of CSTF2tau depletion on protein-coding genes is relatively small and can only be partially explained by the binding pattern. The high number mildly regulated steady-state levels of coding RNAs are most probably caused by a cascade of secondary indirect effects of CSTF2tau depletion. In contrast to the modest effect

on coding genes, CSTF2tau depletion strongly affects the steady-state levels of small non-coding RNAs. Interestingly, a big fraction of those regulated non-coding RNAs is also bound by the CSTF2tau.

As mentioned above, CSTF2tau contributes to 3' end processing of pre-mRNAs. Its function on small non-coding RNAs has not been reported before. As revealed by conCLIP, the binding of CSTF2tau on snRNAs occurs at the 3' end of cDNAs or further downstream. Binding of the CSTF2tau paralog to the snRNA promoter sequences has been reported before [171]. On the other hand, recently it has been found that the other small non-coding RNAs, snoRNAs, can be oligoadenylated [78]. In yeast, the 3'end formation and maturation of snRNAs is accomplished by various pathways, some of them rely on cleavage and polyadenylation. For example, U1 is processed by endonuclease Rnt1, which recognizes the stem loop structure, and ultimately the transcript is further trimmed through the activity of exosome [68]. In contrast, U2 snRNA utilizes the polyadenylation sites located downstream of the stem loop structure to mediate cleavage and maturation. If the Rnt1 cleavage site within the stem loop structure is mutated, longer molecules with a polyadenylation tail accumulate [62]. It was therefore logical to reveal whether the snRNAs are polyadenylated and whether CSTF2tau plays a role in this process. In the context of my thesis it was, for the first time, reported that the depletion of CSTF2tau protein leads to regulation of small nuclear RNAs (Figure 41). Moreover, the steady-state RNA levels of snRNAs were exclusively upregulated upon the depletion of CSTF2tau protein (Table 8). Further I could demonstrate that the fraction of snRNAs is oligoadenylated. Interestingly, the levels of oligoadenylated snRNAs compared to total snRNA levels decreased upon depletion of CSTF2tau (Figure 45).

The intriguing property of a poly(A) tail to determine mRNA longevity on the one hand and a rapid decay on the other, has been emphasized in the introduction (section 1.4). RNAs can be "stably" polyadenylated, and thus stabilized. Oppositely, oligoadenylation of molecules can trigger their degradation [88]. The observed A tails attached to the snRNAs were relatively short, with an average length of 18 nucleotides (Figure 43). Considering the possibility that in this case the oligoadenylated snRNAs are faster degraded, I evaluated the relative stability rates of snRNAs upon transcription actinomycin D inhibition. Interestingly,

CSTF2tau depletion leads to increased stability of snRNAs, as assessed upon transcription halt (Figure 46).

Of particular interest is the position of oligo(A) tails attached to the snRNAs. The oligoadenylated snRNAs, in particular U1, U2, U4 and U5 are truncated and polyadenylated "internally". Previously, the truncated U1 snRNA (U1-tfs) molecules have been described [172]. U1-tfs were lacking the Sm site and unable to form the Sm heptamer [172]. Interestingly, the last nucleotide of U1-tfs is exactly the same, where I detected the oligoadenylated tail attached. Moreover U1-tfs have been shown to be more rapidly degraded and localized primarily to P-bodies [172]. Recently the alternative processing of U2 snRNAs has been described [165]. Yet, the short fragments identified by Mazieres and co-authors correspond to the 3' end products of endonucleolytic cleavage [165]. Interestingly, the position of the 5' most nucleotide is exactly the position of the 3' most nucleotide where the U2 molecule is oligoadenylated, as detected here.

As described above, the binding of the CSTF2tau protein occurs at the 3' end of the snRNAs or further downstream. When taking into consideration the truncation of oligoadenylated snRNAs, it becomes apparent that the protein binds downstream of the endonucleolytic cleavage site. The observation that the depletion of the CSTF2tau protein leads to decrement of polyadenylated fraction suggests that the binding of the protein on the 3'ends of snRNAs promotes polyadenylation. Simultaneously, the stability of total snRNAs increases. This is in line with previous reports that oligoadenylated RNAs are substrates for fast decay [173, 174]. I am thus proposing a model, according to which CSTF2tau promotes the oligoadenylation of snRNAs resulting in a faster degradation of the affected molecules and lowering the levels of total snRNAs (Figure 47). Upon depletion of the CSTF2tau, the proportion of oligoadenylated molecules decreases, leading to a higher stability of the snRNAs and elevated steady-state mRNA levels. Of note, the oligoadenylated snRNAs, observed in the context of this work are shorter than the processed snRNAs.

