Dissertation zur Erlangung des Doktorgrades
der Fakultät für Chemie und Pharmazie
der Ludwig-Maximilians-Universität München

Transcriptome maps of mRNP biogenesis factors define pre-mRNA recognition

Carlo Bäjen
aus
Borna, Deutschland

2014

Erklärung

Diese Dissertation wurde im Sinne von § 7 der Promotionsordnung vom 28. November 2011 von Herrn Professor Dr. Patrick Cramer betreut.

Eidesstattliche Versicherung

Diese Dissertation wurde eigenständig und ohne unerlaubte Hilfe erarbeitet.

Göttingen, den 25.06.2014

.....
Carlo Bäjen

Dissertation eingereicht am 01.07.2014

1. Gutachterin/ 1. Gutachter: Professor Dr. Patrick Cramer

2. Gutachterin/ 2. Gutachter: PD Dr. Dietmar Martin

Mündliche Prüfung am 25.07.2014

To My Family

”The family is the first essential cell of human society.”

Pope John XXIII, 1959

Summary

Biogenesis of eukaryotic mRNAs involves pre-mRNA synthesis by RNA polymerase (Pol) II and co-transcriptional RNA processing, which encompasses 5'-capping, intron splicing, and 3'-RNA cleavage and polyadenylation (3'-processing). The mature mRNA is packaged with RNA-binding proteins (RBPs) into messenger ribonucleoprotein particles (mRNPs) and exported to the cytoplasm where it directs protein synthesis. Factors for mRNP biogenesis are recruited co-transcriptionally by interactions with the C-terminal domain (CTD) of Pol II (Buratowski, 2009; Heidemann and Eick, 2012; Hsin and Manley, 2012) and by interactions with the emerging pre-mRNA transcript (Mandel *et al.*, 2008; Wahl *et al.*, 2009; Chan *et al.*, 2011; Proudfoot, 2011; Darnell, 2013; Mller-McNicoll and Neugebauer, 2013).

Mapping of mRNP biogenesis factors onto pre-mRNA and mature mRNA promises insights into RNA determinants for splicing, 3'-processing, and RNA export, and the coupling between these processes. Biogenesis factors can in principle be mapped onto the transcriptome by *in vivo* protein-RNA cross-linking and immunoprecipitation (CLIP) (Ule *et al.*, 2003). CLIP is based on UV light-induced cross-linking and identifies direct protein-RNA interaction sites after sequencing of the cross-linked RNA regions (Milek *et al.*, 2012). CLIP-based methods could indeed provide transcriptome maps for several human 3'-processing factors (Martin *et al.*, 2012) and mRNA-binding proteins in the yeast *Saccharomyces cerevisiae* (Tuck and Tollervey, 2013). However, mRNP biogenesis factors have not been systematically mapped onto pre-mRNA, likely due to difficulties in trapping short-lived RNAs in cells, and due to the complexity caused by the large variety of pre-mRNA species.

Here I mapped 23 mRNP biogenesis factors onto the newly synthesized yeast transcriptome, providing 10^5–10^6 high-confidence RNA interaction sites per factor. These maps were obtained by photoactivatable-ribonucleoside-enhanced (PAR)-CLIP, which was developed in human cells (Hafner *et al.*, 2010) and recently adopted to yeast (Creamer *et al.*, 2011; Schulz *et al.*, 2013). For data analysis, I helped to develop a computational pipeline based on advanced statistical models and motif searches with XXmotif (Hartmann *et al.*, 2013). The programming and modulation of the pipeline was performed by Phillipp Torkler. I show that PAR-CLIP efficiently captures short-lived pre-mRNA intermediates, and provide deep insights into the *in vivo* RNA-binding preferences of mRNA biogenesis factors. My analysis includes factors implicated in 5'-cap binding, splicing, 3'-processing, and mRNA export. They define conserved interactions between the splicing factors Mud2-Msl5 (U2AF65-BBP) and U1/U2 snRNPs, and pre-mRNA introns. They also identify a unified arrangement of the 3'-processing factors CPF/CPSF and CFIA/CstF at pre-mRNA polyadenylation (pA) sites in yeast and humans, which results from a distinct A/U dinucleotide signature. Furthermore, global analysis of the data indicates that 3'-processing factors have roles in RNA splicing and surveillance, and couple biogenesis events to restrict nuclear export to mature mRNPs.

Publications

Parts of this work contributed to following publications or are in the process of being published:

Schulz D.*, Schwalb B.*, Kiesel A., **Baejen C.**, Torkler P., Gagneur J., Söding J., and Cramer P.
Transcriptome Surveillance by Selective Termination of Noncoding RNA Synthesis.
Cell. 155, 2013 Nov 21;155(5):1075-87. *These authors contributed equally to this work.

D.S. and P.C. conceived and designed the study. D.S. performed ChIP-seq and 4tU-seq. C.B. performed PAR-CLIP. B.S., D.S., J.S., and J.G. designed data analysis. B.S., A.K., P.T., D.S., and C.B. carried out data analysis. D.S. and P.C. wrote the manuscript with input from all authors. P.C. supervised the project.

Baejen C.*, Torkler P.*, Gressel S., Essig K., Söding J., and Cramer P.
Transcriptome maps of mRNP biogenesis factors define pre-mRNA recognition.
Molecular Cell, *under revision*. *These authors contributed equally to this work.

C.B. and P.C. designed the experiments. P.T., J.S., C.B. and P.C. designed data analysis methods. C.B. established protocols and planned experiments. C.B., S.G., and K.E. performed experiments. P.T. carried out data analysis. P.C. and C.B. wrote the manuscript with input from all authors. P.C. and J.S. supervised the work.

Acknowledgement

I would like to express my special appreciation and thanks to my advisor Professor Dr. Patrick Cramer, you have been a tremendous mentor for me. I would like to thank you for encouraging my research and for allowing me to grow as a scientist. Your boundless energy and enthusiasm motivates your PhD students, including me. Your advice on both research as well as being a group leader have been priceless and will beneficially influence me in my further career.

I warmly thank PD Dr. Dietmar Martin for advising this thesis, and his support throughout my studies. I would also like to thank my further committee members (Professor *in spe*) Dr. Katja Strässer and Professor Klaus Förstemann as well as Professor Mario Halic and Professor Karl-Peter Hopfner for your comments and suggestions, thanks to you.

I am deeply grateful to the entire Cramer group from the past, presence and future at the LMU and MPI, respectively. I warmly thank Claudia Buchen and Stefan Benkert for their helpful administrative support. I would like to thank all postdocs, 'fellow sufferer' (PhD students), students and technicians for practical and emotional support. Many thanks to Thomas Fröhlich and the Blum lab, including Sylvia, Stefan, Alex and Helmut. A special thank goes to Heidi Feldmann, for being such a great mentor and friend.

I am deeply grateful to Phillipp Torkler. You have been the Ying to my Yang, the hop in my Maß, and so forth. I guess, we have become a very successful example for team working between different scientific fields and might inflame further collaboration between bioinformaticians and biochemists. In this way, I would also like to thank Phillipp's mentor Johannes Söding. Johannes, I admire your logical and mathematical thoughts. You are a much better biologist and biochemist as you believes.

People come and go, but friends always stay. I have met many people, but some are going to be in my heart for ever. Unfortunately, I cannot mention each of those. Therefore, I would like to embrace the opportunity to thank my two students Katharina and Saskia. You guys helped me a lot and I really enjoyed time with you. Thanks for your efforts. Next, I have to thank Daniel Schulz for being such a great bench neighbor. We did share almost everything in the lab and I thank you for all funny moments. I have to thank Kathrin for all coffee brakes and your friendly ear.
Last but not least, I would like to express my special appreciation and thanks to Christoph Lutz Engel. You have become a unique and valuable friend to me and my family (especially to Bruno). We shared many beds during our lab and Gene Center retreats, and I will never forget my first day as a 'Cramer'.

Many thanks to my whole family. Here, I will thank my parents, Marion and Bernd, for their faith in me and allowing me to be as ambitious as I wanted. It was under their watchful eye that I gained so much drive and an ability to tackle challenges head on. Besides your parental duties, you have become great mentors and friends. Also, I thank my parents-in-law, Silke and Axel, as well as my amazing sisters-in-law Juliane, Sophia and Helena.

I am deeply grateful to my grandparents, Hans-Joachim and Anni. You have supported me invariably during my childhood and studies. I will never forget your tremendous help and love. Dear Opi, rest in peace forever. You have been one of the most influencing person in my life, and I always appreciate this.

Bruno, my little 'Moahn', I honestly thank you for all wakeful nights, well-filled and stinky diapers, and especially all bright-eyed moments, which illuminated some darker times. I am already incredible proud of you and I am looking forward watching you becoming a great boy, man, father and friend.

Finally, and most importantly, I would like to thank my girl and mother of my awesome son. Anna's support, encouragement, quiet patience, and unwavering love were undeniably the bedrock upon which the past four years of my life have been built. Her tolerance of my occasional vulgar moods is a testament in itself of her unyielding devotion, cuteness and love. Anna, I am extremely happy that I have found you and I am going to always need and love you.

Euch allen, danke ich sehr,
Carlo

List of Figures

List of Tables

List of Equations

Table of Contents

1 Introduction

1.1 Not only DNA's messenger

Every known form of life on Earth is based on three major macromolecules: deoxyribonucleic acid (DNA), ribonucleic acid (RNA), and proteins. Francis Crick's 'central dogma of molecular biology', put simply, describes it as follows: "DNA makes RNA, RNA makes proteins, proteins make us" (Crick, 1958). While DNA functions as repository of the cellular information, the genetic code, RNA was long believed to be only DNA's messenger, a simple intermediate between DNA and protein synthesis, obtained from gene transcription. Generally, a nascent RNA is synthesized from a DNA template by a DNA-dependent RNA polymerase (Pol). In bacteria and archaea, all forms of RNA are transcribed by one single polymerase (Zhang *et al.*, 1999). In an eukaryotic cell, RNA transcription is carried out by four polymerases: Pol I, Pol II, Pol III and the mitochondrial RNA polymerase (mitoPol) (Vannini and Cramer, 2012; Schwinghammer *et al.*, 2013). While Pol II transcribes messenger RNA (mRNA) and several small RNAs, polymerase I and III produce ribosomal RNA (rRNA) and transfer RNA (tRNA), respectively (Vannini and Cramer, 2012). Further Pol II-synthesized transcripts are small nuclear RNAs (snRNAs), small nucleolar RNAs (snoRNAs), stable unannotated transcripts (SUTs), cryptic unstable transcripts (CUTs), and non-coding RNAs (Vannini and Cramer, 2012; Schulz *et al.*, 2013).

However, from an evolutionary point of view, RNA is supposed to be the older, primordial molecule that preceded the contemporary DNA- and protein-based life (Woese *et al.*, 1966; Crick, 1968). Especially the discovery of the ability of RNA to catalyze and theoretically replicate itself by forming complex secondary structures (Kruger *et al.*, 1982; Guerrier-Takada *et al.*, 1983; Cech, 1986), inflamed the so called 'RNA world' hypothesis, and therefore resulted in a variety of models proposing the possible existence of ancient "Ribo-organism" that carried out complex, RNA-based metabolism even before higher molecules evolutionarily appeared (Cech, 1986; Gilbert, 1987; Benner *et al.*, 1989).

Due to the capacity of being both genetic material and cellular enzyme (referred to as ribozyme) (Kruger *et al.*, 1982; Guerrier-Takada *et al.*, 1983), RNA was now considered an underestimated, multitalented molecule with still unknown functions. The discovery that modern organisms like Gram-positive bacteria and plants are capable to selectively bind metabolites by using so called riboswitches, additionally supported the hypothesis that ancient regulatory and sensory systems might have been initially based on exclusively RNA molecules (Ellington and Szostak, 1990; Winkler *et al.*, 2002; Mandal and Breaker, 2004; Cochrane and Strobel, 2008; Roth and Breaker, 2009).

Despite the ability of RNA molecules to accomplish sophisticated reactions, especially in large RNA complexes (Ferr-D'Amar and Scott, 2010), modern RNAs commonly operate in concert with RNA-binding proteins (RBPs) and several protein complexes (Spliceosome, Ribosome, etc.), which mainly benefit from direct RNA–RNA interactions as well as the catalytic activity of involved RNAs (Will and Lhrmann, 2011; Moore and Steitz, 2011).

1.2 The complex networks of mRNA's life

The biogenesis of mRNAs in *S. cerevisiae* involves pre-mRNA synthesis by RNA Pol II and co-transcriptional RNA processing, which encompasses 5'-capping, intron splicing, and 3'-end RNA cleavage and polyadenylation (3' processing). The mature mRNA is packaged with RBPs into messenger ribonucleoprotein particles (mRNPs) and exported to the cytoplasm, where it directs protein synthesis. Factors for mRNP biogenesis are recruited co-transcriptionally by interactions with the C-terminal domain (CTD) of the largest subunit of Pol II (Rpb1) (Buratowski, 2009; Heidemann and Eick, 2012; Hsin and Manley, 2012) and by interactions with the emerging pre-mRNA transcript (Mandel *et al.*, 2008; Wahl *et al.*, 2009; Proudfoot, 2011; Darnell, 2013; Mller-McNicoll and Neugebauer, 2013).

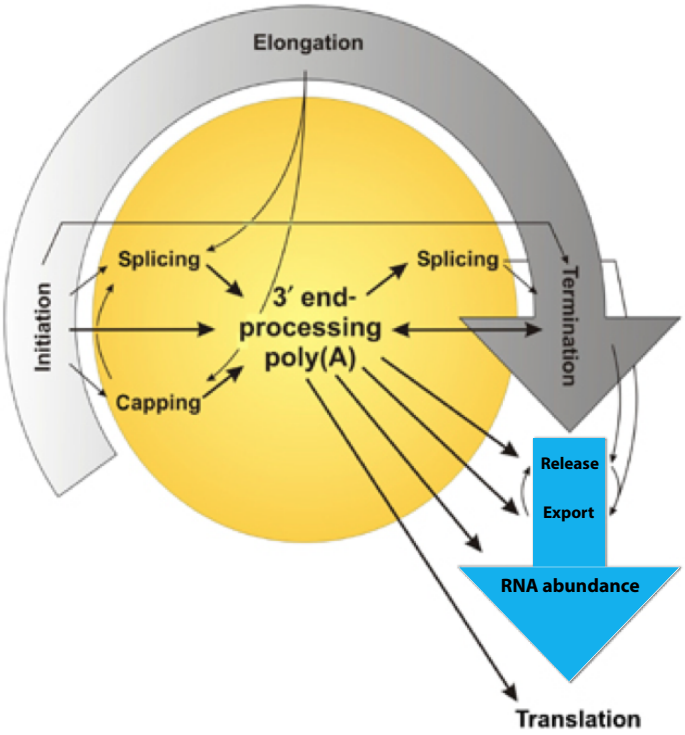

Figure 1.1: **The complex networks of mRNA's life** [adapted from Danckwardt *et al.* (2008)]. All main steps during transcription (grey arrow) are functionally interconnected to mRNA processing (yellow center) and post-transcriptional mechanisms (blue arrow).

1.2.1 The CTD code controls mRNA fate

The CTD is a unique domain composed of heptapeptide repeats with the consensus sequence Tyr1-Ser2-Pro3-Thr4-Ser5-Pro6-Ser7 ($Y_1S_2P_3T_4S_5P_6S_7$) that is conserved from yeast to humans (Buratowski, 2009; Heidemann and Eick, 2012; Hsin and Manley, 2012). The number of repeats depends on the complexity of the organism ranging from 27 repeats in *S. cerevisiae* to 52 in *Homo sapiens* (Chapman *et al.*, 2008). In yeast, more than seven repeats are required for cell viability (West and Corden, 1995).

Table 1.1: **Relevant kinases (”writers”) and phosphatases (”eraser”) for *in vivo* CTD modification**. The capital letter 'P' marks the phoshorylation state of the peptide. Neither a kinase nor phosphatase is known for Thr4, which was consequently excluded from this list.

Peptide	Kinase(s)	Phosphatase(s)	Citations
Tyr1	-	Glc7	Schreieck *et al.* (2014)
Ser2	Bur1, Ctk1	Fcp1	Archambault *et al.* (1997); Patturajan *et al.* (1999)
			Murray *et al.* (2001)
Ser5	Kin28	Rtr1, Ssu72	Mosley *et al.* (2009); Akhtar *et al.* (2009); Xiang *et al.* (2012)
Ser7	Kin28	Ssu72	Akhtar *et al.* (2009); Xiang *et al.* (2012)

Five out of seven residues can be phosphorylated and dephosphorylated by kinases (”writers”) and phosphatases (”eraser”), respectively (Table 1.1) (Kim *et al.*, 2009; Mayer *et al.*, 2010). Importantly, each repeat can be treated separately, resulting in a unique code with 128 theoretical combinations per repeat (2^7) or a total of about 7.85×10^{56} different combinations over the entire yeast CTD (Heidemann and Eick, 2012; Schreieck *et al.*, 2014).

Due to its unique and changeable code (Figure 1.2), the Pol II CTD primarily functions as a general recruiting platform for several proteins (”CTD readers”) that are involved in transcription initiation, elongation and termination, as well as in co-transcriptional RNA processing, export and histone modification (Maxon *et al.*, 1994; Cho *et al.*, 1997; Schroeder *et al.*, 2000; Ng *et al.*, 2003; Buratowski, 2009; Mayer *et al.*, 2010; Kubicek *et al.*, 2012; Meinel *et al.*, 2013).

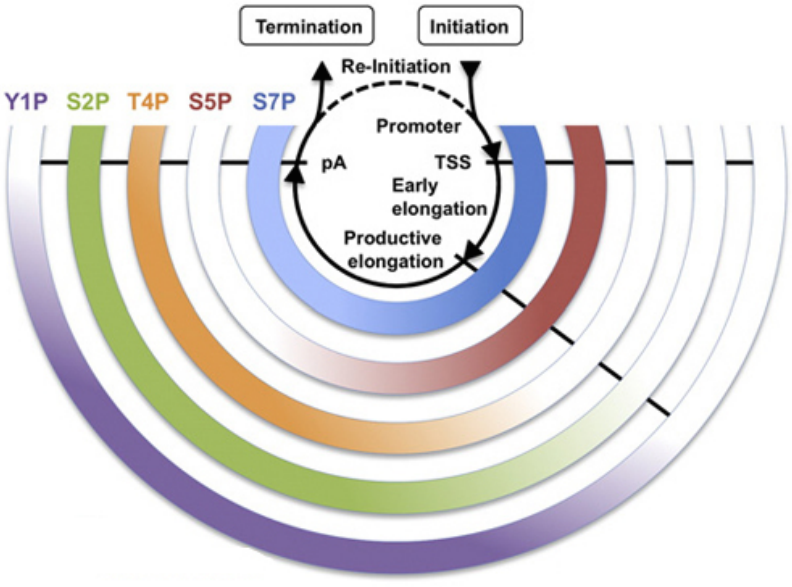

Figure 1.2: **Coordination of the Pol II transcription cycle by the CTD code** [adapted from Heidemann *et al.* (2013)]. Main stages of the Pol II transcription cycle are centered and indicated by the black circle, including the transcription start site (TSS) as well as the cleavage and polyadenylation site (pA). The surrounding arcs symbolize the five modifiable peptides of a CTD repeat in a specific color code. The stronger the color saturation, the stronger is the expected phosphorylation level of the peptide at this state during the transcription cycle.

1.2.2 Initiation of transcription and subsequent 5'-capping

The earliest steps in transcription initiation are the assembly of the pre-initiation complex (PIC) and the recruitment of RNA Pol II to the gene promoter, accompanied by large complexes such as nucleosome remodelers (SWI/SNF) or histone modifiers (INO80/SWR1) (Wilson *et al.*, 1996; Buratowski, 2009; Tosi *et al.*, 2013). Briefly, the TATA box is bound by the TATA-binding protein (TBP), a subunit of the transcription factor II (TFII)D, followed by the recruitment of TFIIA and TFIIB (Buratowski *et al.*, 1989; Thomas and Chiang, 2006). Subsequently, the Mediator complex, which functions as coactivator and connector, delivers Pol II to the emerging PIC, assisted by its direct interaction with the hypophosphorylated CTD (Myers *et al.*, 1998; Bourbon *et al.*, 2004; Kornberg, 2005). The transcription factors IIE and IIF then bind preferentially to Pol II and its CTD, and finally recruit TFIIH (Maxon *et al.*, 1994; Kang and Dahmus, 1995). This last transcription factor functions in two ways: (i) it unwinds the DNA and thus assists in the formation of an open complex (Giardina and Lis, 1993; Kostrewa *et al.*, 2009; Grnberg *et al.*, 2012), and (ii) it phosphorylates the Ser5 and Ser7 residues of the CTD by its kinase subunit Kin28 (Table 1.1 and Figure 1.2) (Akhtar *et al.*, 2009). This specific mark leads to the dissociation of the Mediator complex and allows Pol II to start scanning for the initiator element or transcription start site (TSS) (40–120 bp downstream of the TATA box) (Sgaard and Svejstrup, 2007). This process is referred to as promoter clearance or promoter escape (Luse, 2013).

After reaching the TSS, Pol II starts transcribing the first 12 to 13 nucleotides (nt) until the nascent RNA fragment clashes with TFIIB, which is subsequently released, and therefore triggers the formation of a stable transcription elongation complex (TEC) (Hahn, 2004; Sainsbury *et al.*, 2013). To protect the growing, nascent transcript from degradation, it is being capped within three catalytic reactions (Ghosh and Lima, 2010). Firstly, the γ-phosphate from the 5'-triphosphate is removed by the 5'-triphosphatase Cet1. Secondly, the guanylyltransferase Ceg1 adds an inverted guanylyl group. Finally, the methyltransferase Abd1 adds a methyl group to the N7 atom of the terminal guanine group. Both Ceg1 and Abd1 are previously recruited by the phosphorylated CTD (Figure 1.3) (Cho *et al.*, 1997; Schroeder *et al.*, 2000). Futhermore, Abd1 is hypothesized to have influence on promoter escape in a methyltransferase-independent manner (Schroeder *et al.*, 2004).

To ensure that only correctly capped transcripts pursue the productive elongation phase, nascent transcripts might be qualitatively controlled during the promoter-proximal pausing (Kim *et al.*, 2004a; Mandal *et al.*, 2004; Jiao *et al.*, 2010). Whereas transcripts without or with aberrant cap structures are co-transcriptionally removed by the Rai1-Rat1 decay pathway (Buratowski, 2009; Xiang *et al.*, 2009; Jiao *et al.*, 2010), completed 7-methyl-guanosine (m7G) caps are associate with the cap-binding complex (CBC) that functions in both pre-mRNA splicing and mRNA export (Schwer and Shuman, 1996; Lewis and Izaurralde, 1997; Calero *et al.*, 2002; Mazza *et al.*, 2002). Abd1 and CBC are important for recruitment of the kinases Ctk1 and Bur1 (Table 1.1), which promote elongation and capping enzyme release (Lidschreiber *et al.*, 2013).

However, once the TEC is formed and the 5'-capping checkpoint has been overcome, the initiation-elongation transition takes place, meaning the exchange of initiation factors by elongation factors like Spt4/5/6, which finally introduces the productive elongation phase (Mayer *et al.*, 2010; Martinez-Rucobo *et al.*, 2011; Lidschreiber *et al.*, 2013).

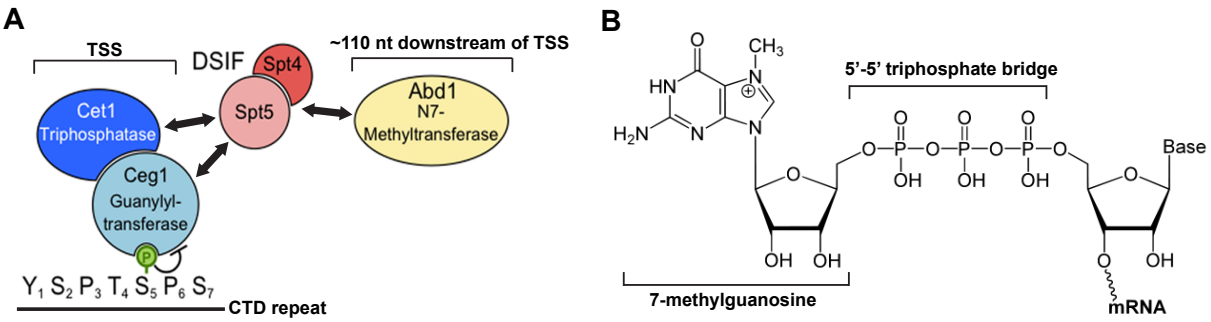

Figure 1.3: **Recruitment and regulation of the capping enzymes**. **A**. The guanylyltransferase Ceg1 binds phosphorylated Ser5 (green) and recruits the triphosphatase Cet1 by direct binding. ChIP experiments showed that Spt5 contributes to stable recruitment of the mRNA capping enzymes Cet1, Ceg1, and Abd1 (Lidschreiber *et al.*, 2013) [adapted from Jeronimo *et al.* (2013)]. **B**. Completed 7-methylguanosine (m7G) cap structure, connected to the 5'-end of mRNA via a 5'-to-5'-triphosphate bridge.

1.2.3 Elongation factors ensure barrier-free transcription

Efficient transcript elongation must overcome several blocks that are intrinsic to Pol II and its chromatinized DNA template. Because the chromatin architecture represents a barrier to the transcribing elongation complex, histone structures have to be displayed. One essential complex that stimulates both transcription elongation and Pol II productivity is the Spt4/5 or yeast DSIF complex (Hartzog and Fu, 2013). After binding to the Pol II clamp domain, Spt5 recruits and binds Spt4 (Hartzog *et al.*, 1998; Martinez-Rucobo *et al.*, 2011; Hartzog and Fu, 2013). This complex formation is a kind of an "initial spark" that enables further reactions. For instance, Spt4/5 directly interacts with the elongation factor Spt6, which reassembles chromatin after Pol II has passed (Figure 1.4) (Hartzog *et al.*, 1998). Furthermore, Spt5 possesses a repetitive C-terminal region (CTR) (Swanson *et al.*, 1991; Zhou *et al.*, 2009). Similar to the CTD, this region functions as platform to recruit further proteins (Zhou *et al.*, 2009). One complex that needs to be recruited by the CTR, is the Pol II-associated factor (PAF) complex (Zhou *et al.*, 2009; Liu *et al.*, 2009). Like Spt4/5, PAF is required to recruit histone H3K4 methyltransferase Set1 and Spt16, members of the COMPASS and FACT complex, respectively (Belotserkovskaya *et al.*, 2003; Krogan *et al.*, 2003; Kim *et al.*, 2004b). Set1 in particular also interacts with phosphorylated Ser5 and enables proper H3K4 methylation (H3K4me) to finally ensure further histone acetylation and a barrier-free passing of the TEC (Smolle and Workman, 2013). The Set1-dependent H3K4me additionally recruits Nrd1, allowing early termination by the Nrd1-Nab3-Sen1 pathway (Figure 1.4) (Terzi *et al.*, 2011).

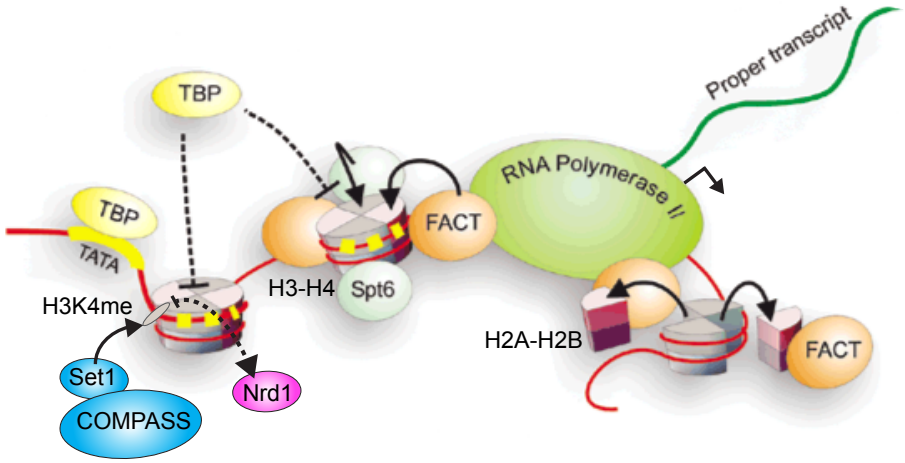

Figure 1.4: **Elongating Pol II associates with Spt6, COMPASS and FACT** [adapted from Carrozza *et al.* (2003)]. Before Pol II reaches a histon, FACT mediates the deposition of the histone H2A–H2B heterodimers. After the transcribing Pol II has passed, both Spt6 and FACT reassemble the histone structure. While Spt6 reinserts histone H3–H4 heterodimers, FACT reconstitutes H2A–H2B. Set1-dependent H3K4 trimethylation (COMPASS) recruits Nrd1 and initiates the Nrd1-Nab3-Sen1 pathway (Terzi *et al.*, 2011).

1.2.4 Spliceosome recruitment and pre-mRNA splicing

Unlike most eukaryotic genomes, the *S. cerevisiae* genome consists of a few introns only (95 % of genes are intronless) (Neuvglise *et al.*, 2011). Compared to the human system, the yeast 'intronome' is smaller on average, primarily located near the 5'-end of a gene, and predominantly limited to one intron per gene (Woolford and Peebles, 1992; Neuvglise *et al.*, 2011; Will and Lhrmann, 2011).

Intron recognition by the spliceosome is the initial step in pre-mRNA splicing and was extensively studied *in vitro* by Will and Lhrmann (2011). Spliceosome recruitment is triggered by intron-specific key sequences as well as histone and CTD modifications during transcription (Morris and Greenleaf, 2000; Shieh *et al.*, 2011).

Briefly, splicing begins with binding of the branch point (BP)-binding protein (BBP) (Msl5) to the BP and binding of U2AF65 (Mud2) to a pyrimidine-rich region between the BP and 3'-splice site (3'-SS), and continues with binding of the U1 nuclear ribonucleoproteins (snRNP) to the 5'-SS (Will and Lhrmann, 2011). The resulting complex E is remodeled, and the small nuclear RNA (snRNA) of the U2 snRNP displaces BBP by base pairing with the BP region, positioning U2 snRNP near the 3'-SS and giving rise to complex A. After the recruitment of the U4/U6.U5 tri-snRNP complex, the non-catalytic A complex is transformed into a pre-catalytic B complex (Boehringer *et al.*, 2004). Due to major rearrangements, the U4/U6 interaction is disrupted and the U6 snRNA replaces the U1 snRNP complex at the 5'-SS. Following, the U1 and U4 snRNPs are dismissed, and the B complex becomes active by catalyzing the first transesterification, performed by U6/U2 snRNPs (Wolf *et al.*, 2009). This consequently leads to the formation of the C complex and the second transesterification (Tseng and Cheng, 2013). Finally, adjoining exons are covalently ligated, and the resulting intron lariat is released together with the bound U2, U5, and U6 snRNP.

1.2.5 mRNP export is deeply linked to 3'-end processing

In *S. cerevisiae*, the 3'-end mRNA processing machinery consists of over 20 different proteins in several subcomplexes (Mandel *et al.*, 2008). Each subcomplex recognizes and binds a specific sequence element within the 3'-untranslated region (3'-UTR). Hrp1 and Nab4, which form the cleavage factor IB (CFIB), are described to initially interact with the AU-rich efficiency elements to facilitate precise identification of the cleavage and polyadenylation (pA) site (Guo *et al.*, 1995; Leeper *et al.*, 2010; Mischo and Proudfoot, 2013). The cleavage factor IA (CFIA) complex recognizes the A-rich positioning element, commonly located 10–30 nt upstream of the pA (Guo *et al.*, 1995; Dichtl and Keller, 2001; Leeper *et al.*, 2010). The main cleavage and polyadenylation factor (CPF) complex comprises a core structure including the poly(A) polymerase Pap1, the polyadenylation factor (PFI), cleavage factor II (CFII) and six additional proteins, termed as APT (Figure 1.5) (Nedea *et al.*, 2003; Mandel *et al.*, 2008). Recruitment of these complexes or single proteins occurs via binding to those specific mRNA elements and the phosphorylated CTD (Section 1.2.1) (Minvielle-Sebastia *et al.*, 1994; Komarnitsky *et al.*, 2000; Meinhart and Cramer, 2004). While the CFIA complex binds the CTD through the conserved CTD interaction domain (CID) of Pcf11 (Sadowski *et al.*, 2003), the CFII complex uses its factor Cft1 to couple elongation and pA site recognition (Dichtl *et al.*, 2002). Additional studies demonstrate further connections of the 3'-end processing to histone modification, mRNA splicing, and/ or mRNA export (Hirose and Manley, 2000).

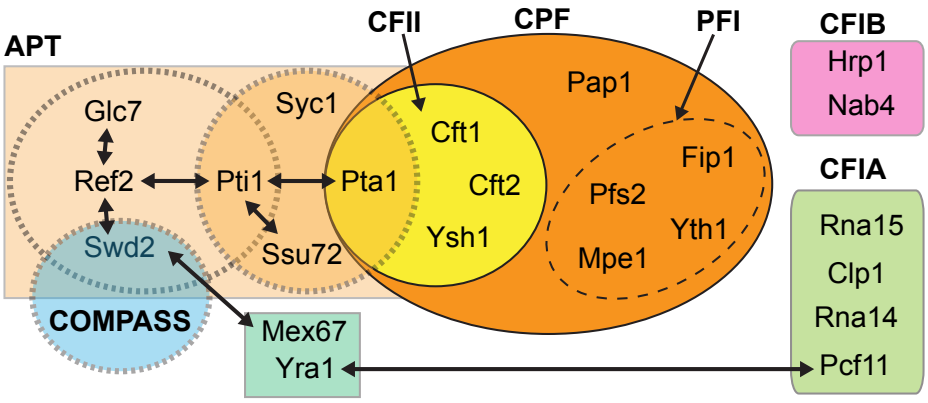

Figure 1.5: **The 3'-end mRNA processing machinery of *S. cerevisiae*** [adapted from Nedea *et al.* (2003)]. The orange and yellow ellipses symbolize the CPF (incl. PFI) and CFII complex, respectively. The rectangle represents the APT complex with its two subcomplexes (dashed circles), connected through Pti1. The factor Pta1 is required for bridging CPF/CFII to APT, and is a component of both complexes. Also Swd2 is found in both the APT and COMPASS complex (blue circle). Recruitment of Swd2 to the APT/CPF is done by Ref2. The cleavage factor IA (green) and the cleavage factor IB (rose) are depicted on the right. Double-headed arrows display known interactions.

Export of mRNA requires packing into messenger ribonucleoprotein particles (mRNPs) (Strsser *et al.*, 2002). These mRNPs are then bound by the THO/TREX complex and exported through the nuclear pore complex (NPC) into the cytosol (Strsser *et al.*, 2002). Briefly, the DEAD-box helicase Sub2, already bound to the nascent transcript, binds Hpr1 and recruits Mex67 by binding its C-terminal

ubiquitin-associated (UBA) domain (Jensen *et al.*, 2001). Mex67 principally acts as key to unlock the nuclear envelop (Segref *et al.*, 1997; Reed and Hurt, 2002). Binding of Mex67 to Yra1 has been proposed to displace Sub2 (Reed and Hurt, 2002). It is assumed that the THO/TREX factors coordinate transcription elongation and 3'-end processing as well as final transcript release and R-loop prevention (Jimeno *et al.*, 2002; Rougemaille *et al.*, 2008; Gmez-Gonzlez *et al.*, 2011).

Mex67 itself has a low intrinsic affinity and therefore requires mRNA-binding adaptors to carry its cargo through the NPC (Rodrguez-Navarro and Hurt, 2011). Protein-protein interactions are mediated by its N-terminal ribonucleoprotein (RNP) domain (Liker *et al.*, 2000), whereas the UBA domain promotes NPC-binding and co-transcriptional recruitment (Dieppois *et al.*, 2006; Hobeika *et al.*, 2007). The phosphorylation state of the Mex67 adapter Npl3 depends on Glc7, a component of the APT subcomplex (Gilbert and Guthrie, 2004). The stability of the APT on the other hand is ensured by correct ubiquitylation of Swd2, controlled by histone H2B modification (COMPASS complex). Ubiquitylated Swd2 promotes correct recruitment of Mex67 to its UBA domain by coupeling the APT subcomplex to the nuclear export machinery (Figure 1.5) (Vitaliano-Prunier *et al.*, 2012). Besides the CFIB factor Hrp1, the Mex67 adaptor and poly(A)-binding protein Nab2 influence cleavage and polyadenylation as well as mRNA packaging and nuclear surveillance (Anderson *et al.*, 1993; Green *et al.*, 2002; Hector *et al.*, 2002; Tuck and Tollervey, 2013). Furthermore, Yra1 demonstrates a crucial and previously unrecognized involvement in coupling the 3'-end maturation with nuclear mRNA export by directly interacting with Pcf11 (Figure 1.5) (Johnson *et al.*, 2009).

1.2.6 Termination of transcription and mRNA decay

During the productive elongation phase, Tyr1 is constantly held in a phosphorylation state (Y1P) to prevent recruitment of the termination machinery (Figure 1.2) (Heidemann *et al.*, 2013). When Pol II approaches the pA site, Tyr1 gets dephoshorylated and thus allows the recruitment of Pcf11, which initiates the subsequent termination process (Mayer *et al.*, 2012). After Pol II has passed the pA site, the RNA is cleaved by the putative endoribonuclease Ysh1, promoted by the CFIA factor Rna15 (Birse *et al.*, 1998). Following, the poly(A) polymerase Pap1 polyadenylates the pre-mRNA, which can then be exported to the cytosol for translation and ultimate degradation (Figure 1.6B).

Termination and release of Pol II is either triggered by destabilization through conformational changes of the Pol II EC after transcribing the pA site ('allosteric model') or caused by the 5'$\rightarrow$3' exonuclease Rat1 (human Xrn2), which interacts with the RNA-helicase Sen1 to degrade the transcript and finally collides with the EC that may consequently induce termination ('torpedo model') (Figure 1.6A) (Logan *et al.*, 1987; Kim *et al.*, 2004c; Kawauchi *et al.*, 2008). In contrast to mRNA, termination of ncRNAs depends on the early termination factor Nrd1 that either can directly bind to mRNA or to the CTD via its CTD interaction domain (CID) (Steinmetz and Brow, 1996; Creamer *et al.*, 2011). Nrd1 preferentially binds RNA motifs, which are enriched in ncRNAs and depleted in mRNAs except in some mRNAs

whose synthesis is controlled by transcription attenuation (Schulz *et al.*, 2013). Depletion of Nrd1 from the nucleus results in Nrd1-unterminated transcripts (NUTs) that can deregulate transcription and disturb promoter directionality (Schulz *et al.*, 2013).

Ultimate degradation is initiated by the deadenylation process performed by the Ccr4-Not complex (Chen and Shyu, 2011). The shortened 3'-end triggers both decapping by Dcp1/Dcp2 and exosome-mediated 3'–5' exonucleolytic decay. The 5'–3' exonucleolytic decay, which is performed by Xrn1, starts right after the complete removal of the cap structure by the decapping complex (Figure 1.6B).

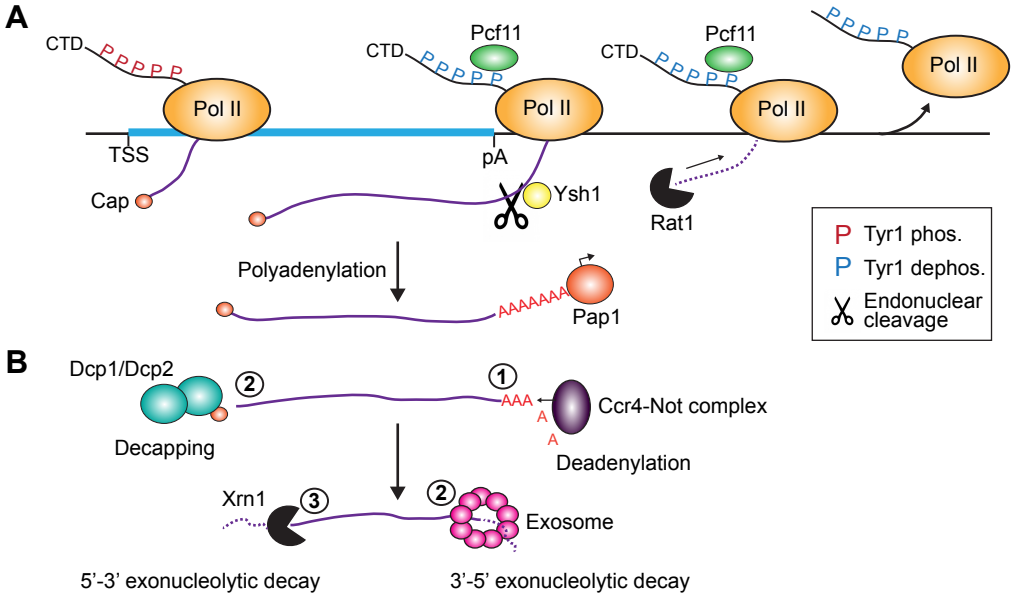

Figure 1.6: **Eukaryotic 'torpedo' termination and decay mechanisms** [adapted from Decker and Parker (2002) and Luo *et al.* (2006)]. **A**. The torpedo model. Endonucleolar cleavage (scissor) at the pA site creates an entry site for the 5'–3' exonuclease Rat1 that degrades until it collides with the elongating Pol II. The direct contact or the short nascent RNA may consequently trigger the release of Pol II. **B**. Final decay of mRNA is initiated by deadenylation by the Ccr4-Not complex, which can then trigger decapping (by Dcp1/Dcp2) and exosome-mediated decay. Two general mRNA decay pathways exist: The 5'–3' exonuclease digestion by Xrn1 (requires decapping) and the 3'–5' exonuclease digestion by the exosome.

