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

Structural and Biochemical Characterization of the human Cleavage Stimulation Factor CstF

Michaela Hartwig

aus

Gräfelfing, Deutschland

2022

Erklärung

Diese Dissertation wurde im Sinne von §7 der Promotionsordnung vom 28. November 2011 von Frau Prof. Elena Conti, PhD betreut.

Eidesstattliche Versicherung

Diese Dissertation wurde eigenständig und ohne unerlaubte Hilfe erarbeitet.

München, den 10.05.2022

Michaela Hartwig

Michaela Hartwig

Dissertation eingereicht am 10.05.2022

1. Gutachterin: Prof. Dr. Elena Conti
2. Gutachter: Prof. Dr. Klaus Förstemann

Mündliche Prüfung am 29.06.2022

CONTENTS

Summary

Polyadenylation of pre-mRNAs is an essential step in maturation of nascent pre-mRNA transcripts. This highly conserved process consists of two essential steps: endonucleolytic cleavage of e tpre-mRNA at the cleavage site, also called poly(A) site, and addition of adenine nucleotides to the upstream cleavage product. Both reactions are mediated by a huge protein machinery, called the human 3'-end processing machinery, which consists of several multi protein complexes. Certain subcomplexes of this machinery are essential to define location of the cleavage site by interacting with a set of three distinct *cis*-elements, called poly(A) signals (PAS), on the RNA transcript. An UGUA-containing sequence element upstream of the poly(A) site is recognized by Cleavage Factor I_m, the most characteristic and very conserved AAUAAA PAS is bound by Cleavage and Polyadenylation Specificity Factor (CPSF) and a G/U-rich *cis*-element downstream the poly(A) site is bound by Cleavage Stimulation Factor (CstF). Hexameric AAUAAA and G/U-rich downstream elements are sufficient to define the cleavage site, which is located between both motifs. UGUA-poly(A) signals are supposed to fine tune positioning of protein factors along poly(A) signals.

Within the last years, many structures of protein complexes of the human 3'-end processing machineries have been solved by either cryo-EM or X-ray crystallography, helping to understand the molecular mechanism of this highly dynamic process. However, overall structure of the CstF complex was not solved yet and basis of its RNA target selection remained unclear. In contrast to the CPSF-AAUAAA interaction, CstF is missing a conserved consensus sequence motif within downstream *cis*-elements. Although several studies have proposed a bipartite sequence pattern consisting of either G/U-rich or U-rich sequence motifs, and determined binding affinities of the RNA binding domain of CstF2, up to date, no data is available about binding affinities of the full-length CstF complex.

In presented experiments, recombinantly purified CstF complex and several subcomplexes were used for biochemical and biophysical studies with a special focus on the RNA binding behavior. Moreover, mutational analysis of the CstF2 RNA recognition motif (RRM) revealed a set of amino acids involved in RNA binding. By mimicking dimeric CstF assembly in a simple setup, CstF2 S17 residue was identified to play a dual role in RNA binding depending on presence of full-length proteins or the single RRM domain. Additionally, a yet unidentified role of C-terminal residues of CstF2 was assigned to RNA binding mechanism, by providing a second strong RNA binding domain due to presence of 17 RG/RGG motifs.

Summary

From a structural point of view, the full-length CstF complex was intensively studied by cryo-Electron Microscopy (cryo-EM). However, due to its dynamic structural arrangement, CstF was too flexible to obtain a high-resolution reconstruction from the full complex. Consequently, the more stable CstF1-CstF3 subcomplex was structurally characterized by cryo-EM and showed a very dynamic behavior of the CstF1 WD40 propellers. I was able to obtain medium resolution reconstructions of the CstF1-CstF3 subcomplex, where I could build a model of a minimal CstF complex based on available structures and AlphaFold (AF) models in combination with data derived from XL-MS. By modelling the minimal CstF, I suggested that dynamic movement of the CstF1 WD40 propellers is indirectly linked to different positions of the CstF2 RRM domains, thereby potentially influencing the RNA binding mechanism.

To sum up, biochemical and structural investigations of the CstF complex provided initial insights into the structural arrangement of its subunits and its highly flexible behavior. Besides that, by presented RNA binding studies, new insights into the RNA sequence selection mechanism were gained and binding affinities for full CstF were determined for different G/U-rich containing RNA ligands. Presented results deliver an anchor for follow-up research to better understand the highly complex mechanism of mRNA polyadenylation.

1. Introduction

1.1 The mRNA life cycle

The central dogma of molecular biology claims, that in flow of genetic information, which is stored in deoxyribonucleic acid (DNA) in the nucleus, ribonucleic acid (RNA) serves as blueprint of the genetic code carried within the DNA. Via so called messenger-RNAs (mRNAs), which are transcribed from a DNA template, genetic information from this DNA template is transported from the nucleus into the cytoplasm, where proteins are synthesized. Generation of mRNA transcripts occurs in a highly regulated process, called transcription. However, to produce matured mRNAs, which can effectively be exported into the cytoplasm and translated into the corresponding amino acid sequences, initial pre-mRNA transcripts have to undergo several nuclear maturation and processing steps, including 5′-end capping, splicing and formation of 3′-ends. All these steps are tightly coupled to the process of transcription and highly regulated in gene expression. After the mRNA maturation process, exported RNAs are translated in the cytoplasm into the corresponding primary protein sequence in a process called translation.