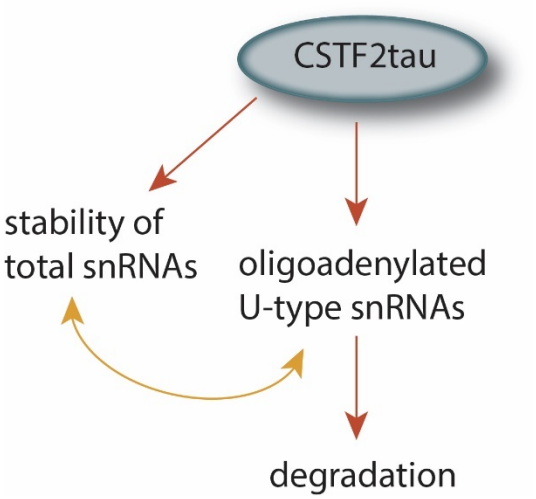


Figure 47 - Proposed model of regulation of oligoadenylation of snRNAs upon CSTF2tau depletion and resulting stability of snRNAs.

My thesis addresses two topics. On one hand it provides the scientific community with a robust, reliable and straight-forward method to explore RNA-protein interactions. This method can be applied to study RNA-protein interactions in different physiological and pathophysiological conditions. On the other hand, it addresses a so far poorly studied topic of snRNA decay mechanism, whereby the level of snRNAs can be fine-tuned. It is possible that the other components of cleavage and polyadenylation machinery also recognize snRNA genes and contribute to their oligoadenylation. Along the same lines it would be interesting to further study, which poly(A) polymerase is involved in oligoadenylation and which decay components are responsible for the decay of oligoadenylated snRNAs.

Another implication of my work is to exploit the conCLIP method to study steady-state and dynamic interactions of other RNA-binding proteins. Recently, a huge number of RNA-binding proteins, which possess enzymatic activities (so called moonlighting proteins [175]) have been described [96, 99, 122, 129]. It is therefore interesting to address the RNA specificities of such proteins, as well as to explore their binding repertoire in health and disease, as exemplified by work of Beckmann and co-authors [129].

There has been numerous reports published, aiming to explain the selection of cleavage and polyadenylation site usage in different cell types and upon various conditions in health and disease [159]. Yet many aspects of regulation remain uncovered, as discussed by

Shi and Manley [38]. The study of dynamic binding of components of cleavage and polyadenylation machinery upon depletion of the parts of this machinery might shed light onto the competitive binding and the regulation of APA upon abundance change of the core regulators. Alternatively, studies addressing dynamic binding, resulting from the differential post-transcriptional modifications or mutations in a RRM motif, may utilize the conCLIP approach, established here.

Literature

1. Consortium, E.P., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.

2. Matlin, A.J., F. Clark, and C.W. Smith, *Understanding alternative splicing: towards a cellular code.* Nat Rev Mol Cell Biol, 2005. **6**(5): p. 386-98.

3. de Klerk, E. and P.A. t Hoen, *Alternative mRNA transcription, processing, and translation: insights from RNA sequencing.* Trends Genet, 2015. **31**(3): p. 128-39.

4. Zhai, L.T. and S. Xiang, *mRNA quality control at the 5' end.* J Zhejiang Univ Sci B, 2014. **15**(5): p. 438-43.

5. Shuman, S., *Capping enzyme in eukaryotic mRNA synthesis.* Prog Nucleic Acid Res Mol Biol, 1995. **50**: p. 101-29.

6. Furuichi, Y. and A.J. Shatkin, *Viral and cellular mRNA capping: past and prospects.* Adv Virus Res, 2000. **55**: p. 135-84.

7. Coller, J. and R. Parker, *Eukaryotic mRNA decapping.* Annu Rev Biochem, 2004. **73**: p. 861-90.

8. Wahl, M.C., C.L. Will, and R. Luhrmann, *The spliceosome: design principles of a dynamic RNP machine.* Cell, 2009. **136**(4): p. 701-18.

9. Levine, A. and R. Durbin, *A computational scan for U12-dependent introns in the human genome sequence.* Nucleic Acids Res, 2001. **29**(19): p. 4006-13.

10. Zarnack, K., et al., *Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements.* Cell, 2013. **152**(3): p. 453-66.

11. Nilsen, T.W. and B.R. Graveley, *Expansion of the eukaryotic proteome by alternative splicing.* Nature, 2010. **463**(7280): p. 457-63.

12. Merkin, J., et al., *Evolutionary dynamics of gene and isoform regulation in Mammalian tissues.* Science, 2012. **338**(6114): p. 1593-9.

13. Proudfoot, N.J., *Ending the message: poly(A) signals then and now.* Genes Dev, 2011. **25**(17): p. 1770-82.

14. Shi, Y., et al., *Molecular architecture of the human pre-mRNA 3' processing complex.* Mol Cell, 2009. **33**(3): p. 365-76.

15. Chang, H., et al., *TAIL-seq: genome-wide determination of poly(A) tail length and 3' end modifications.* Mol Cell, 2014. **53**(6): p. 1044-52.

16. Sachs, A. and E. Wahle, *Poly(A) tail metabolism and function in eucaryotes.* J Biol Chem, 1993. **268**(31): p. 22955-8.

17. Proudfoot, N.J., A. Furger, and M.J. Dye, *Integrating mRNA processing with transcription.* Cell, 2002. **108**(4): p. 501-12.