1.3 Target recognition of RNA-binding proteins (RBPs)

Supposingly, about 600 annotated RBPs, possessing several well-defined RNA-binding domains (RBDs), are encoded in the budding yeast *Saccharomyces cerevisiae* (Hogan *et al.*, 2008). These RBPs associate with sets of RNAs at both a particular stage during the cell cycle (Dreyfuss *et al.*, 2002; Moore, 2005) and at a similar localization (Hogan *et al.*, 2008). Structural diversity of RBPs and target recognition of RBDs are a function of the type, number, and arrangement of RBDs, helping RBPs to attain specificity and high affinity for an (m)RNA sequence (Lunde *et al.*, 2007). Several RBDs are currently recognized including the RNA-recognition motif (RRM), Zinc finger (ZF), K-homology (KH) domain, and double-stranded RBD (dsRBD) (Table 1.2 and Figure 1.7).

Table 1.2: **Topology of the most prominent RNA-binding domains.** The typical RNA-recognition motif (RRM) consists of four anti-parallel beta-sheets and two alpha-helices, whereas a third alpha-helices can be included during RNA binding. Two different types of the K Homology (KH) domain are known, referred to as type I and type II. While type I domains are mainly found in eukaryotes, type II domains predominantly exist in prokaryotes. By far the best-characterized class of zinc fingers is the Cys_2His_2-like fold group (C2H2). Double stranded RNA-binding domains (dsRBDs) recognize secondary structures of RNAs and are mainly found in post-transcriptional gene regulation.

Domain	Amino acids	Topology	Citations
RRM	90	$\beta1$-$\alpha1$-$\beta2$-$\beta3$-$\alpha2$-$\beta4$	Sachs *et al.* (1987)
$KH_{type\text{-}1}$	70	$\beta1$-$\alpha1$-$\alpha2$-$\beta2$-$\beta3$-$\alpha4$	Musco *et al.* (1997)
$KH_{type\text{-}2}$	70	$\alpha1$-$\beta1$-$\beta2$-$\alpha2$-$\alpha3$-$\beta3$	Grishin (2001)
ZF_{C2H2}	30	$[\beta1\text{-}\beta2\text{-}\alpha1 + Zn^{2+}]_2$	Miller *et al.* (1985)
dsRBD	65	$\alpha1$-$\beta1$-$\beta2$-$\beta3$-$\alpha2$	Bycroft *et al.* (1995)

The direct binding of RBDs to RNA targets can be limited by either too short recognition sequences or greater distances between essential target sites. While several RBPs are restricted due to only one RBD, many RBPs consist of multiple domains that can be modulated precisely through either combinations of different kinds of RBDs or a certain number of domain repeats (i.e. the ZF tandem domain, Figure 1.7). Thus, RBPs are able to interact with a much larger RNA surface and therefore increase their specificity as well as affinity for long stretches of continuous or even discontinuous RNA targets (Lunde *et al.*, 2007; Chen and Varani, 2013). Additionally, flexible linker sequences between RBDs enable the recognition of RNA target sites over long distances, even allowing RBDs to bind to separated target sites (Figure 1.8) (Conte *et al.*, 2000). Resulting intra-molecular interactions, including electrostatic interactions, shape complementarity, and hydrogen bonding lead to conformational changes that can significantly increase the binding to adjacent RNA (Conte *et al.*, 2000; Lunde *et al.*, 2007).

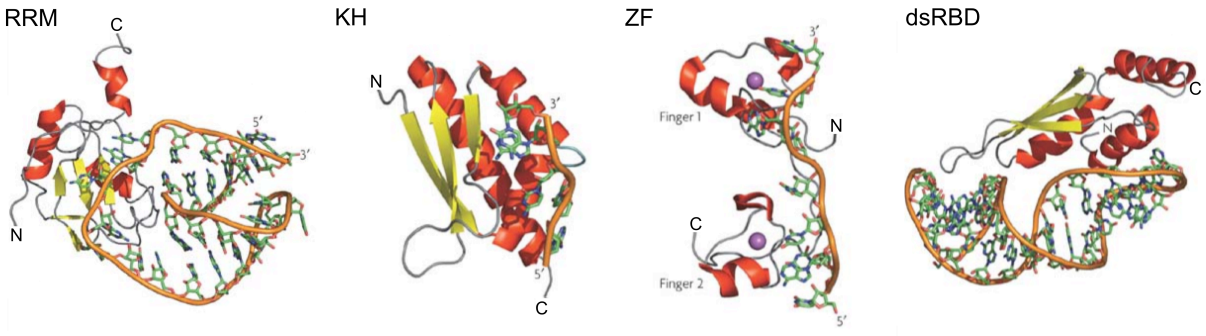

Figure 1.7: **Examples of known RBD structures bound to RNA** [adapted form Lunde *et al.* (2007)]. Depicted are the N-terminal RNA-recognition motif (RRM) of human U1A, the K Homology (KH) domain of Nova-2, two zinc fingers (ZFs) of the AU-rich element binding protein TIS11d, and the dsRBD of yeast's nuclear dsRNA-specific ribonuclease Rnt1 (alias RNase III). Alpha-helices and beta-sheets are colored in red and yellow, respectively. The RNA backbone is depicted as orange ribbon, and the zinc atoms of the ZF structure is highlighted in purple.

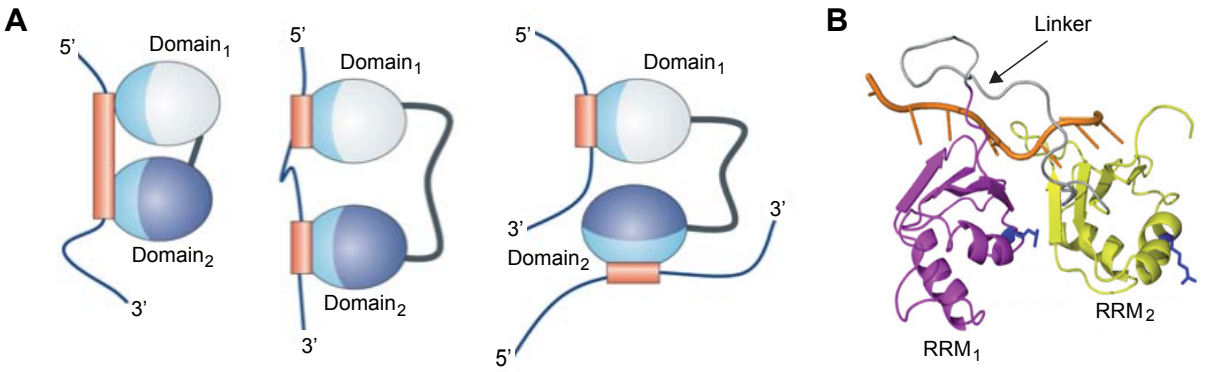

Figure 1.8: **Multi-domain conformations increase RNA-binding affinity** [adapted from Lunde *et al.* (2007) and Mackereth *et al.* (2011)]. **A**. Scheme of modular RNA-binding domains that can be arranged to recognize a long RNA sequence (left), separated target sites (centre), and/or RNAs that belong to different target molecules (right). Both domains are connected by a linker (grey). **B**. Arrangement of the tandem RNA recognition motif (RRM) domains of the human U2 snRNP auxiliary factor (U2AF65) after binding to a polypyrimidine tract (depicted in orange).

1.4 Cross-linking and immunoprecipitation (CLIP)

One of the first approaches to identify target-specific ligands (ssDNA or RNA) *in vitro* came up in the early 1990s, termed as 'Systematic evolution of ligands by exponential enrichment' (SELEX) (Tuerk and Gold, 1990). The first method to localize binding sites of RBPs *in vivo*, referred to as 'cross-linking and immunoprecipitation' (CLIP), was introduced by Ule *et al.* (2003). CLIP uses UV light at 254 nm to cross-link cells or tissues prior to cell lysis to avoid post-lysis interactions (Ule *et al.*, 2003). The energy of the UV_{254} light introduces a covalent bond between the protein of interest and the target RNA. The protein-RNA complex can be isolated through immunoprecipitation and SDS-PAGE. Cross-linked RNAs are trimmed by RNase digestion and bound RNA fragments are released by reverse cross-linking (by proteinase K treatment). Following 3' and 5' adapter ligation, the reverse transcription (RT) is performed to generate cDNA libraries that are finally amplified and sequenced (Ule *et al.*, 2003, 2005). During the RT, deletions are introduced at the cross-link site (Figure 1.9), which are more reliable in comparison to possible point mutations (Zhang and Darnell, 2011). Furthermore, it was shown by Zhang and Darnell (2011) that cross-links exclusively emerge at RNA target sites containing a uridine, indicating that either adenosines, cytidines and/ or guanosines are not 'clipable' or that the reverse transcriptase (RTase) cannot overcome these nucleosides.

1.4.1 CLIP-based methods and their specifications

Nowadays, many variants of the original CLIP protocol from Ule *et al.* (2003) exist. With the development of next-generation sequencing (NGS) platforms, the "high-throughput era" was finally initiated allowing much more complex and sophisticated studies. The combination of CLIP with NGS resulted in 'High-throughput sequencing CLIP' (or HITS-CLIP) that allowed the identification of RNA-binding sites in

a transcriptome-wide manner by comparing clustered sequence reads with specific negative controls (Figure 1.11) (Licatalosi *et al.*, 2008). In 2010, two advanced CLIP variants were introduced by Knig *et al.* (2010) and Hafner *et al.* (2010) referred to as 'Individual-nucleotide resolution CLIP' (iCLIP) and 'Photoactivatable ribonucleoside-enhanced CLIP' (PAR-CLIP), respectively. Compared to the original CLIP, both now ensured a single-nucleotide resolution without the need of negative controls by using two widely different approaches (Figure 1.11). While iCLIP relies on a modified cDNA library preparation based on circularization and frequent reverse transcriptase termination events at the cross-linked site (Knig *et al.*, 2010), PAR-CLIP uses photo-reactive nucleotides (such as 4-thiouridine or 4-thiouracil) and UV light at 365 nm for the cross-linking reaction, which increases the incidence of point mutations at the cross-link sites (Hafner *et al.*, 2010; Creamer *et al.*, 2011; Ascano *et al.*, 2012) (Figure 1.9 and Section 1.4.1). Additionally, iCLIP was shown to be performed with double-tagged RBPs using stringent tandem affinity purification (TAP) instead of an immunoprecipitation, termed as 'Individual nucleotide resolution UV-crosslinking and affinity purification' (iCLAP) (Wang *et al.*, 2010). Another CLIP approach that uses TAP similar to iCLAP is the 'Crosslinking and cDNA analysis' (CRAC) developed from Granneman *et al.* (2009).

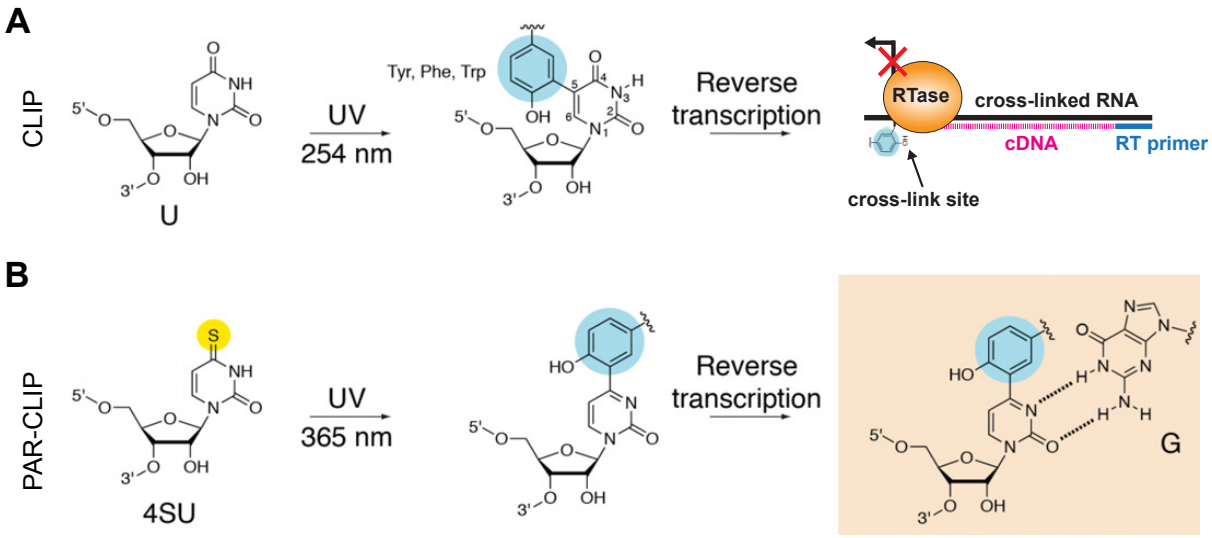

Figure 1.9: **Fundamental distinction between UV cross-linking with 254 and 365 nm** [adapted from Ascano *et al.* (2012)]. UV light cross-linking introduces a covalent cross-link between a nucleoside and an aromatic side chain amino acid (Tyr, Phe, and/or Trp). Whereas the protein is absolutely digested during the CLIP/ PAR-CLIP protocol, the cross-linked adduct remains at the nucleoside. **A**. CLIP uses 254 nm that connects the amino acid to position 5 of the nucleobase uracil (U). During reverse transcription the reverse transcriptase (RTase) stalls at this position, but is also able to skip the cross-link site resulting in a base deletion. **B**. PAR-CLIP uses 365 nm that established a covalent cross-link between the amino acid and the thio group of the 4-thiouridine (4sU). Compared to CLIP, the RTase reads through the position and mistakenly incorporates a guanosine (G) instead of an adenosine (A), which leads to the characteristic T-to-C transitions after cDNA library preparation defining the sites of cross-linking.

1.4.2 Photoactivatable ribonucleoside-enhanced CLIP

Most CLIP-related procedures (Section 1.4.1) use rigorous and stringent washes to biochemically reduce occurring background binding. PAR-CLIP uses photo-reactive ribonucleotides to address this problem (Hafner *et al.*, 2010). These analogues are added to the growth medium, which are then randomly taken up by cells and eventually incorporated into nascent RNAs in a transcriptome-wide manner. Predominantly, the photoactivatable substrates 4-thiouridine (4sU) and 4-thiouracil (4tU) are used for *in vivo* RNA labeling in the human and yeast system, respectively (Ascano *et al.*, 2012; Sun *et al.*, 2012). Similar analogs like 6-thioguanosine (6sG), 5-iodouridine (5iU) and 5-bromouridine (5BrU) have been additionally assayed by Hafner *et al.* (2010) (Figure 1.10), but showed lower cross-linking efficiencies compared to 4sU. Both the efficiency of nucleoside uptake and the potential toxicity varies between cell types (Lozzio and Wigler, 1971). However, incorporation of 4sU or 4tU is restricted to U-containing regions within the transcript.

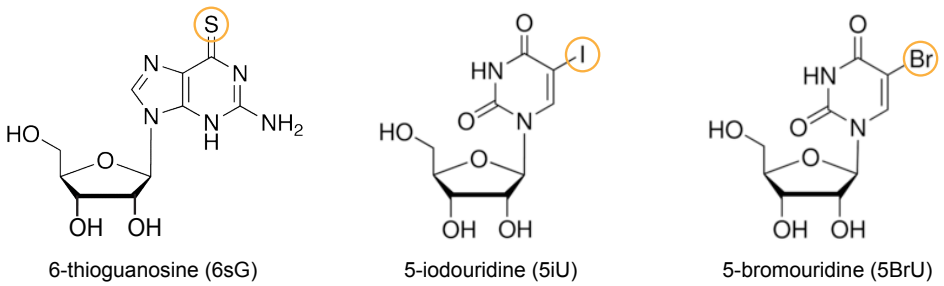

Figure 1.10: **Structures of photoactivatable nucleosides**. The photo-reactive groups are encircled.

Another critical step in the PAR-CLIP procedure is the *in vivo* UV light cross-linking at 365 nm, a long-wavelength where natural nucleotides no longer cross-link. Compared to related CLIP approaches (Section 1.4.1), decreasing the wavelength to 365 nm has three main advantages in relation to 254 nm (CLIP): (i) less UV damage using the same amount of radiation energy (Gaillard and Aguilera, 2008; Ascano *et al.*, 2012), (ii) an improved RNA recovery up to 1000-fold applying photoactivatable nucleosides (Hafner *et al.*, 2010), and (iii) high-resolution binding sites due to PAR-CLIP-specific T-to-C transition, which results from the incorporated base analogue (Hafner *et al.*, 2010; Spitzer *et al.*, 2014). Compared to CLIP, the RTase reads through the site of cross-linking and mistakenly incorporates a guanosine instead of an adenosine, which leads to the characteristic U-to-C transitions during reverse transcription that, when mapped to the genome, manifest themselves as T-to-C mismatches (Figure 1.9B) (Ascano *et al.*, 2012). During the bioinformatics, this transition is used to distinguish between true and false cross-linking events, and enables a much more precise identification of RBP binding sites. Other possible sources of nucleotide mismatches are sequencing errors and differences between the genome sequence of the organisms used in PAR-CLIP experiments and the reference sequence onto which the PAR-CLIP reads are mapped. However, these mismatches can be easily identified and computationally eliminated during the data analysis.

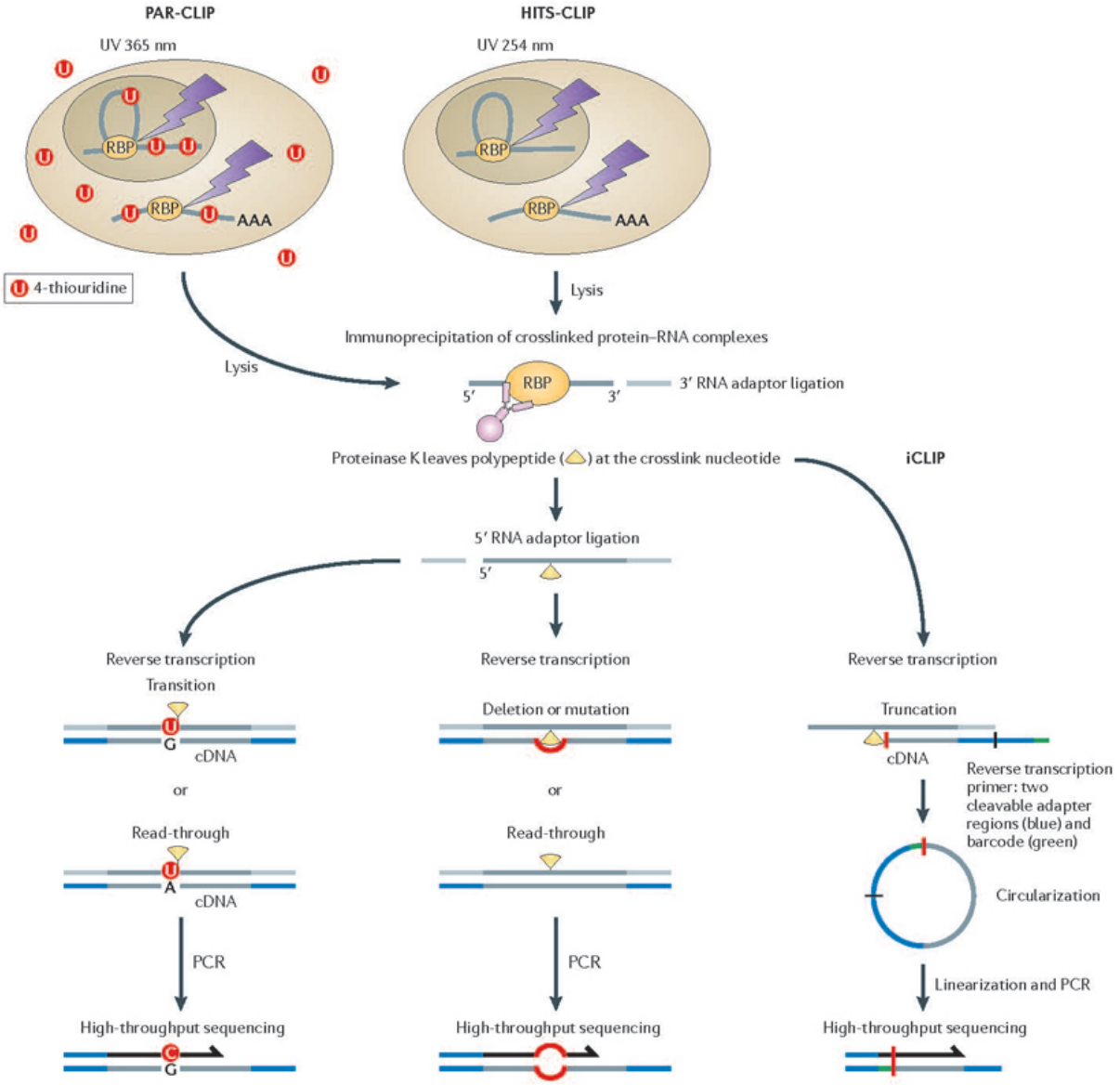

Figure 1.11: **Comparison of the original (HITS-)CLIP with the PAR-CLIP and iCLIP variants** [copied from Knig *et al.* (2011)]. Cells are irradiated with UV light at 254 nm (HITS-CLIP/iCLIP) and 365 nm (PAR-CLIP) that induces the formation of covalent cross-links between RNA-binding protein (RBP) and the RNA (light blue). Photoactivatable ribonucleoside-enhanced (PAR)-CLIP uses photo-reactive ribonucleotides (i.e. 4-thiourdine) to increase cross-linking efficiency (the analogue is depicted as red circle). Cells are lysed, bound RNAs are partially digested, and the protein–RNA complexes are immunoprecipitated. While HITS-CLIP's library preparation is based on a deletion or mutation at the cross-link site (middle panel), individual-nucleotide resolution CLIP (iCLIP) relies on a modified cDNA library preparation based on circularization and frequent reverse transcriptase termination events at the cross-linked site (right panel). Due to the incorporated 4-thiourdine, in PAR-CLIP the reverse transcriptase reads through the site of cross-linking and mistakenly incorporates a guanosine instead of an adenosine, which leads to the characteristic U-to-C transitions during reverse transcription that, when mapped to the genome, manifest themselves as T-to-C mismatches.

1.5 Aims and scope of this thesis

The RNA molecule does not only pass genetic information from DNA to protein, it has crucial roles in biological systems by regulating various biological processes (Ferr-D'Amar and Scott, 2010; Will and Lhrmann, 2011). From an evolutionary point of view, RNA is supposed to be the older, primordial molecule that preceded the contemporary DNA- and protein-based life on Earth (Section 1.1) (Woese *et al.*, 1966; Crick, 1968). Despite the ability of RNA molecules to accomplish sophisticated reactions without protein-based enzymes, modern RNAs are commonly found in association with RNA-binding proteins (Cech, 1986; Moore and Steitz, 2011). Those RBPs associate with RNAs through their RNA-specific binding domains (Section 1.3) (Conte *et al.*, 2000). Especially during mRNA biogenesis, RBPs are involved in pre-mRNA synthesis by RNA polymerase II and co-transcriptional RNA processing (Sections 1.2). Mapping of mRNP biogenesis factors onto pre-mRNA and mature mRNA promises insights into RNA determinants for splicing, 3'-processing, and RNA export, but also into the coupling between these processes. RBPs can in principle be mapped onto the transcriptome by *in vivo* protein-RNA cross-linking and immunoprecipitation (CLIP) (Section 1.4) (Ule *et al.*, 2003), and can indeed provide transcriptome-wide maps for several biogenesis factors.

Consequently, the first aim of this work was the establishment of the recent CLIP technique in the yeast system, referred to as Photoactivatable ribonucleoside-enhanced (PAR)-CLIP (Hafner *et al.*, 2010). It takes advantage of diagnostic T-to-C conversions that occur during reverse transcription as a result of a photo-reactive ribonucleotide being covalently cross-linked to the protein of interest, thus enabling direct protein–RNA interactions to be distinguished from indirect non-cross-linked interactions (Section 1.4.2) (Ascano *et al.*, 2012). For this purpose, several investigations had to be planned and established to finally get a cutting-edge protocol with high resolution (Section 5.1). Especially the RNA labeling and UV light cross-linking represent crucial steps in the PAR-CLIP procedure and were intensively tested and adapted (Section 5.1.1 to 5.1.3).

Because almost no mature computational pipeline for PAR-CLIP data analysis in *S. cerevisiae* was known by this time, a specifically adapted and powerful pipeline for data analysis had to be designed and implemented, allowing faster processing with the highest possible accuracy (Section 5.2). To achieve this goal, parts of this work were performed in cooperation with Phillipp Torkler (AG Söding) and Alexander Graf (AG Blum). To calculate p-values for true cross-linking sites, a null hypothesis had to quantitatively be modeled and tested to finally distinguish between a true T-to-C transition, observed from a real cross-link event, and a false mismatch. This tailored pipeline was supposed to combine both a transparent and powerful tool for analyses of sequencing data obtained from PAR-CLIP experiments and a graphical user interface (GUI) for user-friendly applications.

Not all transcripts are equally expressed during the cell cycle (Xu *et al.*, 2009; Pelechano *et al.*, 2013). In chromatin immunoprecipitation (ChIP) experiments, data is necessarily normalized by using both an input reference and a mock immunoprecipitation control (Mayer *et al.*, 2010). To reduced transcript-to-transcript signal fluctuation, resulting from RNAs with different abundance, a way to normalize CLIP data had to be found and included into the bioinformatics. To achieve this aim, expression levels after RNA labeling and UV treatment should be assayed applying both microarray analysis and RNA-Seq (Section 5.3.1).

However, mRNP biogenesis factors have not been systematically mapped onto pre-mRNAs, likely due to difficulties in trapping short-lived RNAs in cells, and due to the complexity caused by the large variety of pre-mRNA species. Consequently, one initial aim was the clipping and mapping of biogenesis factors onto the newly synthesized yeast transcriptome, including factors that are involved transcription, splicing, and 3'-processing of pre-mRNAs, as well as in the assembly of mature messenger ribonucleoprotein (mRNP) complexes for nuclear export. Resulting data was supposed to provide new insights into mRNP biogenesis and therefore helps to understand how those factors recognize pre-mRNA elements and specific target regions *in vivo*. Thus, the overall impact of the macromolecule RNA and its 'hidden' elements and functions might be decoded, demonstrating the influence and importance of RNAs in coordination with the transcription cycle as well as post-transcriptional processes in the budding yeast *S. cerevisiae*.

2 Materials

2.1 Chemicals

Table 2.1: List of chemicals used in this thesis.

Manufacturer	Chemical	CAS	Order number
Sigma-Aldrich[1]	4-Thiouracil	591-28-6	440736
	4-Thiouridine	13957-31-8	T4509
	6-Thioguanine	154-42-7	A4882
	6-Thioguanosine	345909-25-3	858412
	Adenosine	58-61-7	A9251
	Acetonitrile	75-05-8	34998
	Citric acid	77-92-9	251275
	Cytidine	65-46-3	C122106
	Dithiothreitol	3483-12-3	43815
	Ethanol	64-17-5	32205
	Guanosine	118-00-3	G6752
	NP-40	9036-19-5	NP40S
	Magnesium chloride hexahydrate	7791-18-6	63072
	Sodium acetate	127-09-3	S2889
	Sodium chloride	7647-14-5	S7653
	Sodium deoxycholate	302-95-4	D6750
	Sodium dodecyl sulfate	151-21-3	71736
	Sodium phosphate dibasic	7558-79-4	S3264
	Acetic acid triethylamine	5204-74-0	09748
	Triton X-100	9002-93-1	93443
	Thymidine	50-89-5	T9250
neoLab[2]	Benzamidine hydrochloride	1670-14-0	14830.0005
	Leupeptin hemisulfate	103476-89-7	12005.0025
	Pepstatin A	26305-03-3	11645.0010
	Phenylmethanesulfonyl fluoride	329-98-6	16350.0005

Table 2.1: – continued on next page

[1] Sigma-Aldrich Corporation, St. Louis, MO 63103, USA; website: http://www.sigma-aldrich.com
[2] neoLab Migge Laborbedarf-Vertriebs GmbH, Heidelberg, 69123, Germany; website: http://www.neolab.de

Table 2.1: – continued from previous page

Manufacturer	Chemical	CAS	Order number
Merck[3]	Sodium phosphate dibasic dihydrate	10028-24-7	1065801000
	Potassium dihydrogen phosphate	7778-77-0	1048731000
	Potassium chloride	7447-40-7	1049380500
	Potassium hydroxide	1310-58-3	1050331000
Carl Roth[4]	Dimethyl sulfoxide	67-68-5	7029.1
	Ethylene glycol tetraacetic acid	67-42-5	3054.2
	Glycerol	56-81-5	3783.1
	HEPES	7365-45-9	9105.3
	Ethylenediaminetetraacetic acid	60-00-4	CN06.1
	Skim milk powder	68514-61-4	T145.1
	Tween20®	9005-64-5	9127.1
Serva[5]	Bromophenol Blue Na-salt	34725-61-6	153751

2.2 TAP-tagged strains

The Open Biosystems, Inc. (Thermo Scientific[6]) Yeast-TAP Tagged ORF library was used containing C-terminally tagged proteins of the *Saccharomyces cerevisiae* BY4741 strain. The C-terminal TAP consists of a calmodulin binding peptide, a TEV cleavage site, and two IgG binding domains of *Staphylococcus aureus* protein A. Table 2.2 lists all tagged proteins that were applied in this thesis with systematic name and library coordinate.

Table 2.2: List of TAP-tagged strains used in this thesis.

Tagged protein	Systematic name	Plate	Well
Cbc2	YPL178W	7GS2	E4
Cft2	YLR115W	2GS3	C1
Gbp2	YCL011C	6GS3	D3
Hpr1	YDR138W	4GS4	C3
Hrb1	YNL004W	6GS3	C11

Table 2.2: – continued on next page

[3]Merck KGaA, Darmstadt, 64293, Germany; website: http://www.merckgroup.com
[4]Carl Roth GmbH + Co. KG, Karlsruhe, 76231, Germany; website: http://www.carlroth.de
[5]SERVA Electrophoresis GmbH, Heidelberg, 69123, Germany; website: http://www.serva.de
[6]Fisher Scientific - Germany GmbH, Schwerte, 58239, Germany; website: http://www.thermoscientificbio.com

Table 2.2: – continued from previous page

Tagged protein	Systematic name	Plate	Well
Ist3	YIR005W	12GS4	C7
Luc7	YDL087C	9GS3	G1
Mex67	YPL169C	4GS3	A9
Mpe1	YKL059C	8GS4	A6
Msl5	YLR116W	5GS3	F4
Mud1	YBR119W	8GS3	C11
Mud2	YKL074C	4GS3	G11
Nab2	YGL122C	3GS2	E5
Nab3	YPL190C	1GS2	G12
Nam8	YHR086W	4GS3	H4
Npl3	YDR432W	1GS1	H1
Nrd1	YNL251C	1GS1	D8
Pab1	YER165W	1GS1	D6
Pub1	YNL016W	1GS1	F10
Rna15	YGL044C	6GS2	C2
Snp1	YIL061C	10GS4	A10
Sub2	YDL084W	1GS1	F12
Tho2	YNL139C	1GS4	C2
Yth1	YPR107C	11GS3	A5

2.3 Buffers and Media

Buffers and media used in this thesis are listed in Table 2.3 and Table 2.4, respectively.

Table 2.3: Buffer compositions

Buffer		Composition
PBS buffer (1x)	137 mM	NaCl
	2.7 mM	KCl
	10 mM	Na_2HPO_4
	2 mM	KH_2PO_4
TBE buffer (1x)	89 mM	Tris base
	20 mM	EDTA (pH 8.0)
	8 mM	Boric acid

Table 2.3: – continued on next page

Table 2.3: – continued from previous page

Buffer		Composition
Lysis buffer	50 mM	Tris-HCl (pH 7.5)
	100 mM	NaCl
	0.1 % (v/v)	SDS
	0.5 % (v/v)	NP-40
	0.5 % (v/v)	Na deoxycholate
Wash buffer	50 mM	Tris-HCl (pH 7.5)
	1 M	NaCl
	0.1 % (v/v)	SDS
	0.5 % (v/v)	NP-40
	0.5 % (v/v)	Na deoxycholate
T1 buffer	50 mM	Tris-HCl (pH 7.5)
	2 mM	EDTA
Phosphatase reaction buffer (pH 6.5)	50 mM	Tris-HCl (pH 7.0)
	1 mM	$MgCl_2$
	100 mM	$ZnCl_2$
Phosphatase wash buffer	50 mM	Tris-HCl (pH 7.5)
	20 mM	EGTA
	0.5 % (v/v)	NP-40
Polynucleotide kinase (PNK) buffer	50 mM	Tris-HCl (pH 7.5)
	50 mM	NaCl
	10 mM	$MgCl_2$
Proteinase K buffer	50 mM	Tris-HCl (pH 7.5)
	75 mM	NaCl
	6.25 mM	EDTA
	1 % (v/v)	SDS
SDS-PAGE loading buffer (2x)	100 mM	Tris-HCl (pH 6.8)
	4 mM	EDTA
	20 % (v/v)	Glycerol
	200 mM	DTT
	4 % (v/v)	SDS
	0.02 % (w/v)	Bromophenol blue

Table 2.4: Media compositions

Medium		Composition
YPD medium	1 %	Yeast extract (w/v)
	2 %	Peptone (w/v)
	2 %	Glucose (w/v)
for YPD-Agar	plus 2 %	Agar (w/v)
SC medium	10 mg/l	Adenine, Uracil
	20 mg/l	L-Methionine, L-Histidine HCl, L-Methionine
	50 mg/l	L-Arginine, L-Isoleucine, L-Lysine HCl, L-Tryptophan,
		L-Tyrosine, L-Phenylalanine
	80 mg/l	L-Aspartic acid
	100 mg/l	L-Leucine, L-Threonine, 4-Thiouracil
	140 mg/l	Valine
	6.9 g/l	Yeast Nitrogen Base
	2 %	Glucose (w/v)

2.4 Commercial buffers and Reagent systems (Kits)

Commercial buffers and reagent kits used in this thesis are listed in Table 2.5 and Table 2.6, respectively.

Table 2.5: Commercial buffers.

Manufacturer	Buffer	Order number
Fermentas[7]	T4 PNK Reaction Buffer A	EK0031
	10x TURBO™ DNase buffer	EN0541
Invitrogen[8]	5X First-Strand Buffer	18080-044
NEB[9]	Antarctic Phosphatase Reaction Buffer	M0289S
	Rnal2tr ligase buffer	M0351S
	T4 RNA Ligase 1 Reaction Buffer (10X)	B0204
Sigma-Aldrich[10]	TRI Reagent® RNA Isolation Reagent	T9424

[7] Fermentas GmbH, St. Leon-Rot, 68789, Germany; website: http://www.fermentas.de
[8] Invitrogen GmbH, Karlsruhe, 76131, Germany; website: http://www.invitrogen.com
[9] New England Biolabs GmbH, Frankfurt, 65929, Germany; website: http://www.neb-online.de
[10] Sigma-Aldrich Corporation, St. Louis, MO 63103, USA; website: http://www.sigma-aldrich.com

Table 2.6: Commercial reaction systems (Kits).

Manufacturer	Kit	Order number
Affymetrix[11]	GeneChip® 3' IVT Express Kit	
	GeneChip® Hybridization Kit	
Illumina[12]	MiSeq Reagent Kits v2 (50-cycles)	MS-102-2001
	TruSeq SR Rapid Cluster Kit v3	GD-402-4001
	TruSeq Rapid SBS reagent kit v3 (50-cycles)	FC-402-4002
	cBot Single-Read Cluster Generation Kit	GD-300-1001
	TruSeq SBS reagent kit v2-GA (36-cycles)	FC-104-5001
NuGEN[13]	Encore Complete RNA-Seq Library Systems	0311
	RNA-Seq DR Multiplex System 1-8	0333
PeqLab[14]	KAPAHiFi™ PCR Kit	07-KK2100-01
Qiagen[15]	QIAquick Gel Extraction Kit	28704

2.5 Enzymes

Table 2.7: Enzymes used in this thesis.

Manufacturer	Enzyme	Activity	Order number
Fermentas[16]	RNase T1	1,000 U/ µl	EN0541
	T4 Polynucleotide Kinase	10 U/ µl	EK0031
NEB[17]	Antarctic Phosphatase	5 U/ µl	M0289
	T4 RNA Ligase 2 (K227Q)	200 U/ µl	M0351
	Proteinase K	20 mg/ ml	P8102
	T4 RNA Ligase 1	20 U/ µl	M0204
	Phusion HF DNA Polymerase	20,000U/ µl	M0530
Invitrogen[18]	Bacterial Alkaline Phosphatase	150 U/ µl	18011-015
	RNAse OUT™	40 U/ µl	10777-019
	SuperScript III RT	200 U/ µl	18080-093
	Ambion® TURBO™ DNase	2 U/ µl	AM2238

Table 2.7: – continued on next page

[11] Affymetrix UK Ltd., High Wycombe, HP10 0HH, United Kingdom; website: http://www.affymetrix.com
[12] Illumina, Inc., San Diego, CA, 92101, USA; website: http://www.illumina.com
[13] NuGEN Technologies, Inc., San Carlos, CA 94070, USA; website: http://www.nugen.com
[14] PEQLAB Biotechnologie GMBH, Erlangen, 91052, Germany; website: http://www.peqlab.de
[15] Qiagen N.V., Hilden, 40724, Germany; website: http://www.qiagen.com
[16] Fermentas GmbH, St. Leon-Rot, 68789, Germany; website: http://www.fermentas.de
[17] New England Biolabs GmbH, Frankfurt, 65929, Germany; website: http://www.neb-online.de
[18] Invitrogen GmbH, Karlsruhe, 76131, Germany; website: http://www.invitrogen.com

Table 2.7: – continued from previous page

Manufacturer	Enzyme	Activity	Order number
Sigma-Aldrich[19]	Phosphodiesterase I	≥0.40 U/ vail	P3243

2.6 Oligonucleotide primer

Desalted oligonucleotide primer were mainly obtained from IDT[20]; whereas the common concentration amounted between 25 nmole and 1µmole. For cDNA library preparation the oligonucleotide sequences from the NEXTflex™ Small RNA Sequencing Kit (Bioo Scientific[21], #5132-02) were used. The final Fusion-PCR was performed using the NEXTflex™ Small RNA Barcode Set A (Bioo Scientific, #513301). The entire set of used primer and barcodes respectively is listed in Table 2.8.

Table 2.8: Listed is the designation and sequence of each oligonucleotide primer. Specific modifications at the 5' or 3' end are marked by brackets: [App] = pre-adenylation, [ddC] = Dideoxy-C, and [Phos] = phosphorylation. The position of the actual barcode sequence of the Barcode Fusion Primer is labeled as "Barcode". Used barcode sequences are listed below containing the actual primer sequence (5'-3') as well as the reverse complement (c3'-c5') which is finally sequenced. Furthermore a self-made DNA marker was used for cDNA library preparation. Therefore three ssDNA oligonucleotide PCR templates were designed, resulting in Fusion PCR products with a size of 118 nt, 140 nt and 180 nt. The black square marks the additional insert of the Marker140 and Marker180, respectively.

Designation	Primer sequence
Fwd-Not1-TAP	5' - TTA CAT GTG GGC ATT GAA GC - 3'
Fwd-Nrd1-TAP	5' - AAG ACA TGA GGC CGA AAA TG - 3'
Fwd-Mpt5-TAP	5' - TCA ACC AAA ACG CAT ATC CC - 3'
Rev-universal-TAP	5' - AAC CCG GGG ATC CGT CGA CC - 3'
3' Adapter	5' -[App]- TGG AAT TCT CGG GTG CCA AGG -[ddC]- 3'
5' Adapter	5' -[Phos]- GUU CAG AGU UCU ACA GUC CGA CGA UC - 3'
RT-Primer	5' - CCT TGG CAC CCG AGA TTC CA - 3'
microRNA control	5' -[Phos]- CUCAGGAUGGCGGAGCGGUCU - 3'
Universal Fusion Primer	5' - AAT GAT ACG GCG ACC ACC GAG ATC TAC ACG TTC AGA ...
	... GTT CTA CAG TCC GA- 3'
Barcode Fusion Primer	5' - CAA GCA GAA GAC GGC ATA CGA GA -Barcode- GTG ACT ...
	... GGA GTT CCT TGG CAC CCG AGA ATT CC- 3'
Nextera primer 1	5' - AAT GAT ACG GCG ACC ACC GA - 3'
Nextera primer 2	5' - CAA GCA GAA GAC GGC ATA CGA - 3'

Table 2.8: – continued on next page

[19] Sigma-Aldrich Corporation, St. Louis, MO 63103, USA; website: http://www.sigma-aldrich.com
[20] Integrated DNA Technologies, Inc., Coralville, Iowa 52241 USA; website: http://eu.idtdna.com
[21] Bioo Scientific Corp., Austin, Texas 78744, USA; website: http://www.biooscientific.com/

Table 2.8: – continued from previous page

Designation	Primer sequence		
Barcode 1	5’ - TCGTGAT - 3’	$\longrightarrow$	3’ - ATCACGA - 5’
Barcode 2	5’ -TACATCG - 3’	$\longrightarrow$	3’ - CGATGTA - 5’
Barcode 3	5’ -TGCCTAA - 3’	$\longrightarrow$	3’ - TTAGGCA - 5’
Barcode 4	5’ -TTGGTCA - 3’	$\longrightarrow$	3’ - TGACCAA - 5’
Barcode 5	5’ -TCACTGT - 3’	$\longrightarrow$	3’ - ACAGTGA - 5’
Barcode 6	5’ -TATTGGC - 3’	$\longrightarrow$	3’ - GCCAATA - 5’
Barcode 7	5’ -TGATCTG - 3’	$\longrightarrow$	3’ - CAGATCA - 5’
Barcode 8	5’ -TTCAAGT - 3’	$\longrightarrow$	3’ - ACTTGAA - 5’
Barcode 9	5’ -TCTGATC - 3’	$\longrightarrow$	3’ - GATCAGA - 5’
Barcode 10	5’ -TAAGCTA - 3’	$\longrightarrow$	3’ - TAGCTTA - 5’
Barcode 11	5’ -TGTAGCC - 3’	$\longrightarrow$	3’ - GGCTACA - 5’
Barcode 12	5’ -TTACAAG - 3’	$\longrightarrow$	3’ - CTTGTAA - 5’
Marker118	5’ - CCT TGG CAC CCG AGA ATT CCA ■ GAT CGT CGG ACT GTA ...		
	... GAA CTC TGA AC- 3’		
■ Marker140	5’ - AGC ATG TCA AAT TGA TAA GGC G - 3’		
■ Marker180	5’ - AGC ATG TCA AAT TGA TAA GGC GAT GTA GTC CTT CAA ...		
	... AGT TCG TAA GAC CTC CTG ATT ATG CA- 3’		
RNAdigest-Control	5’ - rCrGrUrArCrGrCrUrGrArA-rUrArGrUrUrUrArArArCrUrGrU - 3’		

2.7 Consumables

Table 2.9: Special consumables that have been used in this thesis.