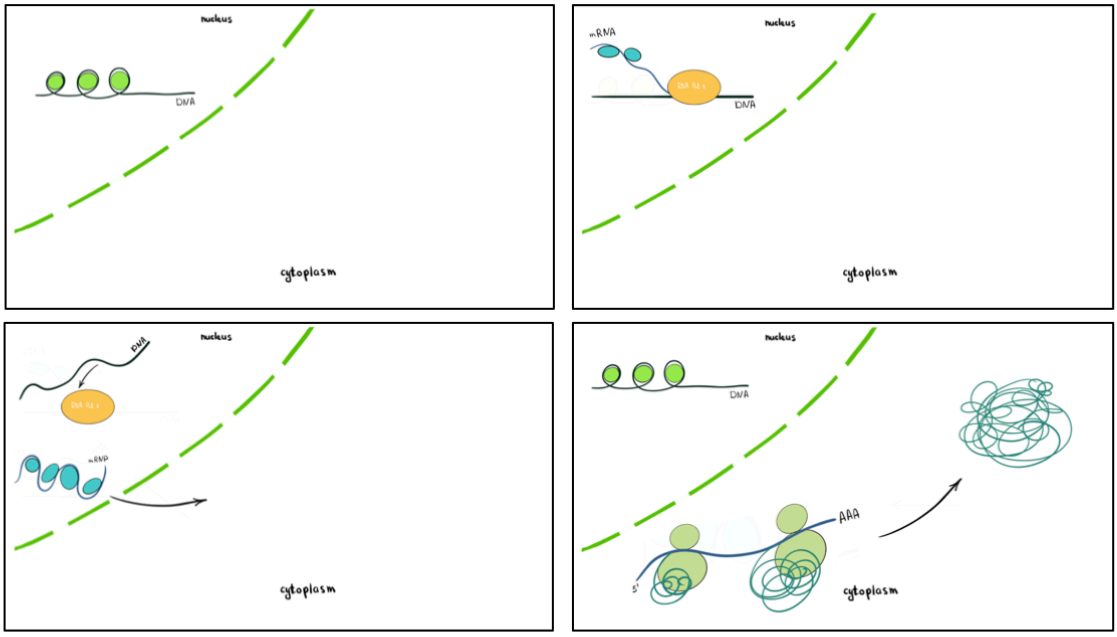


Figure 1. Central dogma of molecular biology. Francis Crick postulated 1957, that the flow of genetic information stored as DNA (top row) is transported via mRNA from the nucleus into the cytoplasm. Top row left: genetic information is stored as DNA in the nucleus. Tow row right: A nascent mRNA transcript is generated from a certain DNA template in a process called transcription. Pre-mRNAs undergo various maturation steps, before a mature mRNA packed as mRNP can be exported into the cytoplasm (bottom row left). Bottom row right: Translation takes place at ribosomes in the cytoplasm. Genetic information carried in the mRNA transcript is translated into corresponding amino acid sequence. Proteins are synthesized and structurally folded based on given mRNA sequence.

1.1.1 Transcription

Generation of pre-mRNA transcripts from a DNA template is the first essential step in gene expression. There are three different RNA polymerases in eukaryotes (RNA pol I-III), but RNA pol II is responsible for generation of most mRNA species including noncoding RNAs (ncRNAs) (Vannini and Cramer 2012). Usually, genes contain a promotor and terminator region to determine start and end of the transcription process. The promotor region is, with some exceptions, located upstream of the transcription start site (TSS). The transcription process is initiated at the promotor sequence by formation of a so-called pre-initiation complex (PIC) consisting of general transcription factors (GTFs) recruiting RNA pol II and the mediator complex (Cramer 2004, Carninci, Sandelin et al. 2006, Fuda, Ardehali et al. 2009, Sikorski and Buratowski 2009, Baumann, Pontiller et al. 2010, Malik and Roeder 2010). After release from the promotor, RNA pol II continues with transcript elongation (Yudkovsky, Ranish et al. 2000). RNA pol II associated with the DNA template, nascent transcript and various protein factors is called the elongating complex (EC).

1.1.2 Co-transcriptional pre-mRNA processing

Coupled to the process of transcription, the RNA product synthesized by RNA pol II undergoes several processing steps, which are essential for the maturation of messenger RNAs (mRNA) in eukaryotes.

1.1.2.1 5'-end capping

The 7-methylguanosine (m^7G) cap at the 5'-ends of eukaryotic mRNAs is very important for splicing, nuclear export of mRNAs and their stability (Ghosh and Lima 2010, Li and Kiledjian 2010). 5′-end capping is the first modification step of pre-mRNA transcripts after transcription initiation (Shatkin 1976, Shatkin and Manley 2000, Merrick 2004, Liu and Kiledjian 2006, Ghosh and Lima 2010). The 5′-cap is not only important for the cap-dependent initiation of protein synthesis, but it also functions as protector from 5′ to 3′ exonucleases (Ramanathan, Robb et al. 2016). As soon as the nascent mRNA transcript is 25 to 30 bases long, the 5′-end of the pre-mRNA emerges from the RNA exit channel of RNA pol II and 5′-capping is initiated (Shatkin and Manley 2000, Moteki and Price 2002). Three enzymatic reactions are needed to convert the 5′-triphosphate of the pre-mRNA into the cap-structure (Shuman 2001, Gu and Lima 2005): RNA triphosphatase (TPase), RNA guanylyltransferase (GTase) and guanine-N7 methyltransferase (guanine N7-MTase). After addition of the m^7G cap structure, the RNA TPase and GTase complex is released from the mRNA transcript in a process coupled to dephosphorylation of Ser5 of the RNA pol II C-terminal domain (CTD) (Schroeder, Schwer et

al. 2000), whereas the RNA MTase remains associated with RNA pol II and travels along the gene (see figure 2 A-D).