18. Hirose, Y. and J.L. Manley, *RNA polymerase II and the integration of nuclear events.* Genes Dev, 2000. **14**(12): p. 1415-29.

19. Maniatis, T. and R. Reed, *An extensive network of coupling among gene expression machines.* Nature, 2002. **416**(6880): p. 499-506.

20. Misra, A., et al., *Global Promotion of Alternative Internal Exon Usage by mRNA 3' End Formation Factors.* Mol Cell, 2015. **58**(5): p. 819-31.

21. Langemeier, J., M. Radtke, and J. Bohne, *U1 snRNP-mediated poly(A) site suppression: beneficial and deleterious for mRNA fate.* RNA Biol, 2013. **10**(2): p. 180-4.

22. Kaida, D., et al., *U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation.* Nature, 2010. **468**(7324): p. 664-8.

23. Derti, A., et al., *A quantitative atlas of polyadenylation in five mammals.* Genome Res, 2012. **22**(6): p. 1173-83.

24. Tian, P., et al., *Tandem alternative polyadenylation events of genes in non-eosinophilic nasal polyp tissue identified by high-throughput sequencing analysis.* Int J Mol Med, 2014. **33**(6): p. 1423-30.

25. Lukiw, W.J. and N.G. Bazan, *Cyclooxygenase 2 RNA message abundance, stability, and hypervariability in sporadic Alzheimer neocortex.* J Neurosci Res, 1997. **50**(6): p. 937-45.

26. Hall-Pogar, T., et al., *Alternative polyadenylation of cyclooxygenase-2.* Nucleic Acids Res, 2005. **33**(8): p. 2565-79.

27. Alt, F.W., et al., *Synthesis of secreted and membrane-bound immunoglobulin mu heavy chains is directed by mRNAs that differ at their 3' ends.* Cell, 1980. **20**(2): p. 293-301.

28. Di Giammartino, D.C., Y. Shi, and J.L. Manley, *PARP1 represses PAP and inhibits polyadenylation during heat shock.* Mol Cell, 2013. **49**(1): p. 7-17.

29. Takagaki, Y., et al., *The polyadenylation factor CstF-64 regulates alternative processing of IgM heavy chain pre-mRNA during B cell differentiation.* Cell, 1996. **87**(5): p. 941-52.

30. Mayr, C. and D.P. Bartel, *Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells.* Cell, 2009. **138**(4): p. 673-84.

31. Proudfoot, N.J. and G.G. Brownlee, *3' non-coding region sequences in eukaryotic messenger RNA.* Nature, 1976. **263**(5574): p. 211-4.

32. Hu, J., et al., *Bioinformatic identification of candidate cis-regulatory elements involved in human mRNA polyadenylation.* RNA, 2005. **11**(10): p. 1485-93.

33. Shi, Y., *Alternative polyadenylation: new insights from global analyses.* RNA, 2012. **18**(12): p. 2105-17.

34. Beaudoing, E., et al., *Patterns of variant polyadenylation signal usage in human genes.* Genome Res, 2000. **10**(7): p. 1001-10.

35. Arhin, G.K., et al., *Downstream sequence elements with different affinities for the hnRNP H/H' protein influence the processing efficiency of mammalian polyadenylation signals.* Nucleic Acids Res, 2002. **30**(8): p. 1842-50.

36. Bagga, P.S., et al., *The G-rich auxiliary downstream element has distinct sequence and position requirements and mediates efficient 3' end pre-mRNA processing through a trans-acting factor.* Nucleic Acids Res, 1995. **23**(9): p. 1625-31.

37. Zarudnaya, M.I., et al., *Downstream elements of mammalian pre-mRNA polyadenylation signals: primary, secondary and higher-order structures.* Nucleic Acids Res, 2003. **31**(5): p. 1375-86.

38. Shi, Y. and J.L. Manley, *The end of the message: multiple protein-RNA interactions define the mRNA polyadenylation site.* Genes Dev, 2015. **29**(9): p. 889-97.

39. Chan, S.L., et al., *CPSF30 and Wdr33 directly bind to AAUAAA in mammalian mRNA 3' processing.* Genes Dev, 2014. **28**(21): p. 2370-80.

40. Schonemann, L., et al., *Reconstitution of CPSF active in polyadenylation: recognition of the polyadenylation signal by WDR33.* Genes Dev, 2014. **28**(21): p. 2381-93.

41. Martin, G., et al., *Genome-wide analysis of pre-mRNA 3' end processing reveals a decisive role of human cleavage factor I in the regulation of 3' UTR length.* Cell Rep, 2012. **1**(6): p. 753-63.

42. Lackford, B., et al., *Fip1 regulates mRNA alternative polyadenylation to promote stem cell self-renewal.* EMBO J, 2014.

43. Gruber, A.R., et al., *Cleavage factor Im is a key regulator of 3' UTR length.* RNA Biol, 2012. **9**(12): p. 1405-12.