Consumables	Manufacturer
Amersham Hyperfilms™ ECL	GE Healthcare
NuPAGE® Novex 4-12 % Bis-Tris Mini Gels	Life Technologies Ltd.
PVDF membrane	Carl Roth GmbH + Co. KG
RNA Clean & Concentrator™5	Zymo Research
2 ml FastPrep tubes + Lids	MP
0.5 ml DNA LoBind Tubes	Eppendorf, Hamburg
1.5 ml DNA LoBind Tubes	Eppendorf, Hamburg
2 ml DNA LoBind Tubes	Eppendorf, Hamburg

2.8 Equipment

Table 2.10: Equipment used in this thesis.

	Designation	Manufacturer
Centrifugation	Centrifuge 5810R	Eppendorf, Hamburg
	Centrifuge 5424R	Eppendorf, Hamburg
	Vacuum centrifuge Concentrator plus	Eppendorf, Hamburg
	Sorvall Evolution RC superspeed	Thermo Fisher Scientific GmbH, Ulm
	Rotor SLC-6000	Thermo Fisher Scientific GmbH, Ulm
Electrophoresis	Power Supply EPS 3501	Amersham Pharmacia Biotech Europe GmbH, Freiburg
	Novex Mini Cell	Invitrogen, Karlsruhe
	PerfectBlue™ Horizontal Gel Systems	PEQLAB Biotechnologie GMBH, Erlangen
	Power Supply Model 200/2.0	Bio-Rad Laboratories GmbH, Munich
	Agilent 2100 Bioanalyzer	Agilent Technologies, Waldbronn
	Experion Electrophoresis Station	Bio-Rad Laboratories, Munich
	Trans-Blot® Turbo™ Transfer System	Bio-Rad Laboratories GmbH, Munich
Genomics	Affymetrix GeneChip® Scanner 3000	Affymetrix, England
	GeneChip® Hybridization Oven 640	Affymetrix, England
	HiSeq 1500	Illumina, Inc., San Diego
	Genome Analyzer IIx	Illumina, Inc., San Diego
HPLC	Programmable Solvent Module 126	Beckman System Cold, USA
	Diode Array Detektor Module 168	Beckman System Cold, USA
	Programmable Solvent Module 126	Beckman System Cold, USA
Incubation	T Professional Basic Gradient	Biometra GmbH, Göttingen
	Thermocycler T3000	Biometra GmbH, Göttingen
	Rotating wheel	Labinco B.V., Netherlands
	Rocking table STS Cat	neoLab, Heidelberg
	Rotamax 120	Heidolph Instruments GmbH & Co. KG, Schwabach
	WTC Binder Incubator	BINDER GmbH, Tuttlingen
	Barnstead Lab-Line Titer Plate Shaker	Thermo Fisher Scientific GmbH, Ulm
	Thermomixer comfort 5436	Eppendorf, Hamburg
	Magnetic stirrer MR 3001	Heidolph, Schwalbach

Table 2.10: – continued on next page

Table 2.10: – continued from previous page

	Designation	Manufacturer
Imaging	White Light Transilluminator TW-26	UVP, Cambridge, England
	Transilluminator	Intas-Science Imaging Instruments GmbH, Göttingen
	Gel Documentation System	Intas-Science Imaging Instruments GmbH, Göttingen
	Epson perfection 3200 photo Scanner	EPSON Deutschland GmbH, Meerbusch
	Image Eraser	GE Healthcare Europe GmbH, Freiburg
	Storm 860	GE Healthcare Europe GmbH, Freiburg
	Microscope DMLS	Leica Mikrosysteme Vertrieb GmbH, Wetzlar
	Exposer Cassette	GE Healthcare Europe GmbH, Freiburg
Measurement	pH-Meter 766 Calimatic	Knick, Berlin
	NanoDrop 1000	PEQLAB Biotechnologie GmbH, Erlangen
	AC 210 P Analytical balance MC1	Sartorius, Göttingen
	LC 2200 P Analytical balance MC1	Sartorius, Göttingen
	Dial-O-gram balance	Ohaus Europe GmbH, Swiss
	Qubit fluorometer	Invitrogen, Karlsruhe
	BioPhotometer	Eppendorf, Hamburg
Miscellaneous	Experion Priming Station	Bio-Rad Laboratories, Munich
	Vortex Mixer	neoLab, Heidelberg
	Microwave	Siemens AG, Munich
	Crosslinker Bio-Link BLX-365	PEQLAB Biotechnologie GMBH, Erlangen
Lysis	FastPrep-24	MP Biomedicals Europe, France
	Bioruptor Next Generation Sonication	Diagenode, Lige
	Mixer Mill MM 400	Retsch, Haan
	Bioruptor Water Cooler	Diagenode, Lige

3 Biochemical methods

3.1 Cultivation of *Saccharomyces cerevisiae*

Cultivation of *S. cerevisiae* BY4741 strains was performed at 30 °C and 160 rpm using either Yeast Extract Peptone Dextrose (YPD) or Synthetic complete (SC) medium supplemented with 2 % glucose. Therefore, cells were plated on YPD agar and cultivated at 30 °C for 1-2 days. One colony was subsequently used to inoculate a 30 ml YPD pre-culture. Cell density was photometrically determined using a spectrometer at 600 nm. One optical density unit (OD_{600}) corresponds to ~2.5 x 10^7 yeast cells.

3.2 TAP-tag validation by western blot analysis

Yeast strains containing TAP-tagged genes were tested for expression of the correct tagged protein by western blotting. Therefore, the cell lysate of a 50 ml YDP culture was diluted 1:100 in lysis buffer, mixed with one volume of 2x SDS loading buffer and incubated for 3 min at 95 °C. Subsequently, the sample was spun for 30 sec at 500 rpm, immediately loaded and run on a pre-cast NuPAGE 4-12 % Bis-Tris gel (Invitrogen). Following SDS-PAGE, samples were blotted onto a 0.2 µm polyvinylidene difluoride (PVDF) membrane provided in the Trans-Blot® Turbo™ Transfer Pack by using the Trans-Blot® Turbo™ Transfer System (Bio-Rad Laboratories, Inc.). The membrane was then blocked at room temperature for 30 min in 20 ml PBS buffer with 5 % non-fat dry milk and 0.1 % Tween20. Five microliter of the primary PAP antibody (Sigma-Aldrich) were diluted in 10 ml PBS buffer + 2 % non-fat dry and incubated with the membrane for another hour. After three brief washing steps with 20 ml PBS buffer each, the membrane was incubation in 20 ml PBS + 0.1 % Tween20 for 10 min, and finally rinsed with fresh tap water. Antibody detection was performed using Pierce enhanced chemiluminescence (ECL) Western blotting substrate (Thermo Scientific) in combination with Amersham Hyperfilm™ ECL (GE Healthcare).

3.3 PAR-CLIP

3.3.1 RNA labeling with 4-thiouracil

S. cerevisiae cells expressing the TAP-tagged protein were grown at 30 °C and 160 rpm from OD_{600} of 0.1 to OD_{600} of 0.5 in one liter of CSM minimal medium (Formedium) supplemented with 10 mg/l uracil, 100 μM 4-thiouracil (4tU) and 2 % glucose. After reaching OD_{600} of 0.5, another 900 μM 4tU were added and cells were grown further for 4 h (OD_{600} of 1.3 till 1.6).

3.3.2 UV light crosslinking

Following RNA labeling, cells were harvested at 3,000 rpm and 30 °C for 5 min, resuspended in 30 ml of ice-cold Phosphate-buffered saline (PBS) buffer, and immediately UV-crosslinked. Therefore, living cells were transferred onto a sterile 15 cm tissue culture plate, irradiated on ice with 365 nm UV light in a Bio-Link BLX-365 (Vilber Lourmat), applying an energy dose of 10 till 12 J/cm^2 and a continuous shaking at 50 rpm. Subsequently, cells were collected and stuck cells were dislodged by washing plate with additional 10 ml of PBS buffer. The pooled sample was spun at 3,000 rpm for 3 min and harvested cells were either directly lysed (as described in Section 3.3.3) or shock-frozen (in liquid nitrogen) and stored at -80 °C.

3.3.3 Cell lysis and Sonication

Fresh or frozen yeast pellets were resuspended in a total of 4 ml ice-cold Lysis buffer containing pre-mixed protease inhibitors (1 mM Leupeptin, 2 mM Pepstatin A, 100 mM Phenylmethylsulfonyl fluoride, 280 mM Benzamidine) and separated into two 2 ml FastPrep tubes. Pre-cooled Zirconia beads (0.5 mm ⌀ , already stored at -20 °C) were added until the tube was filled almost completely and cells were lysed 8 times for 40 sec in the FastPrep-24 with cooling on ice for 2 min in between. Samples were then pooled into 1.5 ml TPX microtubes and furthermore solubilized for 4 min by sonication using the Bioruptor™ UCD-200 (Diagenode, Inc.) at high intensity and 30 sec on/off intervals. Treated cell lysate was subsequently cleared by centrifugation at 13,000 rpm and 4 °C for 30 min. The upper phase was carefully collected and used for further steps, i.e. western blotting (Section 3.2) or immunoprecipitation (Section 3.3.4). Alternatively, cell lysate was shock-frozen (in liquid nitrogen) and stored at -80 °C.

3.3.4 Immunoprecipitation

Initially, ten milligram Protein G Dynabeads® [~330 µl] (Invitrogen) were washed twice in 400 µl PBS buffer. Beads were then incubated with 10 µg rabbit serum IgGs (Sigma) per mg beads in 1 ml PBS buffer for 45 min at room temperature on a rotating wheel. Unbound antibodies were removed through trine washing with 1 ml of PBS buffer. Co-immunoprecipitation of tagged proteins to IgG-conjugated beads was performed on a rotating wheel for 2 h at 4 °C (cold room). Beads were collected in a low-binding tube for at least 5 min and subsequently washed twice in 400 µl Wash buffer. Beads were resuspended in 400 µl T1 buffer and stored on ice until proceeding (see Section 3.4).

3.4 Data acquisition

3.4.1 Partial RNase digest and Phosphorylation

After adding T1 buffer containing 50 U of RNase T1 per ml, the bead suspension was incubated for 20 min at 25 °C and 400 rpm. Beads were then washed twice in T1 buffer and once in Phosphatase reaction buffer. For dephosphorylation, Antarctic phosphatase reaction buffer (NEB) with 1 U/ml of Antarctic phosphatase and 1 U/ml of RNase OUT (Invitrogen) were added and the suspension was incubated at 37 °C for 30 min and 800 rpm. Beads were subsequently washed once in Phosphatase wash buffer and twice in Polynucleotide kinase (PNK) buffer. The phosphorylation reaction was performed in T4 PNK reaction buffer A (Fermentas) with a final concentration of 1 U/ml T4 PNK and 1 U/ml RNase OUT with either 1 mM ATP per ml (”cold labeling”) or 0.5 mCi γ-32-P-ATP per ml (”hot labeling”). The reaction mix was incubated for 30 min at 37 °C and 800 rpm, finally washed four times in 400 µl PNK buffer, and stored on ice until proceeding (see Section 3.4.2).

3.4.2 On-bead adapter ligation

For 3’ adaptor ligation, beads were resuspended in T4 RNA ligase buffer (NEB) containing 10 U/ml T4 RNA ligase 2 (K227Q) (NEB), 3’ Adaptor [5 mM] (IDT), 1 U/ml RNase OUT (Invitrogen), and 15 % (w/v) PEG 8000. The bead suspension was incubated for at least 18 h at 16 °C and 400 rpm. Beads were then washed four times in PNK buffer to remove unligated adapters. For 5’ adaptor ligation, beads were resuspended in T4 RNA ligase buffer containing 2 U/ml T4 RNA ligase 1 (NEB), 10 mM 5’ Adaptor (IDT), 1mM ATP, 1 U/ml RNase OUT (Invitrogen), 5 % (v/v) DMSO, and 10 % (v/v) PEG 8000. This reaction mix was incubate for 3.5 h at 16 °C followed by another 30 min at 37 °C. Beads were subsequently washed twice in PNK buffer and proteinase K buffer.

3.4.3 RNA recovery and Ethanol precipitation

Following adapter ligations, magnetic beads were boiled in 90 µl Proteinase K buffer for 5 min at 95 °C to eluted the RNA-protein complexes. The buffer was transferred into a low-binding tube and the procedure was repeated. After pooling both, 1.5 mg/ml Proteinase K (NEB) were added and digestion was performed for 2 h at 55 °C. The released RNA was recovered by acidic phenol/chloroform extraction using one volume Roti-Aqua Rhenol (Roth) and 1/5 volume chloroform, followed by an overnight ethanol precipitation at -20 °C, supported by addition of 1 µl GlycoBlue (Invitrogen) and 100 mM of the Reverse transcription (RT) primer (IDT). Subsequently, precipitated RNA was recovered for 1 h at 15,000 rpm and 4 °C. The RNA pellet was washed in 800 µl of 75 % ethanol and carefully air-dried for approximately 10 till 15 min. Recovered RNA was either stored at - 80 °C (for at least 3 months) or dissolved in a total of 12 µl DNase/ RNase-free dH_2O and directly used for First-Strand cDNA Synthesis (Section 3.4.4).

3.4.4 Reverse transcription

After adding one microliter of 10 mM dNTP mix (Fermentas) to previously precipitated RNA (see Section 3.4.3), the sample was heated at 95 °C for 1 min, and then immediately incubated at 50 °C for 5 min. Reverse transcription (RT) was initially performed at 44 °C for 1 h in First-Strand cDNA Synthesis cocktail (Invitrogen) containing 1x First-Strand Buffer, 5 mM DTT, 2 units/µl RNase OUT, and 10 units/µl SuperScript® III RTase. Subsequently, the temperature was increased to 55 °C and incubation was proceeded for another 30 min. The RT reaction was finally stopped by heating at 95 °C for 5 min and samples were stored on ice or at -20 °C until amplification.

3.5 Barcoded library generation and Sequencing

3.5.1 Fusion PCR

Initial amplification of *de-novo* transcribed cDNA was performed using the Phusion HF master mix (NEB) containing 0.2 mM of each dNTP and 20 units/ml phusion polymerase. Illumina-specific adapter input and barcoding was done using 250 nM of the NEXTflex universal primer in combination with an equimolar barcode primer (listed in Table 2.8). Applied PCR cycle conditions are listed in the following Table 3.1.

Table 3.1: Cycle conditions for initial cDNA amplification (Fusion PCR)

	Step	**Temperature**	**Time**
1	Initial denaturing	98 °C	120 sec
2	Denaturing	98 °C	15 sec
3	Annealing	60 °C	30 sec
4	Elongation	72 °C	25 sec
Repeat steps 2-4 (29 times)			
5	Final elongation	72 °C	5 min
6	Hold	10 °C	paused

After PCR, amplified cDNA (aDNA) was purified and size-selected on a precast 4% High-ReSolution agarose E-Gel® (Invitrogen) using a self-made size marker. Amplificates were gel-isolated and purified applying the QIAquick MinElute Gel Extraction Kit (Qiagen) according to the manufacturer's instructions.

3.5.2 One-Cycle-PCR

Concatemers and other PCR artifacts in the previously generated aDNA were eliminated through an additional PCR cycle, referred to as One-Cycle-PCR. Therefore, the KAPAHiFi™ PCR Kit (Peqlab Biotechnologie GmbH) was used containing a final concentration of 0.3 mM dNTP mix, 0.02 units/µl KAPAHiFi™ DNA polymerase, and 200 nM of Nextera primer 1 and 2 in 1x KAPAHiFi™ buffer. For proper ampli-

fication, 20-40 ng of purified aDNA were used. After denaturing the sample at 94 °C for 200 sec, the primer annealing was carried out at 55 °C for 30 sec, followed by an extended elongation step at 72 °C for 240 sec. PCR products were subsequently purified using AMPure XP beads (Beckman Coulter, Inc.) according to a standard protocol and finally eluted in 10 µl 10 mM Tris-HCl (pH 7.2).

3.5.3 Illumina sequencing

Following quantitation on an Agilent DNA 1000 Chip using a Agilent 2100 Bioanalyzer (Agilent Technologies, Inc.) according to the manufacturers instructions, cluster generation was performed on Illumina's Cluster Station using TruSeq SR Cluster Kit v5-CS-GA containing the Flow Cell v4. Read and index sequencing for multiplexed runs was performed using either Illumina's Genome Analyzer IIx or HiSeq 1500 sequencer. All steps regarding the deep sequencing were done by Stefan Krebs (LAFUGA, Gene Center Munich); according to the manufacturer's instructions.

3.6 Global expression profiling

3.6.1 Microarray analysis

Yeast cells were labeled with different concentrations of 4-Thiouracil and subsequently UV-irradiated (as described in Sections 3.3.1 and 3.3.2). Cells were harvested, resuspended in RiboPure lysis reagents (Invitrogen), and disrupted by bead beating at 4 °C using silica-zirconia beads (Roth) and the FastPrep-24 (MP). Total RNA was extracted by acid phenol/chloroform extraction using Roti-Phenol (Carl Roth) and Ethanol precipitated. DNase I treatment was performed using the RiboPure Yeast Kit (Invitrogen). The concentration and purity of isolated RNAs were determined applying the Experion RNA StdSens Analysis Kit (Bio-Rad). Target preparation on Affymetrix' industry-standard, 3'-expression arrays were performed using the GeneChip 3' IVT Express Kit (Affymetrix) according to the manufacturer's instructions.

3.6.2 RNA-Seq for global RNA abundance normalization

Yeast cells were treated as for PAR-CLIP using the identical labeling conditions and a UV-light (365 nm) energy dose of 1 J/cm^2 (Sections 3.3.1 and 3.3.2). After bead beating, total RNA was isolated by acid phenol/chloroform extraction using Roti-Phenol (Carl Roth), and purified and concentrated using the RNA Clean Concentrator-5 (Zymo Research). Purified RNA was depleted of ribosomal RNAs using Ribo-Zero rRNA removal kit (Epicenter). The resulting rRNA-depleted RNA was used for multiplexed RNA-Seq library preparation using the NuGEN Encore Complete RNA-Seq Library Systems. Libraries were qualified on an Agilent Bioanalyzer 2100 (Agilent Technologies) and sequenced on an Illumina MiSeq machine.

3.7 Determination of 4tU-incorporation level

Quantification of the nucleotide composition of total RNA after the labeling with 4-thiouracil using enzymatic digestion and HPLC analysis was essentially as described (Andrus and Kuimelis, 2001). Therefore, 50 ml cultures were grown at same conditions as descibed in section 3.3.1. Cells were harvested at 3,000 rpm for 5 min at 4 °C, equilibrated to a comparable amount of cells, and washed with 1 ml of ice-cold PBS buffer.

3.7.1 Isolation of total RNA

Total RNA was isolated by treating 4tU-labeled cells in 1 ml TRI reagent (Sigma-Aldrich) at 25 °C for 5 min. After adding 200 µl chloroform and approximately 500 µl Zircona beads, the samples were lysed by milling at 30 Hz and 4 °C for 10 min using the mixer mill MM 400 (Retsch). Samples were then spun at 15,000 rpm for 15 min, and the aqueous phase, containing the total RNA, was transferred into a new low-binding tube, already supplemented with 1 µl GlycoBlue (Invitrogen). Precipitation was performed with 500 µl 2-propanol at 25 °C for 15 min. The RNA was recovered at 15,000 rpm and 4 °C for 30 min. Pelleted RNA was then washed with 1 ml of 75 % Ethanol (EtOH), air-dried for approximately 10 till 15 min, and subsequently dissolved in 30 µl dH_2O at 55 °C for 10 min. RNA was purified and concentrated using the RNA Clean Concentrator-5 to obtain high quality total RNA ($A_{260/A280} > 1.9$, $A_{260/230} > 1.8$). The RNA concentration was finally measured by using the NanoDrop 1000 spectrophotometer.

3.7.2 Enyzmatic ribonucleoside hydrolysis

Twenty microgram of isolated and purified RNA were used for subsequent dephosphorylation and enzymatic hydrolysis to single ribonucleosides. For that reason, the RNA was incubated at 37 °C for 18 h in a 40 l digestion mixture consisting of 18 mM MgCl, 47 mM Tris-HCl (pH 7.5), 0.045 U Snake Venom Phosphodiesterase (PDE), and 0.23 U Bacterial Alkaline Phosphatase (BAP). Remained RNA fragments were then twice precipitated at -80 °C for 15 min using 3 M sodium acetate (pH 5.2) and ice-cold absolute ethanol with subsequent spinning at 15,000 rpm and 20 °C for 5 min. The supernantant was then transfered into a new 1.5 ml tube and evaporated to complete dryness using a SpeedVac (Eppendorf) at 45 °C for approximately 1 h. Dried samples were finally dissolved in 60 µl HPLC buffer A, consisting of 1 % acetonitrile and 50 mM triethylammonium acetate.

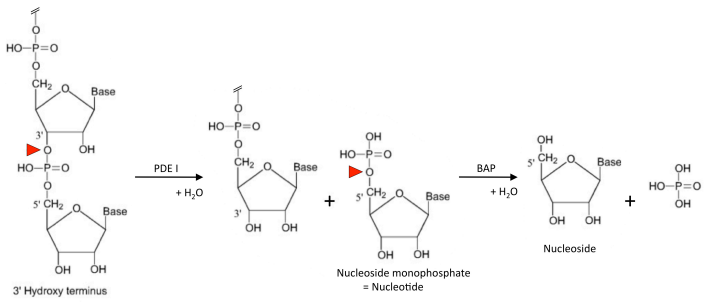

Figure 3.1: Enzymatic reactions of RNA digestion for HPLC analyses. Phosphodiesterase I (PDE) from *Crotalus adamanteus* venom catalyzes the exonucleolytic cleavage of phosphodiester bonds in the 3'- to 5'-direction (marked by red triangel). Resulting nucleoside monophosphates are dephosphorylated by Bacterial Alkaline Phosphatase (BAP), which finally leads to single nucleosides that are analyzed by HPLC.

3.7.3 HPLC analysis

Following the enyzmatic hydrolysis, the obtained ribonucleotide mix was separated on a Supelco Discovery C18 reverse phase column with bonded phase silica 5 µM particles (250 x 4.6 mm) connected to the Programmable Solvent Module 126 (Beckman System Gold). Sample injection (commonly 50 µl) and detection was done using an Autosampler 507 and Diode Array Detektor Module 168, respectively. Ribonucleotide mix separation was performed applying an isocratic gradient with 90 % Acetonitrile (HPLC buffer B), a flow rate of 1 ml per minute, and a maximal pressure of 6 kpsi. After each run, the column was washed in Buffer B for 15 min and finally equilibrated in Buffer A for 5 min.

3.7.4 Calculation of incorporation levels

Determination of integrated areas was done using the software 32 Karat (version 7.0) at 260 nm, 330 nm, and a bandwidth of 5 nm per wavelength. Before calculation of the base composition, each base-specific integrated area under the curve (AUC) was divided by its appropriate cofactor to get an extinction-corrected AUC (cAUC). Therefore the molar extinction coefficient (ε) of each nucleoside at a specific concentration (c) had to be determined for both wavelengths in a 1.2 cm cuvette (= d) (Equation 3.1). To calculate the cofactor, each ε was normalized by dividing it by the smallest value (Equation 3.2). The substitution ratio was finally calculated by dividing the cAUC of 4sU by the sum of both the cAUC of rU and r4sU at 260 nm and 330 nm, respectively.

$$\varepsilon\ [\mathrm{M^{-1}cm^{-1}}] = \frac{E}{c\ [\mathrm{M}] \times d\ [\mathrm{cm}]} \tag{3.1}$$

$$\mathrm{Cofactor} = \frac{\varepsilon_{1\ldots4}}{\varepsilon_{\mathrm{min}}} \tag{3.2}$$

4 Bioinformatical analyses

4.1 Sequencing data quality control and mapping

A script for pre-processing of sequencing data obtained from the Illumina GaIIx or Illumina HiSeq machine was designed and written. The script calls widely used NGS-software and custom scripts with parameter settings adapted to PAR-CLIP analysis. Initially, adapter sequences are first trimmed from the raw sequencing files. The quality filter then discards all reads containing unidentified nucleotides (N), Phreds scores below 30, reads shorter than 15 nt, or reads that are flagged by Illumina's internal chastity filter. Quality-trimmed reads are aligned to the *S. cerevisiae* genome (sacCer3, version 64.1.1) using the short read aligner Bowtie (version 0.12.7) (Langmead, 2010) with a maximum of one mismatch and taking unique matches only (options: -q -p 4 -S -sam -nohead -v 1 -e 70 -l 28 -y -a -m 1 -best -strata -phred33 -quals). The resulting SAM files are then converted into BAM and PileUp files using SAMTools (Li *et al.*, 2009).

4.2 Calculation of P-values and false discovery rates for factor binding sites

The P-value calculation was performed for any genomic T site, given the total number of reads covering the site ("coverage") and given the fraction of these reads which show the T-to-C mismatch. Owing to the exquisite sensitivity of the experimental PAR-CLIP procedure, a very stringent P-value cut-off of 0.005 and a minimum coverage threshold of 4 was set.

To estimate the false discovery rate (FDR) at threshold P < 0.005 for each PAR-CLIP experiment, the number of T sites with T-to-C mismatches among the reads, N_{T-C}, and the number of C sites with C-to-T transitions among the reads, N_{C-T}, was counted. With the latter, the N_{FP}, the part of N_{T-C} that was not due to crosslinking, which yielded FDR was estimated (Equation 4.1).

$$FDR = \frac{NFP}{NT-C} \tag{4.1}$$

4.3 Computation of occupancy profiles

For true cross-linking sites passing our stringent thresholds, the PAR-CLIP-induced T-to-C transitions strongly dominate over the contributions by sequencing errors and SNPs. Therefore, for any given T site in the transcriptome, the number of reads showing the T-to-C transition, N_{TC}, is proportional to the occupancy of the factor on the RNA. But it will also be proportional to the concentration of RNAs covering

the T site. This concentration was estimated from the RNA-Seq read coverage measured under comparable conditions (3.6.2). To reduce noise, the read coverage $N_{RNA-Seq}$ was estimated by smoothing with a running mean of window size 30, therefore:

$$\text{Factor occupancy} \propto \frac{N_{TC}}{N_{RNA-seq}} \tag{4.2}$$

The unscaled factor occupancy was then smoothed, again with a running mean over 30 nucleotides. To fix the scale, a 97 % quantile of occupancy for the called sites (with P-value < 0.005) was set to 100 %. Hence an occupancy was obtained, relative to the 97 % quantile, but it will be a reasonable estimate for absolute occupancy under the assumption that the best 3 % of binding sites are nearly fully occupied on each transcript and that cross-linking efficiency for a given factor does not depend strongly on sequence context. The latter assumption may be violated in genomic regions depleted of T residues.

4.4 Derivation of precise TSS and pA gene annotations

To annotate TSS and pA sites, the recent TIF-Seq data from (Pelechano et al., 2013) was used which yielded much sharper sequence features around TSS and pA sites than the previous TSS and pA annotation data from (Xu et al., 2009) and also greatly improved the resolution of many PAR-CLIP-derived occupancy peaks with respect to these features. Therefore, the original S1_TIFs.txt with around 1.8 million transcript start and end positions (unique transcript isoforms) was downloaded using only selected entries with the annotation ”covering one intact ORF”, yielding approximately one million transcripts. For each gene identifier the most abundant transcript isoform under the YPD condition was picked. Finally, TSS and pA site were annotated for each gene according to those of the most dominant isoforms, giving precise TSS and pA positions for 5578 gene transcripts.

4.5 Occupancy profiles for all genes or introns

All transcripts from the filtered TIF-Seq annotation (see above) were sorted by length and aligned at their TSS. Smoothed occupancies were binned in cells of 20 nucleotide positions times 10 transcripts to avoid aliasing effects due to limited resolution of the plots. The color code displays the occupancy of the PAR-CLIPped factor (with the 97 % quantile of these bins scaled to 1). All introns (obtained form SGD annotation) with lengths between 150 and 600 nucleotides were aligned at the 5’-splice site and the occupancy of each intron was displayed without binning in either x or y direction.

4.6 Motif searches with XXmotif results

To find binding motifs for the investigated factors, the 2000 binding sites with the highest occupancies were selected. Sequence regions ±25 nucleotides around the cross-linked position to XXmotif to de novo motif discovery were submitted, using parameters "-negSet -zoops -merge-motif-threshold LOW -max-match-positions 10". The negative set submitted to XXmotif was composed of 1000 regions of 51 nt length randomly selected from the yeast transcriptome and carrying no significant binding site.

4.7 Calculation of the 'splicing index'

A selection of 245 verified introns out of the SGD yeast annotation (64.1.1) was used to construct a sequence file containing exon-intron (EI), intron-exon (IE) and exon-exon (EE) junctions. Pre-processed sequencing data were mapped to these sequences, and for each of the I = 245 introns, the read counts at the exact EI, IE and EE junctions (N^{EI}, N^{IE}, N^{EE}, respectively) were used to calculate the log-ratio of spliced and unspliced mRNA, which was defined here as 'splicing index' (SI):

$$SI = \log_2 \frac{2\sum_{i=1}^{I} N_i^{EE}}{\sum_{i=1}^{I}(N_i^{EI} + N_i^{IE})} \tag{4.3}$$

4.8 Calculation of the 'processing index'

Read counts N^{down} downstream of a pA site can only occur from pre-mRNAs, $N^{down} = N^{prem}$, whereas read counts N^{up} upstream of a pA site are a mixture of mature mRNA counts N^{mat} and pre-mRNA counts N^{prem}. Therefore, $N^{up} = N^{mat} + N^{prem}$. For increased robustness with regard to different transcript isoforms and uncertainties in the exact location of pA sites, N^{up} and N^{down} were computed as average of the read counts within 50 nt upstream and downstream NPM of the pA site over all G gene transcripts, respectively. The 'processing index' (PI) was defined as follows:

$$PI = \log_2 \left(\frac{1}{G} \frac{\sum_{i=1}^{G} N_i^{PM}}{\sum_{i=1}^{G} N_i^{M}} \right) \tag{4.4}$$

4.9 Binding profile correlation matrix

For each factor f and all transcripts t between 300 and 5000 nt length, the occupancies in the region between the TSS and the pA site were rescaled to an equal length of 300 bins. In this way, each transcript has a resized profile $p^{f,t}$, where $p^{f,t}{}_i$ is the occupancy of factor f at transcript t at location bin i $\in$ {1,...,300}. Next, the mean occupancy per transcript was calculated and assigned each $p^{f,t}$ to one of 10 equal-sized

quantiles (deciles). For each of these 10 deciles d the resized profiles $p^{f,t}$ were summed up to obtain the whole decile average occupancies which resulted in averaged binding shapes $p^{f,d}{}_i$ for each factor for each decile d. For each pair of factors f and f' and each decile d, the Pearson correlation was computed between their binding profile shapes $p^{f,d}{}_i$ and $p^{f',d}{}_i$ as a measure of the similarity of their binding profiles (Equation 4.5).

$$cor(f,f') = \frac{\sum_I (p_i^{f,d} - \overline{p}_i^{f,d})(p_i^{f',d} - \overline{p}_i^{f',d})}{\sqrt{\sum_i (p_i^{f,d} - \overline{p}_i^{f,d})^2 \sqrt{\sum_i (p_i^{f',d} - \overline{p}_i^{f',d})^2}}} \tag{4.5}$$

4.10 Total co-occupancy matrix

To calculate the tendency of pairs of factors to co-occupy similar subsets of transcripts, the pairwise Pearson correlations of their total occupancies $z^{f,t}$ over all transcripts t was computed as follows:

$$cor(f,f') = \frac{\sum_t (z^{f,t} - \overline{z}^{f,t})(z^{f',t} - \overline{z}_i^{f',t})}{\sqrt{\sum_t (z^{f,t} - \overline{z}_i^{f,t})^2 \sqrt{\sum_t (z^{f',t} - \overline{z}^{f',t})^2}}} \tag{4.6}$$

Furthermore, noise was reduced by weighting up/ down the contribution of binding sites to the total occupancy of a transcript that were typical/ atypical of the binding location of the factor within a transcript. The averaged binding profile $p^f{}_i$ of each factor was calculated and from it the normalized weights $w^f{}_i = p^f{}_i / \sum_i p^f$. These weighted the occupancies along each transcript according to how likely they are to be functional. The weighted total occupancy $z^{f,t}$ of a factor f at transcript is therefore computed as follows:

$$z^{f,t} = \sum_i^{|p^{f,g}|} p_i^{f,t} w_i^f \tag{4.7}$$

4.11 Local co-occupancy map

To calculate the tendency of pairs of factors A and B to bind locations in the transcriptome near to each other, the average occupancy of factor B within $\pm$12 nt of occupancy peaks of factor A (unsmoothed occupancy data) was computed. To suppress statistical noise, only peaks of A above the 75 % quantile of all peaks occupancys of A were selected. The average occupancy of B was divided by the background occupancy of B, which was calculated by averaging the occupancy of B within 25 nt windows out of 2000 randomly selected positions in the transcriptome.

5 Results and Discussion

5.1 A high resolution PAR-CLIP procedure for *S. cerevisiae*

We used PAR-CLIP to study direct RNA-protein interactions on a global scale at high-resolution (Methods). For this purpose, we adapted the PAR-CLIP protocol, originally developed by Tom Tuschl and colleagues (Hafner *et al.*, 2010), to the yeast system as previously described (Creamer *et al.*, 2011; Tuck and Tollervey, 2013). Briefly, photoactivatable-ribonucleoside-enhanced (PAR)-CLIP relies on an *in vivo* incorporation of a photoreactive ribonucleoside analog into nascent RNA transcripts that enhances the UV light cross-linking efficiency applying a less energetic wavelength of 365 nm (Hafner *et al.*, 2010). Before the *in vivo* cross-linking introduces a covalent bond between the protein of interest and the applied base analogue, it is crucial to incorporate as much as possible of a photoactivatable ribonucleoside into the nascent RNA without perturbing the cell or changing expression levels. Due to the inability of *Saccharomyces cerevisiae* to incorporate the photoactivatable ribonucleoside 4-thiouridine (4sU) or similar analogs (6-thioguanosine, 5-iodouridine or 5-bromouridine), the nucleobase analogue 4-thiouracil (4tU) was used instead (Hafner *et al.*, 2010; Creamer *et al.*, 2011; Sun *et al.*, 2012; Tuck and Tollervey, 2013).

Consequently, several investigations were planned and established to finally get a cutting-edge protocol with high resolution. First, we measured the growth changes during 4tU-labeling in *S. cerevisiae* and determined resulting *in vivo* incorporation levels of 4sU by HPLC analysis (Hafner *et al.*, 2010; Andrus and Kuimelis, 2001). Second, different culture media, a wide variety of base analogue concentrations, and various labeling periods have been assayed and compared to determine the optimal condition for subsequent PAR-CLIP experiments in *S. cerevisiae*. Third, we successfully optimized further crucial steps including cross-linking efficiency and cell lysis as well as data acquisition.

5.1.1 Labeling efficiency depends on 4tU concentration and labeling time

We measured the maximal absorbance (λ_{max}) of each applied nucleoside and calculated their specific molar extinction coefficients (Table 5.1). The ribonucleotides Adenosine (rA) and Uridine (rU) as well as the deoxyribonucleotide Thymidine (dT) had their λ_{max} at around 260 nm, whereas Cytidine (rC) and Guanosine (rG) showed a shift to 271 nm and 253 nm (UVC), respectively. The base analogue 4sU absorbed maximally around 345 nm (UVA), but also displayed an absorbance at ~250 nm. Importantly, an absorbance above 300 nm was only observed for 4sU. Following, the absorbance at 260 nm and 330 nm for each nucleoside was measured and arithmetically averaged using different molar concentrations (10 μM, μ100 M and 1 mM). Individual molar extinction coefficients and resulting cofactors were subsequently calculated (as described in Section 3.7.4). While rC showed the lowest coefficient for the common nucleosides, an almost doubled value was assayed for rA, resulting in a cofactor of 1.0 and 1.98 for rC and rA, respectively (Table 5.1). As expected, molar extinction coefficients at 330 nm were only determined for the base analogue 4sU.

Table 5.1: **Experimentally determined molar extinction coefficients of all applied nucleosides and 4-thiouridine (4sU) for 260 nm and 330 nm**. We measured the molar extinction at three concentrations at 25 ℃ and neutral pH and calculated the individual coefficients for both wavelengths (Section 3.7.4). The average error of calculated extinction coefficients was around 7 % under these conditions. Nucleosides are arranged ascendingly according to their assayed cofactors.

		Molar extinction coefficients		
Nucleoside (analogue)	**Wavelength (λ_{max})**	**260**	**330 nm**	**Cofactor**
Cytidine (rC)	271 nm	7,110	0	1.00
Thymidine (dT)	263 nm	8,520	0	1.20
Uridine (rU)	262 nm	9,055	0	1.27
Guanosine (rG)	253 nm	11,530	0	1.62
Adenosine (rA)	260 nm	14,070	0	1.98
4-thiouridine (4sU)	334 nm	2,410	16,700	2.35

Furthermore, the retention time of each nucleoside (analogue) was determined using an isocratic gradient on a HPLC reverse phase column (Section 3.7.3). For this purpose, we analyzed a defined concentration of each substance separately to identify the individual nucleoside-specific retention time with variances of 0.4–0.8 minutes (Figure 5.1A). The previously investigated cofactors (Table 5.1) were validated using an equimolar mixture of all common nucleosides. For normalization, we corrected the integrated areas under the curves (AUC) of each nucleoside via the determined cofactors (AUC $\rightarrow$ cAUC). With this the nucleoside composition was determined to be 20 $\pm$ 0.8 % for each nucleoside (Figure 5.1B). Validation of the entire HPLC procedure including the initial enyzmatic ribonucleoside hydrolysis (Section 3.7.2) was performed by digesting a twenty-four nucleotide long RNA oligonucleotide primer with known composition ($C_4G_5A_7U_8$). Although each nucleoside was correctly separated in respect to the previously assayed retention times (Figure 5.1A), the calculated occurrence of rG and rU did not completely agree with the expected distribution (Table 5.2), which might be due to an incomplete RNA hydrolysis.

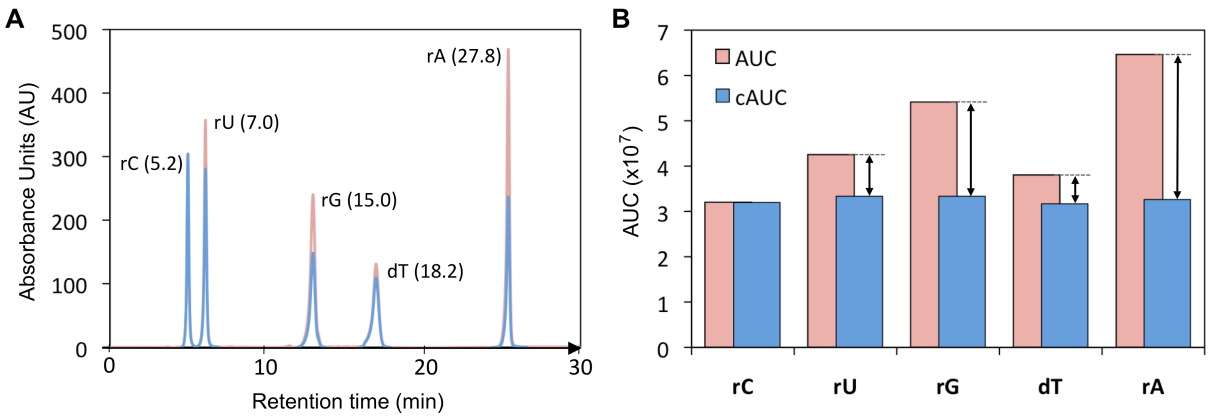

Figure 5.1: **An equimolar nucleoside mixture enables methode validation**. **A**. Equimolar nucleoside chromatogram with measured (red) and corrected extinctions (blue). **B**. Comparison of uncorrected and corrected AUCs.

Table 5.2: **Calculation of ribonucleoside composition of a 24-nt RNA oligonucleotide primer ($C_4G_5A_7U_8$) after RNA digestion and quantitative HPLC analysis.** The integrated area under the curve for each separated ribonucleoside was measured and corrected applying the determined cofactors. The resulting occurrences (in percent and base counts) are compared to the actual composition.