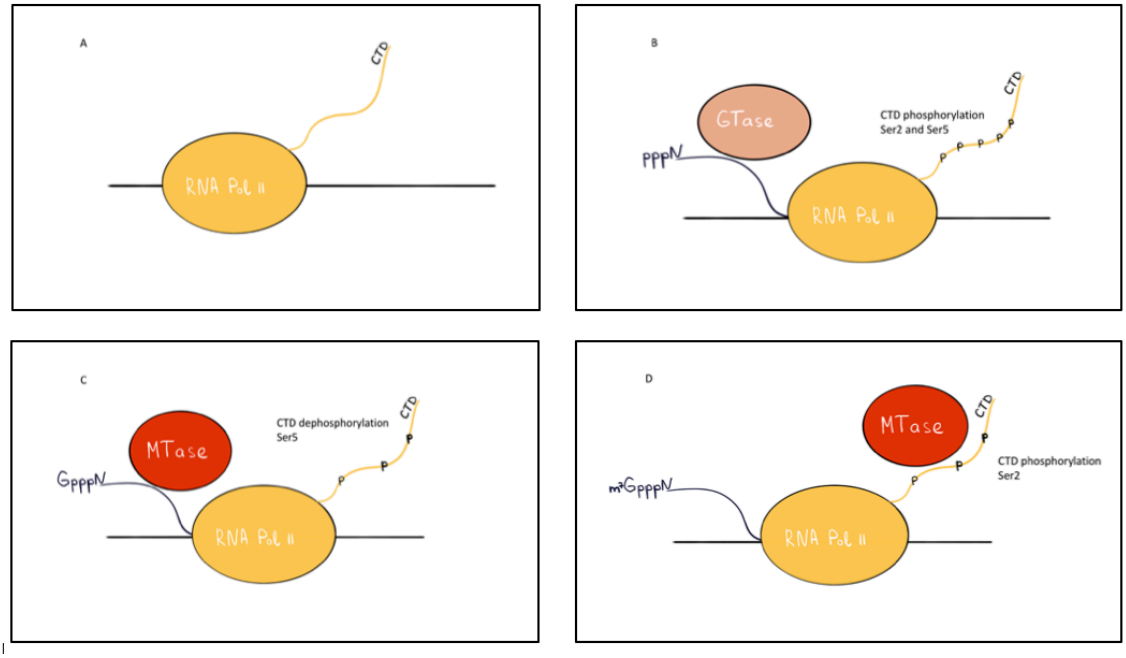


Figure 2. Model of co-transcriptional 5'-end capping. A) 5'-end capping occurs co-transcriptional. B) Co-transcriptional 5'-capping is initiated by recruitment of the capping machinery by phosphorylation of the RNA pol II CTD at position Serine2 (Ser2) and Serine5 (Ser5). TPase hydrolyzes the 5' triphosphate end to a diphosphate (step not shown) immediately followed by addition of a GMP to the diphosphate end by the GTase. C) MTase (red circle) methylates 5' guanine at position guanine N7 and due to loss of Ser5 phosphorylation, GTase dissociates from the elongating complex. D) Ser2 phosphorylation maintains the CTD-MTase interaction

1.1.2.2 Splicing

Splicing of pre-mRNA transcripts is an essential step in gene expression to remove non-coding regions (introns), which are interspersing with protein coding regions (exons) of a gene (Berget, Moore et al. 1977, Chow, Gelinas et al. 1977). Besides the very precise canonical splicing mechanism, some mammalian genes are also subject to alternative splicing events, resulting in the production of alternative mRNA isoforms (Pan, Shai et al. 2008). This process creates another regulatory level of gene expression and drives proteome diversity. The splicing reaction is mediated by a mega-Dalton ribonucleoprotein (RNP) machinery called the spliceosome (Brody and Abelson 1985, Frendewey and Keller 1985, Yan, Wan et al. 2019). The spliceosome consists of several uridine-rich small nuclear RNP (snRNP) particles named U1, U2. U4, U5 and U6, which are surrounded by various protein factors, so-called splicing factors (Wahl, Will et al. 2009). A splicing cycle consists of three phases: spliceosome assembly and activation, the actual splicing reaction and disassembly of the spliceosomal machinery (Yan, Wan et al. 2019). The location of the actual splicing reaction at introns is

defined by three sequence motifs, the so-called 5′-splice site (SS), 3'-SS and branch site (BS) (Will and Luhrmann 2011).

1.1.2.3 Transcription termination and 3'-end formation

Evidence, that transcription termination is tightly coupled to 3'-end processing of pre-mRNA transcripts, was obtained from several studies (Moore and Proudfoot 2009). In detail, transcription termination depends on recognition of the poly(A) site and must therefore occur co-transcriptionally (Nagaike and Manley 2011). The formation of 3′-ends is one of the fundamental steps in maturation of initial pre-mRNAs. This step is not only coupled to transcription termination, but also tightly connected to splicing, mRNA export, translation and mRNA stability (Colgan and Manley 1997, Zhao, Hyman et al. 1999, Mandel, Bai et al. 2008, Moore and Proudfoot 2009). Defects in 3′-end processing can affect cell growth and dysfunctional polyadenylation might lead to diseases like thalassemia and lysosomal storage disorder (Higgs, Goodbourn et al. 1983, Orkin, Cheng et al. 1985, Gieselmann, Polten et al. 1989, Zhao, Hyman et al. 1999, Danckwardt, Hentze et al. 2008). The efficiency of 3′-end processing and the diversity of RNA isoforms can be regulated by the usage of alternative poly(A) sites and length control of the poly(A) tail (Di Giammartino, Nishida et al. 2011, Shi, Kirwan et al. 2012, Tian and Manley 2013). Alternative polyadenylation is not only a regulator of transcript levels in different cell types, but also involved in various diseases (Ji, Lee et al. 2009, Mayr and Bartel 2009). Most of eukaryotic pre-mRNA transcripts, except histone pre-mRNA, are processed in a tightly coupled two-step mechanism (Dominski and Marzluff 1999): First, RNA is cleaved at a specific site, second, poly(A) polymerase adds a tail of 200-250 adenines (poly(A) tail) (Wahle and Ruegsegger 1999, Zhao, Hyman et al. 1999). These two steps require a multitude of proteins, that form a huge machinery or a dynamic set of protein complexes, which are highly regulated and allow for cross-talk with other steps of gene expression (Moore and Proudfoot 2009).

1.1.2.3.1 Transcription termination

Transcription termination is closely connected to cleavage and polyadenylation of the nascent mRNA transcript. As soon as the transcribing RNA pol II passes the 3′-end *cis*-elements, which serve as poly(A) signals and are required for transcription termination, and reaches the terminator region located downstream of the poly(A) signals, endonucleolytic cleavage of mRNA is triggered (Zaret and Sherman 1982, Whitelaw and Proudfoot 1986, Logan, Falck-Pedersen et al. 1987). Two models exist to describe the 3′-end processing coupled transcription termination. The so-called ‘allosteric model’ is based on conformational changes of RNA pol II as a consequence of dissociating transcription factors after passing the 3′-end

poly(A) signal (Logan, Falck-Pedersen et al. 1987). This leads to transcription termination without forward translocation of the enzyme (Licatalosi, Geiger et al. 2002, Kim, Ahn et al. 2004, Kim, Krogan et al. 2004, Zhang, Fu et al. 2005, Zhang and Gilmour 2006, Epshtein, Cardinale et al. 2007). The second model, the 'torpedo model', relies on endonucleolytic pre-mRNA cleavage during 3′-end processing (Connelly and Manley 1988). It is supposed, that the 5′ to 3′ exonuclease Xrn2 catches up with elongating RNA pol II and uses the 5′-phosphate generated by RNA cleavage at the poly(A) site as entry point, thereby leading to transcription termination and dissociation of RNA pol II (Kim, Krogan et al. 2004, West, Gromak et al. 2004)