44. Yao, C., et al., *Overlapping and distinct functions of CstF64 and CstF64tau in mammalian mRNA 3' processing.* RNA, 2013. **19**(12): p. 1781-90.

45. de Vries, H., et al., *Human pre-mRNA cleavage factor II(m) contains homologs of yeast proteins and bridges two other cleavage factors.* EMBO J, 2000. **19**(21): p. 5895-904.

46. Di Giammartino, D.C., et al., *RBBP6 isoforms regulate the human polyadenylation machinery and modulate expression of mRNAs with AU-rich 3' UTRs.* Genes Dev, 2014. **28**(20): p. 2248-60.

47. Yang, X.C., et al., *FLASH, a proapoptotic protein involved in activation of caspase-8, is essential for 3' end processing of histone pre-mRNAs.* Mol Cell, 2009. **36**(2): p. 267-78.

48. Narita, T., et al., *NELF interacts with CBC and participates in 3' end processing of replication-dependent histone mRNAs.* Mol Cell, 2007. **26**(3): p. 349-65.

49. Schaufele, F., et al., *Compensatory mutations suggest that base-pairing with a small nuclear RNA is required to form the 3' end of H3 messenger RNA.* Nature, 1986. **323**(6091): p. 777-81.

50. Yang, X.C., et al., *A complex containing the CPSF73 endonuclease and other polyadenylation factors associates with U7 snRNP and is recruited to histone pre-mRNA for 3'-end processing.* Mol Cell Biol, 2013. **33**(1): p. 28-37.

51. Kolev, N.G., et al., *Conserved motifs in both CPSF73 and CPSF100 are required to assemble the active endonuclease for histone mRNA 3'-end maturation.* EMBO Rep, 2008. **9**(10): p. 1013-8.

52. Friend, K., A.F. Lovejoy, and J.A. Steitz, *U2 snRNP binds intronless histone pre-mRNAs to facilitate U7-snRNP-dependent 3' end formation.* Mol Cell, 2007. **28**(2): p. 240-52.

53. Youngblood, B.A., P.N. Grozdanov, and C.C. MacDonald, *CstF-64 supports pluripotency and regulates cell cycle progression in embryonic stem cells through histone 3' end processing.* Nucleic Acids Res, 2014.

54. Szczepinska, T., et al., *DIS3 shapes the RNA polymerase II transcriptome in humans by degrading a variety of unwanted transcripts.* Genome Res, 2015. **25**(11): p. 1622-33.

55. Kishore, S., et al., *Insights into snoRNA biogenesis and processing from PAR-CLIP of snoRNA core proteins and small RNA sequencing.* Genome Biol, 2013. **14**(5): p. R45.

56. Zhang, Y., L. Yang, and L.L. Chen, *Life without A tail: new formats of long noncoding RNAs.* Int J Biochem Cell Biol, 2014. **54**: p. 338-49.

57. Okamura, M., H. Inose, and S. Masuda, *RNA Export through the NPC in Eukaryotes.* Genes (Basel), 2015. **6**(1): p. 124-49.

58. Peart, N., et al., *Non-mRNA 3' end formation: how the other half lives.* Wiley Interdiscip Rev RNA, 2013. **4**(5): p. 491-506.

59. Chen, J. and E.J. Wagner, *snRNA 3' end formation: the dawn of the Integrator complex.* Biochem Soc Trans, 2010. **38**(4): p. 1082-7.

60. Ciliberto, G., et al., *Formation of the 3' end on U snRNAs requires at least three sequence elements.* EMBO J, 1986. **5**(11): p. 2931-7.

61. Hernandez, N. and A.M. Weiner, *Formation of the 3' end of U1 snRNA requires compatible snRNA promoter elements.* Cell, 1986. **47**(2): p. 249-58.

62. Morlando, M., et al., *Functional analysis of yeast snoRNA and snRNA 3'-end formation mediated by uncoupling of cleavage and polyadenylation.* Mol Cell Biol, 2002. **22**(5): p. 1379-89.

63. Shchepachev, V., et al., *Human Mpn1 promotes post-transcriptional processing and stability of U6atac.* FEBS Lett, 2015. **589**(18): p. 2417-23.

64. Tani, T. and Y. Ohshima, *mRNA-type introns in U6 small nuclear RNA genes: implications for the catalysis in pre-mRNA splicing.* Genes Dev, 1991. **5**(6): p. 1022-31.

65. Mroczek, S. and A. Dziembowski, *U6 RNA biogenesis and disease association.* Wiley Interdiscip Rev RNA, 2013. **4**(5): p. 581-92.

66. Albrecht, T.R. and E.J. Wagner, *snRNA 3' end formation requires heterodimeric association of integrator subunits.* Mol Cell Biol, 2012. **32**(6): p. 1112-23.

67. Baillat, D. and E.J. Wagner, *Integrator: surprisingly diverse functions in gene expression.* Trends Biochem Sci, 2015. **40**(5): p. 257-64.