Nucloside	Integrated area	Corrected area	Occurrence	Calculated	Actual
Cytidine (rC)	447753	447753	17.4 %	4.18	4
Guanosine (rG)	760241	468804	18.2 %	4.37	5
Adenosine (rA)	1430582	722917	28.0 %	6.72	7
Uridine (rU)	1195693	938860	36.4 %	8.74	8

Based on this findings, we adapted the original protocol from Hafner *et al.* (2010) to our experimental conditions and initially measured *in vivo* incorporation rates after RNA labeling in YPD medium supplemented with 4-thiouracil. To achieve this, different concentrations of 4tU were tested and labeling time were assayed. Resulting incorporation rates are summarized in (Figures 5.2 and 5.4). No significant incorporation was detected after 6 min labeling using $\leq$1 mM 4tU (Figure 5.2 and 5.3A). In contrast, 5 mM 4tU, as used for the cDTA approach showed an incorporation level of 0.305 % (Sun *et al.*, 2012). Within the first 24 min the incorporation rates ranged from 0.097 % (100 μM 4tU) to 0.810 % (5 mM 4tU). The measured incorporation rate for 100 μM 4tU at 24 min deviated from the series and might be an analytical error (Figure 5.2). Rates above 1 % were only achieved after 48 min labeling with the highest concentration of 4tU (5 mM) and between 96 and 192 min using 500 μM or 1 mM 4tU. The highest amount of 4sU (with 2.7 %) was measured after labeling with 5 mM 4tU for 192 min (Figure 5.3B). This led to the conclusion that both, the extended labeling time as well as the final concentration of 4-thiouracil largely contribute to the labeling efficiency. Nevertheless, the uptake of 4tU as well as the subsequent incorporation of 4sU into nascent transcripts triggers a nucleolar stress response (Burger *et al.*, 2013). Consequently, we monitored growth (defects) and global expression levels in subsequent experiments (see Section 5.1.2).

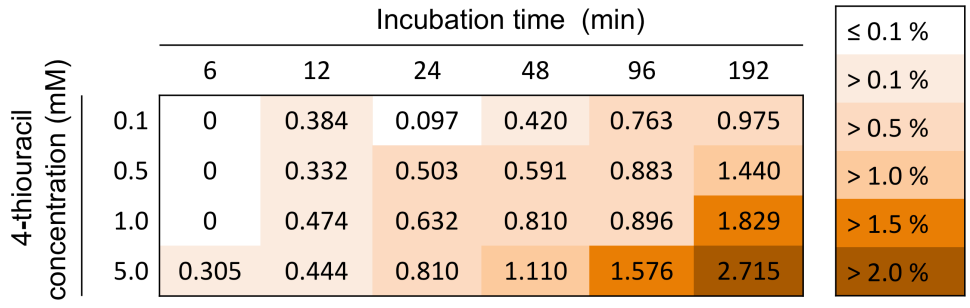

Figure 5.2: **Levels of 4sU incorporation measured by HPLC**. Four final concentrations of 4-thiouracil (100 μM, 500 μM, 1 mM and 5 mM) were used for labeling. Six different time points were chosen to measure 4sU incorporation, whereas the labeling time was doubled at each point. The matrix lists the percentage levels of incorporation, color code indicate by the right.

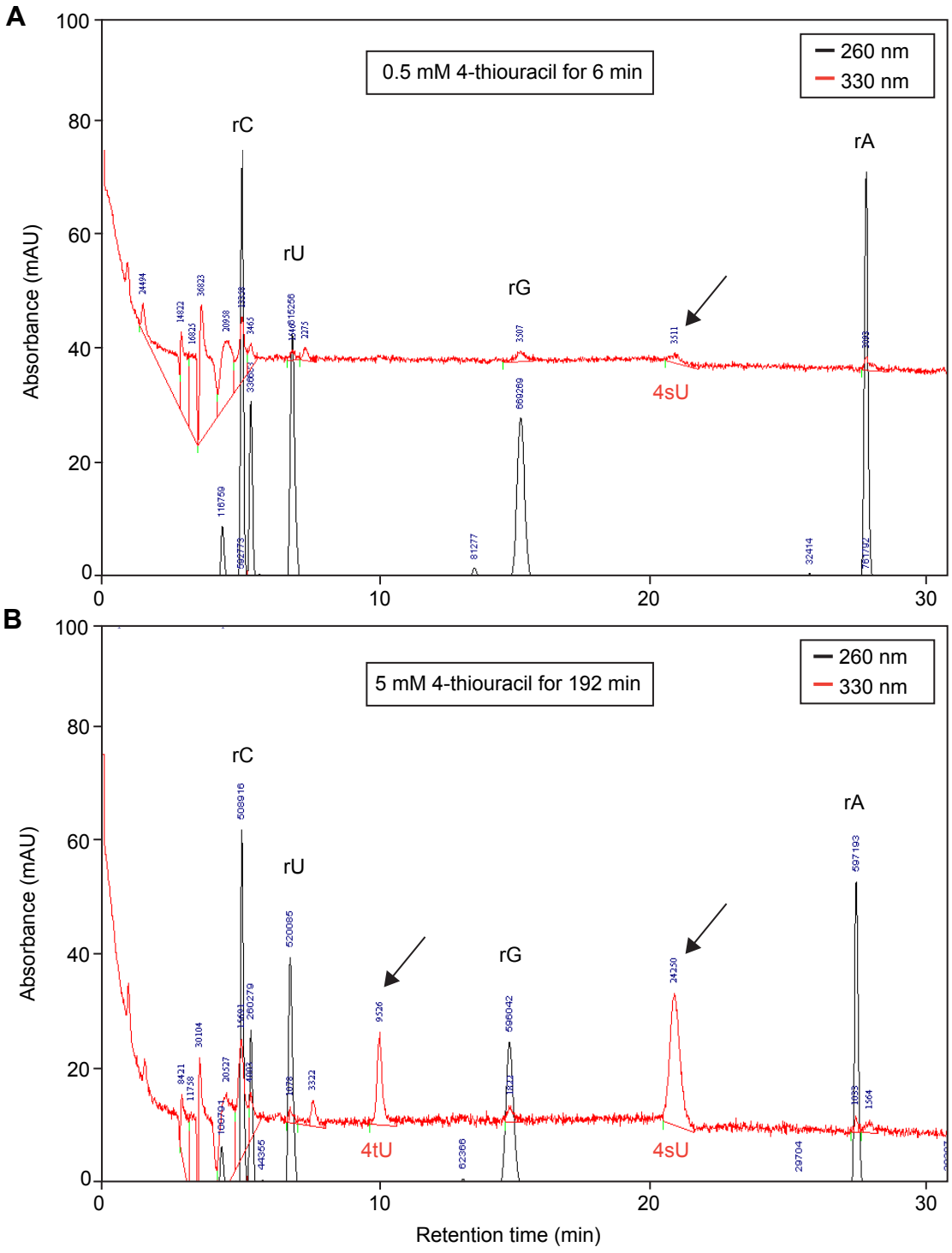

Figure 5.3: **Chromatograms after HPLC analysis**. Common ribonucleosides (rC, rU, rG and rA) were detected at 260 nm (black line). The analogue (4tU/4sU) was measured at 330 nm (red line). Each ribonucleoside showed a distinct retention time that matched with the run of the 'clean', equimolar mixture (Figure 5.1A). Corresponding peaks are labeled; the 4tU (10.2 min) and 4sU (21.1 min) peaks are indicated by an arrow. **A**. An almost no detectable 4sU peak was identified after 6 min labeling using 500 μM 4tU, which led to an incorporation rate around zero as listed in Figure 5.2. **B**. Highest concentration of 4sU or 4tU were achieved after labeling with 5 mM for 192 min.

5.1.2 Labeling conditions influence growth and amounts of cross-link sites

After the addition of 4tU into the YDP, changes in growth during RNA labeling were evaluated by measuring the OD_{600} at the same time points (6, 12, 24, 48, 96, and 192 min) (Section 5.1.1 and Section 5.1.1). No growth defects were observed within the first 24 min of labeling, but a reduced growth was monitored after 48 min (Figure 5.4A). After 192 min the untreated cells had an OD_{600} of 2.2 $\pm$ 0.12, whereas cells treated with 5 mM 4tU showed a significant growth defect with a final OD_{600} of 1.5 $\pm$ 0.05. The lowest growth change in relation to the untreated samples was observed for the all samples treated with a final concentration of 100 μM 4tU. Cells that were incubated with either 0.5 mM or 1 mM 4tU showed a similar OD_{600} of ~1.6 after 192 min labeling, suggesting that both concentration disturbing the cell equally.

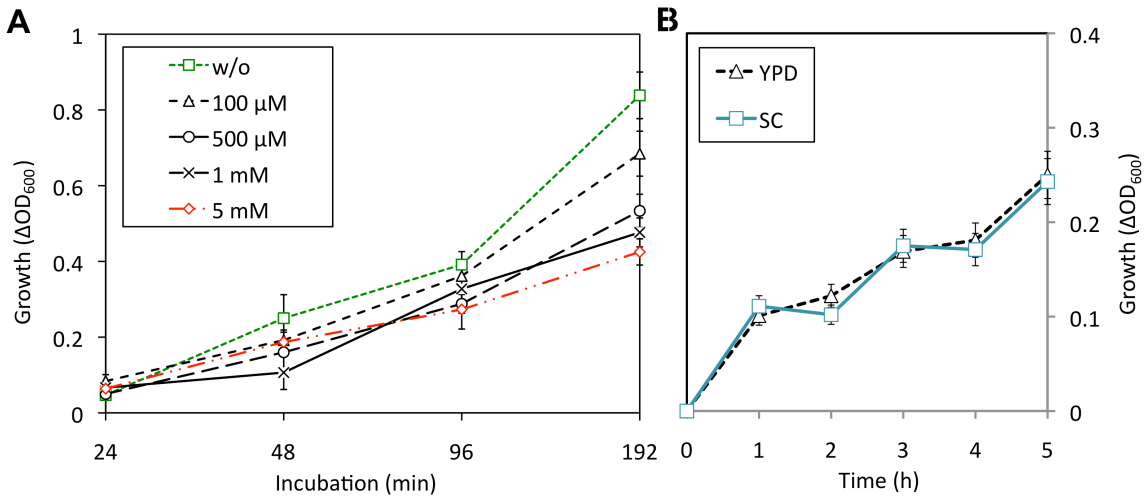

Figure 5.4: **Growth changes during RNA labeling with 4-thiouracil**. **A**. Growth curves show changes in growth at progressive time points. The green line corresponds to the untreated samples, whereas the red line stands for the highest concentration of the base analogue 4tU (with 5 mM). **B**. Growth curves compare cultures incubated with a total of 1 mM 4tU in either YPD medium or Synthetic complete (SC) medium supplemented with 10 mg/l uracil (50 %) and 100 μM 4-thiouracil.

Initially, 4tU-labeling was performed in YPD medium that commonly consists of 2 % peptone and 1 % yeast extract (Table 2.4). The exact composition of both components is mainly manufacturer-dependent and most variable. Due to this undefined character, an alternative culture medium for 4tU labeling was tested. For this purpose, we chose Synthetic complete (SC) medium containing a complete supplement mixture of amino acids and vitamins (Formedium). We supplemented the SC medium with 100 μM 4-thiouracil, 2 % glucose, and only 10 mg/l uracil (50 % of the original amount). The defined character of SC as well as the reduction of uracil was intended to increase the 4tU uptake and 4sU incorporation.

For this purpose, we changes the labeling strategy by incubating cells from OD_{600} of 0.1 to 0.5 in SC before then the concentration was raised to 1 mM 4tU by adding another 900 μM. Surprisingly, no obvious distinctions in growth were observed between YPD and SC medium (Figure 5.4B). Due to this finding, we again measured incorporation levels after performing the 4tU-labeling in SC medium (as previously described).

Even though assayed incorporation levels in SC medium were only slightly higher in comparison to YPD (Figure 5.5A), the amount of high-resolution cross-link sites was remarkably increased in SC using the identical dose of UV light (Figure 5.5B–D).

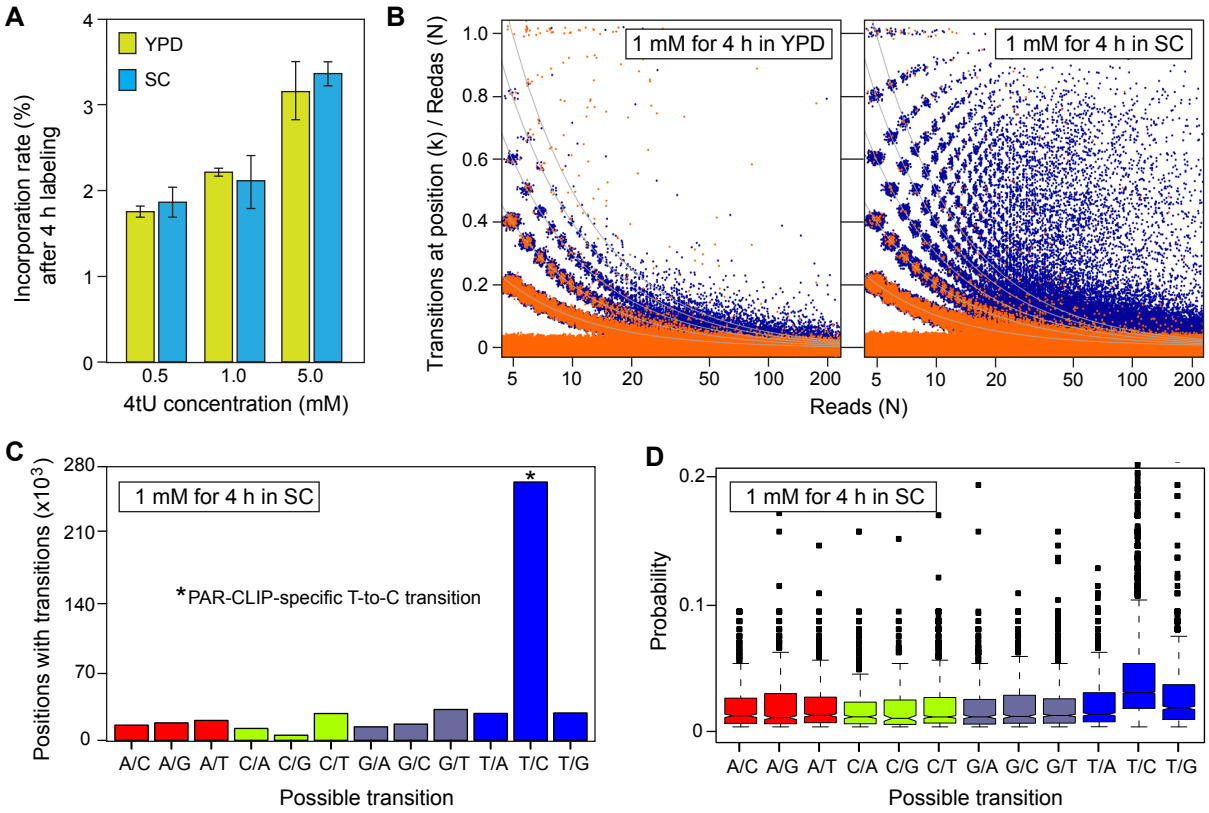

Figure 5.5: **Incorporation quality is increased by defined Synthetic complete (SC) medium using the example of Nrd1**. **A**. Comparison of incorporations rates after 4tU treatment in both YPD and SC medium using three different concentrations (500 μM, 1 mM, and 5 mM). **B**. Discrete probability distribution of correct cross-link sites (blue dots) after labeling in YPD (left) and SC medium (right) compared to other possible transitions resulting from undefined events like UV light damage (orange dots). **C**. Counts of all possible base transitions and their corresponding nucleotide distribution probabilities (**D**).

5.1.3 Cross-linking efficiency depends on UV light dose

Another critical step in the PAR-CLIP procedure is the *in vivo* UV light cross-linking at 365 nm. Compared to related approaches like individual-nucleotide resolution CLIP (referred to as iCLIP) (Knig *et al.*, 2010), decreasing the wavelength to 365 nm has three main advantages in relation to 254 nm (iCLIP): (i) less UV damage using the same amount of radiation energy (Ascano *et al.*, 2012), (ii) an improved RNA recovery up to 1000-fold applying a photoactivatable nucleosides (Hafner *et al.*, 2010), and (iii) high-resolution binding sites due to PAR-CLIP-specific T-to-C transition, which results from the incorporated base analogue (Hafner *et al.*, 2010; Spitzer *et al.*, 2014). Even though it is known that the radiation energy of 254 nm does not break phosphodiester bonds, it is believed that UV_{254} might disrupt the disulfide bond, which is established during UV cross-linking (Correia *et al.*, 2012).

To ensure a maximal cross-linking efficiency, several doses of UV light at 365 nm were tested after 2 h labeling using 1 mM 4tU. The tested protein (Nrd1-TAP) was immunoprecipitated and cross-linked RNAs were labeled radioactively using γ-32-P-ATP, and subsequently analyzed via SDS-PAGE (as described in Section 3.4.1). Instead of a sharp radiating band, a smear displaying the protein of interest with different sizes of covalently bound RNA transcripts was observed. Knowing the molecular weight of the analyzed TAP-tagged protein as well as the mean weight of a RNA nucleotide (0.32 kDa), the size distribution of bound fragments before the RNase treatment could be calculated.

The lowest radiation signal was observed after UV cross-linking using 0.5 J/cm^2, whereas the signal intensities accumulated with increasing energy doses (Figure 5.6A). In contrast, less than 0.5 J/cm^2 are needed to obtain adequate cross-links in human cell lines (Hafner *et al.*, 2010; Farazi *et al.*, 2014). Therefore, clear differences between yeast and thin-layered cell lines were demonstrated. This can be explained by the thick cell wall, the spherical shape, and the non-adherent growth of yeast cells, which all together leads to an increased absorption of UV light. In order to determine cell viability after cross-linking, UV light treated cell were again incubated in YPD. Strong growth defects compared to untreated cells were observed after cross-linking with energy dose above 1 J/cm^2 (Figure 5.6B). Similar results were obtained for ChIP experiments after cross-linking for 20 min at 20℃ using 1 % formaldehyde (data not shown).

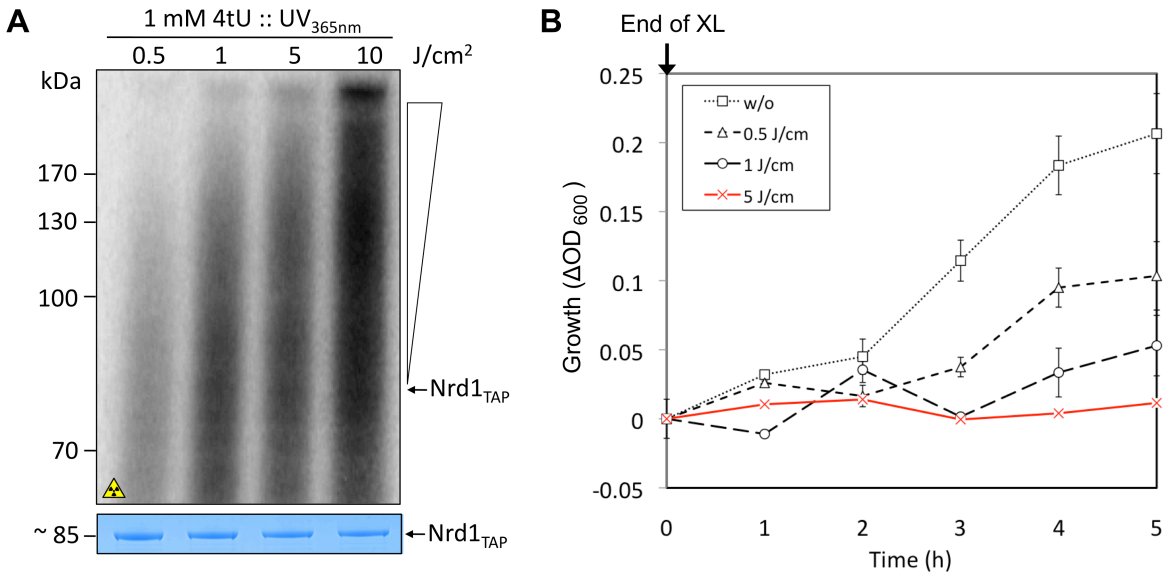

Figure 5.6: **Cross-linking efficiency depends on UV$_{365}$ light dose**. Four different doses were tested: 0.5, 1, 5 and 10 J/cm^2. Depicted are the effects of increased UV light exposure using the example of Nrd1-TAP (size: ~85 kDa). **A**. The autoradiogram of a SDS-PAGE gel exemplifies the enhanced yield of cross-linked, radioactively labeled RNA (top). Same gel after coomassie staining (below). **B**. Viability decreases during the UV light treatment and results in dysfunctional or not viable cells after the *in vivo* cross-linking (XL).

5.1.4 Yeast cells require harsher lysis approaches than mammalian cells

While mammalian cell lines can be easily lyzed by the addition of a sufficient amount of detergents (NP-40, Triton X-100, etc.), yeast cells require a more robust, physical treatment. Consequently, we optimized the cell lysis by testing different approaches. Achieved lysis efficiencies were determined photometrically (OD_{800}) and by light microscopy. Firstly, we tested different lysis buffers containing various concentrations of Triton X-100, SDS and/ or NP-40. Because we observed poor lysis effects, we tried to transform cells into spheroplasts before the detergent-based treatment. For this purpose, we assayed three different enzymes with lytic activity against living yeast cell walls. Each enzyme was tested at 37 °C with a final concentration of 50 U per ml cell suspension. No spheroplast formation was observed after the treatment with Glusulase, a commercial preparation containing both the β-glucuronidase as well as sulfatase (Figure 5.7A). After the incubation with Lyticase, nearly 20 % of cells had lost their cell wall. However, highest rates of spheroplast formation with almost 70 % effected cells was obtained after the treatment with Zymolyase (Figure 5.7B).

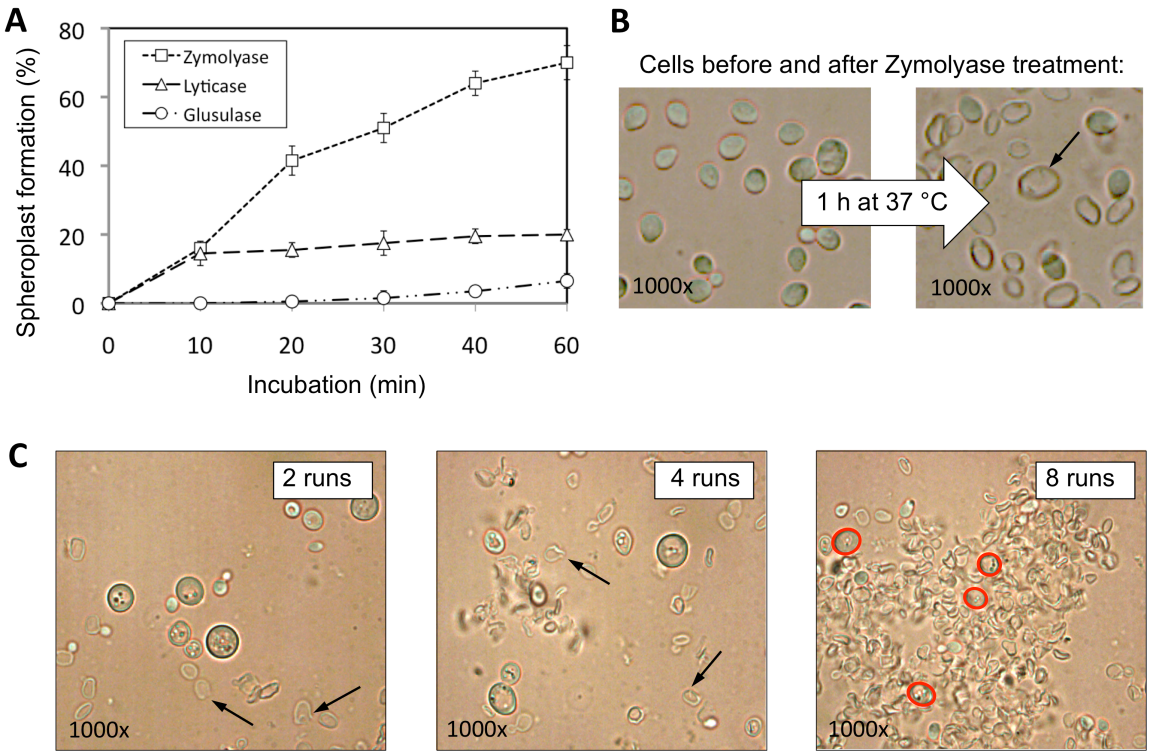

Figure 5.7: **Cell lysis efficiency after enzymatic and mechanical treatment. A**. Progress of spheroplast formation using three different lytic active enzymes preparations (50 U/ ml cell suspension) was measured at OD_{800}. Reactions were performed at 25 and 37 °C, and subsequently averaged. **B**. Comparison of untreated and Zymolyase-treated cells (after 60 min). Intact, cell wall-containing cells are darker, while spheroplasts appear more transparent. The arrow indicates an already cracked spheroplast. **C**. Microscopies of mechanical lyzed yeast cells after using the advanced FastPrep-24 homogenizer for two, four and eight runs of 40 sec each. Disrupted cells are arrow-marked, while unlyzed and intact cells are highlighted in red.

Despite these results, we were not able to immunoprecipitate the proteins of interest from cells that have been treated with either Lyticase or Zymolyase anymore. This might be either a result of the extended incubation at 37°C, which results in stress-induced expression changes, or due to 'polluted' enzyme stocks containing several unknown ingredients (i.e. various proteases). No information concerning additional ingredients was supplied by the manufacturer.

Finally, we tested three mechanical approaches that have been proven to effectively disrupt cells. By using the high-pressure french press system TS 0.75 (Constant Systems Ltd.), a lysis efficiency of ~80 % was achieved after the first 5 min. Despite the excellent lysis properties this device was finally impractical due to extreme frothing effects and a loss of material of about 30 %. In case of the two-dimensional mixer mill MM 400 (Retsch GmbH), comparable results were obtained after at least 90 min of milling. Even in the cold room, samples had to be additionally cooled down every 15 min to avoid thermal overheating. Consequently, we were looking for a variant with comparable lysis capabilities without sample loss and denaturation effects. An appropriate device was finally found in the FastPrep-24 homogenizer (MP Biomedicals) that led to a lysis efficiency of $\geq$80 % within eight runs of 40 sec each (Figure 5.7C).

5.1.5 Improper RNase treatment impairs cDNA library preparation

In order to ensure an optimal fragment size for subsequent adapter ligation, we assayed the RNA fragmention by RNase T1 digestion. Initially, we applied the same RNase T1 concentration as used by Hafner *et al.* (2010) for the first digestion reaction. Treated RNA segments were radioactively labeled using γ-32-P-ATP (as described in Section 3.4.1), and the immunoprecipitated protein (Nrd1-TAP) with bound and labeled transcripts was separated by SDS-PAGE. In comparison to untreated fragments (Figure 5.8A, left lane), the RNase-digested samples already displayed strong degradation effects, even without a subsequent RNase treatment (Figure 5.8A, right lane). Hafner *et al.* (2010) used two RNase treatments to reach an adequate fragment size, whereas the initial digestion is firstly applied to pre-digest bound transcripts. Due to the observed effects of the first treatment, we decided to abolish this step and searched for alternatives. Because sonication is commonly used for DNA fragmentation during ChIP-chip experiments (Mayer *et al.*, 2010) (Figure 5.8B, upper gel), we assayed the shearing effects of the acoustic cavitation on total RNA size applying Diagenode's Bioruptor ultrasonicator. Even though small RNAs seemed to be resistant to the ultrasonic treatment (Figure 5.8B, lower gel), rRNAs showed clear fragmentation effects with increasing sonication cycles (Figure 5.8B, middle gel). To avoid RNA and protein damage by the acoustic cavitation, we finally chose 4 cycles of sonication and substituted the initial RNase digestion by this mechanical lysis approach. Pre-fragmented RNA transcripts were subsequently RNase-treated as performed in the original PAR-CLIP protocol. Due to the general overexpression of proteins in mammalian cell lines (Hafner *et al.*, 2010; Spitzer *et al.*, 2014; Farazi *et al.*, 2014), the yield of cross-linked binding partners seems to be highly increased and therefore enhances the separation of RNA segments bound by the protein of interest. While Hafner *et al.* (2010) added RNase T1 to a final concentration of 100 U/μl, we deemed an 2,000-fold reduction (50 U/ml) appropriate. This tremendous

reduction might be explained by the decreased amount of immoprecipitated protein resulting from our not overexpressed approach. Nevertheless, the combination of both the sonication (4 cycles) and optimized RNase treatment (50 U/ml for 20 min at 25℃), led to an averaged Gaussian-like fragment size distribution around 40 nucleotides (Figure 5.9C), perfectly suitable for an Illumina sequencing platform (GAIIx, MiSeq or HiSeq).

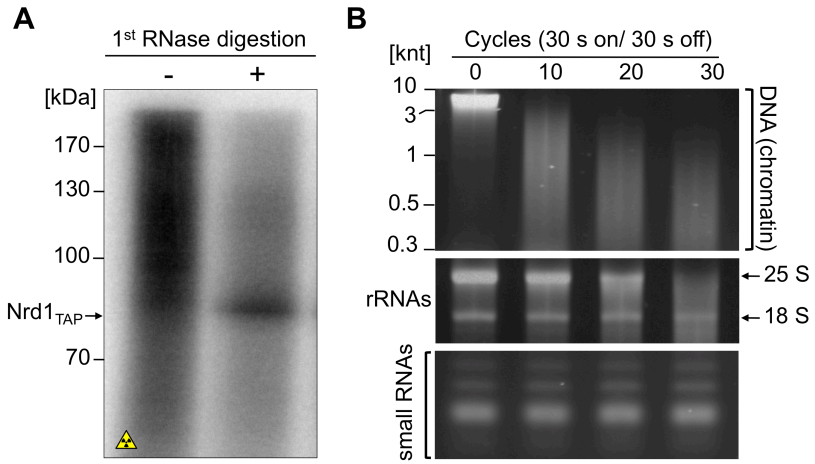

Figure 5.8: **Sonication of cross-linked RNA to ensure fragmentation is more controllable than the initial RNase digestion. A**. The first and initial RNase T1 treatment from the original PAR-CLIP protocol (Hafner *et al.*, 2010) digests cross-linked RNA most intensively and disables the following cDNA library preparation. Autoradiogram of immunoprecipitated Nrd1-TAP samples, run on a SDS-PAGE, shows size of cross-linked RNA after 1 h IP in lysis buffer with (+) or without (-) RNase T1 (1 U/μl). **B**. Shearing effects of sonification on different classes of RNAs. Diagenode's Bioruptor ultrasonicator was used to shear total RNA using 10, 20 and 30 cycles of sonication at high intensity and 30 sec on/off intervals. Different classes of RNAs were isolated separately and run on a high-resolution 18%-Urea-PAGE gel: mRNAs (top), rRNAs (middle), and small RNAs (bottom).

5.1.6 Optimized library preparation improves data outcome

After we ensured a proper fragment size of bound transcripts (previous Section 5.1.5), a new protocol for data acquisition in order to improve both the quantity and quality of cDNA libraries had to be established and optimized. In the original PAR-CLIP protocol from Hafner *et al.* (2010), adapter ligation and size selection are performed by applying multiple radioactive labeling procedures and subsequent gel purifications (Hafner *et al.*, 2008). We applied this protocol intensively, but were not able to produce a single, 'sequenceable' cDNA library. This might have two explanations: (i) we were not able to recover enough RNA segments due to the limited amount of proteins (as discussed before), and/ or (ii) we lost too much material during the two ligation and purification steps.

The original protocol for cDNA library preparation from Hafner *et al.* (2008) is an enormous and quite sophisticated procedure using a multitude of radioactive material. Due to this fact, we searched for a commercial "all-in-one package" for RNA library preparation that contains all required enzymes, buffers, and oligonucleotide primer. An adequate product containing even the specialized pre-adenylated 3' adapter,

was finally found in the NEXTflex Small RNA Sequencing Kit v1 (Bioo Scientific). This complex and expensive oligonucleotide modification was introduced to prevent ligation to the blocked 5' adaptor by using a truncated RNA ligase. However, we observed a strongly increased adapter dimer formation resulting in a DNA band at 118 bp (Figure 5.9A, left lane). After extraction and sequencing of the actual library above that size (~130 till 160 bp), roughly 40 % of sequenced reads still contained only adapter sequences. We assumed that especially the modified ligation adaptors might have been defective in terms of integrity and/ or purity, which consequently leads to unwanted ligations and a dramatically reduced data outcome.

Finally, the entire library preparation procedure was renewed and intensively optimized. To circumvent adapter dimer formation, the ligation protocol was modified and converted to a so called "on-bead ligation" procedure (Methods). In comparison to successive ligations that are separately performed in the same reaction tube and/ or buffer conditions, here referred to as "mixed ligation", the varying buffers and enzymes, as well as oligonucleotides are removed much more efficiently. As a consequence, no PCR product pointing out an adapter dimer was obtained following the on-bead ligation and associated washing steps (5.9A, right lane). We additionally validated the tremendous improvement with the BioAnalyzer by comparing both libraries obtained from either mixed or on-bead ligation (Figure 5.9B). While the mixed ligation of the NEXTflex kit resulted in a pervasive dimer peak at ~118 bp, the library obtained from the on-bead ligation showed a clear distribution between 120 and 220 bp with a maximum at ~160 bp and, lacked the dimer band (Figure 5.9B).

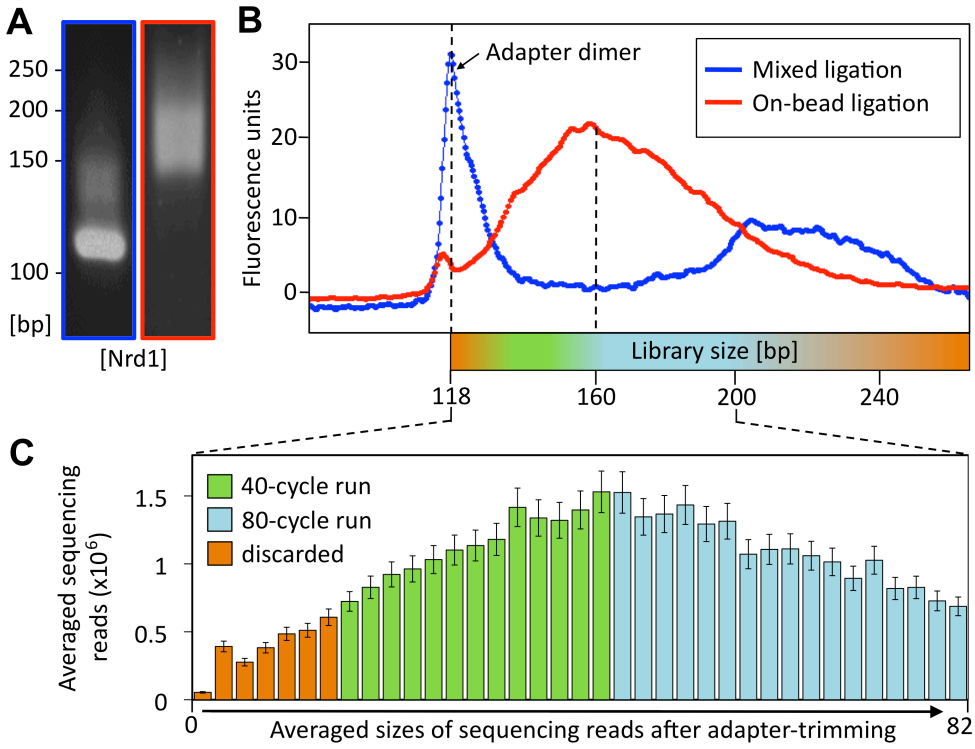

Figure 5.9: **Optimized data acquisition improves library size and yield. A**. Nrd1-TAP cDNA library using a commercial preparation kit (left) and an optimized protocol (right). **B**. On-bead ligation minimizes adapter dimer formation. **C**. Averaged fragment sizes after adapter-trimming provide an insight into size distribution of previously bound RNA fragment.

Importantly, not only the size of the final library (after the adapter-trimming) was optimized (Figure 5.9C), we also noticed a multiple increased yield that might be due to an approximately 6-fold reduction in terms of the adapter formation (Figure 5.9B). These changes and optimizations finally ensured an increased number of specific, information-containing reads, and therefore improved and facilitated the data outcome and further data analyses, respectively.

5.2 An advanced computational pipeline for PAR-CLIP data

We designed a processing pipeline for sequencing data obtained from PAR-CLIP experiments, which have been sequenced on either an Illumina GAIIx or HiSeq machine. For this purpose, we combined open source tools with mainly self-written R and python scripts and termed our pipeline 'CLiPAR' (Figure 5.10). The programming and modulation of the pipeline was predominantly performed by Phillipp Torkler and Alexander Graf. Used parameter settings for the pre-processing pipeline, including the quality trimming and mapping (Figure 5.10), have been individually adapted to our PAR-CLIP procedure to finally ensure the highest possible accuracy. First we removed remaining adapter sequences that have already become $\leq$5 % due to the optimized library preparation (as previously discussed).

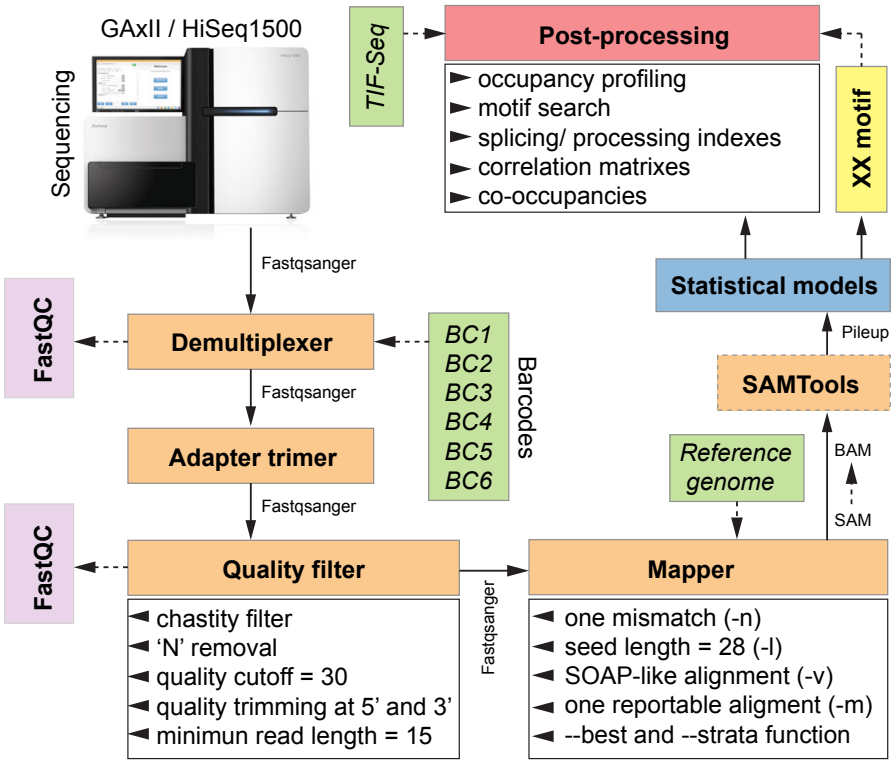

Figure 5.10: **Schematic overview of the CLiPAR pipeline.** Raw data was obtained from an Illumina platform (GAxII or HiSeq). The pre-processing block contains several moduls (orange) used for demultiplexing, adapter trimming, quality-filtering, mapping and format conversion. The tool "FastQC" (purple) was applied for initial data evaluation. Following pre-processing the actual modeling (blue) and post-processing (red) took place. For motif discovery the external tool "XX motif" was used (Hartmann *et al.*, 2013) (yellow). Required data input like the reference genome or the annotations are coloured green.

To characterize the quality of our sequenced reads, we determined the Phred quality score by base calling each nucleotide. Surprisingly, most data sets had already a very high accuracy by predominantly showing mean sequence qualities above 20 (Figure 5.11A), corresponding to a base call accuracy of 99 %. Due to the excellent data quality of the performed PAR-CLIP experiments, we were finally able to set the quality filter threshold to 30, meaning a minimal probability of one incorrect base call within 1000 nucleotides (= 99.9 % accuracy) (Figure 5.11B).

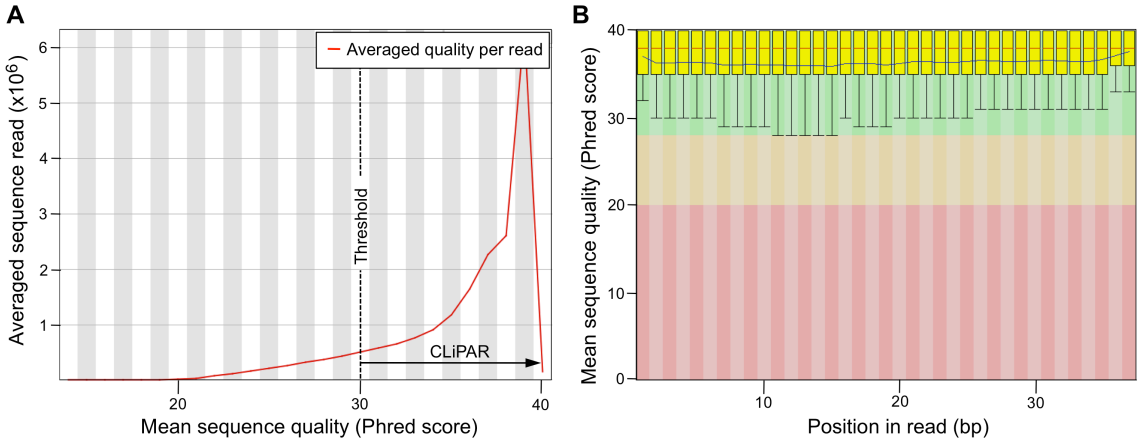

Figure 5.11: **Phred scores enable validation of data quality.** (**A**.) Averaged Quality score distributions over all sequenced reads. (**B.**) Phred scores across all bases after quality filtering using the 'CLiPAR' pre-processing pipeline.