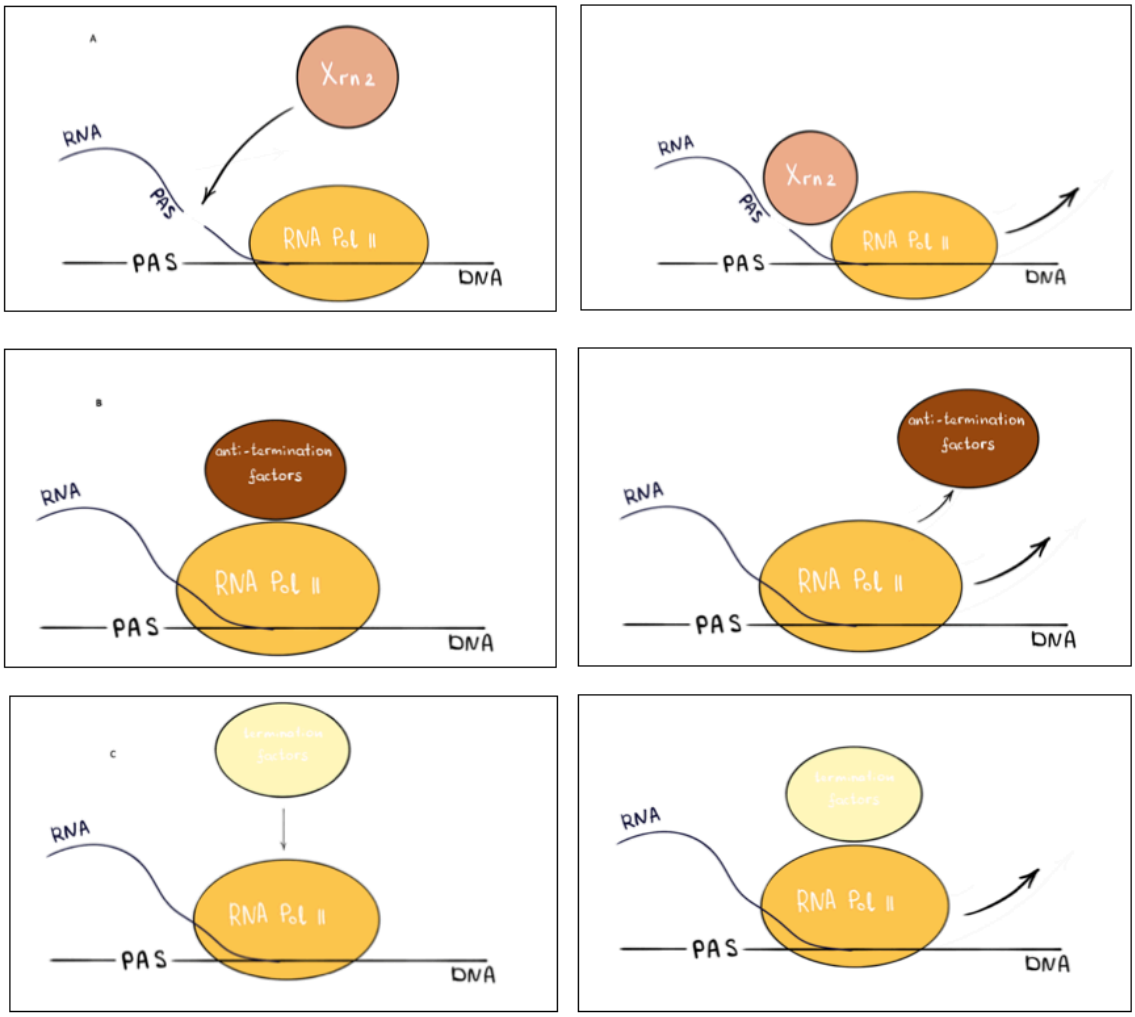


Figure 3. Current models for poly(A) site dependent transcription termination (adopted from Rosonia et al., 2006). A) The 'torpedo model' is based on endonucleolytic cleavage of RNA transcripts, which creates an entry site for exonuclease Xrn2. Xrn2 degrades the downstream cleavage product and RNA pol II-mediated transcription on the remaining RNA is terminated. B and C) The 'allosteric model' is based on RNA pol II undergoing conformational changes upon passing of poly(A) signals (PAS). Release of RNA pol II from the DNA template is mediated by the dissociation of anti-termination factors (B) or association of termination factors (C).

1.1.2.3.2 Nuclear cleavage and polyadenylation

Polyadenylation of the 3′-end of pre-mRNAs is a co-transcriptional occurring process, discovered in early studies in nuclear extract of calf thymus (Edmonds and Abrams 1960). Addition of a poly(A) tail to the 3′-untranslated regions (3′-UTR) of synthesized pre-mRNA transcripts happens to almost all prokaryotic and eukaryotic pre-mRNAs, except replication-dependent histone pre-mRNAs (see paragraph 1.1.2.3.5). The process of mRNA cleavage and polyadenylation (CPA) is tightly coupled to transcription termination and mediated by a multi-protein machinery, the so called 3′-end processing complex. The human 3′-end processing machinery consists of several multi-protein complexes, which include Cleavage and Polyadenylation Specificity Factor (CPSF), Cleavage Stimulation Factor (CstF), Cleavage Factors I (CF I_m) and II (CF II_m) and the scaffold protein Symplekin (SYMPK) (Figure 4).