68. Allmang, C., et al., *Functions of the exosome in rRNA, snoRNA and snRNA synthesis.* EMBO J, 1999. **18**(19): p. 5399-410.

69. Kugel, J.F. and J.A. Goodrich, *In new company: U1 snRNA associates with TAF15.* EMBO Rep, 2009. **10**(5): p. 454-6.

70. Crea, F., et al., *Integrated analysis of the prostate cancer small-nucleolar transcriptome reveals SNORA55 as a driver of prostate cancer progression.* Mol Oncol, 2016. **10**(5): p. 693-703.

71. Dieci, G., M. Preti, and B. Montanini, *Eukaryotic snoRNAs: a paradigm for gene expression flexibility.* Genomics, 2009. **94**(2): p. 83-8.

72. Kishore, S. and S. Stamm, *The snoRNA HBII-52 regulates alternative splicing of the serotonin receptor 2C.* Science, 2006. **311**(5758): p. 230-2.

73. Petfalski, E., et al., *Processing of the precursors to small nucleolar RNAs and rRNAs requires common components.* Mol Cell Biol, 1998. **18**(3): p. 1181-9.

74. Carneiro, T., et al., *Depletion of the yeast nuclear exosome subunit Rrp6 results in accumulation of polyadenylated RNAs in a discrete domain within the nucleolus.* Mol Cell Biol, 2007. **27**(11): p. 4157-65.

75. Grzechnik, P. and J. Kufel, *Polyadenylation linked to transcription termination directs the processing of snoRNA precursors in yeast.* Mol Cell, 2008. **32**(2): p. 247-58.

76. Falaleeva, M., et al., *Dual function of C/D box small nucleolar RNAs in rRNA modification and alternative pre-mRNA splicing.* Proc Natl Acad Sci U S A, 2016.

77. Kishore, S., et al., *The snoRNA MBII-52 (SNORD 115) is processed into smaller RNAs and regulates alternative splicing.* Hum Mol Genet, 2010. **19**(7): p. 1153-64.

78. Berndt, H., et al., *Maturation of mammalian H/ACA box snoRNAs: PAPD5-dependent adenylation and PARN-dependent trimming.* RNA, 2012. **18**(5): p. 958-72.

79. Cook, A., et al., *Structural biology of nucleocytoplasmic transport.* Annu Rev Biochem, 2007. **76**: p. 647-71.

80. Herold, A., et al., *TAP (NXF1) belongs to a multigene family of putative RNA export factors with a conserved modular architecture.* Mol Cell Biol, 2000. **20**(23): p. 8996-9008.

81. Katahira, J., et al., *Adaptor Aly and co-adaptor Thoc5 function in the Tap-p15-mediated nuclear export of HSP70 mRNA.* EMBO J, 2009. **28**(5): p. 556-67.

82. Leung, E. and J.D. Brown, *Biogenesis of the signal recognition particle.* Biochem Soc Trans, 2010. **38**(4): p. 1093-8.

83. Ohno, M., et al., *PHAX, a mediator of U snRNA nuclear export whose activity is regulated by phosphorylation.* Cell, 2000. **101**(2): p. 187-198.

84. McCloskey, A., et al., *hnRNP C tetramer measures RNA length to classify RNA polymerase II transcripts for export.* Science, 2012. **335**(6076): p. 1643-6.

85. Bjork, P. and L. Wieslander, *Nucleocytoplasmic mRNP export is an integral part of mRNP biogenesis.* Chromosoma, 2011. **120**(1): p. 23-38.

86. Dolken, L., et al., *High-resolution gene expression profiling for simultaneous kinetic parameter analysis of RNA synthesis and decay.* RNA, 2008. **14**(9): p. 1959-72.

87. Ghosh, S. and A. Jacobson, *RNA decay modulates gene expression and controls its fidelity.* Wiley Interdiscip Rev RNA, 2010. **1**(3): p. 351-61.

88. Slomovic, S., et al., *Addition of poly(A) and poly(A)-rich tails during RNA degradation in the cytoplasm of human cells.* Proc Natl Acad Sci U S A, 2010. **107**(16): p. 7407-12.

89. Jacobson, A., *Regulation of mRNA decay: decapping goes solo.* Mol Cell, 2004. **15**(1): p. 1-2.

90. Hoefig, K.P. and V. Heissmeyer, *Degradation of oligouridylated histone mRNAs: see UUUUU and goodbye.* Wiley Interdiscip Rev RNA, 2014. **5**(4): p. 577-89.

91. Smith, J.E. and K.E. Baker, *Nonsense-mediated RNA decay--a switch and dial for regulating gene expression.* Bioessays, 2015. **37**(6): p. 612-23.

92. Bremer, K.A., A. Stevens, and D.R. Schoenberg, *An endonuclease activity similar to Xenopus PMR1 catalyzes the degradation of normal and nonsense-containing human beta-globin mRNA in erythroid cells.* RNA, 2003. **9**(9): p. 1157-67.