Quality-trimmed reads were then mapped to the *S. cerevisiae* genome using the short read aligner Bowtie (Langmead, 2010) with PAR-CLIP-adapted parameters (see Section 4.1). Chemical cross-links between the PAR-CLIPped protein and the RNA-incorporated 4tU lead to U-to-C transitions during reverse transcription that, when mapped to the genome, manifest themselves as T-to-C mismatches (Hafner *et al.*, 2010). However, other possible sources of nucleotide mismatches are sequencing errors and differences between the genome sequence of the organisms used in PAR-CLIP experiments and the reference sequence onto which the PAR-CLIP reads are mapped. To calculate P-values for true cross-linking sites, a null hypothesis had to quantitatively modeled, i.e. the probability that the T-to-C mismatches observed in reads covering a certain T nucleotide in the genome are not caused by cross-links between the immunoprecipitated factor and the RNA, but are due to the other sources of mismatches. This null model distribution could finally be estimated by fitting a two-component binomial mixture distribution to the frequency of the other 11 possible mutations (as previously shown, Figure 5.5D). The first binomial component models the sequencing errors, while the second component models SNPs (Methods). Luckily, the cross-linking sites that passed our very stringent thresholds strongly dominated over the contributions by sequencing errors and SNPs, and were assessed being 'true' induced T-to-C transitions resulted from the performed PAR-CLIP experiment. A rough overview of the CLiPAR pipeline is given in Figure 5.10, depicting the main process flow with its modules.

Before occupancy profiles and correlations for all genes or introns could be computed, most precise annotations for transcriptions start sites (TSS), splice sites (SS) and polyadenylation sites (pA) had to be used. Initially, we applied the TSS and pA annotations obtained from tiling array analyses by Xu *et al.* (2009) (Figure 5.12A). Even though these annotations already provided high-resolution data and deep insights, we were still able to greatly improve the resolution of many PAR-CLIP-derived occupancies by using the recent annotations derived from transcript isoform sequencing (TIF-Seq) (Pelechano *et al.*, 2013). Because RNA-Seq and its variants demonstrate a much broader dynamic range compared to micro and tilling arrays, a more precise detection of low abundance transcripts or their isoforms, as well as the identification of genetic variants in a single-nucleotide resolution can be accomplished as used for various CLIP experiments (Ule *et al.*, 2003; Hafner *et al.*, 2010; Pelechano *et al.*, 2013; Tuck and Tollervey, 2013; Spitzer *et al.*, 2014). We finally picked the most abundant transcript isoform (Section 4.4), giving precise TSS and pA positions for 5578 gene transcripts, and were consequently be able to see much sharper sequence features around both the TSS and pA sites (Figure 5.12B).

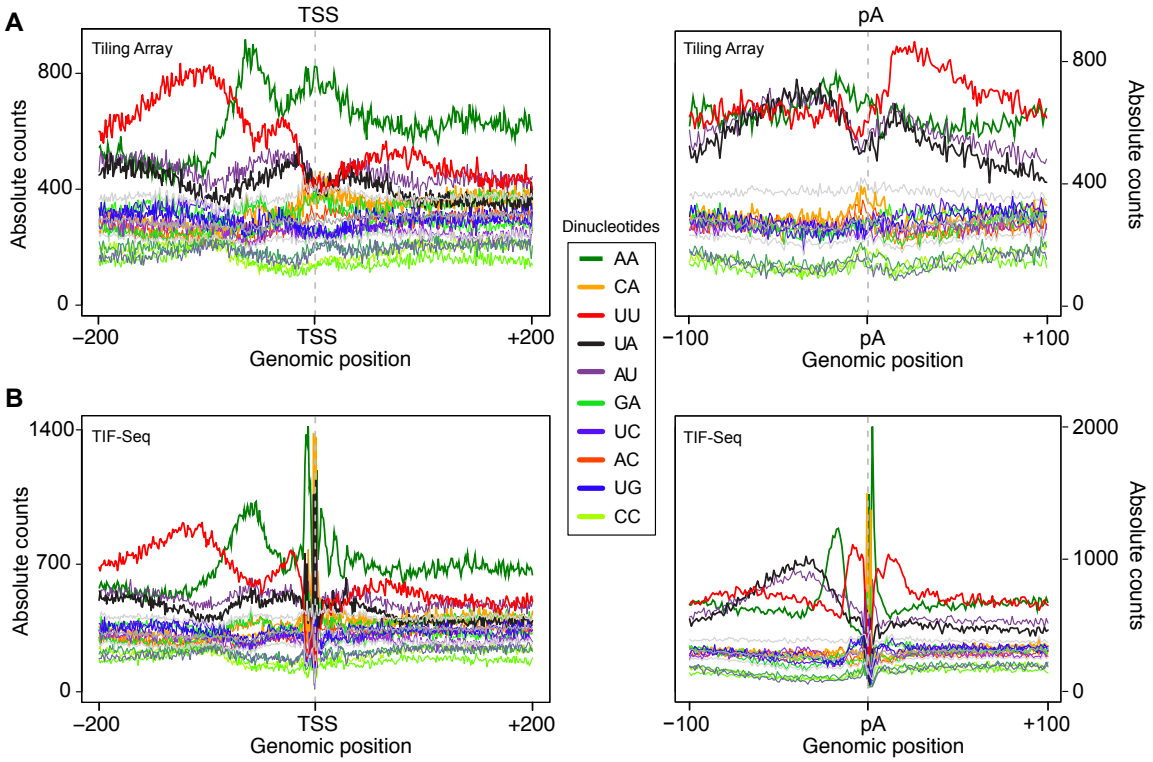

Figure 5.12: **Comparison of TSS and pA annotations from tiling array and TIF-Seq data.** We used the recent TIF-Seq data from Pelechano *et al.* (2013) (panel **B**) which yielded much sharper sequence features around TSS and pA sites than the previous annotation from Xu *et al.* (2009) (panel **A**).

Through CLiPAR, we were now able to analyze our PAR-CLIP data in high-resolution and in a much more individualized and comprehensible way in comparison to other, already published computational approaches for PAR-CLIP analyses like CLIPZ, PARalyzer, wavClusteR and miRTarCLIP, etc. (Khorshid *et al.*, 2011; Corcoran *et al.*, 2011; Sievers *et al.*, 2012; Chou *et al.*, 2013). Whereas the CLIPZ server

(at the Swiss Institute of Bioinformatics) comes with a graphical user interface (GUI) (Khorshid *et al.*, 2011), most available tools still require advanced knowledge in bioinformatics, especially in the statistical computing environment R as well as in Bioconductor or the Comprehensive R Archive Network (CRAN) packages. To solve this issue, we additionally wrote wrapper to include our pipeline into Galaxy, an open, web-based platform with an user-friendly GUI (Goecks *et al.*, 2010). Thus, it can be ensured that even user with less bioinformatical background are able to use our pipeline up to a certain point.

However, all introduced computational approaches have their limitations and do not offer an overall solution. While the wavClusteR package for instance require an already mapped and pre-processed input file, generated by SAMTools (Li *et al.*, 2009; Sievers *et al.*, 2012), the entire analysis pipeline of miRTarCLIPs is limited to microRNA target site detection (Chou *et al.*, 2013). A recently tailored analysis tool, dubbed as PARalyzer, firstly includes an interaction site identification by using a novel motif-finding algorithm (Georgiev *et al.*, 2010; Corcoran *et al.*, 2011). Compared to our pipeline, PARalyzer still requires further efforts and computational skills to pre- and even post-process the issued data individually. As an important issue of bioinformatic tools remains the accessibility of the source code(s). The CLIPZ pipeline for instance mainly remains unknown and confirms its actual ”blackbox” character. We addressed this issue by offering two possibilities (Figure 5.13): (i) a default variant using PAR-CLIP-specific settings or (ii) the ability to manually change each value individually.

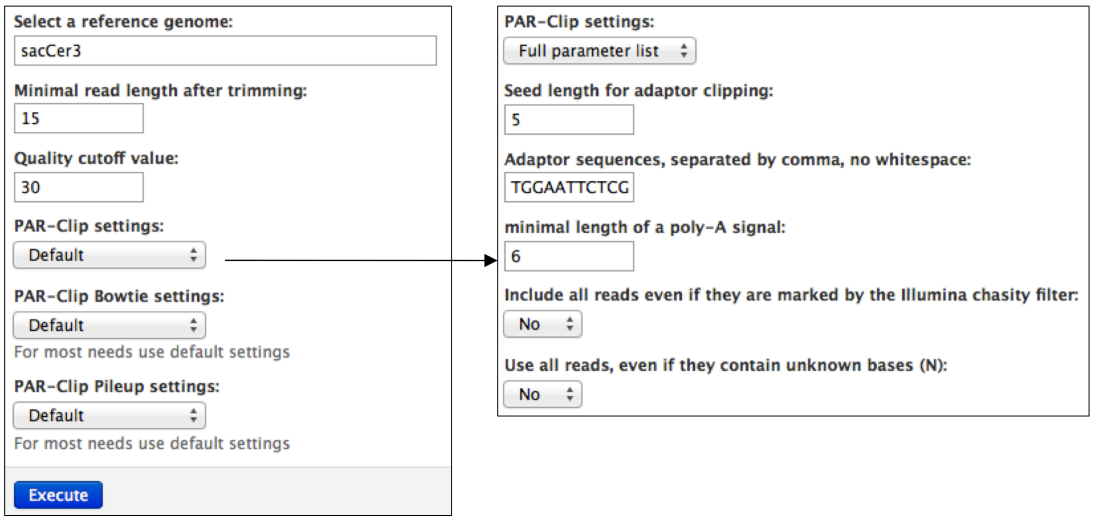

Figure 5.13: **Screenshot of CLiPAR, integrated into Galaxy's graphical user interface.** Depicted are the main menu with 'Default' settings (left) and the after choosing the 'Full parameter list' of the initial 'PAR-CLIP settings' (right).

Even though we did not solve every problem, we could at least present a transparent and powerful tool for analyses of sequencing data obtained from PAR-CLIP experiments (Torkler *et al.*, *in preparation*). With the development and usage of our new, cutting-edge computational pipeline, both the quantity and the quality of the data outcome was dramatically improved, as it will be demonstrated in Section 5.3.

5.3 Transcriptome maps of mRNP biogenesis factors define pre-mRNA recognition

To map mRNP biogenesis factors over cellular RNA at high resolution, we optimized the PAR-CLIP protocol and obtained high RNA labeling efficiencies with 4-thiouracile (4tU) in exponentially growing yeast cells (Methods). We found conditions that led to high reproducibility between biological replicates (Figure 5.14A) and enabled high 4tU incorporation levels of ~2 % (Andrus and Kuimelis, 2001) without significant changes in cellular mRNA abundance (Figure 5.14B and 5.16A).

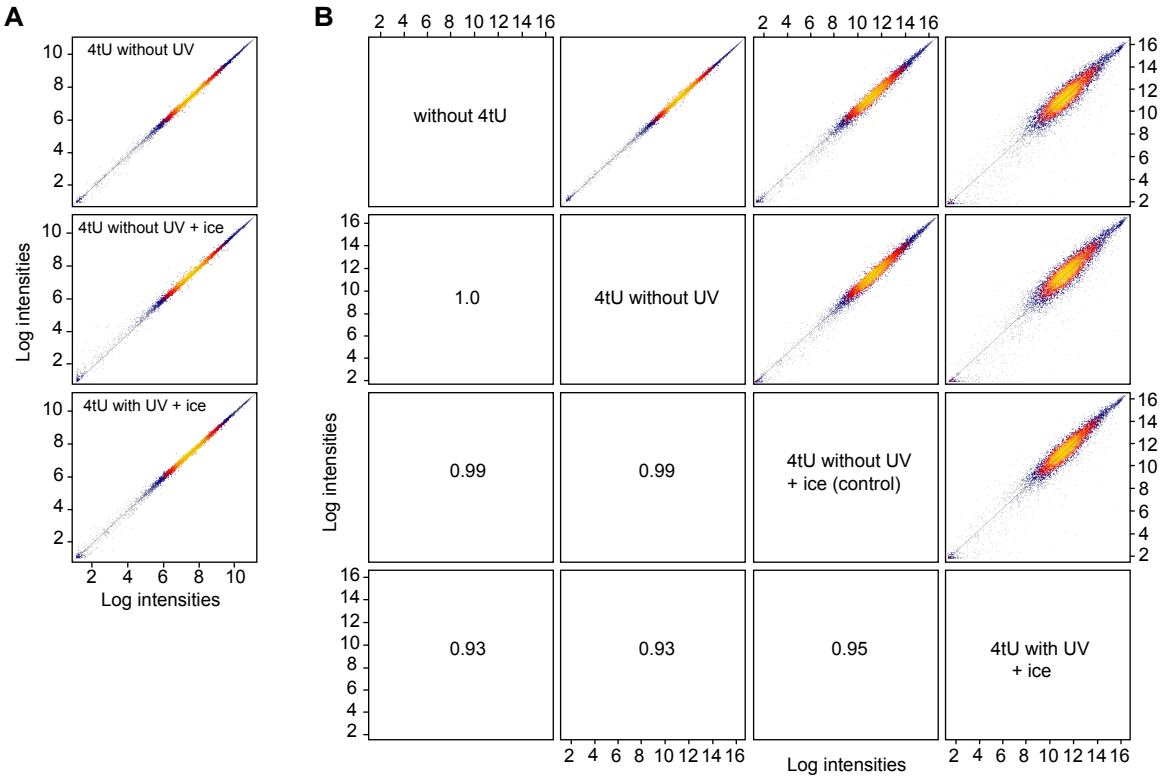

Figure 5.14: **4tU labeling and UV-treatment leave gene expression levels nearly unchanged**. **A**. Correlation of expression levels of between pairs of biological replicates with same treatment: without 4tU labeling after 4tU-labeling, and after subsequent UV-light treatment with an energy dose of 1 J/cm^2. **B**. Correlation of expression levels between cells after the various treatment steps during the PAR-CLIP procedure.

We also developed a computational pipeline for data analysis that uses a statistical model to compute P-values for factor binding sites (Section 5.2 and Figure 5.10). The pipeline also analyzes the cross-linked region with the motif discovery tool XXmotif (Hartmann *et al.*, 2013) and detects short motif preferences. For each factor, we found between 25,000 and 800,000 high-confidence protein-RNA binding sites with a P-value below 5×10^{-3}, which corresponds to low false discovery rates of 0.18–3.5 % (Table 5.3, Methods). We applied the optimized PAR-CLIP protocol to 23 mRNP biogenesis factors that showed reproducible signals (Table 5.3).

Table 5.3: **mRNP biogenesis factors analyzed by PAR-CLIP in this work** [RRM = RNA recognition motif; ZF = zinc finger domain].

Biogenesis event	Factor/ subunit	Complex	RNA-binding domain	PAR-CLIP XL sites	False discovery rate [%]
Capping	Cbc2	CBC	RRM	98,034	0.178
Splicing	Luc7	U1 snRNP	ZF	93,261	1.035
	Mud1	U1 snRNP	RRM	99,384	1.918
	Nam8	U1 snRNP	RRM	151,813	1.675
	Snp1	U1 snRNP	RRM	25,493	0.447
	Ist3	U2 snRNP	RRM	66,003	3.184
	Mud2	BBP-U2AF65	RRM	801,430	1.769
	Msl5	BBP-U2AF65	ZN	476,370	1.961
3'-Processing	Rna15	CFIA	RRM	582,756	3.463
	Mpe1	CPF	ZF	122,500	2.262
	Yth1	CPF (PFI)	ZF	59,049	3.432
	Cft2	CPF (CFII)	-	189,866	1.723
	Pab1	-	RRM	233,513	2.052
	Pub1	-	RRM	371,902	1.332
Export	Hpr1	THO/TREX	-	249,887	1.913
	Tho2	THO/TREX	-	400,965	1.064
	Sub2	TREX	-	228,620	1.085
	Mex67	TREX	-	288,579	1.010
	Yra1	Export adaptor	RRM	400,156	0.681
	Nab2	Export adaptor	ZF	283,606	2.413
	Npl3	Export adaptor	RRM	770,240	1.282
	Hrb1	SR-like	RRM	395,402	0.976
	Gbp2	SR-like	RRM	65,692	0.182

These include the cap-binding complex (CBC) subunit Cbc2 and components of the splicing machinery, namely the yeast homologs of the branch point (BP)-binding protein BBP (Msl5) and U2AF65 (Mud2), and subunits of the snRNPs U1 (Luc7, Mud1, Nam8/Mud15, Snp1) and U2 (Ist3/Snu17). Factors in the 3'-processing machinery included the Rna15 subunit of cleavage factor (CF) IA, and three subunits of the cleavage and polyadenylation factor (CPF), Mpe1, Yth1 (CPF subcomplex PFI), and Cft2/Ydh1 (CPF subcomplex CFII). We also included nine proteins implicated in mRNP export, namely subunits of the THO/TREX complex (Hpr1, Tho2, Sub2), the export factor Mex67, and its putative mRNA adaptors Nab2, Npl3 (also known as Nop3 or Nab1), and Yra1/She11, and the SR-like factors Gbp2 and Hrb1. Finally, we studied the poly(A)- and poly(U)-binding proteins Pab1 and Pub1 that regulate mRNP export and stability (Mangus *et al.*, 2003). Together these data map the protein-RNA interaction landscape underlying mRNP biogenesis (Figure 5.15 and Supplementary Figure S1).

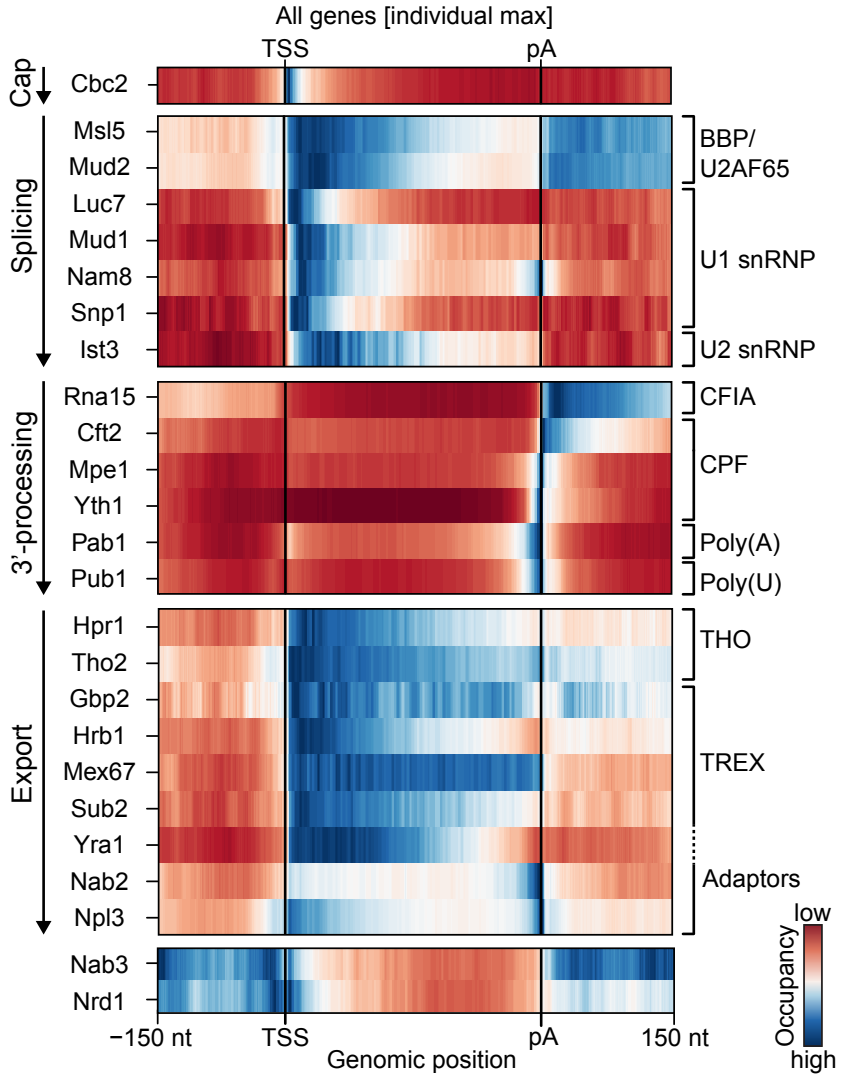

Figure 5.15: **Transcript-averaged occupancy profiles of mRNP biogenesis factors.** Individual ORF-Ts were scaled such that their TSSs and pA sites coincide.

5.3.1 RNA abundance normalization reveals capped transcripts

PAR-CLIP cross-links for the CBC subunit Cbc2 clustered at the 5'-ends of mRNAs as expected, but often extended for several hundred nucleotides (nt) downstream (Figure 5.16C). We found however that Cbc2 binding appeared much more focused at mRNA 5'-ends after the data were corrected for RNA abundance (Figure 5.16D) measured by RNA-Seq under the same experimental conditions (Figure 5.16B). We estimated relative occupancies of the cross-linked factors along mRNAs by dividing the frequency of T-to-C transitions by the RNA-Seq signal at this site (Methods). The normalization reduced the transcript-to-transcript signal fluctuation, led to an even distribution of estimated occupancy over RNAs with different abundance (Figure 5.16E), and abolished a weak artificial correlation of PAR-CLIP signals with RNA levels (Figure 5.16F). The resulting distribution of transcript-averaged occupancy profiles was very similar between strongly and weakly bound transcripts (Figures 5.16 and 5.17).

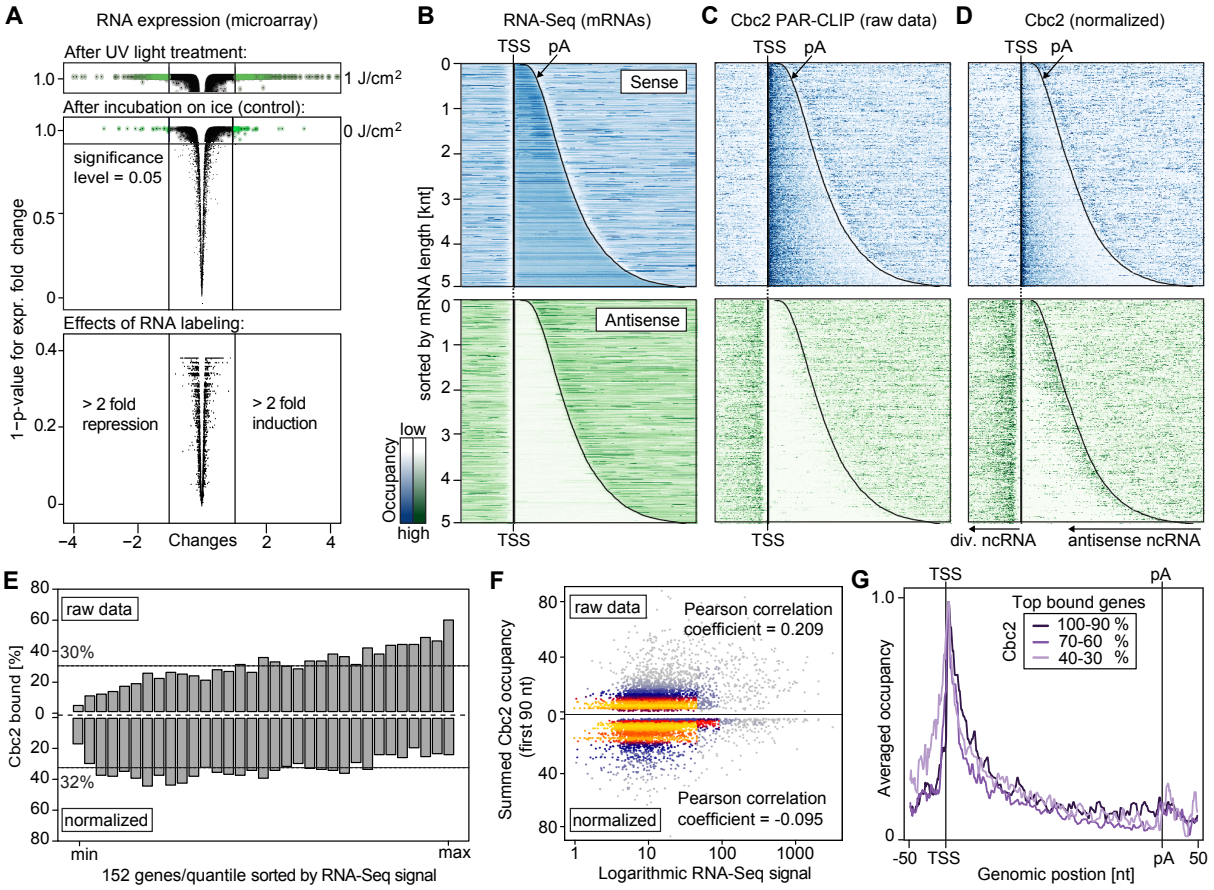

Figure 5.16: **PAR-CLIP measurements with RNA abundance normalization estimate factor occupancies over the yeast transcriptome. A**. 4-thiouracil (4tU) labeling has only a very minor effect on cellular mRNA levels. Vulcano plots of expression fold changes for mRNAs measured by Affymetrix microarrays show that only few mRNAs significantly change their abundance due to RNA labeling, incubation on ice, and UV light exposure. **B**. Smoothed Cbc2 RNA-Seq data in sense (blue) and antisense (green) direction for all open reading frame-containing transcribed regions (ORF-Ts). ORF-Ts are sorted by length and aligned at their transcription start site (TSS). **C**. Smoothed, raw Cbc2 RNA-binding strength as measured by the number of PAR-CLIP T-to-C transitions per U site in sense (blue) and antisense (green) direction for all ORF-Ts sorted by length and aligned at their TSS. **D**. Normalization of PAR-CLIP signals reduces noise. Cbc2 occupancy as estimated by dividing the number of T-to-C transitions for each U site by the RNA-Seq signal at the corresponding genomic position in sense (blue) and antisense (green) direction for all ORF-Ts. **E**. Normalization of PAR-CLIP signals enables interpretation as occupancy profiles. Whereas raw PAR-CLIP binding strength (shown in C) strongly depends on mRNA level, normalized occupancies (shown in D) are independent of mRNA levels. Y-axis: percentage of transcripts whose mean occupancy within the first 90 nt of a transcript is larger than the average of this mean over all ORF-Ts. **F**. Normalization abolishes the dependence of estimated occupancy on mRNA level. Pearson correlation between mRNA level and the PAR-CLIP binding strength in the first 90 nt of each ORF-T before (top) and after (bottom) RNA abundance normalization. **G**. Cbc2-binding profiles are independent of factor occupancy. Transcript-averaged Cbc2 occupancy for three mRNA level classes [100–90 %, 70–60 %, and 40–30 % expression quantile].

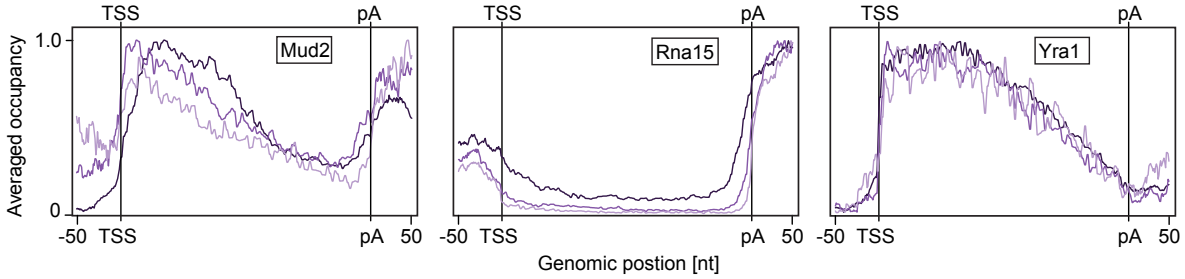

Figure 5.17: **Occupancy profiles are independent of factor occupancy**. Transcript-averaged occupancy for three expression level classes [100–90 %, 70–60 %, and 40–30 % expression quantile] of Mud2, Rna15, and Yra1. This demonstrates that occupancy profiles are reliable even at lowly occupied genes.

The normalization thus leads to realistic profiles and prevents misinterpretation due to systematic over-representation of abundant transcripts. In the normalized data, strongest binding of Cbc2 was observed within the first ~90 nt downstream of the transcription start site (TSS) within the 5'-untranslated region (5'-UTR) of mRNAs (Figures 5.15 to 5.17). The normalization also enhanced Cbc2 signals on ncRNA transcripts (Figure 5.16C–D, green panels), facilitating the detection of capped ncRNAs (Figure 5.17 and Supplementary Figure S1). Widespread Cbc2 binding was observed at the 5'-end of divergent ncRNA transcripts that emerged from bidirectional promoters antisense to mRNAs. Cbc2 sites were found starting at ~120 nt upstream of the TSS of the sense transcript, with the peak of Cbc2 cross-linking at ~250 nt (Figure 5.18, upper panel). This is consistent with the presence of two distinct Pol II initiation complexes for sense and divergent transcription from bidirectional promoters (Rhee and Pugh, 2012), and indicates that divergent transcripts are capped before they associate with the Nrd1 complex that triggers their degradation (Jensen *et al.*, 2013; Mischo and Proudfoot, 2013; Schulz *et al.*, 2013). Cbc2 also cross-linked to antisense RNA 100–300 nt upstream of the polyadenylation (pA) site, identifying capped antisense ncRNAs at the 3'-ends of many genes (Figure 5.18, lower panel).

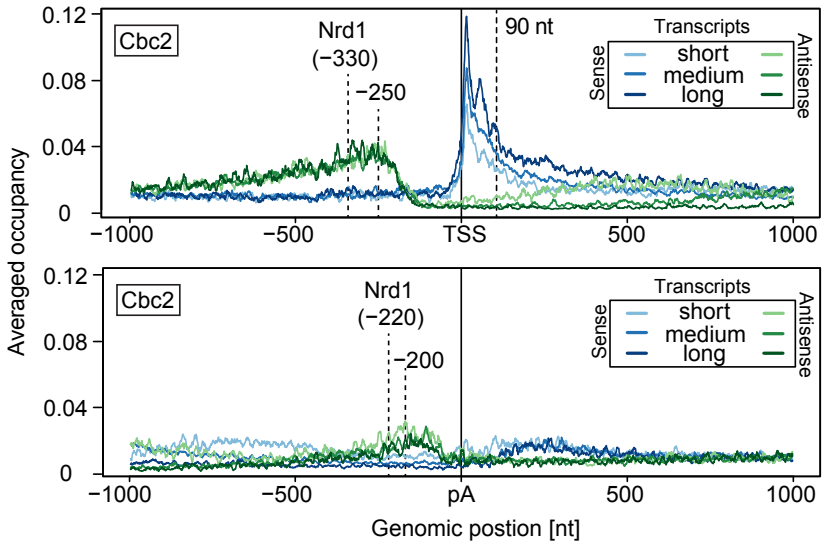

Figure 5.18: **Transcript-averaged Cbc2 occupancies in sense (blue) and antisense (green) directions**. Occupancies are centered at the TSS (upper panel) and the polyadenylation site (pA) (lower panel), for short [0-1 kb], medium [1-2 kb], and long [2-5 kb] transcripts.

We also identified Cbc2-binding sites in cryptic unstable transcripts (CUTs) and stable untranslated transcripts (SUTs) (Wery *et al.*, 2011), with stronger signals for CUTs (Figure 5.19A–B). The Cbc2 data also enabled comparison with the recent CRAC-based mapping of Cbc1, the other subunit of CBC (Tuck and Tollervey, 2013). Both Cbc1 and Cbc2 showed RNA interactions at the 5'-ends of transcripts, cross-validating the studies. However, the PAR-CLIP protocol and normalization procedure used here apparently led to more focused signals at RNA 5'-ends and enhanced signals for short-lived RNAs and RNAs with low abundance, prompting us to use it for an investigation of factors involved in the recognition of pre-mRNA elements.

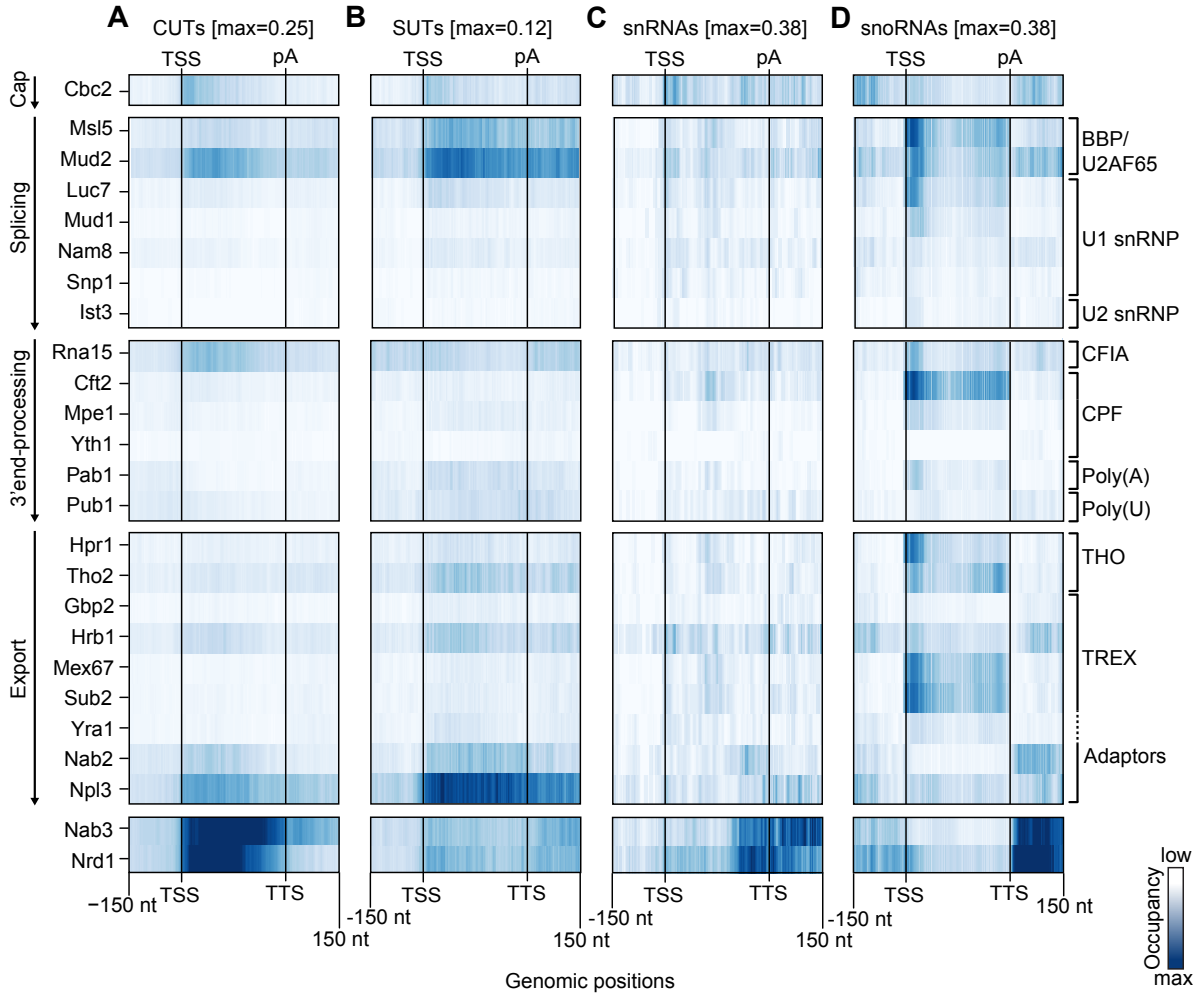

Figure 5.19: **Overview of occupancy profiles of all investigated proteins on non-coding RNAs**. For each factor, transcript class-averaged occupancies aligned at TSS and scaled to coincide at their transcription termination sites (TTS). Occupancies for each factor were divided by the maximum attained over any transcript, including all ORF-Ts. **A**. Occupancy profiles on cryptic unstable transcripts (CUTs), **B**. on stable unnnotaed transcripts (SUTs), **C**. small nuclear RNAs (snRNAs), and **D**. small nucleolar RNAs (snoRNAs).

5.3.2 Conserved recognition of pre-mRNA introns

Intron recognition is the initial step in pre-mRNA splicing and was extensively studied in vitro (Will and Lhrmann, 2011). It begins with binding of BBP to the branch point (BP) and binding of U2AF65 to a pyrimidine-rich region between the BP and 3'-splice site (3'-SS), and continues with binding of the U1 snRNP to the 5'-SS. The resulting complex E (Figure 5.20) is then remodeled, and U2 snRNA displaces BBP by base pairing with the BP region, positioning U2 snRNP near the 3'-SS and giving rise to complex A (Figure 5.20).

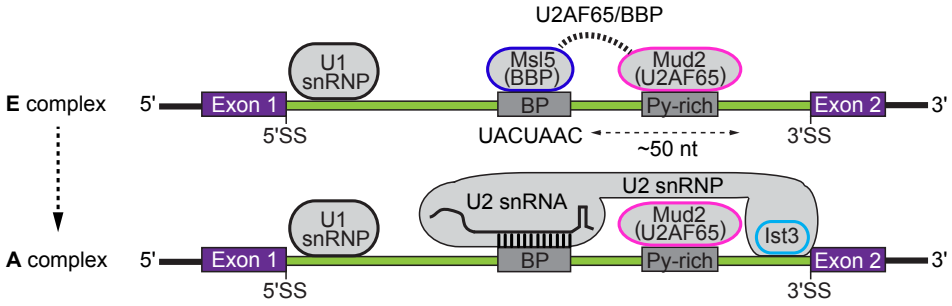

Figure 5.20: **Model of factors recognizing an intron during formation of E and A complexes**.

The protein-RNA interactions underlying intron recognition have not been systematically analyzed *in vivo*. Despite the rapid degradation of introns *in vivo*, our protocol could capture intron sequences bound by splicing factors involved in intron recognition (Figures 5.21A-B and 5.22). Cross-linking signals for the BBP homologue Msl5 and the U2AF65 homologue Mud2 spanned entire introns and showed peaks near the 5'-SS and the 3'-SS, respectively (Figure 5.21D). The BP motif UACUAAC was detected around Mud2- and Msl5-bound sites in intron-containing genes (Figures 5.21A and 5.22) and is generally located within ~50 nt upstream of the 3'-SS (Figure 5.21E). When we averaged occupancy profiles after aligning introns at the BP, Msl5 displayed a peak on the BP (Figure 5.21F), consistent with binding of yeast Msl5 to the BP *in vivo*. Mud2 and Ist3 peaked 15 nt and 27 nt downstream of the BP, respectively (Figure 5.21F). Thus we could resolve binding of the U2AF65 homolog Mud2 to a pyrimidine/U-rich region that was defined *in vitro* in the human system (Mackereth *et al.*, 2011). These results agree with *in vitro*-derived functions of the Msl5-Mud2 complex in BP recognition (Berglund *et al.*, 1997), and in bridging between the BP and U1 snRNP at the 5'-SS (Abovich and Rosbash, 1997). Msl5 and Mud2 cross-linked also to intron-less RNAs (Figures 5.15 and 5.19), consistent with scanning of RNAs for U-rich regions by the U2AF65-BBP complex. Cross-links of U1 snRNP subunits peaked ~17 nt downstream of the 5'-SS (Figure 5.21D). Motif searches around cross-linking peaks ($\pm$25 nt) detected the consensus 5'-SS sequence GUAUGU in Luc7, Mud1, Nam8, and Snp1 data (Figures 5.21A and 5.22). As expected, cross-link sites of U1 snRNP subunits were not significantly enriched around the BP (Figure 5.23). The U2 subunit Ist3 cross-linked mainly ~10 nt upstream of the 3'-SS (Figure 5.21D).

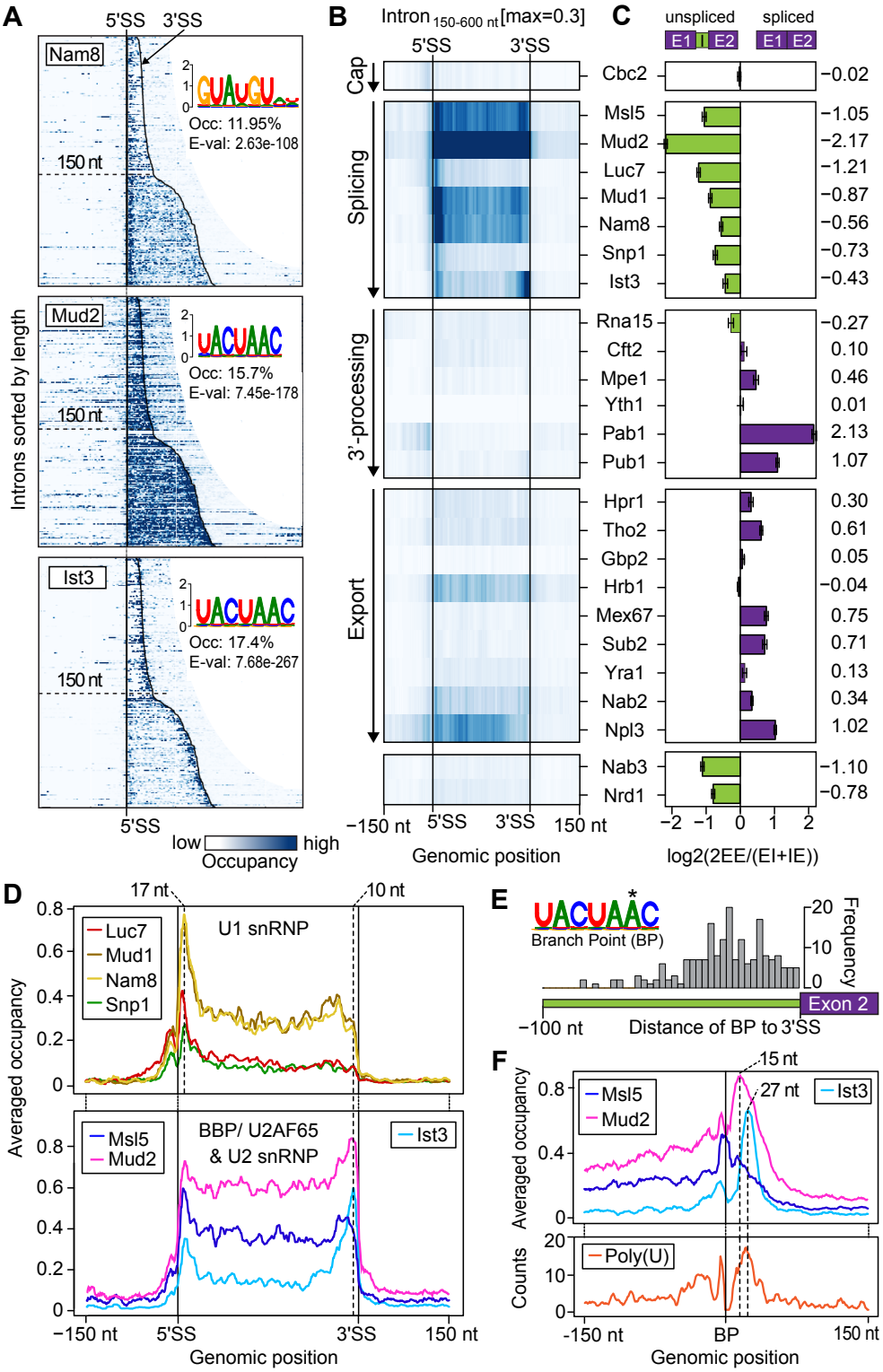

Figure 5.21: **Conserved recognition of pre-mRNA introns *in vivo*.** **A**. Normalized and smoothed occupancy profiles of U1 subunits Nam8, Mud2 (human U2AF65) and U2 subunit Ist3 around introns of up to 600 nt length. Introns were sorted by length and aligned at their 5'-splice site (5'-SS). **B**. Transcript-averaged occupancy profiles of all factors around introns between 150 and 600 nt length. **C**. Splicing factors show high affinity for unspliced RNAs. Splicing indices (Methods) indicating the degree of preference for spliced versus unspliced RNAs for all factors. **D**. Intron-averaged factor occupany profiles show binding of U1 snRNP near the 5'-splice site and binding of the U2 snRNP and the commitment complex (BBP/U2AF65) over the entire intron with a peak at the 3'-splice site (3'-SS). **E**. The branch point (BP) lies within 50 nt upstream of the 3'-SS. Distance distribution of the branch point (BP) motif from the 3'-SS. **F**. Yeast Msl5 (human BBP) binds the BP in vivo, whereas Mud2 (U2AF65) and U2 snRNP (Ist3) bind downstream of the BP. Transcript-averaged occupancy profiles of Msl5, Mud2, and Ist3, centered at the BP (top), compared to the poly(U) distribution over the same region (bottom).