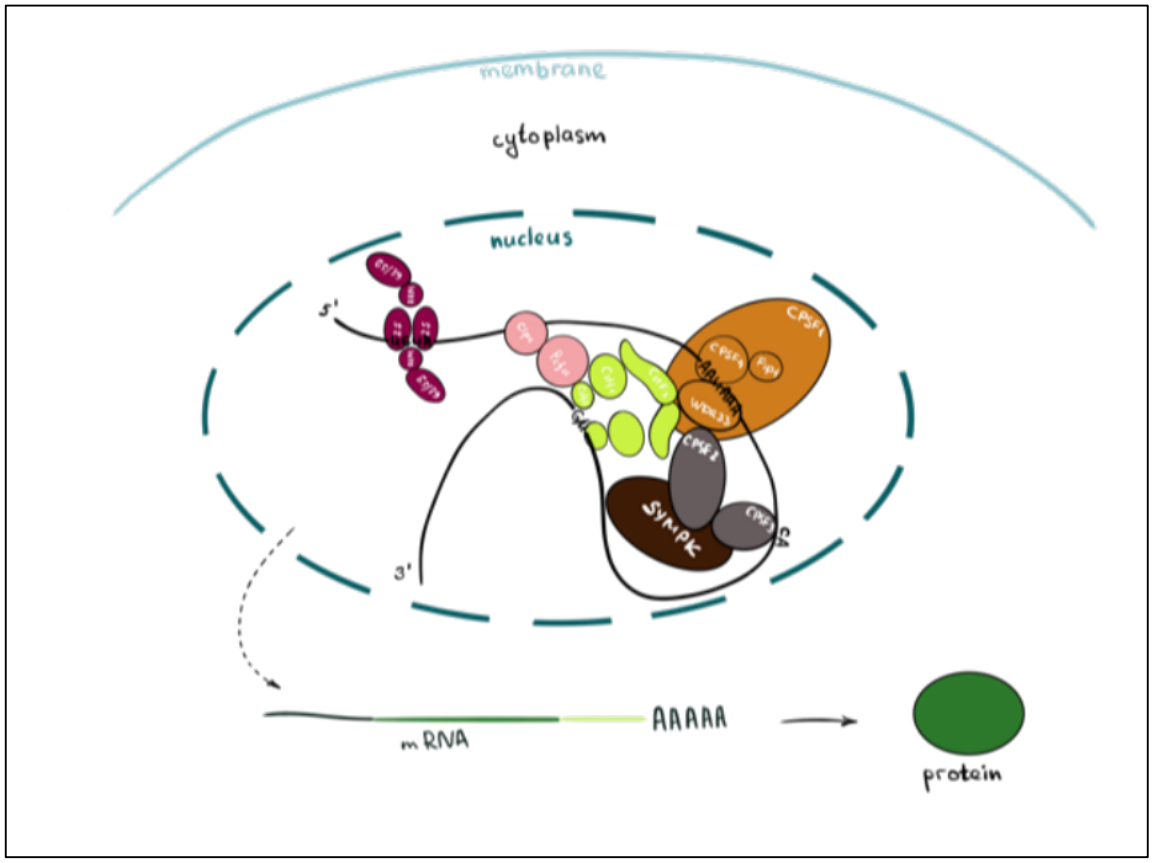


Figure 4. Nuclear polyadenylation machinery. Proteins of the human 3'-end processing machinery are assembled in the nucleus on a pre-mRNA transcript containing three sequence elements (polyA signals; PAS). UGUA: upstream element – AAUAAA: conserved hexamer – G/U: GU-rich downstream element. The cleavage site is depicted by the conserved CA dinucleotide. Purple: CF I_m – rose: CF II_m – orange: mPSF – grey: mCF – brown: Symplekin – green: CstF. Matured mRNAs containing poly(A) tails are exported into the cytoplasm.

CPA is initiated by recruitment of the human 3′-end cleavage and polyadenylation machinery to the so-called poly(A) site by means of short conserved sequence motifs in the 3′-UTR of pre-mRNAs in defined distance from each other. In humans, the poly(A) site is defined by at least two poly(A) signals (Nunes, Li et al. 2010) and therefore located between a highly conserved hexameric AAUAAA sequence motif and a G/U-rich sequence motif downstream of the poly(A) site. The actual poly(A) site has no consensus sequence, but is characterized by a conserved CA sequence (Figure 5). Additionally, a set of UGUA-rich sequence elements located upstream of the AAUAAA poly(A) signal helps to define the strength of a poly(A) site, meaning the frequency of its selection. Besides that, these sequence elements serve as additional platform for proteins of the 3′-end processing machinery, which will be introduced in detail in paragraph 1.2 (Danckwardt, Kaufmann et al. 2007, Hall-Pogar, Liang et al. 2007).

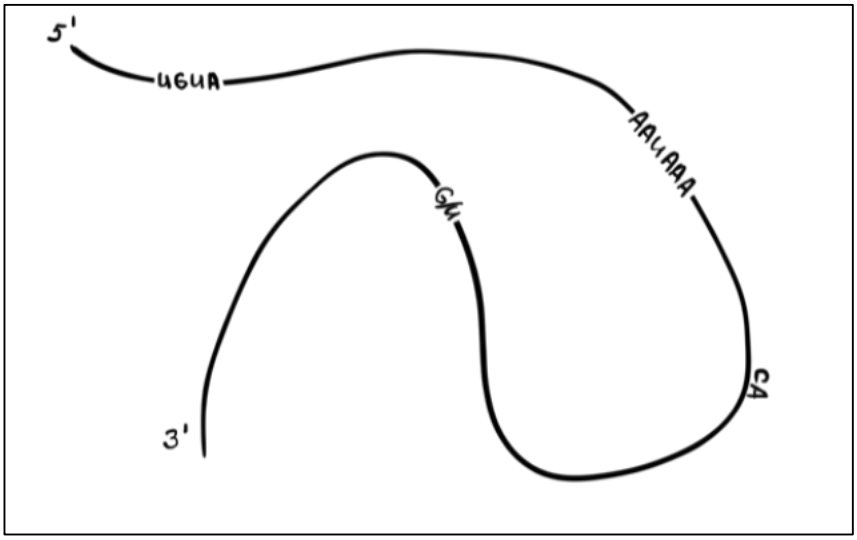


Figure 5. Poly(A) signals in 3'-UTRs of human pre-mRNAs. 3'-UTRs of genes contain a set of poly(A) signals (PAS) in distinct distance to each other defining location of the poly(A) site (CA). Upstream elements UGUA are located 40-100 nt upstream of the cleavage site (CA). Very conserved hexameric AAUAAA poly(A) signal is located 10-15 nt upstream of the cleavage site. G/U-rich downstream elements are located within 30 nt downstream of the poly(A) site.

Recognition of poly(A) signals on the pre-mRNA starts with recruitment of the CPSF complex to the hexameric AAUAAA consensus motif and simultaneously of the CstF complex to G/U-rich downstream elements (DSEs) (Takagaki, MacDonald et al. 1992, Takagaki and Manley 1994, Takagaki and Manley 1997, Takagaki and Manley 2000, Shi, Di Giammartino et al. 2009, Chan, Huppertz et al. 2014, Schonemann, Kuhn et al. 2014). Binding of CF I_m to the UGUA sequence elements of pre-mRNA was thought to stabilize the CPSF-RNA interaction (Coseno, Martin et al. 2008, Yang, Gilmartin et al. 2010). With help of this network of protein-RNA interactions, the endonuclease CPSF3 is positioned at the cleavage site (Mandel, Kaneko et al. 2006) to perform endonucleolytic cleavage of the RNA transcript (Ryan, Calvo et al. 2004, Dominski, Yang et al. 2005).