93. Eulalio, A., et al., *Deadenylation is a widespread effect of miRNA regulation.* RNA, 2009. **15**(1): p. 21-32.

94. Khabar, K.S., *Hallmarks of cancer and AU-rich elements.* Wiley Interdiscip Rev RNA, 2016.

95. Lukong, K.E., et al., *RNA-binding proteins in human genetic disease.* Trends Genet, 2008. **24**(8): p. 416-25.

96. Castello, A., et al., *Insights into RNA Biology from an Atlas of Mammalian mRNA-Binding Proteins.* Cell, 2012. **149**(6): p. 1393-1406.

97. Baltz, Alexander G., et al., *The mRNA-Bound Proteome and Its Global Occupancy Profile on Protein-Coding Transcripts.* Molecular Cell, 2012. **46**(5): p. 674-690.

98. Hentze, M.W. and T. Preiss, *The REM phase of gene regulation.* Trends Biochem Sci, 2010. **35**(8): p. 423-6.

99. Castello, A., et al., *RNA-binding proteins in Mendelian disease.* Trends Genet, 2013. **29**(5): p. 318-27.

100. Elden, A.C., et al., *Ataxin-2 intermediate-length polyglutamine expansions are associated with increased risk for ALS.* Nature, 2010. **466**(7310): p. 1069-75.

101. Buratti, E. and F.E. Baralle, *TDP-43: gumming up neurons through protein-protein and protein-RNA interactions.* Trends Biochem Sci, 2012. **37**(6): p. 237-47.

102. Lagier-Tourenne, C. and D.W. Cleveland, *Rethinking ALS: the FUS about TDP-43.* Cell, 2009. **136**(6): p. 1001-4.

103. Strong, M.J. and K. Volkening, *TDP-43 and FUS/TLS: sending a complex message about messenger RNA in amyotrophic lateral sclerosis?* FEBS J, 2011. **278**(19): p. 3569-77.

104. Honda, D., et al., *The ALS/FTLD-related RNA-binding proteins TDP-43 and FUS have common downstream RNA targets in cortical neurons.* FEBS Open Bio, 2013. **4**: p. 1-10.

105. Fujioka, Y., et al., *FUS-regulated region- and cell-type-specific transcriptome is associated with cell selectivity in ALS/FTLD.* Sci Rep, 2013. **3**: p. 2388.

106. Kim, K.Y., et al., *Significance of molecular markers in survival prediction of oral squamous cell carcinoma.* Head Neck, 2012. **34**(7): p. 929-36.

107. Denkert, C., et al., *Expression of the ELAV-like protein HuR in human colon cancer: association with tumor stage and cyclooxygenase-2.* Mod Pathol, 2006. **19**(9): p. 1261-9.

108. Mrena, J., et al., *Cyclooxygenase-2 is an independent prognostic factor in gastric cancer and its expression is regulated by the messenger RNA stability factor HuR.* Clin Cancer Res, 2005. **11**(20): p. 7362-8.

109. Wang, J., et al., *The expression of RNA-binding protein HuR in non-small cell lung cancer correlates with vascular endothelial growth factor-C expression and lymph node metastasis.* Oncology, 2009. **76**(6): p. 420-9.

110. Wang, J., et al., *Cytoplasmic HuR expression correlates with angiogenesis, lymphangiogenesis, and poor outcome in lung cancer.* Med Oncol, 2011. **28 Suppl 1**: p. S577-85.

111. Denkert, C., et al., *Expression of the ELAV-like protein HuR is associated with higher tumor grade and increased cyclooxygenase-2 expression in human breast carcinoma.* Clin Cancer Res, 2004. **10**(16): p. 5580-6.

112. Lopez de Silanes, I., A. Lal, and M. Gorospe, *HuR: post-transcriptional paths to malignancy.* RNA Biol, 2005. **2**(1): p. 11-3.

113. Darnell, R., *CLIP (cross-linking and immunoprecipitation) identification of RNAs bound by a specific protein.* Cold Spring Harbor protocols, 2012. **2012**: p. 1146-60.

114. Ule, J., et al., *CLIP: a method for identifying protein-RNA interaction sites in living cells.* Methods (San Diego, Calif.), 2005. **37**: p. 376-86.

115. Ule, J., et al., *CLIP identifies Nova-regulated RNA networks in the brain.* Science, 2003. **302**(5648): p. 1212-5.

116. Musunuru, K. and R.B. Darnell, *Paraneoplastic neurologic disease antigens: RNA-binding proteins and signaling proteins in neuronal degeneration.* Annu Rev Neurosci, 2001. **24**: p. 239-62.

117. Darnell, R.B. and J.B. Posner, *Paraneoplastic syndromes involving the nervous system.* N Engl J Med, 2003. **349**(16): p. 1543-54.

118. Darnell, R.B., *Developing global insight into RNA regulation.* Cold Spring Harb Symp Quant Biol, 2006. **71**: p. 321-7.

119. Meisenheimer, K.M. and T.H. Koch, *Photocross-linking of nucleic acids to associated proteins.* Crit Rev Biochem Mol Biol, 1997. **32**(2): p. 101-40.