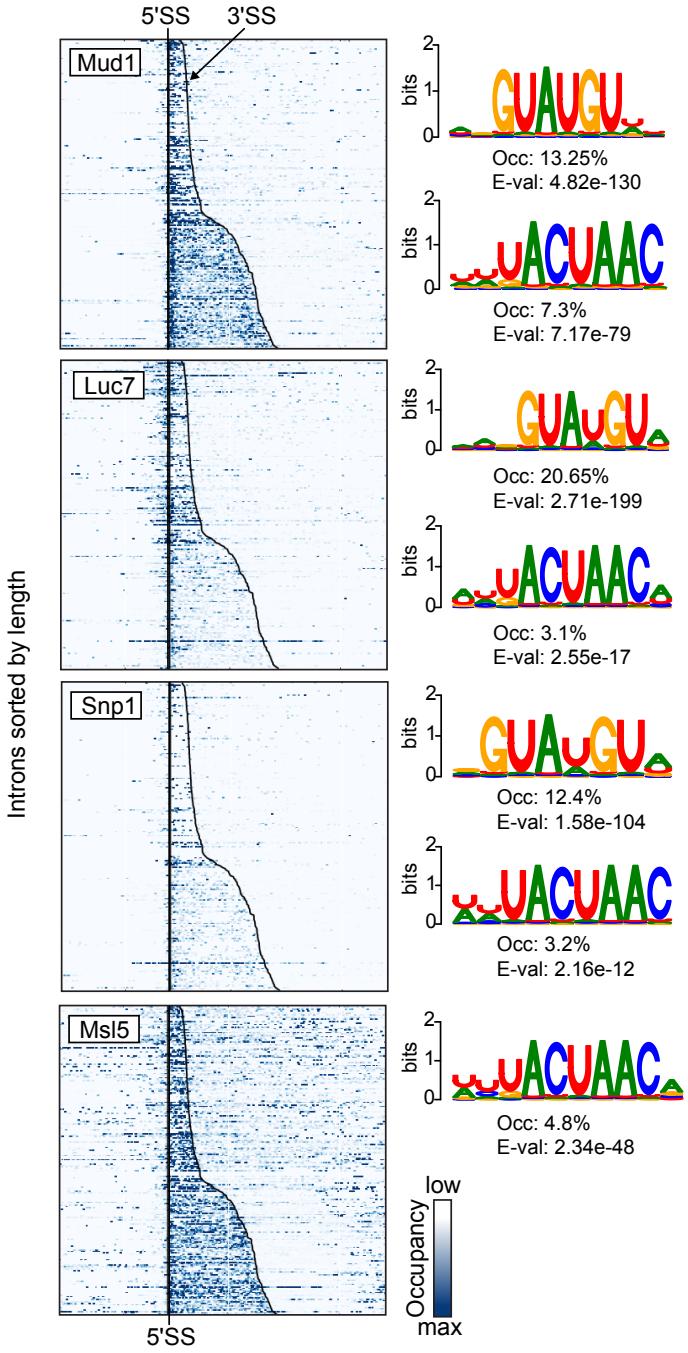

Figure 5.22: **Occupancy of splicing factors around introns.** Occupancy profiles of the U1 snRNP splicing factors Mud1, Luc7, and Snp1, and the BBP/Msl5 derived from PAR-CLIP experiments for all intronsEach line represents an intron, and introns are sorted by length and aligned at their 5'-SS. Motifs found by XXmotif to be enriched ±20 nt around the cross-linking sites are shown next to the factors around which they are enriched.

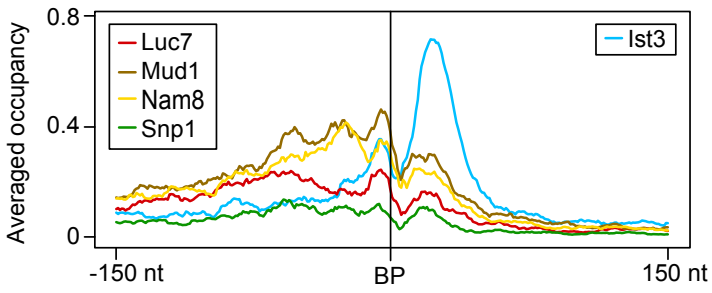

Figure 5.23: **Occupancy of splicing factors around the branch point (BP)**. Average occupancy profiles for the U1 snRNP Luc7, Mud1, Nam8 and Snp1, and the U2 snRNP Ist3 around the BP.

These results agree with the *in vitro*-derived binding of U1 and U2 snRNPs near the 5'-SS and the 3'-SS, respectively (Will and Lhrmann, 2011). The splice site RNA motifs were apparently responsible for recruitment of U1 and U2 snRNPs, because their subunits generally did not cross-link to intron-less RNAs (compare Figure 5.15 and Supplementary Figure S1). To investigate the order of factor binding to introns, we calculated a 'splicing index' (Figure 5.21C, Supplementary Figure S2, and Methods) (Schneider *et al.*, 2012). All splicing factors obtained negative splicing indices, demonstrating preferential binding to unspliced RNA. The strongest preference for unspliced over spliced RNA was obtained for Mud2, the weakest for Ist3. Thus our in vivo data support the two-state model of intron recognition derived from *in vitro* studies (Figure 5.20).

5.3.3 Unified recognition of pre-mRNA polyadenylation sites

In human cells, recognition of the pA site involves several RNA sequence elements that are bound by the cleavage and polyadenylation specificity factor (CPSF) complex (Mandel *et al.*, 2008; Chan *et al.*, 2011; Proudfoot, 2011). The CPSF subunit CPSF-160 recognizes the pA signal (PAS) sequence AAUAAA upstream of the pA site. Subunits CPSF-100 and CPSF-30 bind neighboring U-rich sequences, and subunit CPSF-73 cleaves the RNA (Mandel *et al.*, 2008; Chan *et al.*, 2011). Homologous subunits are found in the yeast CPSF counterpart CPF, which also contains additional proteins, such as Mpe1 (Vo *et al.*, 2001). After extensive trials we could map CPF subunits Cft2/Ydh1 (CPSF-100), Yth1 (CPSF-30), and Mpe1 onto transient pre-mRNA (Figure 5.24A). Cft2 cross-linked to regions flanking the pA site, consistent with binding near the cleavage site detected in vitro (Dichtl and Keller, 2001). Yth1 showed a peak ~17 nt upstream of the pA site, consistent with *in vitro* results (Barabino *et al.*, 2000), and with localization of its human counterpart CPSF-30 in vivo (Martin *et al.*, 2012). Mpe1 gave rise to a peak ~6 nt upstream the pA site, explaining why it is an essential factor required for 3'-processing (Vo *et al.*, 2001). Although Cft1/Yhh1 (CPSF-160) and Ysh1 (CPSF-73) did not show PAR-CLIP signals, these data locate the yeast CPSF counterpart CPF at the pA site *in vivo* and define many of its subunit-RNA interactions. Human CPSF is assisted by the CstF complex, which binds to pre-mRNA downstream of CPSF (Mandel *et al.*, 2008; Chan *et al.*, 2011). However, the yeast CstF counterpart CFIA is believed to bind upstream of the CPSF counterpart CPF (Mandel *et al.*, 2008; Chan *et al.*, 2011), and this model is based on *in vitro* evi-

dence that the CFIA subunit Rna15 binds upstream of the pA site (Gross and Moore, 2001). In contrast, we observed very strong cross-linking of the CFIA subunit Rna15 downstream of the pA site *in vivo*, with a peak at ~16 nt (Figure 5.24A).

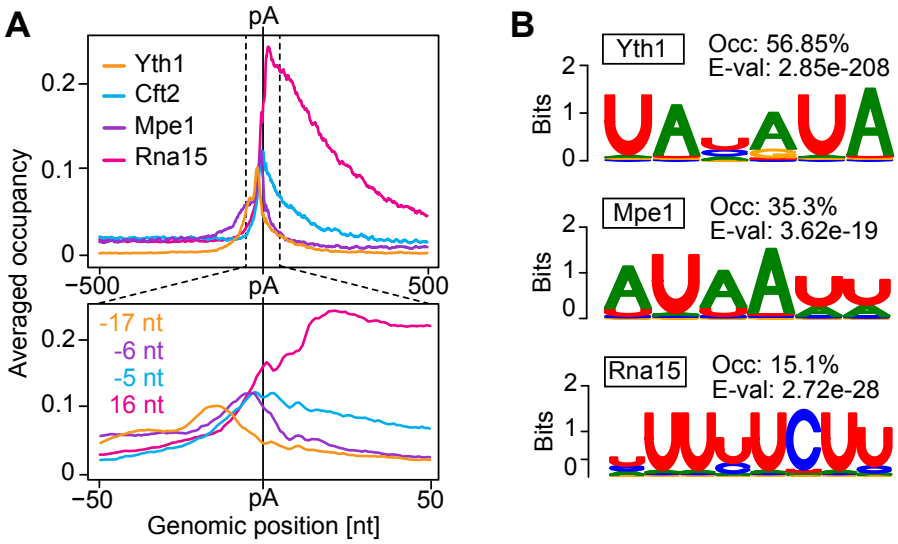

Figure 5.24: **Binding preferences of Rna15 and CPF subunits Cft2, Mpe1, and Yth1)**. **A**. Averaged occupancy profiles of Rna15 and CPF subunits, aligned at the pA site show that, in vivo, CPF binds at the pA site while CFIA binds downstream. **B**. RNA motifs enriched in a window of ±25 nt around the cross-linked sites with fraction of occurrence and XXmotif E-value.

These results agree with an alternative *in vitro* study (Dichtl and Keller, 2001), and with binding of the human Rna15 homologue CstF64 downstream of the pA site in vivo (Martin *et al.*, 2012). Thus CFIA is located downstream, rather than upstream, of the pA site and CPF, consistent with the position of the human CstF complex downstream of the pA site and downstream of the CPF counterpart CPSF. These results lead to a unified model for pA site recognition by the two conserved 3'-processing complexes bound to pre-mRNA (Figure 5.25C).

5.3.4 Definition and decoration of mRNA 3'-ends

To investigate how the pA site is defined in the pre-mRNA sequence, we searched for sequence motifs around cross-linking peaks. Peaks for Yth1 and Mpe1 often contained the motif UAUAUA and AUAAUU, respectively, whereas Rna15 often bound at the motif UUUUCUU (Figure 5.24B). Cft2, Mpe1, and Yth1 preferred RNA sites containing U/A-rich tetramer sequences, whereas Rna15 bound regions that were enriched with the A-less tetrameric motifs UUUU and UCUC (Figure 5.25A). Although these motifs are often absent from pA regions, a systematic analysis revealed a characteristic, conserved signature of RNA dinucleotides around pA sites (Figure 5.25B). The pA sites are strongly enriched with dinucleotides UC (at position +1 downstream of the cleavage site), CA/AA (+2), AA/UC (+3), CA/AA (+4), and AA (+5). Regions with strong UU bias flank pA sites on both sides (-15 to -2 and +6 to at least +25). Further upstream, a region with marked AA bias (-25 to -15), transitions into a region with enrichment for AU/UA

dinucleotides (-90 to -25). The distinct A/U signature at the pA site apparently directs binding of CPF subunits upstream and around pA sites and binding of Rna15 downstream of pA sites, because these factors exhibit corresponding sequence preferences (Figure 5.25A). In some yeast mRNAs, the A-rich upstream region contains a positioning element that may bind Cft1 and may correspond to the human polyadenylation signal (Guo and Sherman, 1996), and a UA-rich efficiency element (Guo *et al.*, 1995) that may bind CFIB/Hrp1 (Kessler *et al.*, 1997). These elements are however dispensable for RNA cleavage (Dichtl and Keller, 2001), indicating that the A/U signature, rather than specific sequence elements, underlies pA site recognition. A similar A/U dinucleotide signature can explain the previously described bias for A and U nucleotides around human pA sites (Martin *et al.*, 2012) and matches the conserved arrangement of 3'-processing factors.

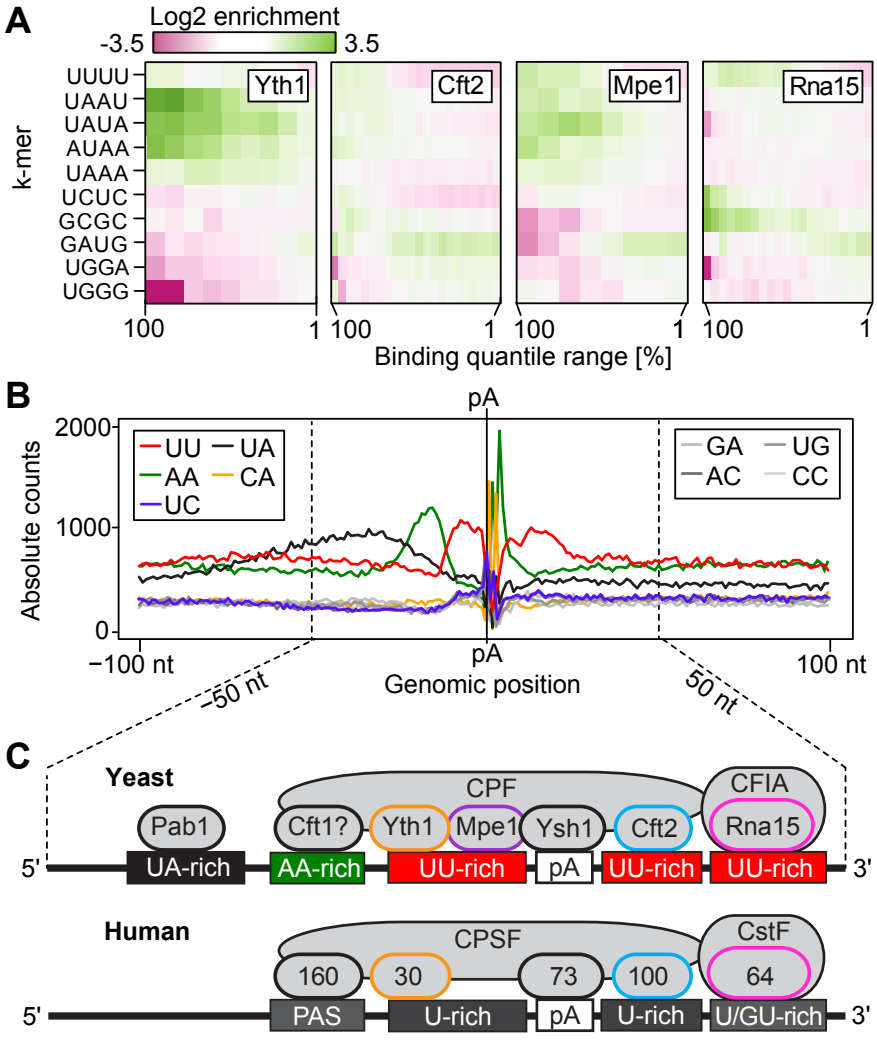

Figure 5.25: **Unified model for polyadenylation (pA) site recognition *in vivo*.** **A**. 3'-processing factors have distinct tetramer-binding preferences. Log-odd scores for enrichments of selected tetramers (y-axis) for bins of binding sites ranging from 100 % occupancy to 1 % occupancy (x-axis). **B**. 'A/U signature': Frequency of UU, AA, UA and CA dinucleotides around the pA site. The AU frequency profile is very similar to the UA profile. **C**. Unified model for pA site recognition in *S. cerevisiae* and human by the two major, conserved 3'-processing complexes CPF (CPSF) and CFIA (CstF) bound on pre-mRNA.

Additional data showed that the region upstream of pA sites bind Pab1 and Pub1 (Figure 5.26). Both factors gave rise to cross-linking near the 3'-end of mRNAs (Figure 5.26A, C). Pab1 bound upstream of the pA site to the sequence motif UAUAUA (Figure 5.26A-C) as described (Riordan *et al.*, 2011; Tuck and Tollervey, 2013). Pub1 occupied both UA-rich regions in the 3'-UTR as described (Vasudevan and Peltz, 2001; Duttagupta *et al.*, 2005) but also poly(U) tracts (Figure 5.26B), and also bound upstream of the open reading frame (ORF) in the 5'-UTR (Figure 5.27) as described (Cui *et al.*, 1995; Ruiz-Echevarra *et al.*, 1998; Ruiz-Echevarra and Peltz, 2000). Pub1 and Pab1 were generally depleted from the translated ORF (Figures 5.26 and 5.27), consistent with a mainly cytoplasmic location of these factors.

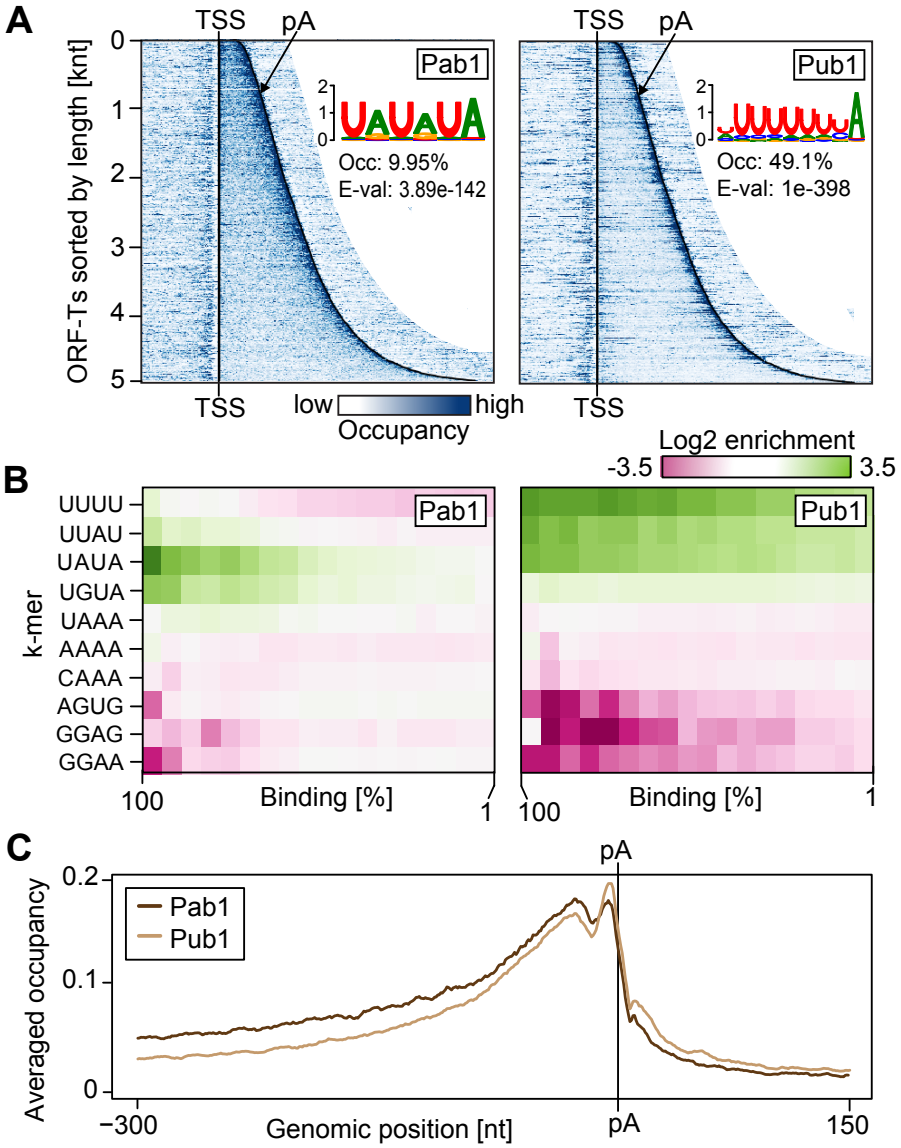

Figure 5.26: **Pab1 and Pub1 bind UA-and U-rich sequences at mRNA 3'-ends**. **A**. Occupancy profiles of the "poly(A)-binding protein" Pab1 and the "poly(U)-binding protein" Pub1 derived from PAR-CLIP data in sense direction for all ORF-Ts with motifs that were found enriched around binding sites (± 25 bp). **B**. Pab1 and Pub1 bind to U/A-rich sequences. Log2 enrichment of selected 4-mer motifs around Pab1 (left) and Pub1 (right) binding sites compared to unbound sequence regions, analyzed within 18 equal-sized bins of occupancy quantiles between 100 % and 1 % site occupancy (x-axis). **C**. Averaged occupancy profiles of Pab1, Pub1 and Yth1 derived from PAR-CLIP data in sense direction for all ORF-Ts, centered at the pA site of all ORF-Ts.

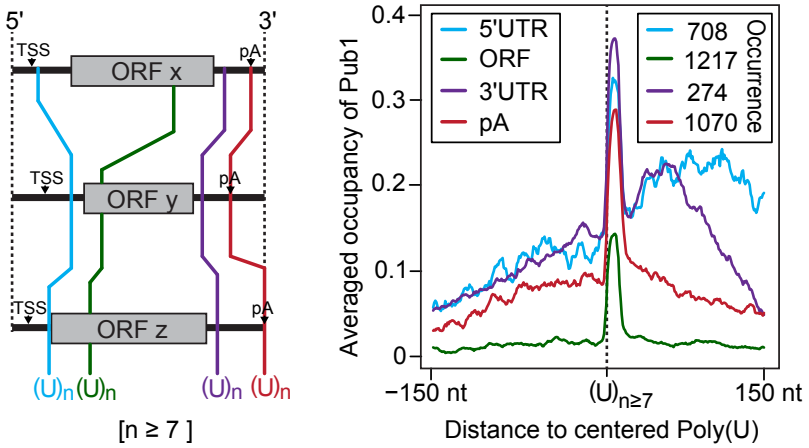

Figure 5.27: **Pub1 preferentially binds poly(U)$_{n\geq 7}$ tracts near the pA site**. Average occupancy profiles of Pub1 around Poly(U)$_{n\geq 7}$ tracts within the 5'-UTR, ORF, 3'-UTR , or near pA sites.

Taken together, these data may be explained as follows. The two major 3'-processing complexes CPF and CFIA preferentially locate their target regions on the pre-mRNA around the pA site via a distinct A/U dinucleotide signature, causing RNA cleavage, polyadenylation, and release of 3'-processing factors, which enables complete decoration of the mRNA 3'-end with Pab1 and Pub1.

5.3.5 Transcription-coupled mRNP export

In our current view, mRNA export begins with the recruitment of the THO/TREX complex during Pol II elongation (Strsser *et al.*, 2002; Luna *et al.*, 2012). Mature mRNA is then exported from the nucleus by the heterodimeric export factor Mex67-Mtr2 (Segref *et al.*, 1997; Grter *et al.*, 1998). Mex67 uses mRNA adaptor proteins such as Nab2, Npl3, and Yra1 (Iglesias *et al.*, 2010; Stewart, 2010; Hackmann *et al.*, 2011; Rodrguez-Navarro and Hurt, 2011). PAR-CLIP analysis revealed similar distributions of the THO subunits Tho2 and Hpr1 over mRNAs (Figure 5.28A and Supplementary Figure S3) and no mRNA preferences, indicating that the THO complex is a general factor associated with Pol II transcripts.

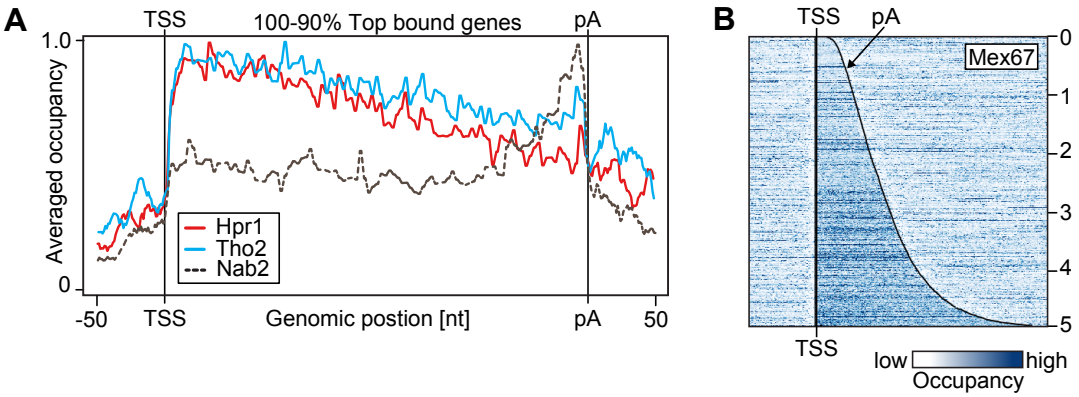

Figure 5.28: **THO complex subunits Tho2 and Hpr1 bind equally. A**. ORF-T-averaged occupancy profiles for Nab2 and the THO complex subunits Tho2 and Hpr1, derived from PAR-CLIP experiments for top-bound ORF-Ts (100–90 %). **B**. ORF-T-averaged occupancy profiles for Mex67.

Tho2 gave stronger signals, consistent with its role in THO complex recruitment (Chvez *et al.*, 2000; Gewartowski *et al.*, 2012). Mex67 bound RNA in vivo (Figure 5.28B), explaining how it remains bound to mRNA after release of adaptor proteins. Mex67 did not show any preference for RNA motifs, consistent with its function as a general export factor, and consistent with data obtained by CRAC (Tuck and Tollervey, 2013). The export adaptors Nab2, Npl3, and Yra1 showed different cross-linking patterns, indicating specific, non-redundant functions (Figure 5.29). The number of mRNAs bound by two or three export adaptors simultaneously was limited (Figure 5.29), showing that these factors exhibit mRNA preferences, as suggested from purification of mRNAs associated with Yra1 (Hieronymus and Silver, 2003). Yra1 occupancy decreased before the pA site, whereas Npl3 also showed cross-linking at 3'-ends, consistent with its influence on pA site choice (Bucheli *et al.*, 2007; Deka *et al.*, 2008).

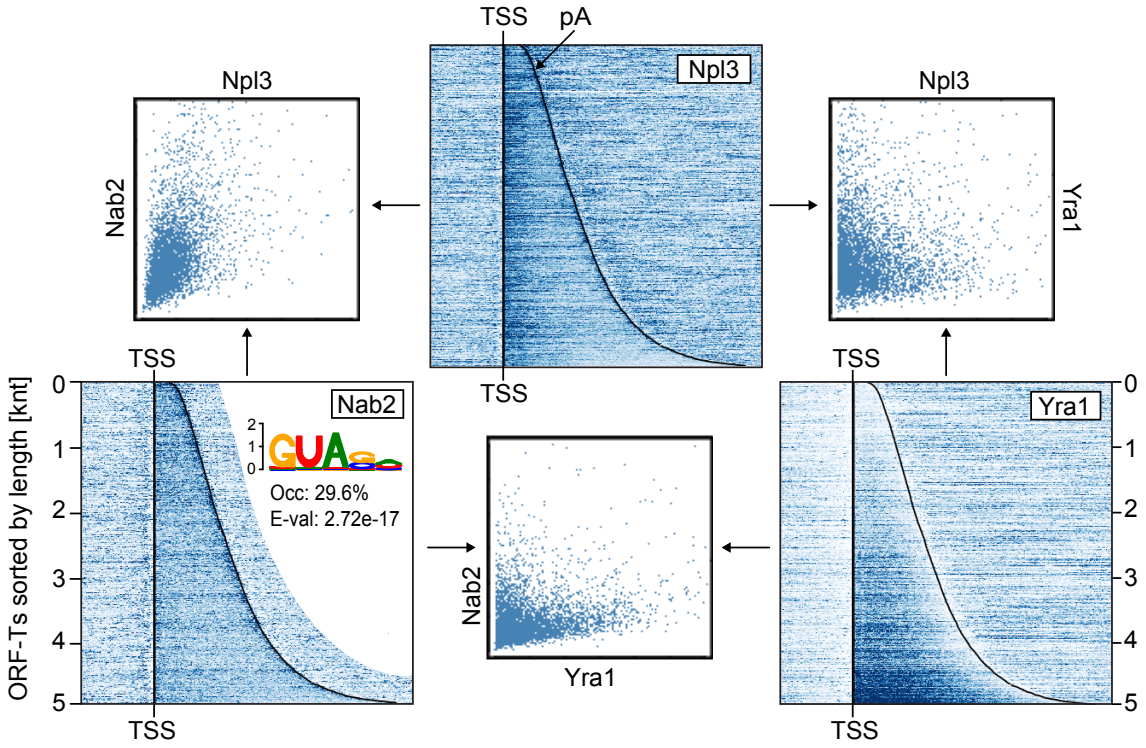

Figure 5.29: **Export adaptors differ in their mRNA-binding preference.** Pairwise correlation scatter plots for occupancies of Yra1, Npl3, and Nab2 on whole ORF-Ts, together with occupancy profiles over all ORF-Ts.

Whereas Nab2 preferentially bound short mRNAs (Figure 5.29), Yra1 and Sub2 preferred long mRNAs (Figure 5.30A). Nab2 occupancy was also stronger at the 3'-ends of ORF-Ts as described (Figure 5.28A) (Tuck and Tollervey, 2013), consistent with its known influence on 3'-processing (Anderson *et al.*, 1993; Green *et al.*, 2002; Hector *et al.*, 2002; Tuck and Tollervey, 2013). Nab2 sites were enriched for the motif GUAG (Figure 5.29) as described (Riordan *et al.*, 2011). Thus these data revealed preferences of components of the mRNA export machinery for RNAs with specific sequences and lengths.

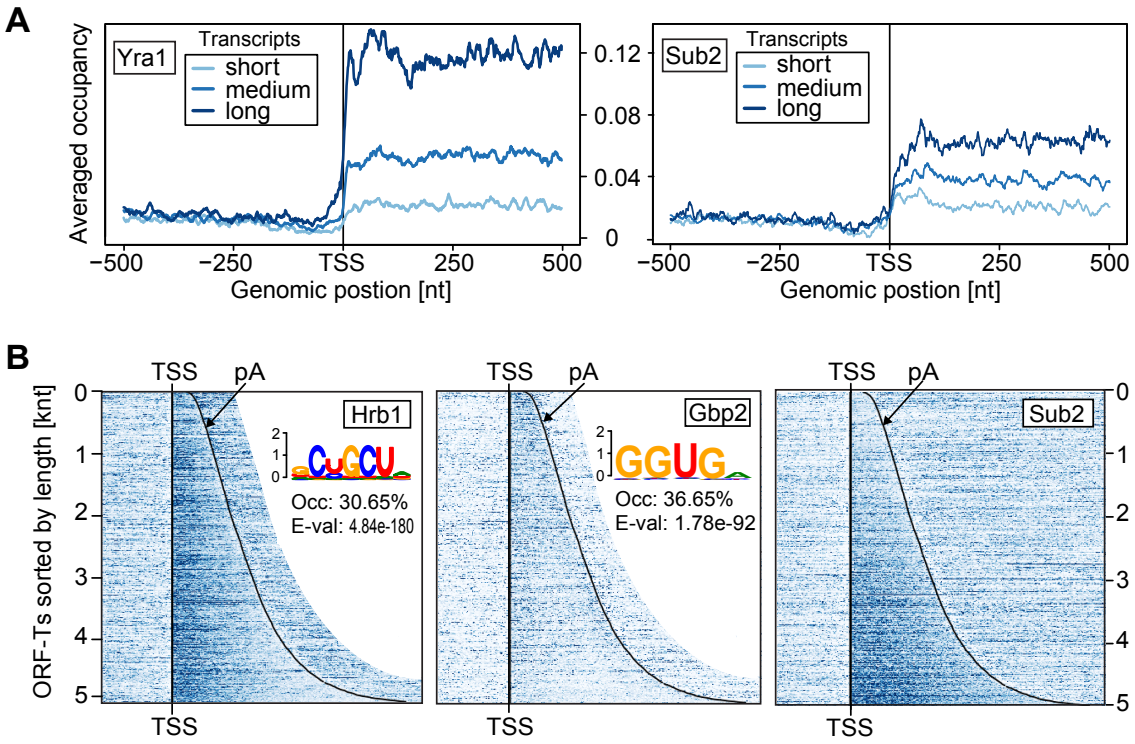

Figure 5.30: **Export adaptor-binding depends on mRNA-length. A**. Transcript-averaged Yra1 and Sub2 occupancies in sense directions, centered at the TSS, for short [0–1 kb], medium [1–2 kb], and long [2–5 kb] transcripts. **B**. Occupancy profiles for SR-like proteins Hrb1 and Gbp2 and for Sub2 for all ORF-Ts.

5.3.6 Global analysis links splicing to 3'-processing

We now subjected all PAR-CLIP data to a global analysis (Figures 5.31 and 5.32). In addition to the splicing index (Figure 5.21C and Supplementary Figure S2), we introduced a 'processing index' (Methods) that measures whether factors preferentially bind uncleaved or cleaved mRNA (Supplementary Figure S4). A plot of splicing versus processing indices (Figure 5.31A) indicates how the composition of protein-RNA complexes is remodeled during mRNP biogenesis (Figure 5.31B).

We further calculated for each pair of factors the Pearson correlation coefficient of the total weighted occupancies over whole transcripts (Figure 5.32, Methods). This measures the extent to which factors co-occupy the same transcripts. We further measured the extent to which two factors co-localize in a window of 25 nt around binding sites (Figure 5.34, Methods). Finally, we computed for each pair of factors the Pearson correlations between their averaged occupancy profiles, to measure the shape similarity of binding profiles (Figure 5.33, Methods). The global analysis provided evidence for an ancient link between splicing and 3'-processing. Splicing factors fell into two groups when sorted by their processing indices (Figures 5.21C and 5.31A). The splicing factors Mud2, Msl5, Snp1, and Luc7 preferentially bound uncleaved RNA, whereas other splicing factors preferred cleaved RNA (Figure 5.31). Mud2 and Msl5 profiles were correlated with those of 3'-processing factors Rna15 and Cft2, and Nam8 correlated with Mpe1, Pab1, and Pub1 (Figure 5.33). Also, Mud2, Msl5, and Nam8 cross-linked near the pA site (Figure 5.15). Nam8 tended to co-localize with Pub1, whereas Mud2 and Msl5 co-occupied transcripts

with Hpr1, Hrb1, Nab2, and Npl3, and they co-localized with Rna15 (Figures 5.32 and 5.34). Indeed, Rna15 preferentially bound unspliced mRNAs (Figures 5.21C and 5.31A), but also showed the lowest processing index (Figure 5.31A and Supplementary Figure S4), confirming its early binding to pre-mRNA (Guo and Sherman, 1996; Leeper *et al.*, 2010).

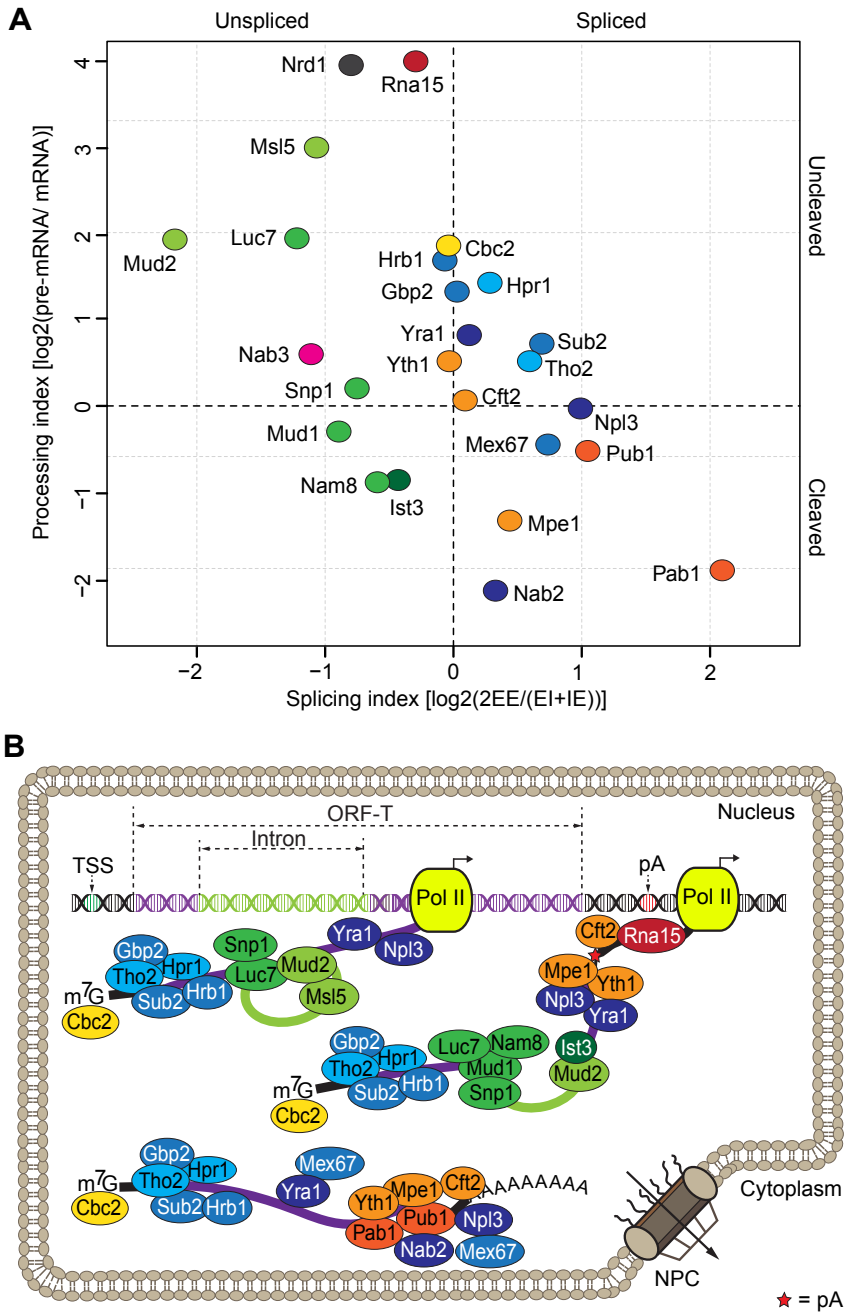

Figure 5.31: **Global analysis reveals links between splicing, 3'-processing, and mRNP export. A**. 'Splicing index' and 'processing index' for all analyzed factors (yellow: capping; orange/red: 3' processing; green: splicing; blue: export; black/pink: RNA surveillance). **B**. Model for mRNP biogenesis resulting from PAR-CLIP-based occupancy measurements.

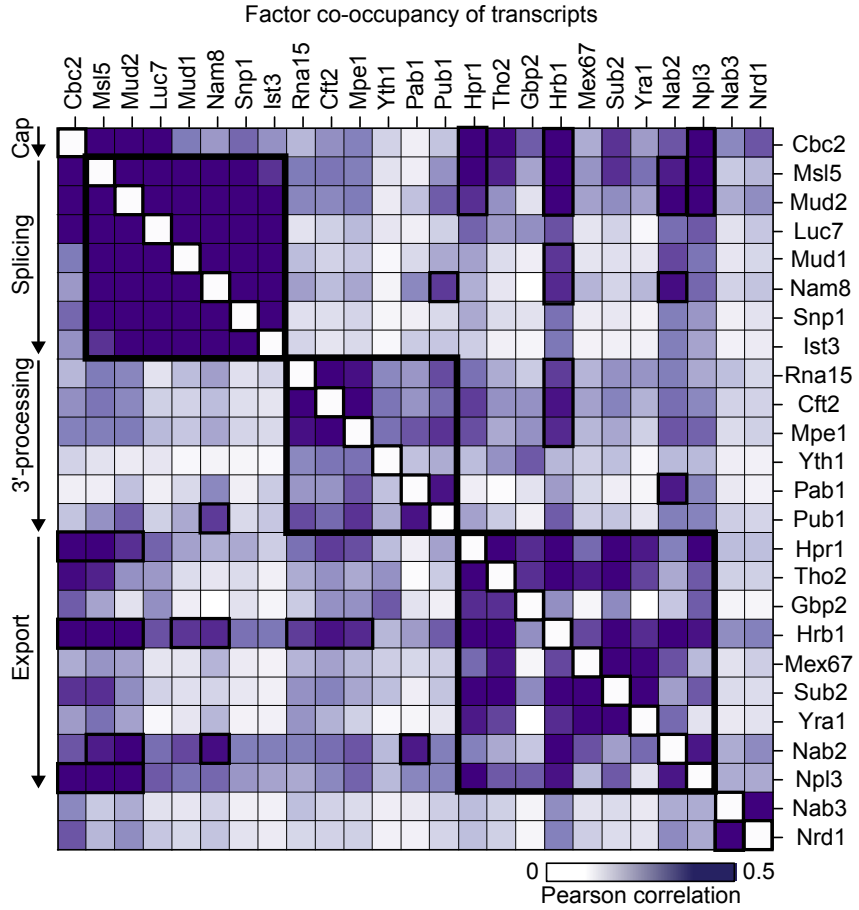

Figure 5.32: **Factor co-occupancy of transcripts.** Pairwise Pearson correlation coefficient of the total weighted occupancies over entire transcripts for all factors.