After cleavage, poly(A) polymerase (PAP) is recruited to the mRNA by the CPSF complex (Christofori and Keller 1988, Takagaki, Ryner et al. 1988, Wahle 1991) and addition of a poly(A) tail by PAP to the 3′-end of the upstream cleavage product is initiated in a slow mode (Bienroth, Keller et al. 1993). The emerging poly(A) tail is bound by PABPN1 molecules, which interact with CPSF to further stimulate PAP in a way that polyadenylation is switched to a fast processive mode until a length of 200-250 adenines is reached. Length control of poly(A) tails is achieved by loss of the cooperative stimulation of PAP by the CPSF complex and PABPN1, which is necessary for fast and processive polyadenylation. Upon a length of 250 nucleotides, processive poly(A) tail elongation is aborted and switched to a slow distributive manner (Wahle 1995, Kuhn, Gundel et al. 2009). According to the current understanding, the growing poly(A) tail bound by PABPN1 was shown to form a circular arrangement, thereby folding back and maintaining the CPSF – PAP interaction necessary for processive elongation. This interaction is disrupted once a length of approx. 250 nucleotides is reached. Therefore, PABPN1 seems crucial for length control (Kuhn, Gundel et al. 2009). Poly(A) tails are supposed to be involved in the formation of export-competent mRNPs together with several proteins recruited to the RNA transcript (Chen, Li et al. 1999, Apponi, Leung et al. 2010).

1.1.2.3.3 Cytoplasmic polyadenylation

Cytoplasmic polyadenylation is a regulatory process which activates silenced RNA transcripts containing a short poly(A) tail by elongating the poly(A) tail in the cytoplasm (Belloc and Mendez 2008, Radford, Meijer et al. 2008, Villalba, Coll et al. 2011, Weill, Belloc et al. 2012). Thereby, protein expression is increased by the translational activation of silenced mRNAs containing short poly(A) tails. This process is mainly involved in oocyte maturation and cell cycle progression and was observed in early embryos of many species (Paris, Osborne et al. 1988, Bilger, Fox et al. 1994, Salles, Lieberfarb et al. 1994). Like canonical nuclear polyadenylation, this step relies on *cis*-acting sequences on the RNA, which have been identified by studies in *Xenopus* oocytes (Belloc and Mendez 2008, Radford, Meijer et al. 2008). One of the sequence motifs is the conserved hexameric AAUAAA sequence. The second sequence element, the so-called cytoplasmic polyadenylation element (CPE), is U-rich and recognized by the conserved cytoplasmic polyadenylation element binding protein (CPEB) (McGrew and Richter 1990, Richter 2007, Coll, Villalba et al. 2010, Villalba, Coll et al. 2011). The distance between both sequence motifs on the RNA has regulatory effects on cytoplasmic polyadenylation (Pique, Lopez et al. 2008). Usually, a poly(A) tail minimum length of 85 nucleotides is required for PABP-mediated translation (Abaza and Gebauer 2008, Jackson, Hellen et al. 2010). Cytoplasmic polyadenylation of RNA transcripts with a short poly(A) tail is achieved by CPEB, the cytoplasmic form of the CPSF (CyPSF) complex, Symplekin and the cytoplasmic poly(A) polymerase Gld-2 (Barnard, Ryan et al. 2004), which is recruited to the

mRNAs by CyPSF (Wang, Eckmann et al. 2002, Barnard, Ryan et al. 2004). CPEB recruits CyPSF and Symplekin and stabilizes the interaction between CyPSF and the AAUAAA sequence element (Mendez, Murthy et al. 2000, Dickson, Thompson et al. 2001, Gavin, Bosche et al. 2002). Cytoplasmic poly(A) polymerase Gld-2 is then recruited to the complex for poly(A) tail elongation.

1.1.2.3.4 Alternative polyadenylation

Polyadenylation of mRNA transcripts is an essential step in mRNA maturation and consists of the endonucleolytic cleavage of pre-mRNA at the poly(A) site and addition of a 200-250 nucleotide long poly(A) tail. Previous studies showed that eukaryotic genes can have more than one poly(A) site and that their usage leads to formation of different mRNA isoforms (Proudfoot 2011, Xia, Donehower et al. 2014). This widespread phenomenon is called alternative polyadenylation (APA) and occurs mostly in 3′-UTRs of eukaryotic mRNAs (Derti, Garrett-Engele et al. 2012, Yeh and Yong 2016). The process of APA is a regulatory mechanism determining mRNA stability, localization and function and is implicated in various diseases (Millevoi and Vagner 2010, Erson-Bensan and Can 2016). Besides that, APA is important for transcriptome complexity and diversity by generating isoforms differing in their coding sequence and 3′-UTRs. This is a consequence of the location of alternative poly(A) sites, which can be located in the 3′-UTR (Figure 6, left panel) of genes, leading to transcripts with different 3′-UTRs (3′-UTR-APA). Second, alternative poly(A) sites can also be located within the coding region of genes (named CR-APA; exons or introns; Figure 6) and thereby produce transcript isoforms with different coding regions (Di Giammartino, Nishida et al. 2011, Tian and Manley 2013).