120. Spitzer, J., et al., *PAR-CLIP (Photoactivatable Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation): a step-by-step protocol to the transcriptome-wide identification of binding sites of RNA-binding proteins.* Methods Enzymol, 2014. **539**: p. 113-61.

121. Ascano, M., et al., *Identification of RNA-protein interaction networks using PAR-CLIP.* Wiley Interdiscip Rev RNA, 2012. **3**(2): p. 159-77.

122. Kwon, S.C., et al., *The RNA-binding protein repertoire of embryonic stem cells.* Nat Struct Mol Biol, 2013. **20**(9): p. 1122-1130.

123. Konig, J., et al., *Protein-RNA interactions: new genomic technologies and perspectives.* Nat Rev Genet, 2011. **13**(2): p. 77-83.

124. Hauer, C., et al., *Improved binding site assignment by high-resolution mapping of RNA-protein interactions using iCLIP.* Nat Commun, 2015. **6**: p. 7921.

125. Wang, Z., et al., *iCLIP predicts the dual splicing effects of TIA-RNA interactions.* PLoS biology, 2010. **8**: p. e1000530.

126. Bohnsack, M.T., et al., *Prp43 bound at different sites on the pre-rRNA performs distinct functions in ribosome synthesis.* Mol Cell, 2009. **36**(4): p. 583-92.

127. Wang, T., Y. Xie, and G. Xiao, *dCLIP: a computational approach for comparative CLIP-seq analyses.* Genome Biol, 2014. **15**(1): p. R11.

128. Tollervey, J.R., et al., *Characterizing the RNA targets and position-dependent splicing regulation by TDP-43.* Nat Neurosci, 2011. **14**(4): p. 452-8.

129. Beckmann, B.M., et al., *The RNA-binding proteomes from yeast to man harbour conserved enigmRBPs.* Nat Commun, 2015. **6**: p. 10127.

130. Rauschenberger, K., et al., *A non-enzymatic function of 17beta-hydroxysteroid dehydrogenase type 10 is required for mitochondrial integrity and cell survival.* EMBO Mol Med, 2010. **2**(2): p. 51-62.

131. Janicke, A., et al., *ePAT: a simple method to tag adenylated RNA to measure poly(A)-tail length and other 3' RACE applications.* RNA, 2012. **18**(6): p. 1289-95.

132. Huppertz, I., et al., *iCLIP: protein-RNA interactions at nucleotide resolution.* Methods, 2014. **65**(3): p. 274-87.

133. Bolger, A.M., M. Lohse, and B. Usadel, *Trimmomatic: a flexible trimmer for Illumina sequence data.* Bioinformatics, 2014. **30**(15): p. 2114-20.

134. Andrews, S. *FastQC A Quality Control tool for High Throughput Sequence Data*. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

135. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-9.

136. Neph, S., et al., *BEDOPS: high-performance genomic feature operations.* Bioinformatics, 2012. **28**(14): p. 1919-20.

137. Anders, S., P.T. Pyl, and W. Huber, *HTSeq--a Python framework to work with high-throughput sequencing data.* Bioinformatics, 2015. **31**(2): p. 166-9.

138. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features.* Bioinformatics, 2010. **26**(6): p. 841-2.

139. Lovci, M.T., et al., *Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges.* Nat Struct Mol Biol, 2013. **20**(12): p. 1434-42.

140. Heinz, S., et al., *Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities.* Mol Cell, 2010. **38**(4): p. 576-89.

141. Thorvaldsdottir, H., J.T. Robinson, and J.P. Mesirov, *Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration.* Brief Bioinform, 2013. **14**(2): p. 178-92.

142. Robinson, J.T., et al., *Integrative genomics viewer.* Nat Biotechnol, 2011. **29**(1): p. 24-6.

143. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 2010. **26**(1): p. 139-40.

144. McCarthy, D.J., Y. Chen, and G.K. Smyth, *Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation.* Nucleic Acids Res, 2012. **40**(10): p. 4288-97.

145. Anders, S., A. Reyes, and W. Huber, *Detecting differential usage of exons from RNA-seq data.* Genome Res, 2012. **22**(10): p. 2008-17.

146. Aznarez, I., et al., *A systematic analysis of intronic sequences downstream of 5' splice sites reveals a widespread role for U-rich motifs and TIA1/TIAL1 proteins in alternative splicing regulation.* Genome Res, 2008. **18**(8): p. 1247-58.

147. Le Guiner, C., et al., *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.* J Biol Chem, 2001. **276**(44): p. 40638-46.

148. Gueydan, C., et al., *Identification of TIAR as a protein binding to the translational regulatory AU-rich element of tumor necrosis factor alpha mRNA.* J Biol Chem, 1999. **274**(4): p. 2322-6.

149. Danckwardt, S., et al., *p38 MAPK controls prothrombin expression by regulated RNA 3' end processing.* Mol Cell, 2011. **41**(3): p. 298-310.