These results indicate that the machineries for splicing and 3'-processing interact in yeast, as inferred by genetics (Chanfreau *et al.*, 1996), although it is currently believed that such an interaction is restricted to mammels (Shi *et al.*, 2009; Martinson, 2011; Proudfoot, 2011). Indeed, 3'-processing may assist in splicing, like in human cells (Kyburz *et al.*, 2006), but splicing apparently does not influence 3'-processing, because unspliced and spliced transcripts recruit 3'-processing factors to a similar extent.

5.3.7 Transcript surveillance and fate

The global analysis also elucidated how nuclear export is restricted to mature mRNPs. First, export factors preferred spliced over unspliced mRNA, and generally did not bind uncleaved RNA (Figure 5.21C and 5.31A). The highest splicing index was found for Pab1, which binds mature mRNA (Brune *et al.*, 2005), whereas the lowest splicing index was found for Mud2, which is expected to initiate intron recognition (Will and Lhrmann, 2011). Second, binding profiles for export factors except Nab2 differed from those of 3'-processing factors (Figure 5.33), reflecting selection of 3'-processed mRNAs by export factors. Indeed, Mex67 preferred binding to mRNAs lacking the pre-mRNA 3'-tail (Figure 5.31A). Third, the SR proteins Gbp2 and Hrb1 (Windgassen and Krebber, 2003) overlapped with THO/TREX subunits, and Hrb1 tended to bind the same transcripts as the Mud2-Msl5 complex (Figure 5.32).

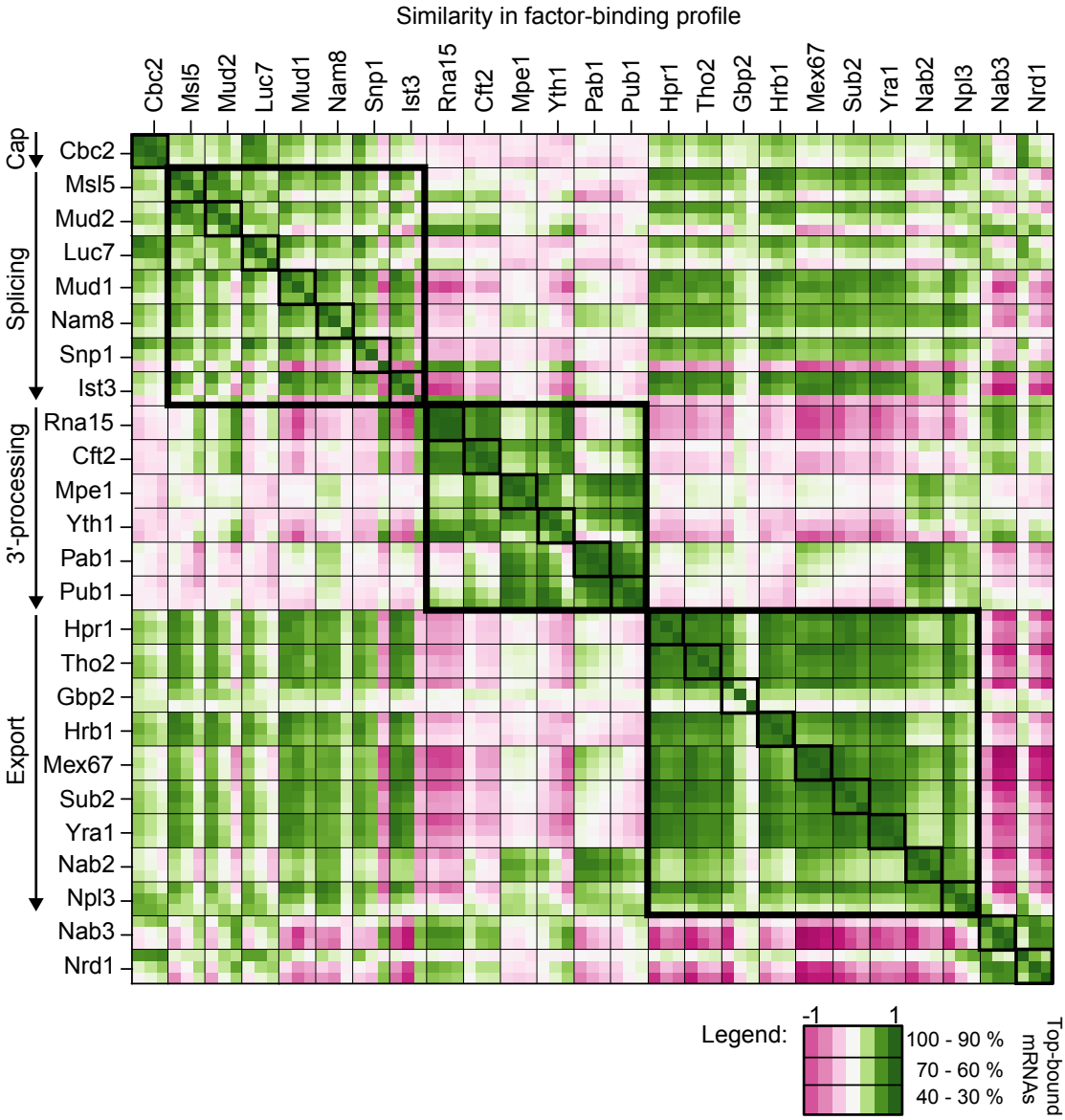

Figure 5.33: **Similarity matrix of factor-binding profiles** shows blocks of functionally linked factors with similar occupancy profiles along transcripts. For each pair of factors, the matrix shows the color-coded Pearson correlations between the occupancy profiles averaged over ORF-Ts in occupancy quantile ranges 100 %–90 %, 70 %–60 % and 40 %–30 %. Each cell corresponding to a pair of factors thus shows 3x3 color coded Pearson correlations. In cells that are entirely green in all 3x3 subcells, profiles are similar to each other across all three quantile ranges of occupancy. Note that export factors profiles show similarity to splicing factor profiles.

This is consistent with a role of Gbp2 and Hrb1 in restricting mRNA export to spliced transcripts (Hackmann *et al.*, 2014). Gbp2 and Hrb1 showed distinct RNA-binding motifs (Figure 5.30B), and Hrb1 co-localized with splicing factors Luc7 and Snp1 (Figure 5.34), consistent with a role in splicing (Shen and Green, 2006; Kress *et al.*, 2008; Will and Lhrmann, 2011). A subset of 3'-processing factors also showed occupancy profiles that were similar to those of RNA surveillance factors Nrd1 and Nab3 (Figure 5.33). Rna15 co-localized with Nrd1 and Nab3 on transcripts (Figure 5.34), and cross-linked to aberrant divergent ncRNAs (Supplementary Figure S1). This indicates that some 3'-processing factors are part of the RNA surveillance machinery that terminates and degrades aberrant RNAs, as predicted by genetics (Mischo and Proudfoot, 2013). Nrd1 and Nab3 co-localized with Cbc2 (Figure 5.34), and preferentially bound uncleaved pre-mRNA, in accordance with their role in triggering early termination of transcription. These observations are consistent with a general nuclear RNA surveillance pathway and suggest that during RNA synthesis a transient surveillance/3'-processing complex takes a decision whether a transcript is subjected to degradation or to polyadenylation and nuclear export.

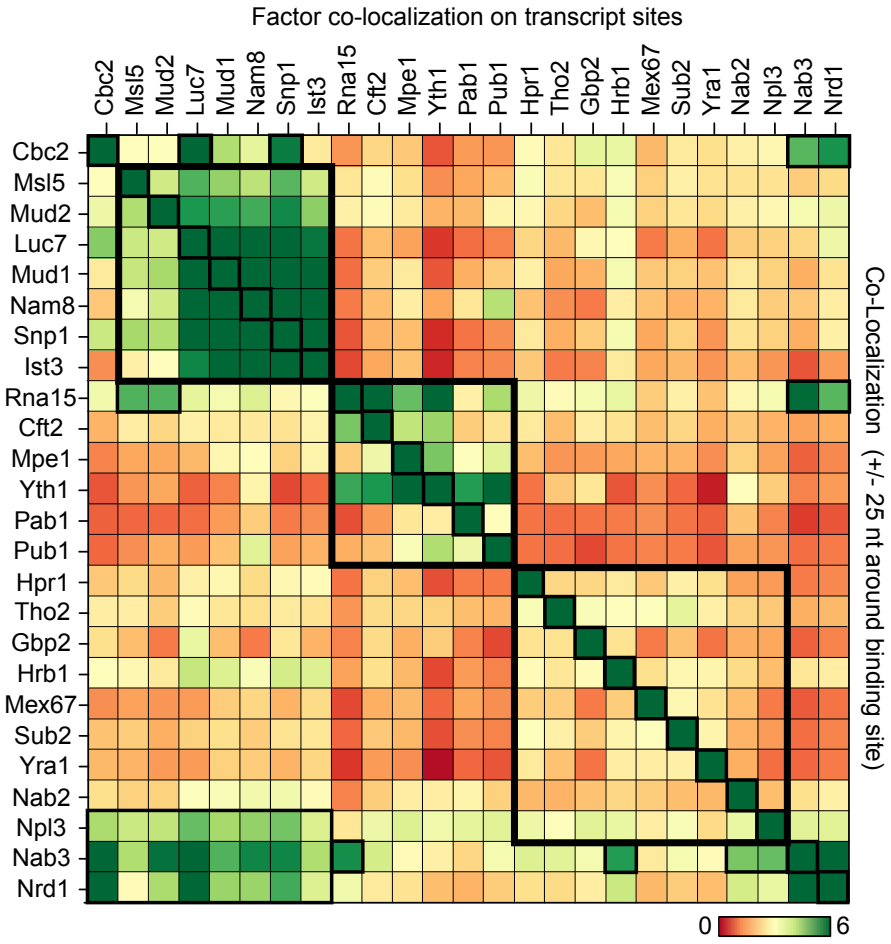

Figure 5.34: **Co-localization** of factor A (rows) within ± 50 nt of strong binding sites of factor B (columns), analyzed over all ORF-Ts.

6 Conclusion and Outlook

6.1 Conclusion

Here we report high-confidence transcriptome maps for 23 protein factors involved in mRNP biogenesis in the eukaryotic model system *S. cerevisiae*. We demonstrate that PAR-CLIP efficiently captures short-lived unspliced and uncleaved pre-mRNAs. This allowed mapping of splicing factors onto short-lived introns and of 3'-processing factors within regions downstream of the pA site, which are rapidly removed and degraded in cells. The distribution of factors over various pre-mRNA species that result from events during mRNP biogenesis enabled integration of the data into a model for mRNP biogenesis based on factor occupancy. Yeast is ideally suited for studying pre-mRNA recognition because yeast genes contain only single introns, thus limiting pre-mRNA complexity.

Most notable insights from our data include the observation of intron recognition by the Mud2-Msl5 (human U2AF65-BBP) and the snRNPs U1 and U2 *in vivo*, a unified, conserved arrangement of the two major 3'-processing complexes CPF and CFIA (human CPSF and CstF) at the pA site, and links of the 3'-processing machinery to RNA splicing and nuclear RNA surveillance. An analysis of the RNA sequences underlying the cross-linked sites recovered known splicing motifs, revealed a characteristic A/U dinucleotide signature around the pA site, defined eight specific RNA motifs bound by biogenesis factors, of which three were new, and showed that most factors exhibited binding preferences for certain RNA tetrameric motifs. These results support the emerging concept that RNA-binding factors generally show binding preferences, whereas DNA-binding factors exhibit binding specificities. In particular, 3'-processing factors detect a strong signature of A/U dinucleotides flanking pA sites, but do not bind an extended, highly conserved sequence motif that could be detected by standard motif discovery tools. To achieve high target specificity, multiple interactions of RNA-binding subunits within a functional complex are often required. In addition, factors often interact with other proteins, such as the Pol II CTD. Synergistic factor binding is evident within the machineries for splicing and 3'-processing and explains how these machines locate sites in pre-mRNA despite a scarcity of motifs and poor sequence conservation. It also explains how mRNA, which is restricted in its sequence due to its coding nature, can evolve to specifically bind multifactor complexes.

Finally, the data provide new insights into the mRNP life cycle and a resource for further studies. A global analysis of the data revealed that processes involved in mRNP biogenesis are more tightly coupled than generally thought. An ancient link between 3'-processing and splicing apparently coordinates both processes and generates mature mRNPs that are selected for nuclear export. In particular, we observed direct RNA interactions of splicing factors at the pA site and a differential distribution of splicing factors on pre-mRNAs before and after RNA 3'-cleavage.

6.2 Future perspectives

Although this work greatly enhances our knowledge about principles underlying mRNP biogenesis and (post-)transcriptional regulation, a lot of important questions are still unanswered. For instance, it will be important to work out how 3'-processing may influence spliceosome dynamics and, more generally, how the composition of protein-RNA complexes is remodeled during mRNP biogenesis. Here, some of the ideas and group-internal **projects** shall be mentioned:

Similar to the "Nrd1 project" (Schulz *et al.*, 2013), experiments could be performed to further investigate the Rat1-Rai1 pathway in yeast. For this purpose, the exonucleases Rat1 and Rai1, as well as various interacting partners (Rtt103, Yth1, etc.), will be chipped and clipped. Additionally, 4tU-Sequencing before and after anchor-away will be performed and correlated with PAR-CLIP results. To observe effects on transcription, Pol II ChIP-seq profiles are necessary after the anchor-away experiments, especially in case of Rat1. Previous PAR-CLIP experiments showed binding of Spt5 downstream of pA site, and might have revealed an unknown connection to Rat1, Rai1 and/or Rtt103. Consequently, 4tU-Sequencing should also be performed after the yeast DSIF complex (Spt4/5) was anchor-awayed. Spt5 comprises KOW domains that might be responsible for RNA-binding. Deletions of these Spt5 KOW domains might lead to termination defects that can be measured.

The PAF complex was first identified in yeast as a Po II-associated factor. PAF comprises five subunits (Paf1, Ctr9, Leo1, Rtf1, and Cdc73) and interacts with the TBP, Spt4/5, and FACT (Carrozza *et al.*, 2003; Zhou *et al.*, 2009; Liu *et al.*, 2009). ChIP experiments revealed the presence of the PAF complex at both promoter and coding regions of transcriptionally active genes. However, some of the PAF subunits might bind RNA directly, and CLIP profiles might give new insights into the PAF-associated network.

The CFIA subcomplex is partly recruited by the Spt5 CTR and partly by RNA (Swanson *et al.*, 1991; Zhou *et al.*, 2009). It would be interesting to characterize these interactions further. Which subunit of CFIA physically interacts with the CTR could be discovered by fluorescence anisotropy measurements of CFIA subunits with a fluorescently labeled CTR peptide. The RNA-binding of CFIA could be characterized by PAR-CLIP to identify specific RNA sequences or regions that are bound by this subcomplex. This study could finally be enlarged by including the remaining factors and subcomplexes of the 3'-end processing machinery.

One important step will be the switch to the other model system like *Schizosaccharomyces pombe* or *Homo sapiens*. For that reason, it will be necessary to adapt both our PAR-CLIP protocol and computational pipeline to the respective system. In comparison with baker's yeast, both organisms

allow a much more intense study of the splicing machinery, especially the process of alternative splicing. In humans, almost all multi-exonic genes are alternatively spliced, which greatly increases the biodiversity of proteins that can be encoded by the genome. However, genome-wide analysis of alternative splicing remains as a challenging task. Now, we might have a powerful tool to study these complexes and interactions in high resolution.

To address all these questions experimentally, it would be beneficial to automate parts of the ChIP-Seq, RNA-Seq and PAR-CLIP protocol using an automated pipetting robot. For this purpose, different approaches should be tested before being implemented into the control and feedback systems of such an automated station.

Futhermore, following projects will be shared within group-external **collaborations**:

Our PAR-CLIP data might be useful to develop a new motif-discovery tool, based on k-mer analyses and variations, or to optimize already published approaches [with Johannes Söding, Max Planck Institute for Biophysical Chemistry, Göttingen, Germany].

Within this study, we did already clip seven factors of the splicing machinery that are involved in branch point recognition as well as in the U1 and U2 snRNP formation. To get a better overview of steps combining the intronome with the splicing machinery, additional factors of the spliceosome should be clipped and correlated to each other [with Reinhard Lührmann, Max Planck Institute for Biophysical Chemistry, Göttingen, Germany].

One additional project will be the clipping of ribosomal proteins in yeast to estimate the transcription rates on miRNA-like RNA targets upon induction of their expression [with Mihaela Zavolan, The Center of Molecular Life Sciences, University of Basel, Switzerland].

Our PAR-CLIP approach might also be a key method to pursue linkages between metabolism and gene regulation by networks of RNA-binding enzymes (REMs) [with Matthias Hentze, European Molecular Biology Laboratory, Heidelberg, Germany].

It was shown that Set1 and H3K4me3 works in a repressive manner (for coding genes) through promotion of 3'-end antisense transcription of a subset of genes (Margaritis *et al.*, 2012). Here, our PAR-CLIP protocol might provide new insights into the factor binding affinity [with Frank Holstege, University Medical Center Utrecht, Utrecht, Netherlands].

References

Abovich, N. and Rosbash, M. (1997). Cross-intron bridging interactions in the yeast commitment complex are conserved in mammals. Cell, 89(3):403–412.

Akhtar, M. S., Heidemann, M., Tietjen, J. R., Zhang, D. W., Chapman, R. D., Eick, D., and Ansari, A. Z. (2009). TFIIH kinase places bivalent marks on the carboxy-terminal domain of RNA polymerase II. Mol Cell, 34(3):387–393.

Anderson, J. T., Wilson, S. M., Datar, K. V., and Swanson, M. S. (1993). NAB2: a yeast nuclear polyadenylated RNA-binding protein essential for cell viability. Mol Cell Biol, 13(5):2730–2741.

Andrus, A. and Kuimelis, R. G. (2001). Base composition analysis of nucleosides using HPLC. Curr Protoc Nucleic Acid Chem, Chapter 10:Unit 10.6.

Archambault, J., Chambers, R. S., Kobor, M. S., Ho, Y., Cartier, M., Bolotin, D., Andrews, B., Kane, C. M., and Greenblatt, J. (1997). An essential component of a C-terminal domain phosphatase that interacts with transcription factor IIF in Saccharomyces cerevisiae. Proc Natl Acad Sci U S A, 94(26):14300–14305.

Ascano, M., Hafner, M., Cekan, P., Gerstberger, S., and Tuschl, T. (2012). Identification of RNA-protein interaction networks using PAR-CLIP. Wiley Interdiscip Rev RNA, 3(2):159–177.

Barabino, S. M., Ohnacker, M., and Keller, W. (2000). Distinct roles of two Yth1p domains in 3'-end cleavage and polyadenylation of yeast pre-mRNAs. EMBO J, 19(14):3778–3787.

Belotserkovskaya, R., Oh, S., Bondarenko, V. A., Orphanides, G., Studitsky, V. M., and Reinberg, D. (2003). FACT facilitates transcription-dependent nucleosome alteration. Science, 301(5636):1090–1093.

Benner, S. A., Ellington, A. D., and Tauer, A. (1989). Modern metabolism as a palimpsest of the RNA world. Proc Natl Acad Sci U S A, 86(18):7054–7058.

Berglund, J. A., Chua, K., Abovich, N., Reed, R., and Rosbash, M. (1997). The splicing factor BBP interacts specifically with the pre-mRNA branchpoint sequence UACUAAC. Cell, 89(5):781–787.

Birse, C. E., Minvielle-Sebastia, L., Lee, B. A., Keller, W., and Proudfoot, N. J. (1998). Coupling termination of transcription to messenger RNA maturation in yeast. Science, 280(5361):298–301.

Boehringer, D., Makarov, E. M., Sander, B., Makarova, O. V., Kastner, B., Lhrmann, R., and Stark, H. (2004). Three-dimensional structure of a pre-catalytic human spliceosomal complex B. Nat Struct Mol Biol, 11(5):463–468.

Bourbon, H.-M., Aguilera, A., Ansari, A. Z., Asturias, F. J., Berk, A. J., Bjorklund, S., Blackwell, T. K., Borggrefe, T., Carey, M., Carlson, M., Conaway, J. W., Conaway, R. C., Emmons, S. W., Fondell, J. D., Freedman, L. P., Fukasawa, T., Gustafsson, C. M., Han, M., He, X., Herman, P. K., Hinnebusch, A. G., Holmberg, S., Holstege, F. C., Jaehning, J. A., Kim, Y.-J., Kuras, L., Leutz, A., Lis, J. T., Meisterernest, M., Naar, A. M., Nasmyth, K., Parvin, J. D., Ptashne, M., Reinberg, D., Ronne, H., Sadowski, I., Sakurai, H., Sipiczki, M., Sternberg, P. W., Stillman, D. J., Strich, R., Struhl, K., Svejstrup, J. Q., Tuck, S., Winston, F., Roeder, R. G., and Kornberg, R. D. (2004). A unified nomenclature for protein subunits of mediator complexes linking transcriptional regulators to RNA polymerase II. Mol Cell, 14(5):553–557.

Brune, C., Munchel, S. E., Fischer, N., Podtelejnikov, A. V., and Weis, K. (2005). Yeast poly(A)-binding protein Pab1 shuttles between the nucleus and the cytoplasm and functions in mRNA export. RNA, 11(4):517–531.

Bucheli, M. E., He, X., Kaplan, C. D., Moore, C. L., and Buratowski, S. (2007). Polyadenylation site choice in yeast is affected by competition between Npl3 and polyadenylation factor CFI. RNA, 13(10):1756–1764.

Buratowski, S. (2009). Progression through the RNA polymerase II CTD cycle. Mol Cell, 36(4):541–546.

Buratowski, S., Hahn, S., Guarente, L., and Sharp, P. A. (1989). Five intermediate complexes in transcription initiation by RNA polymerase II. Cell, 56(4):549–561.

Burger, K., Mhl, B., Kellner, M., Rohrmoser, M., Gruber-Eber, A., Windhager, L., Friedel, C. C., Dlken, L., and Eick, D. (2013). 4-thiouridine inhibits rRNA synthesis and causes a nucleolar stress response. RNA Biol, 10(10).

Bycroft, M., Grnert, S., Murzin, A. G., Proctor, M., and St Johnston, D. (1995). NMR solution structure of a dsRNA binding domain from Drosophila staufen protein reveals homology to the N-terminal domain of ribosomal protein S5. EMBO J, 14(14):3563–3571.

Calero, G., Wilson, K. F., Ly, T., Rios-Steiner, J. L., Clardy, J. C., and Cerione, R. A. (2002). Structural basis of m7GpppG binding to the nuclear cap-binding protein complex. Nat Struct Biol, 9(12):912–917.

Carrozza, M. J., Kusch, T., and Workman, J. L. (2003). Repairing nucleosomes during transcription. Nat Struct Biol, 10(11):879–880.

Cech, T. R. (1986). A model for the RNA-catalyzed replication of RNA. Proc Natl Acad Sci U S A, 83(12):4360–4363.

Chan, S., Choi, E.-A., and Shi, Y. (2011). Pre-mRNA 3'-end processing complex assembly and function. Wiley Interdiscip Rev RNA, 2(3):321–335.

Chanfreau, G., Noble, S. M., and Guthrie, C. (1996). Essential yeast protein with unexpected similarity to subunits of mammalian cleavage and polyadenylation specificity factor (CPSF). Science, 274(5292):1511–1514.

Chapman, R. D., Heidemann, M., Hintermair, C., and Eick, D. (2008). Molecular evolution of the RNA polymerase II CTD. Trends Genet, 24(6):289–296.

Chen, C.-Y. A. and Shyu, A.-B. (2011). Mechanisms of deadenylation-dependent decay. Wiley Interdiscip Rev RNA, 2(2):167–183.

Chen, Y. and Varani, G. (2013). Engineering RNA-binding proteins for biology. FEBS J, 280(16):3734–3754.

Cho, E. J., Takagi, T., Moore, C. R., and Buratowski, S. (1997). mRNA capping enzyme is recruited to the transcription complex by phosphorylation of the RNA polymerase II carboxy-terminal domain. Genes Dev, 11(24):3319–3326.

Chou, C.-H., Lin, F.-M., Chou, M.-T., Hsu, S.-D., Chang, T.-H., Weng, S.-L., Shrestha, S., Hsiao, C.-C., Hung, J.-H., and Huang, H.-D. (2013). A computational approach for identifying microRNA-target interactions using high-throughput CLIP and PAR-CLIP sequencing. BMC Genomics, 14 Suppl 1:S2.

Chvez, S., Beilharz, T., Rondn, A. G., Erdjument-Bromage, H., Tempst, P., Svejstrup, J. Q., Lithgow, T., and Aguilera, A. (2000). A protein complex containing Tho2, Hpr1, Mft1 and a novel protein, Thp2, connects transcription elongation with mitotic recombination in Saccharomyces cerevisiae. EMBO J, 19(21):5824–5834.

Cochrane, J. C. and Strobel, S. A. (2008). Riboswitch effectors as protein enzyme cofactors. RNA, 14(6):993–1002.

Conte, M. R., Grne, T., Ghuman, J., Kelly, G., Ladas, A., Matthews, S., and Curry, S. (2000). Structure of tandem RNA recognition motifs from polypyrimidine tract binding protein reveals novel features of the RRM fold. EMBO J, 19(12):3132–3141.

Corcoran, D. L., Georgiev, S., Mukherjee, N., Gottwein, E., Skalsky, R. L., Keene, J. D., and Ohler, U. (2011). PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. Genome Biol, 12(8):R79.

Correia, M., Neves-Petersen, M. T., Jeppesen, P. B., Gregersen, S., and Petersen, S. B. (2012). UV-light exposure of insulin: pharmaceutical implications upon covalent insulin dityrosine dimerization and disulphide bond photolysis. PLoS One, 7(12):e50733.

Creamer, T. J., Darby, M. M., Jamonnak, N., Schaughency, P., Hao, H., Wheelan, S. J., and Corden, J. L. (2011). Transcriptome-wide binding sites for components of the Saccharomyces cerevisiae non-poly(A) termination pathway: Nrd1, Nab3, and Sen1. PLoS Genet, 7(10):e1002329.

Crick, F. H. (1958). On protein synthesis. Symp Soc Exp Biol, 12:138–163.

Crick, F. H. (1968). The origin of the genetic code. J Mol Biol, 38(3):367–379.

Cui, Y., Hagan, K. W., Zhang, S., and Peltz, S. W. (1995). Identification and characterization of genes that are required for the accelerated degradation of mRNAs containing a premature translational termination codon. Genes Dev, 9(4):423–436.

Danckwardt, S., Hentze, M. W., and Kulozik, A. E. (2008). 3' end mRNA processing: molecular mechanisms and implications for health and disease. EMBO J, 27(3):482–498.

Darnell, Jr, J. E. (2013). Reflections on the history of pre-mRNA processing and highlights of current knowledge: a unified picture. RNA, 19(4):443–460.

Decker, C. J. and Parker, R. (2002). mRNA decay enzymes: decappers conserved between yeast and mammals. Proc Natl Acad Sci U S A, 99(20):12512–12514.

Deka, P., Bucheli, M. E., Moore, C., Buratowski, S., and Varani, G. (2008). Structure of the yeast SR protein Npl3 and Interaction with mRNA 3'-end processing signals. J Mol Biol, 375(1):136–150.

Dichtl, B., Blank, D., Sadowski, M., Hbner, W., Weiser, S., and Keller, W. (2002). Yhh1p/Cft1p directly links poly(A) site recognition and RNA polymerase II transcription termination. EMBO J, 21(15):4125–4135.

Dichtl, B. and Keller, W. (2001). Recognition of polyadenylation sites in yeast pre-mRNAs by cleavage and polyadenylation factor. EMBO J, 20(12):3197–3209.

Dieppois, G., Iglesias, N., and Stutz, F. (2006). Cotranscriptional recruitment to the mRNA export receptor Mex67p contributes to nuclear pore anchoring of activated genes. Mol Cell Biol, 26(21):7858–7870.

Dreyfuss, G., Kim, V. N., and Kataoka, N. (2002). Messenger-RNA-binding proteins and the messages they carry. Nat Rev Mol Cell Biol, 3(3):195–205.

Duttagupta, R., Tian, B., Wilusz, C. J., Khounh, D. T., Soteropoulos, P., Ouyang, M., Dougherty, J. P., and Peltz, S. W. (2005). Global analysis of Pub1p targets reveals a coordinate control of gene expression through modulation of binding and stability. Mol Cell Biol, 25(13):5499–5513.

Ellington, A. D. and Szostak, J. W. (1990). In vitro selection of RNA molecules that bind specific ligands. Nature, 346(6287):818–822.

Farazi, T. A., Leonhardt, C. S., Mukherjee, N., Mihailovic, A., Li, S., Max, K. E. A., Meyer, C., Yamaji, M., Cekan, P., Jacobs, N. C., Gerstberger, S., Bognanni, C., Larsson, E., Ohler, U., and Tuschl, T. (2014). Identification of the RNA recognition element of the RBPMS family of RNA-binding proteins and their transcriptome-wide mRNA targets. RNA.

Ferr-D'Amar, A. R. and Scott, W. G. (2010). Small self-cleaving ribozymes. Cold Spring Harb Perspect Biol, 2(10):a003574.

Gaillard, H. and Aguilera, A. (2008). A novel class of mRNA-containing cytoplasmic granules are produced in response to UV-irradiation. Mol Biol Cell, 19(11):4980–4992.

Georgiev, S., Boyle, A. P., Jayasurya, K., Ding, X., Mukherjee, S., and Ohler, U. (2010). Evidence-ranked motif identification. Genome Biol, 11(2):R19.

Gewartowski, K., Cullar, J., Dziembowski, A., and Valpuesta, J. M. (2012). The yeast THO complex forms a 5-subunit assembly that directly interacts with active chromatin. Bioarchitecture, 2(4):134–137.

Ghosh, A. and Lima, C. D. (2010). Enzymology of RNA cap synthesis. Wiley Interdiscip Rev RNA, 1(1):152–172.

Giardina, C. and Lis, J. T. (1993). DNA melting on yeast RNA polymerase II promoters. Science, 261(5122):759–762.

Gilbert, W. (1987). The exon theory of genes. Cold Spring Harb Symp Quant Biol, 52:901–905.

Gilbert, W. and Guthrie, C. (2004). The Glc7p nuclear phosphatase promotes mRNA export by facilitating association of Mex67p with mRNA. Mol Cell, 13(2):201–212.

Goecks, J., Nekrutenko, A., Taylor, J., and , G. T. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol, 11(8):R86.

Granneman, S., Kudla, G., Petfalski, E., and Tollervey, D. (2009). Identification of protein binding sites on U3 snoRNA and pre-rRNA by UV cross-linking and high-throughput analysis of cDNAs. Proc Natl Acad Sci U S A, 106(24):9613–9618.

Green, D. M., Marfatia, K. A., Crafton, E. B., Zhang, X., Cheng, X., and Corbett, A. H. (2002). Nab2p is required for poly(A) RNA export in Saccharomyces cerevisiae and is regulated by arginine methylation via Hmt1p. J Biol Chem, 277(10):7752–7760.

Grishin, N. V. (2001). KH domain: one motif, two folds. Nucleic Acids Res, 29(3):638–643.

Gross, S. and Moore, C. L. (2001). Rna15 interaction with the A-rich yeast polyadenylation signal is an essential step in mRNA 3'-end formation. Mol Cell Biol, 21(23):8045–8055.

Grnberg, S., Warfield, L., and Hahn, S. (2012). Architecture of the RNA polymerase II preinitiation complex and mechanism of ATP-dependent promoter opening. Nat Struct Mol Biol, 19(8):788–796.

Grter, P., Tabernero, C., von Kobbe, C., Schmitt, C., Saavedra, C., Bachi, A., Wilm, M., Felber, B. K., and Izaurralde, E. (1998). TAP, the human homolog of Mex67p, mediates CTE-dependent RNA export from the nucleus. Mol Cell, 1(5):649–659.

Guerrier-Takada, C., Gardiner, K., Marsh, T., Pace, N., and Altman, S. (1983). The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. Cell, 35(3 Pt 2):849–857.

Guo, Z., Russo, P., Yun, D. F., Butler, J. S., and Sherman, F. (1995). Redundant 3' end-forming signals for the yeast CYC1 mRNA. Proc Natl Acad Sci U S A, 92(10):4211–4214.

Guo, Z. and Sherman, F. (1996). 3'-end-forming signals of yeast mRNA. Trends Biochem Sci, 21(12):477–481.

Gmez-Gonzlez, B., Garca-Rubio, M., Bermejo, R., Gaillard, H., Shirahige, K., Marn, A., Foiani, M., and Aguilera, A. (2011). Genome-wide function of THO/TREX in active genes prevents R-loop-dependent replication obstacles. EMBO J, 30(15):3106–3119.

Hackmann, A., Gross, T., Baierlein, C., and Krebber, H. (2011). The mRNA export factor Npl3 mediates the nuclear export of large ribosomal subunits. EMBO Rep, 12(10):1024–1031.

Hackmann, A., Wu, H., Schneider, U.-M., Meyer, K., Jung, K., and Krebber, H. (2014). Quality control of spliced mRNAs requires the shuttling SR proteins Gbp2 and Hrb1. Nat Commun, 5:3123.

Hafner, M., Landgraf, P., Ludwig, J., Rice, A., Ojo, T., Lin, C., Holoch, D., Lim, C., and Tuschl, T. (2008). Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing. Methods, 44(1):3–12.

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, Jr, M., Jungkamp, A.-C., Munschauer, M., Ulrich, A., Wardle, G. S., Dewell, S., Zavolan, M., and Tuschl, T. (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell, 141(1):129–141.

Hahn, S. (2004). Structure and mechanism of the RNA polymerase II transcription machinery. Nat Struct Mol Biol, 11(5):394–403.

Hartmann, H., Guthhrlein, E. W., Siebert, M., Luehr, S., and Sding, J. (2013). P-value-based regulatory motif discovery using positional weight matrices. Genome Res, 23(1):181–194.

Hartzog, G. A. and Fu, J. (2013). The Spt4-Spt5 complex: a multi-faceted regulator of transcription elongation. Biochim Biophys Acta, 1829(1):105–115.

Hartzog, G. A., Wada, T., Handa, H., and Winston, F. (1998). Evidence that Spt4, Spt5, and Spt6 control transcription elongation by RNA polymerase II in Saccharomyces cerevisiae. Genes Dev, 12(3):357–369.

Hector, R. E., Nykamp, K. R., Dheur, S., Anderson, J. T., Non, P. J., Urbinati, C. R., Wilson, S. M., Minvielle-Sebastia, L., and Swanson, M. S. (2002). Dual requirement for yeast hnRNP Nab2p in mRNA poly(A) tail length control and nuclear export. EMBO J, 21(7):1800–1810.

Heidemann, M. and Eick, D. (2012). Tyrosine-1 and threonine-4 phosphorylation marks complete the RNA polymerase II CTD phospho-code. RNA Biol, 9(9):1144–1146.

Heidemann, M., Hintermair, C., Vo, K., and Eick, D. (2013). Dynamic phosphorylation patterns of RNA polymerase II CTD during transcription. Biochim Biophys Acta, 1829(1):55–62.

Hieronymus, H. and Silver, P. A. (2003). Genome-wide analysis of RNA-protein interactions illustrates specificity of the mRNA export machinery. Nat Genet, 33(2):155–161.

Hirose, Y. and Manley, J. L. (2000). RNA polymerase II and the integration of nuclear events. Genes Dev, 14(12):1415–1429.

Hobeika, M., Brockmann, C., Iglesias, N., Gwizdek, C., Neuhaus, D., Stutz, F., Stewart, M., Divita, G., and Dargemont, C. (2007). Coordination of Hpr1 and ubiquitin binding by the UBA domain of the mRNA export factor Mex67. Mol Biol Cell, 18(7):2561–2568.

Hogan, D. J., Riordan, D. P., Gerber, A. P., Herschlag, D., and Brown, P. O. (2008). Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system. PLoS Biol, 6(10):e255.

Hsin, J.-P. and Manley, J. L. (2012). The RNA polymerase II CTD coordinates transcription and RNA processing. Genes Dev, 26(19):2119–2137.

Iglesias, N., Tutucci, E., Gwizdek, C., Vinciguerra, P., Von Dach, E., Corbett, A. H., Dargemont, C., and Stutz, F. (2010). Ubiquitin-mediated mRNP dynamics and surveillance prior to budding yeast mRNA export. Genes Dev, 24(17):1927–1938.

Jensen, T. H., Boulay, J., Rosbash, M., and Libri, D. (2001). The DECD box putative ATPase Sub2p is an early mRNA export factor. Curr Biol, 11(21):1711–1715.

Jensen, T. H., Jacquier, A., and Libri, D. (2013). Dealing with pervasive transcription. Mol Cell, 52(4):473–484.

Jeronimo, C., Bataille, A. R., and Robert, F. (2013). The writers, readers, and functions of the RNA polymerase II C-terminal domain code. Chem Rev, 113(11):8491–8522.

Jiao, X., Xiang, S., Oh, C., Martin, C. E., Tong, L., and Kiledjian, M. (2010). Identification of a quality-control mechanism for mRNA 5'-end capping. Nature, 467(7315):608–611.

Jimeno, S., Rondn, A. G., Luna, R., and Aguilera, A. (2002). The yeast THO complex and mRNA export factors link RNA metabolism with transcription and genome instability. EMBO J, 21(13):3526–3535.

Johnson, S. A., Cubberley, G., and Bentley, D. L. (2009). Cotranscriptional recruitment of the mRNA export factor Yra1 by direct interaction with the 3' end processing factor Pcf11. Mol Cell, 33(2):215–226.

Kang, M. E. and Dahmus, M. E. (1995). The photoactivated cross-linking of recombinant C-terminal domain to proteins in a HeLa cell transcription extract that comigrate with transcription factors IIE and IIF. J Biol Chem, 270(40):23390–23397.

Kawauchi, J., Mischo, H., Braglia, P., Rondon, A., and Proudfoot, N. J. (2008). Budding yeast RNA polymerases I and II employ parallel mechanisms of transcriptional termination. Genes Dev, 22(8):1082–1092.

Kessler, M. M., Henry, M. F., Shen, E., Zhao, J., Gross, S., Silver, P. A., and Moore, C. L. (1997). Hrp1, a sequence-specific RNA-binding protein that shuttles between the nucleus and the cytoplasm, is required for mRNA 3'-end formation in yeast. Genes Dev, 11(19):2545–2556.

Khorshid, M., Rodak, C., and Zavolan, M. (2011). CLIPZ: a database and analysis environment for experimentally determined binding sites of RNA-binding proteins. Nucleic Acids Res, 39(Database issue):D245–D252.

Kim, H.-J., Jeong, S.-H., Heo, J.-H., Jeong, S.-J., Kim, S.-T., Youn, H.-D., Han, J.-W., Lee, H.-W., and Cho, E.-J. (2004a). mRNA capping enzyme activity is coupled to an early transcription elongation. Mol Cell Biol, 24(14):6184–6193.

Kim, M., Ahn, S.-H., Krogan, N. J., Greenblatt, J. F., and Buratowski, S. (2004b). Transitions in RNA polymerase II elongation complexes at the 3' ends of genes. EMBO J, 23(2):354–364.

Kim, M., Krogan, N. J., Vasiljeva, L., Rando, O. J., Nedea, E., Greenblatt, J. F., and Buratowski, S. (2004c). The yeast Rat1 exonuclease promotes transcription termination by RNA polymerase II. Nature, 432(7016):517–522.

Kim, M., Suh, H., Cho, E.-J., and Buratowski, S. (2009). Phosphorylation of the yeast Rpb1 C-terminal domain at serines 2, 5, and 7. J Biol Chem, 284(39):26421–26426.

Komarnitsky, P., Cho, E. J., and Buratowski, S. (2000). Different phosphorylated forms of RNA polymerase II and associated mRNA processing factors during transcription. Genes Dev, 14(19):2452–2460.

Kornberg, R. D. (2005). Mediator and the mechanism of transcriptional activation. Trends Biochem Sci, 30(5):235–239.

Kostrewa, D., Zeller, M. E., Armache, K.-J., Seizl, M., Leike, K., Thomm, M., and Cramer, P. (2009). RNA polymerase II-TFIIB structure and mechanism of transcription initiation. Nature, 462(7271):323–330.

Kress, T. L., Krogan, N. J., and Guthrie, C. (2008). A single SR-like protein, Npl3, promotes pre-mRNA splicing in budding yeast. Mol Cell, 32(5):727–734.

Krogan, N. J., Dover, J., Wood, A., Schneider, J., Heidt, J., Boateng, M. A., Dean, K., Ryan, O. W., Golshani, A., Johnston, M., Greenblatt, J. F., and Shilatifard, A. (2003). The Paf1 complex is required for histone H3 methylation by COMPASS and Dot1p: linking transcriptional elongation to histone methylation. Mol Cell, 11(3):721–729.