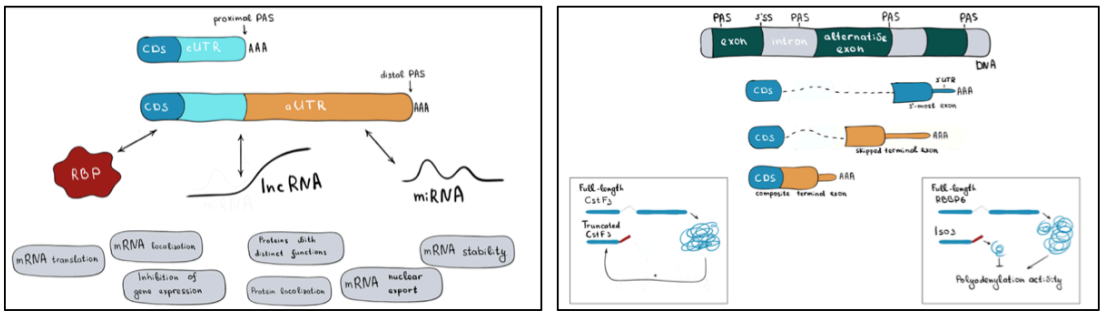


Figure 6. Two different forms of alternative polyadenylation (APA) depending on the location of alternative poly(A) site (Tian and Manley 2017). **Left: 3'-UTR-APA**. The alternative poly(A) site (termed as distal PAS) is located in the 3' UTR of a gene, leading to two different mRNA isoforms with different 3' UTRs. The proximal PAS produces a short mRNA isoform containing the canonical UTR (cUTR), whereas APA produces a long isoform containing the alternative UTR (aUTR). Interaction of RNA binding proteins (RBP) or different RNA species (miRNAs, lncRNAs) can have many functional consequences. **Right: CR-APA.** The alternative poly(A) site is located in the coding region, leading to production of mRNA isoforms with different exon content and therefore different coding sequences and 3' UTRs. mRNA isoforms are shown with containing exons and splicing events are indicated as dashed lines. First isoform is the canonical isoform produced by usage of the PAS located in the last 3' most exon. Bottom right: CR-APA of Rbbp6 results in an isoform acting as dominant-negative regulator (Iso3).

Bottom left: CR-APA of CstF3 leads to formation of a truncated protein with no distinct function. By activating usage of upstream PAS, full-length protein creates a feedback-loop leading to increased levels of truncated protein.

The process of APA involves most proteins of the canonical 3′-end processing machinery required for cleavage and polyadenylation. Recruitment of the APA machinery to alternative poly(A) sites is initiated by recognition of the UGUA region by CF I_m. Following assembly of CPSF and CstF at the CTD of RNA pol II, this complex is translocated with RNA pol II until recognition of the hexameric AAUAAA sequence element by the CPSF complex. The CstF complex now switches to binding of G/U downstream sequences and cleavage reaction at the alternative poly(A) site by CPSF3 is initiated. Associated PAP initiates the addition of adenosine nucleotides and PABPN1 proteins bind to the elongating poly(A) tail to continue APA until it is aborted in a PABPN1-dependent manner (Venkataraman, Brown et al. 2005, Ren, Zhang et al. 2020). There are a few drivers and regulators of APA among proteins of the canonical CPA machinery. Previous studies showed that upon depletion of Fip1, the usage of alternative poly(A) sites led to loss of self-renewal capabilities in mouse embryonic stem cells (ESCs) (Lackford, Yao et al. 2014). The CstF2 subunit of CstF was implicated in various cancer types and found to be a regulator of 3′-UTR shortening (Shell, Hesse et al. 2005, Hwang, Park et al. 2016). Additionally, CF I_m was found to be involved in alternative poly(A) site selection by looping out proximal PAS due to binding of two UGUA binding motifs upstream of an alternative poly(A) site (Venkataraman, Brown et al. 2005, Yang, Gilmartin et al. 2010, Yang and Doublie 2011). Besides that, it was supposed that CF I_m can inhibit proximal poly(A) site selection by recognizing non-optimal binding sites on the pre-mRNA and thereby suppressing recruitment of the CPSF complex (Martin, Gruber et al. , Masamha and Wagner 2018, Zhu, Wang et al. 2018). CF I_m knockdown was shown to globally influence the selection of alternative PAS and thereby increasing gene expression and transcript stability (Kubo, Wada et al. 2006, Weng, Ko et al. 2019). Recent studies showed, that CF I_m participates in APA regulation by binding to so-called enhancer elements at a poly(A) site (Zhu, Wang et al. 2018) and thereby acts as an activator for 3′-end processing of pre-mRNAs.

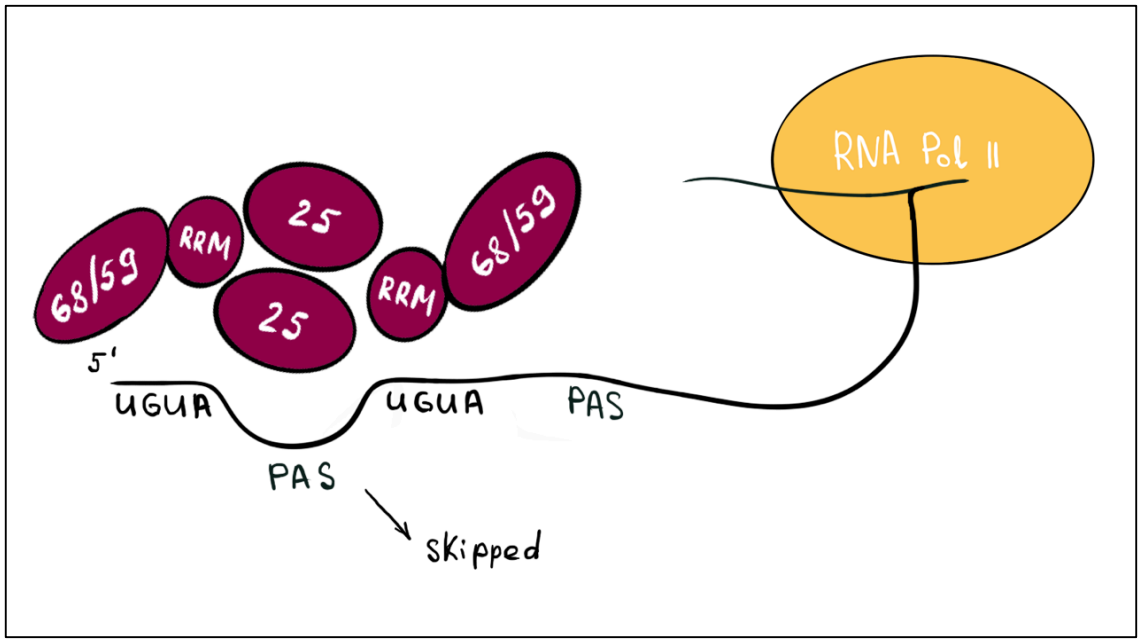


Figure 7. Model for regulation of APA by CF I_m (Tian and Manley 2017). A) CF I_m recognizes two UGUA u elements (USEs) flanking the canonical poly(A) site, therefore looping out and skipping of the canonical poly(A) signals (PAS),