150. Kishore, S., et al., *A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins.* Nat Methods, 2011. **8**(7): p. 559-64.

151. Ina Huppertz a, b., Jan Attig a,b, Andrea D'Ambrogio a,b, Laura E. Easton b, Christopher R. Sibley a,b, and M.T.b. Yoichiro Sugimoto b, c, Julian König a,b,d,⇑, Jernej Ule, *iCLIP: Protein–RNA interactions at nucleotide resolution.* Methods, 2013.

152. Hashimshony, T., et al., *CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification.* Cell Rep, 2012. **2**(3): p. 666-73.

153. Zhang, C. and R.B. Darnell, *Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data.* Nat Biotechnol, 2011. **29**(7): p. 607-14.

154. Islam, S., et al., *Quantitative single-cell RNA-seq with unique molecular identifiers.* Nat Methods, 2014. **11**(2): p. 163-6.

155. Flynn, R.A., et al., *Dissecting noncoding and pathogen RNA-protein interactomes.* RNA, 2014.

156. Dass, B., et al., *Loss of polyadenylation protein tauCstF-64 causes spermatogenic defects and male infertility.* Proc Natl Acad Sci U S A, 2007. **104**(51): p. 20374-9.

157. Yao, C., et al., *Transcriptome-wide analyses of CstF64-RNA interactions in global regulation of mRNA alternative polyadenylation.* Proc Natl Acad Sci U S A, 2012. **109**(46): p. 18773-8.

158. Ruepp, M.D., et al., *Interactions of CstF-64, CstF-77, and symplekin: implications on localisation and function.* Mol Biol Cell, 2011. **22**(1): p. 91-104.

159. Ogorodnikov, A., Y. Kargapolova, and S. Danckwardt, *Processing and transcriptome expansion at the mRNA 3' end in health and disease: finding the right end.* Pflugers Arch, 2016.

160. Li, W., et al., *Systematic profiling of poly(a)+ transcripts modulated by core 3' end processing and splicing factors reveals regulatory rules of alternative cleavage and polyadenylation.* PLoS Genet, 2015. **11**(4): p. e1005166.

161. Lestrade, L. and M.J. Weber, *snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs.* Nucleic Acids Res, 2006. **34**(Database issue): p. D158-62.

162. Thomas, P.D., et al., *PANTHER: a library of protein families and subfamilies indexed by function.* Genome Res, 2003. **13**(9): p. 2129-41.

163. Shepard, P.J., et al., *Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq.* RNA, 2011. **17**(4): p. 761-72.

164. Jenal, M., et al., *The poly(A)-binding protein nuclear 1 suppresses alternative cleavage and polyadenylation sites.* Cell, 2012. **149**(3): p. 538-53.

165. Mazieres, J., et al., *Alternative processing of the U2 small nuclear RNA produces a 19-22nt fragment with relevance for the detection of non-small cell lung cancer in human serum.* PLoS One, 2013. **8**(3): p. e60134.

166. Li, Q., Y. Uemura, and Y. Kawahara, *Cross-linking and immunoprecipitation of nuclear RNA-binding proteins.* Methods Mol Biol, 2015. **1262**: p. 247-63.

167. Sauliere, J. and H. Le Hir, *CLIP-seq to discover transcriptome-wide imprinting of RNA binding proteins in living cells.* Methods Mol Biol, 2015. **1296**: p. 151-60.

168. Dobin, A. and T.R. Gingeras, *Mapping RNA-seq Reads with STAR.* Curr Protoc Bioinformatics, 2015. **51**: p. 11 14 1-11 14 19.

169. Modic, M., J. Ule, and C.R. Sibley, *CLIPing the brain: studies of protein-RNA interactions important for neurodegenerative disorders.* Mol Cell Neurosci, 2013. **56**: p. 429-35.

170. West, S. and N.J. Proudfoot, *Human Pcf11 enhances degradation of RNA polymerase II-associated nascent RNA and transcriptional termination.* Nucleic Acids Res, 2008. **36**(3): p. 905-14.

171. O'Reilly, D., et al., *Human snRNA genes use polyadenylation factors to promote efficient transcription termination.* Nucleic Acids Res, 2014. **42**(1): p. 264-75.

172. Ishikawa, H., et al., *Identification of truncated forms of U1 snRNA reveals a novel RNA degradation pathway during snRNP biogenesis.* Nucleic Acids Res, 2014. **42**(4): p. 2708-24.

173. Eckmann, C.R., C. Rammelt, and E. Wahle, *Control of poly(A) tail length.* Wiley Interdiscip Rev RNA, 2011. **2**(3): p. 348-61.

174. Harnisch, C., et al., *Oligoadenylation of 3' decay intermediates promotes cytoplasmic mRNA degradation in Drosophila cells.* RNA, 2016. **22**(3): p. 428-42.

175. Huberts, D.H. and I.J. van der Klei, *Moonlighting proteins: an intriguing mode of multitasking.* Biochim Biophys Acta, 2010. **1803**(4): p. 520-5.