Kruger, K., Grabowski, P. J., Zaug, A. J., Sands, J., Gottschling, D. E., and Cech, T. R. (1982). Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. Cell, 31(1):147–157.

Kubicek, K., Cerna, H., Holub, P., Pasulka, J., Hrossova, D., Loehr, F., Hofr, C., Vanacova, S., and Stefl, R. (2012). Serine phosphorylation and proline isomerization in RNAP II CTD control recruitment of Nrd1. Genes Dev, 26(17):1891–1896.

Kyburz, A., Friedlein, A., Langen, H., and Keller, W. (2006). Direct interactions between subunits of CPSF and the U2 snRNP contribute to the coupling of pre-mRNA 3' end processing and splicing. Mol Cell, 23(2):195–205.

Knig, J., Zarnack, K., Luscombe, N. M., and Ule, J. (2011). Protein-RNA interactions: new genomic technologies and perspectives. Nat Rev Genet, 13(2):77–83.

Knig, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D. J., Luscombe, N. M., and Ule, J. (2010). iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. Nat Struct Mol Biol, 17(7):909–915.

Langmead, B. (2010). Aligning short sequencing reads with Bowtie. Curr Protoc Bioinformatics, Chapter 11:Unit 11.7.

Leeper, T. C., Qu, X., Lu, C., Moore, C., and Varani, G. (2010). Novel protein-protein contacts facilitate mRNA 3'-processing signal recognition by Rna15 and Hrp1. J Mol Biol, 401(3):334–349.

Lewis, J. D. and Izaurralde, E. (1997). The role of the cap structure in RNA processing and nuclear export. Eur J Biochem, 247(2):461–469.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and , . G. P. D. P. S. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25(16):2078–2079.

Licatalosi, D. D., Mele, A., Fak, J. J., Ule, J., Kayikci, M., Chi, S. W., Clark, T. A., Schweitzer, A. C., Blume, J. E., Wang, X., Darnell, J. C., and Darnell, R. B. (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature, 456(7221):464–469.

Lidschreiber, M., Leike, K., and Cramer, P. (2013). Cap completion and C-terminal repeat domain kinase recruitment underlie the initiation-elongation transition of RNA polymerase II. Mol Cell Biol, 33(19):3805–3816.

Liker, E., Fernandez, E., Izaurralde, E., and Conti, E. (2000). The structure of the mRNA export factor TAP reveals a cis arrangement of a non-canonical RNP domain and an LRR domain. EMBO J, 19(21):5587–5598.

Liu, Y., Warfield, L., Zhang, C., Luo, J., Allen, J., Lang, W. H., Ranish, J., Shokat, K. M., and Hahn, S. (2009). Phosphorylation of the transcription elongation factor Spt5 by yeast Bur1 kinase stimulates recruitment of the PAF complex. Mol Cell Biol, 29(17):4852–4863.

Logan, J., Falck-Pedersen, E., Darnell, Jr, J., and Shenk, T. (1987). A poly(A) addition site and a downstream termination region are required for efficient cessation of transcription by RNA polymerase II in the mouse beta maj-globin gene. Proc Natl Acad Sci U S A, 84(23):8306–8310.

Lozzio, C. B. and Wigler, P. W. (1971). Cytotoxic effects of thiopyrimidines. J Cell Physiol, 78(1):25–32.

Luna, R., Rondn, A. G., and Aguilera, A. (2012). New clues to understand the role of THO and other functionally related factors in mRNP biogenesis. Biochim Biophys Acta, 1819(6):514–520.

Lunde, B. M., Moore, C., and Varani, G. (2007). RNA-binding proteins: modular design for efficient function. Nat Rev Mol Cell Biol, 8(6):479–490.

Luo, W., Johnson, A. W., and Bentley, D. L. (2006). The role of Rat1 in coupling mRNA 3'-end processing to transcription termination: implications for a unified allosteric-torpedo model. Genes Dev, 20(8):954–965.

Luse, D. S. (2013). Promoter clearance by RNA polymerase II. Biochim Biophys Acta, 1829(1):63–68.

Mackereth, C. D., Madl, T., Bonnal, S., Simon, B., Zanier, K., Gasch, A., Rybin, V., Valcrcel, J., and Sattler, M. (2011). Multi-domain conformational selection underlies pre-mRNA splicing regulation by U2AF. Nature, 475(7356):408–411.

Mandal, M. and Breaker, R. R. (2004). Gene regulation by riboswitches. Nat Rev Mol Cell Biol, 5(6):451–463.

Mandal, S. S., Chu, C., Wada, T., Handa, H., Shatkin, A. J., and Reinberg, D. (2004). Functional interactions of RNA-capping enzyme with factors that positively and negatively regulate promoter escape by RNA polymerase II. Proc Natl Acad Sci U S A, 101(20):7572–7577.

Mandel, C. R., Bai, Y., and Tong, L. (2008). Protein factors in pre-mRNA 3'-end processing. Cell Mol Life Sci, 65(7-8):1099–1122.

Mangus, D. A., Evans, M. C., and Jacobson, A. (2003). Poly(A)-binding proteins: multifunctional scaffolds for the post-transcriptional control of gene expression. Genome Biol, 4(7):223.

Margaritis, T., Oreal, V., Brabers, N., Maestroni, L., Vitaliano-Prunier, A., Benschop, J. J., van Hooff, S., van Leenen, D., Dargemont, C., Gli, V., and Holstege, F. C. P. (2012). Two distinct repressive mechanisms for histone 3 lysine 4 methylation through promoting 3'-end antisense transcription. PLoS Genet, 8(9):e1002952.

Martin, G., Gruber, A. R., Keller, W., and Zavolan, M. (2012). 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, 1(6):753–763.

Martinez-Rucobo, F. W., Sainsbury, S., Cheung, A. C. M., and Cramer, P. (2011). Architecture of the RNA polymerase-Spt4/5 complex and basis of universal transcription processivity. EMBO J, 30(7):1302–1310.

Martinson, H. G. (2011). An active role for splicing in 3'-end formation. Wiley Interdiscip Rev RNA, 2(4):459–470.

Maxon, M. E., Goodrich, J. A., and Tjian, R. (1994). Transcription factor IIE binds preferentially to RNA polymerase IIa and recruits TFIIH: a model for promoter clearance. Genes Dev, 8(5):515–524.

Mayer, A., Heidemann, M., Lidschreiber, M., Schreieck, A., Sun, M., Hintermair, C., Kremmer, E., Eick, D., and Cramer, P. (2012). CTD tyrosine phosphorylation impairs termination factor recruitment to RNA polymerase II. Science, 336(6089):1723–1725.

Mayer, A., Lidschreiber, M., Siebert, M., Leike, K., Sding, J., and Cramer, P. (2010). Uniform transitions of the general RNA polymerase II transcription complex. Nat Struct Mol Biol, 17(10):1272–1278.

Mazza, C., Segref, A., Mattaj, I. W., and Cusack, S. (2002). Large-scale induced fit recognition of an m(7)GpppG cap analogue by the human nuclear cap-binding complex. EMBO J, 21(20):5548–5557.

Meinel, D. M., Burkert-Kautzsch, C., Kieser, A., O'Duibhir, E., Siebert, M., Mayer, A., Cramer, P., Sding, J., Holstege, F. C. P., and Strer, K. (2013). Recruitment of TREX to the transcription machinery by its direct binding to the phospho-CTD of RNA polymerase II. PLoS Genet, 9(11):e1003914.

Meinhart, A. and Cramer, P. (2004). Recognition of RNA polymerase II carboxy-terminal domain by 3'-RNA-processing factors. Nature, 430(6996):223–226.

Milek, M., Wyler, E., and Landthaler, M. (2012). Transcriptome-wide analysis of protein-RNA interactions using high-throughput sequencing. Semin Cell Dev Biol, 23(2):206–212.

Miller, J., McLachlan, A. D., and Klug, A. (1985). Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes. EMBO J, 4(6):1609–1614.

Minvielle-Sebastia, L., Preker, P. J., and Keller, W. (1994). RNA14 and RNA15 proteins as components of a yeast pre-mRNA 3'-end processing factor. Science, 266(5191):1702–1705.

Mischo, H. E. and Proudfoot, N. J. (2013). Disengaging polymerase: terminating RNA polymerase II transcription in budding yeast. Biochim Biophys Acta, 1829(1):174–185.

Moore, M. J. (2005). From birth to death: the complex lives of eukaryotic mRNAs. Science, 309(5740):1514–1518.

Moore, P. B. and Steitz, T. A. (2011). The roles of RNA in the synthesis of protein. Cold Spring Harb Perspect Biol, 3(11):a003780.

Morris, D. P. and Greenleaf, A. L. (2000). The splicing factor, Prp40, binds the phosphorylated carboxyl-terminal domain of RNA polymerase II. J Biol Chem, 275(51):39935–39943.

Mosley, A. L., Pattenden, S. G., Carey, M., Venkatesh, S., Gilmore, J. M., Florens, L., Workman, J. L., and Washburn, M. P. (2009). Rtr1 is a CTD phosphatase that regulates RNA polymerase II during the transition from serine 5 to serine 2 phosphorylation. Mol Cell, 34(2):168–178.

Murray, S., Udupa, R., Yao, S., Hartzog, G., and Prelich, G. (2001). Phosphorylation of the RNA polymerase II carboxy-terminal domain by the Bur1 cyclin-dependent kinase. Mol Cell Biol, 21(13):4089–4096.

Musco, G., Kharrat, A., Stier, G., Fraternali, F., Gibson, T. J., Nilges, M., and Pastore, A. (1997). The solution structure of the first KH domain of FMR1, the protein responsible for the fragile X syndrome. Nat Struct Biol, 4(9):712–716.

Myers, L. C., Gustafsson, C. M., Bushnell, D. A., Lui, M., Erdjument-Bromage, H., Tempst, P., and Kornberg, R. D. (1998). The Med proteins of yeast and their function through the RNA polymerase II carboxy-terminal domain. Genes Dev, 12(1):45–54.

Mller-McNicoll, M. and Neugebauer, K. M. (2013). How cells get the message: dynamic assembly and function of mRNA-protein complexes. Nat Rev Genet, 14(4):275–287.

Nedea, E., He, X., Kim, M., Pootoolal, J., Zhong, G., Canadien, V., Hughes, T., Buratowski, S., Moore, C. L., and Greenblatt, J. (2003). Organization and function of APT, a subcomplex of the yeast cleavage and polyadenylation factor involved in the formation of mRNA and small nucleolar RNA 3'-ends. J Biol Chem, 278(35):33000–33010.

Neuvglise, C., Marck, C., and Gaillardin, C. (2011). The intronome of budding yeasts. C R Biol, 334(8-9):662–670.

Ng, H. H., Robert, F., Young, R. A., and Struhl, K. (2003). Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity. Mol Cell, 11(3):709–719.

Patturajan, M., Conrad, N. K., Bregman, D. B., and Corden, J. L. (1999). Yeast carboxyl-terminal domain kinase I positively and negatively regulates RNA polymerase II carboxyl-terminal domain phosphorylation. J Biol Chem, 274(39):27823–27828.

Pelechano, V., Wei, W., and Steinmetz, L. M. (2013). Extensive transcriptional heterogeneity revealed by isoform profiling. Nature, 497(7447):127–131.

Proudfoot, N. J. (2011). Ending the message: poly(A) signals then and now. Genes Dev, 25(17):1770–1782.

Reed, R. and Hurt, E. (2002). A conserved mRNA export machinery coupled to pre-mRNA splicing. Cell, 108(4):523–531.

Rhee, H. S. and Pugh, B. F. (2012). Genome-wide structure and organization of eukaryotic pre-initiation complexes. Nature, 483(7389):295–301.

Riordan, D. P., Herschlag, D., and Brown, P. O. (2011). Identification of RNA recognition elements in the Saccharomyces cerevisiae transcriptome. Nucleic Acids Res, 39(4):1501–1509.

Rodrguez-Navarro, S. and Hurt, E. (2011). Linking gene regulation to mRNA production and export. Curr Opin Cell Biol, 23(3):302–309.

Roth, A. and Breaker, R. R. (2009). The structural and functional diversity of metabolite-binding riboswitches. Annu Rev Biochem, 78:305–334.

Rougemaille, M., Dieppois, G., Kisseleva-Romanova, E., Gudipati, R. K., Lemoine, S., Blugeon, C., Boulay, J., Jensen, T. H., Stutz, F., Devaux, F., and Libri, D. (2008). THO/Sub2p functions to coordinate 3'-end processing with gene-nuclear pore association. Cell, 135(2):308–321.

Ruiz-Echevarra, M. J., Gonzlez, C. I., and Peltz, S. W. (1998). Identifying the right stop: determining how the surveillance complex recognizes and degrades an aberrant mRNA. EMBO J, 17(2):575–589.

Ruiz-Echevarra, M. J. and Peltz, S. W. (2000). The RNA binding protein Pub1 modulates the stability of transcripts containing upstream open reading frames. Cell, 101(7):741–751.

Sachs, A. B., Davis, R. W., and Kornberg, R. D. (1987). A single domain of yeast poly(A)-binding protein is necessary and sufficient for RNA binding and cell viability. Mol Cell Biol, 7(9):3268–3276.

Sadowski, M., Dichtl, B., Hbner, W., and Keller, W. (2003). Independent functions of yeast Pcf11p in pre-mRNA 3' end processing and in transcription termination. EMBO J, 22(9):2167–2177.

Sainsbury, S., Niesser, J., and Cramer, P. (2013). Structure and function of the initially transcribing RNA polymerase II-TFIIB complex. Nature, 493(7432):437–440.

Schneider, C., Kudla, G., Wlotzka, W., Tuck, A., and Tollervey, D. (2012). Transcriptome-wide analysis of exosome targets. Mol Cell, 48(3):422–433.

Schreieck, A., Easter, A. D., Etzold, S., Wiederhold, K., Lidschreiber, M., Cramer, P., and Passmore, L. A. (2014). RNA polymerase II termination involves C-terminal-domain tyrosine dephosphorylation by CPF subunit Glc7. Nat Struct Mol Biol, 21(2):175–179.

Schroeder, S. C., Schwer, B., Shuman, S., and Bentley, D. (2000). Dynamic association of capping enzymes with transcribing RNA polymerase II. Genes Dev, 14(19):2435–2440.

Schroeder, S. C., Zorio, D. A. R., Schwer, B., Shuman, S., and Bentley, D. (2004). A function of yeast mRNA cap methyltransferase, Abd1, in transcription by RNA polymerase II. Mol Cell, 13(3):377–387.

Schulz, D., Schwalb, B., Kiesel, A., Baejen, C., Torkler, P., Gagneur, J., Soeding, J., and Cramer, P. (2013). Transcriptome surveillance by selective termination of noncoding RNA synthesis. Cell, 155(5):1075–1087.

Schwer, B. and Shuman, S. (1996). Conditional inactivation of mRNA capping enzyme affects yeast pre-mRNA splicing in vivo. RNA, 2(6):574–583.

Schwinghammer, K., Cheung, A. C. M., Morozov, Y. I., Agaronyan, K., Temiakov, D., and Cramer, P. (2013). Structure of human mitochondrial RNA polymerase elongation complex. Nat Struct Mol Biol, 20(11):1298–1303.

Segref, A., Sharma, K., Doye, V., Hellwig, A., Huber, J., Lhrmann, R., and Hurt, E. (1997). Mex67p, a novel factor for nuclear mRNA export, binds to both poly(A)+ RNA and nuclear pores. EMBO J, 16(11):3256–3271.

Shen, H. and Green, M. R. (2006). RS domains contact splicing signals and promote splicing by a common mechanism in yeast through humans. Genes Dev, 20(13):1755–1765.

Shi, Y., Di Giammartino, D. C., Taylor, D., Sarkeshik, A., Rice, W. J., Yates, 3rd, J. R., Frank, J., and Manley, J. L. (2009). Molecular architecture of the human pre-mRNA 3' processing complex. Mol Cell, 33(3):365–376.

Shieh, G. S., Pan, C.-H., Wu, J.-H., Sun, Y.-J., Wang, C.-C., Hsiao, W.-C., Lin, C.-Y., Tung, L., Chang, T.-H., Fleming, A. B., Hillyer, C., Lo, Y.-C., Berger, S. L., Osley, M. A., and Kao, C.-F. (2011). H2B ubiquitylation is part of chromatin architecture that marks exon-intron structure in budding yeast. BMC Genomics, 12:627.

Sievers, C., Schlumpf, T., Sawarkar, R., Comoglio, F., and Paro, R. (2012). Mixture models and wavelet transforms reveal high confidence RNA-protein interaction sites in MOV10 PAR-CLIP data. Nucleic Acids Res, 40(20):e160.

Smolle, M. and Workman, J. L. (2013). Transcription-associated histone modifications and cryptic transcription. Biochim Biophys Acta, 1829(1):84–97.

Spitzer, J., Hafner, M., Landthaler, M., Ascano, M., Farazi, T., Wardle, G., Nusbaum, J., Khorshid, M., Burger, L., Zavolan, M., and Tuschl, T. (2014). 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, 539:113–161.

Steinmetz, E. J. and Brow, D. A. (1996). Repression of gene expression by an exogenous sequence element acting in concert with a heterogeneous nuclear ribonucleoprotein-like protein, Nrd1, and the putative helicase Sen1. Mol Cell Biol, 16(12):6993–7003.

Stewart, M. (2010). Nuclear export of mRNA. Trends Biochem Sci, 35(11):609–617.

Strsser, K., Masuda, S., Mason, P., Pfannstiel, J., Oppizzi, M., Rodriguez-Navarro, S., Rondn, A. G., Aguilera, A., Struhl, K., Reed, R., and Hurt, E. (2002). TREX is a conserved complex coupling transcription with messenger RNA export. Nature, 417(6886):304–308.

Sun, M., Schwalb, B., Schulz, D., Pirkl, N., Etzold, S., Larivire, L., Maier, K. C., Seizl, M., Tresch, A., and Cramer, P. (2012). Comparative dynamic transcriptome analysis (cDTA) reveals mutual feedback between mRNA synthesis and degradation. Genome Res, 22(7):1350–1359.

Swanson, M. S., Malone, E. A., and Winston, F. (1991). SPT5, an essential gene important for normal transcription in Saccharomyces cerevisiae, encodes an acidic nuclear protein with a carboxy-terminal repeat. Mol Cell Biol, 11(6):3009–3019.

Sgaard, T. M. M. and Svejstrup, J. Q. (2007). Hyperphosphorylation of the C-terminal repeat domain of RNA polymerase II facilitates dissociation of its complex with mediator. J Biol Chem, 282(19):14113–14120.

Terzi, N., Churchman, L. S., Vasiljeva, L., Weissman, J., and Buratowski, S. (2011). H3K4 trimethylation by Set1 promotes efficient termination by the Nrd1-Nab3-Sen1 pathway. Mol Cell Biol, 31(17):3569–3583.

Thomas, M. C. and Chiang, C.-M. (2006). The general transcription machinery and general cofactors. Crit Rev Biochem Mol Biol, 41(3):105–178.

Tosi, A., Haas, C., Herzog, F., Gilmozzi, A., Berninghausen, O., Ungewickell, C., Gerhold, C. B., Lakomek, K., Aebersold, R., Beckmann, R., and Hopfner, K.-P. (2013). Structure and subunit topology of the INO80 chromatin remodeler and its nucleosome complex. Cell, 154(6):1207–1219.

Tseng, C.-K. and Cheng, S.-C. (2013). The spliceosome catalyzes debranching in competition with reverse of the first chemical reaction. RNA, 19(7):971–981.

Tuck, A. C. and Tollervey, D. (2013). A transcriptome-wide atlas of RNP composition reveals diverse classes of mRNAs and lncRNAs. Cell, 154(5):996–1009.

Tuerk, C. and Gold, L. (1990). Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science, 249(4968):505–510.

Ule, J., Jensen, K., Mele, A., and Darnell, R. B. (2005). CLIP: a method for identifying protein-RNA interaction sites in living cells. Methods, 37(4):376–386.

Ule, J., Jensen, K. B., Ruggiu, M., Mele, A., Ule, A., and Darnell, R. B. (2003). CLIP identifies Nova-regulated RNA networks in the brain. Science, 302(5648):1212–1215.

Vannini, A. and Cramer, P. (2012). Conservation between the RNA polymerase I, II, and III transcription initiation machineries. Mol Cell, 45(4):439–446.

Vasudevan, S. and Peltz, S. W. (2001). Regulated ARE-mediated mRNA decay in Saccharomyces cerevisiae. Mol Cell, 7(6):1191–1200.

Vitaliano-Prunier, A., Babour, A., Hrissant, L., Apponi, L., Margaritis, T., Holstege, F. C. P., Corbett, A. H., Gwizdek, C., and Dargemont, C. (2012). H2B ubiquitylation controls the formation of export-competent mRNP. Mol Cell, 45(1):132–139.

Vo, L. T., Minet, M., Schmitter, J. M., Lacroute, F., and Wyers, F. (2001). Mpe1, a zinc knuckle protein, is an essential component of yeast cleavage and polyadenylation factor required for the cleavage and polyadenylation of mRNA. Mol Cell Biol, 21(24):8346–8356.

Wahl, M. C., Will, C. L., and Lhrmann, R. (2009). The spliceosome: design principles of a dynamic RNP machine. Cell, 136(4):701–718.

Wang, Z., Kayikci, M., Briese, M., Zarnack, K., Luscombe, N. M., Rot, G., Zupan, B., Curk, T., and Ule, J. (2010). iCLIP predicts the dual splicing effects of TIA-RNA interactions. PLoS Biol, 8(10):e1000530.

Wery, M., Kwapisz, M., and Morillon, A. (2011). Noncoding RNAs in gene regulation. Wiley Interdiscip Rev Syst Biol Med, 3(6):728–738.

West, M. L. and Corden, J. L. (1995). Construction and analysis of yeast RNA polymerase II CTD deletion and substitution mutations. Genetics, 140(4):1223–1233.

Will, C. L. and Lhrmann, R. (2011). Spliceosome structure and function. Cold Spring Harb Perspect Biol, 3(7).

Wilson, C. J., Chao, D. M., Imbalzano, A. N., Schnitzler, G. R., Kingston, R. E., and Young, R. A. (1996). RNA polymerase II holoenzyme contains SWI/SNF regulators involved in chromatin remodeling. Cell, 84(2):235–244.

Windgassen, M. and Krebber, H. (2003). Identification of Gbp2 as a novel poly(A)+ RNA-binding protein involved in the cytoplasmic delivery of messenger RNAs in yeast. EMBO Rep, 4(3):278–283.

Winkler, W., Nahvi, A., and Breaker, R. R. (2002). Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. Nature, 419(6910):952–956.

Woese, C. R., Dugre, D. H., Saxinger, W. C., and Dugre, S. A. (1966). The molecular basis for the genetic code. Proc Natl Acad Sci U S A, 55(4):966–974.

Wolf, E., Kastner, B., Deckert, J., Merz, C., Stark, H., and Lhrmann, R. (2009). Exon, intron and splice site locations in the spliceosomal B complex. EMBO J, 28(15):2283–2292.

Woolford, Jr, J. and Peebles, C. L. (1992). RNA splicing in lower eukaryotes. Curr Opin Genet Dev, 2(5):712–719.

Xiang, K., Manley, J. L., and Tong, L. (2012). An unexpected binding mode for a Pol II CTD peptide phosphorylated at Ser7 in the active site of the CTD phosphatase Ssu72. Genes Dev, 26(20):2265–2270.

Xiang, S., Cooper-Morgan, A., Jiao, X., Kiledjian, M., Manley, J. L., and Tong, L. (2009). Structure and function of the 5'–>3' exoribonuclease Rat1 and its activating partner Rai1. Nature, 458(7239):784–788.

Xu, Z., Wei, W., Gagneur, J., Perocchi, F., Clauder-Mnster, S., Camblong, J., Guffanti, E., Stutz, F., Huber, W., and Steinmetz, L. M. (2009). Bidirectional promoters generate pervasive transcription in yeast. Nature, 457(7232):1033–1037.

Zhang, C. and Darnell, R. B. (2011). Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. Nat Biotechnol, 29(7):607–614.

Zhang, G., Campbell, E. A., Minakhin, L., Richter, C., Severinov, K., and Darst, S. A. (1999). Crystal structure of Thermus aquaticus core RNA polymerase at 3.3 A resolution. Cell, 98(6):811–824.

Zhou, K., Kuo, W. H. W., Fillingham, J., and Greenblatt, J. F. (2009). Control of transcriptional elongation and cotranscriptional histone modification by the yeast BUR kinase substrate Spt5. Proc Natl Acad Sci U S A, 106(17):6956–6961.

List of Abbreviations

4sU 4-thiouridine
4tU 4-thiouracil
5BrU 5-bromouridine
5iU 5-iodouridine
6sG 6-thioguanosine
6tG 6-thioguanine
A ..... Adenosine
aDNA ..... amplified cDNA
App ..... Pre-adenylation
AUC ..... Area under the curve
BAM ..... binary SAM
BAP ..... Bacterial alkaline phosphatase
BBP ..... BP-binding protein
BP ..... Branch point
bp ..... Base pairs
C ..... Cytosine
cAUC ..... corrected AUC
CBC ..... Cap-binding complex
cDNA ..... complementary DNA
ChIP ..... Chromatin immunoprecipitation
CID ..... CTD interaction domain
CLIP ..... Cross-linking and immunoprecipitation
CPF ..... Cleavage and polyadenylation factor
CPSF ..... Cleavage and polyadenylation specificity factor
CRAC ..... UV crosslinking and analysis of cDNAs
CRAC ..... Crosslinking and cDNA analysis
CstF ..... Cleavage stimulation factor
CTD ..... C-terminal domain
CTR ..... C-terminal region
CUT ..... Cryptic unstable transcript
ddC ..... Dideoxy-C
DNA ..... Deoxyribonucleic acid
dNTP ..... Deoxyribonucleotide triphosphates
dsRBD ..... Double stranded RNA-binding domain
dsRNA ..... Double stranded RNA
dT ..... Thymidine
DTT ..... Dithiothreitol
ECL ..... Enhanced chemiluminescence
EE ..... Exon-exon
EI ..... Exon-intron
EtOH ..... Ethanol
FDR ..... False discovery rate

G ..... Guanosine
GTF ..... General transcription factor
HITS-CLIP ..... High-throughput sequencing CLIP
HPLC ..... High-performance liquid chromatography
iCLAP ..... Individual-nucleotide resolution UV-crosslinking and affinity purification
iCLIP ..... Individual-nucleotide resolution CLIP
IE ..... Intron-exon
IgG ..... Immunglobuline G
IP ..... Immunoprecipitation
kb ..... Kilobases
m7G 7-methyl-guanosine
mitoPol ..... mitochondrial Polymerase
mRNA ..... messenger RNA
mRNP ..... messenger Ribonucleoprotein particle
ncRNA ..... non-coding RNA
NGS ..... Next-Generation-Sequencing
NPC ..... Nuclear pore complex
nt ..... Nucleotide
NUT ..... Nrd1-unterminated transcripts
OD ..... Optical density
ORF ..... Open reading frame
pA ..... Cleavage and polyadenylation site
PAF ..... Pol II-associated factor
PAGE ..... Polyacrylamide gel electrophoresis
PAR ..... Photoactivatable-ribonucleoside-enhanced
PAS ..... pA signal
PBS ..... Phosphate-buffered Saline
PCR ..... Polymerase chain reaction
PDE ..... Phosphodiesterase
PI ..... Processing index
PNK ..... Polynucleotide kinase
Pol ..... Polymerase
Pro ..... Proline
PVDF ..... Polyvinylidene difluoride
rA ..... Adenosine
RBD ..... RNA-binding domain
RBP ..... RNA-binding protein
rC ..... Cytidine
rG ..... Guanosine
RNA ..... Ribonuclein acid
RNP ..... Ribonucleoprotein
RRM ..... RNA recognition motif
rRNA ..... ribosomal RNA
RT ..... Rerverse transcription
RTase ..... Reverse transcriptase

rU ..... Uridine
SAM ..... Sequence Alignment/Map
SC ..... Synthetic complete
SDS ..... Sodium dodecyl sulfate
SELEX ..... Systematic evolution of ligands by exponential enrichment
Seq ..... Sequencing
Ser ..... Serien
SGD ..... Saccharomyces genome database
SI ..... Slicing index
snoRNA ..... small nucleolar RNA
SNP ..... Single nucleotide polymorphism
snRNA ..... small nuclear RNA
snRNP ..... small nuclear Ribonucleic particle
SS ..... Splice site
ssDNA ..... single stranded DNA
SUT ..... Stable utranslated transcript
TAP ..... Tandem affinity purification
TBE ..... TRIS-Borat-EDTA
TBP ..... TATA-binding protein
TEC ..... Transcription elongation complex
TEV ..... Tobacco etch virus
TF ..... Transcription factor
Thr ..... Threonine
TIF ..... Transcript isoform
tRNA ..... transfer RNA
TSS ..... Transcription start sits
Tyr ..... Tyrosine
U ..... Units
UBA ..... Ubiquitin-associated
UTR ..... Untranslated region
UV ..... Ultraviolet
v/v ..... Volume/volume
w/v ..... Weight/volume
YPD ..... Yeast extract peptone dextrose
ZF ..... Zinc finger

Appendix

Supplementary Figures

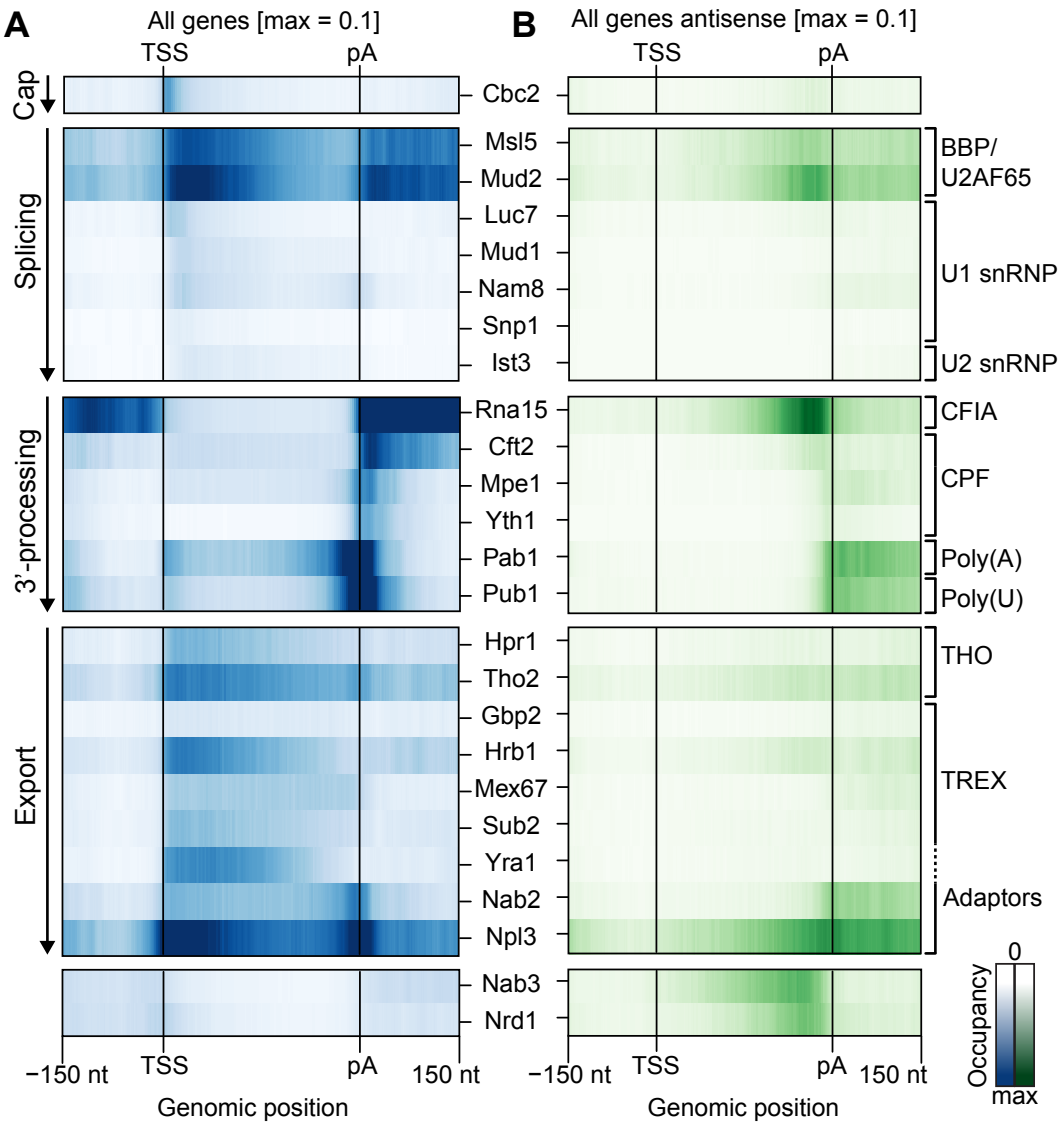

Figure S1: **Overview of occupancy profiles of all investigated proteins on ORF-Ts.** Smoothed occupancy profiles around all ORF-Ts were aligned at their TSS, length-scaled such that their pA sites coincided, and the occupancies averaged over all transcripts. **A**. Occupancy profiles on sense strand, i.e., on the proper mRNA. **B**. Occupany on the transcripts antisense to the annotated mRNA direction. Note the high occupancy of early termination factors Nab3 and Nrd1, termination factor Rna15, splicing factor Mud2, and export adaptor Npl3 on antisense transcripts.

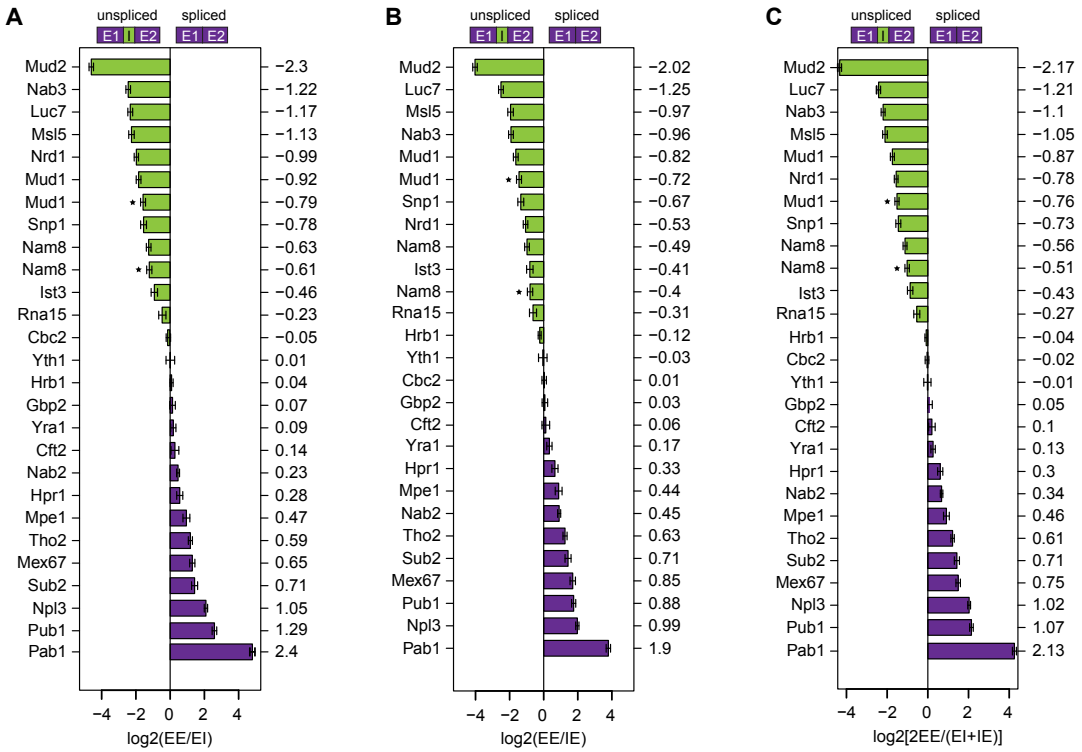

Figure S2: **The splicing index is robust with respect to using the coverage of coverage exon-intron or intron-exon junctions. A**. Splicing index calculated using coverage of exon-intron (EI) junctions instead of arithmetic mean of EI and IE junctions covered. **B**. Splicing index calculated using coverage of inton-exon (IE) junctions instead of arithmetic mean of EI and IE junctions covered. **C**. Splicing index calculated using arithmetic mean of EI and IE junctions covered.

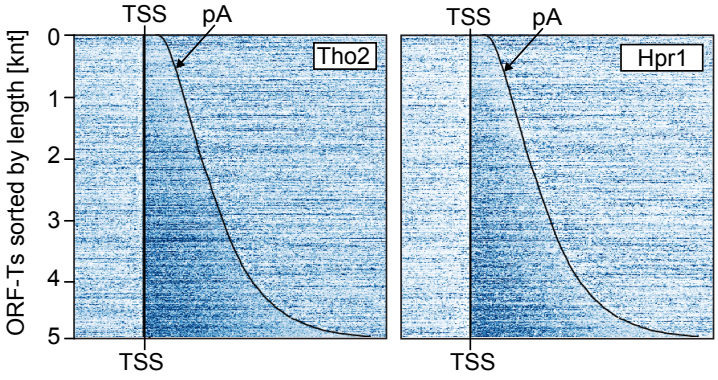

Figure S3: **Occupancy profiles of TREX complex components Tho2 and Hpr1 around ORF-Ts aligned at their TSS and sorted by length.**

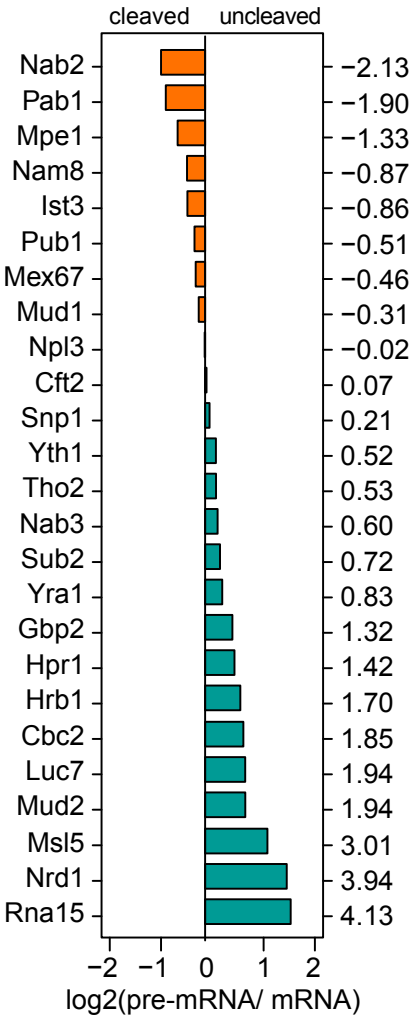

Figure S4: **Processing indices of all investigated factors, sorted by processing index values.**

Curriculum vitae

Persönliche Informationen

GEBURTSTAG	23. Oktober 1984
GEBURTSORT	Borna, Deutschland
NATIONALITÄT	deutsch
FAMILIENSTAND	ledig, ein Sohn (Bruno)
WOHNHAFT	Höfestieg 7 in 37077 Göttingen
E-MAILADRESSE	carlo.baejen@gmx.de

Wissenschaftlicher Werdegang

AB 07/2014	**Wissenschaftlicher Mitarbeiter** Molekularbiologie	
MIT PATRICK CRAMER	Max-Planck-Institut für biophysikalische Chemie	Göttingen
10/2010 – 06/2014	**Doktorand** Systembiochemie	
MIT PATRICK CRAMER	Genzentrum, Ludwig-Maximilians-Universität	München
02/2010 – 08/2010	**Gastwissenschaftler** RNA Biologie	
MIT THOMAS TUSCHL	Rockefeller University	New York City, USA
03/2008 – 08/2008	**Diplomand** Mikrobielle Physiologie	
MIT ROLAND MÜLLER	Helmholtz-Zentrum für Umweltforschung	Leipzig
07/2007 – 09/2007	**Hilfswissenschaftler** Ökogenetik	
08/2006 – 02/2007	**Forschungspraktikant** Molekularer Biotechnologie	
MIT UTA BREUER	Helmholtz-Zentrum für Umweltforschung	Leipzig

Schulische und akademische Ausbildung

10/2010 – 07/2014	**Dr. rer. nat.** Biochemie	
PROMOTION	Ludwig-Maximilians-Universität	
PHD-PROGRAMM	*NanoBioTechnology*, Center for NanoScience	München
09/2008 – 10/2010	**M.Sc. in Engineering** Biotechnologie	
MASTERSTUDIUM	Management Center Innsbruck	Innsbruck, Österreich

09/2004 – 08/2008	**Diplom-Ingenieur (FH)** Umwelttechnik	
INGENIEURSSTUDIUM	Hochschule Mittweida	Mittweida
08/1995 – 06/2003	**Allgemeine Hochschulreife** (Abitur)	
SEKUNDARSTUFE II	Johann-Gottfried-Seume-Gymnasium	Grimma

Weiterbildung (Zertifikate)

10/2012 – 01/2014	**Vertriebsingenieur (IHK)**	
FERNLEHRGANG	Industrie- und Handelskammer und SGD	Darmstadt
10/2011 – 02/2012	**Projektmanagement**	
SEMESTERKURS	Ludwig-Maximilians-Universität	München
10/2010 – 11/2011	**Gewerblicher Rechtsschutz** (Patentingenieur)	
FERNSTUDIENKURS	FernUniversität Hagen	Hagen
10/2010 – 05/2011	**Selbst-, Führungs- und Lehrkompetenzen**	
SEMESTERKURSE	Center for Leadership and People Management	München
06/2009 – 09/2009	**Advanced English Course**	
SPRACHREISE	Kaplan Aspect English College	Cairns, Australien

——————————- **Göttingen**, den 25.07.2014