1.1.2.3.5 3'-end processing of histone pre-mRNA

Metazoan replication-dependent histone pre-mRNAs are the only exception known so far, that do not undergo the classical cleavage and polyadenylation step. Instead, they contain a very conserved stem loop at their 3′-ends, which is crucial to regulate their synthesis in cell cycle (Marzluff, Wagner et al. 2008, Pirngruber and Johnsen 2010). In contrast to the canonical 3′-end processing of pre-mRNAs, there is only one endonucleolytic cleavage step necessary to form matured 3′-ends and to release the transcripts of replication-dependent histone genes from the DNA template (Pandey, Chodchoy et al. 1990). Histone pre-mRNA 3′-ends are formed by a special processing machinery, which recognizes certain sequence elements on histone mRNAs. The cleavage site is located between a stem loop (SL) and another distinct sequence element, the so-called histone-downstream-element (HDE) (Dominski and Marzluff 1999). The HDE is located around 15 nucleotides downstream of the cleavage site and is characterized by a high content of purines. Recent studies solved the structure of an active histone 3′-end processing machinery, whereas the molecular mechanism still remains unclear (Sun, Zhang et al. 2020).

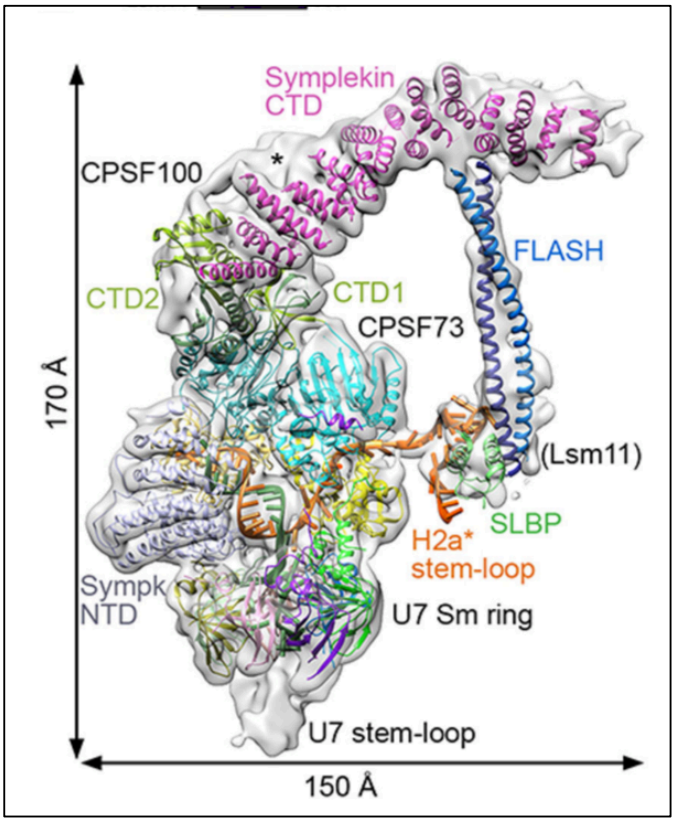


Figure 8. Cryo-EM structure of human histone 3'-end cleavage complex (Sun, Zhang et al. 2020). After lowpass filtering the EM map to 8 Å, density for the FLASH coiled coils and the SLBP are visible. CPSF2 is termed CPSF100 and CPSF3 is termed CPSF73 in this figure. Sympk: Symplekin

The histone pre-mRNA cleavage complex, termed HCC, consists of the endonuclease CPSF3, CPSF2, Symplekin and CstF2, which are also part of the canonical 3′-end cleavage and polyadenylation machinery (Kolev and Steitz 2005, Sun, Zhang et al. 2020). 3′-end cleavage of histone pre-mRNA transcripts is initiated by the assembly of an active complex based on U7 snRNP. FLASH and Lsm1 (Yang, Xu et al. 2011) are necessary for recruitment of HCC, containing the endonuclease CPSF3 (Sun, Zhang et al. 2020). This pre-assembled complex recognizes the pre-mRNA and defines the cleavage site (Sun, Zhang et al. 2020). The cleavage reaction is performed by CPSF3 (Dominski, Yang et al. 2005, Dominski, Yang et al. 2005, Dominski and Marzluff 2007). Immediately after cleavage, the processing machinery dissociates from the downstream cleavage product, which is degraded in 5′ to 3′ direction by the 5′-exonuclease activity of CPSF3 (Yang, Sullivan et al. 2009), to reassemble for a new processing cycle (Walther, Wittop Koning et al. 1998, Sun, Zhang et al. 2020).

1.1.3 mRNA export

The initial mRNA transcript undergoes several processing steps including 5′-capping, splicing and polyadenylation, as described above, before the matured ribonucleoprotein complex (mRNP) can be exported to the cytoplasm to be translated into the corresponding protein sequence (Singh, Pratt et al. 2015, Heath, Viphakone et al. 2016, Stewart 2019).
The traffic between nucleus and cytoplasm occurs for various molecules via the nuclear pore complex (NPC). Many cellular RNAs (tRNA, miRNA) travel through the NPC by help of certain transport receptors, in case of mRNAs, the importin/karyopherin-β receptor family (Hetzer and Wente 2009, Strambio-De-Castillia, Niepel et al. 2010). Nuclear export of mRNAs, which is conserved from yeast to humans, is not only mediated via the importin/karyopherin-β transport receptor, but requires the dimeric export factor Tap-p15 (Nxf1-Nxt1) (Kohler and Hurt 2007, Terry, Shows et al. 2007, Stewart 2010, Tutucci and Stutz 2011) (Santos-Rosa, Moreno et al. 1998, Hurt, Strasser et al. 2000). Although both proteins of the Tap-p15 dimer are able to bind RNA, additional factors are needed to specifically select mRNA targets (Segref, Sharma et al. 1997, Santos-Rosa, Moreno et al. 1998, Katahira, Strasser et al. 1999). Among these factors, the conserved transcription-export complex (TREX) is involved in mRNA target selection via the Tap-p15 heterodimer. The human TREX complex consists of the THO complex (Thoc1, Thoc2, Thoc3, Thoc5, Thoc6, Thoc7), DEXD/H-box helicase Uap56 and AlyRef (Reed and Cheng 2005, Rodriguez-Navarro and Hurt 2011, Tutucci and Stutz 2011, Chanarat, Burkert-Kautzsch et al. 2012, Katahira 2012). AlyRef directly interacts with the Tap-p15 dimer to function as an export adaptor (Strasser and Hurt 2000, Rodrigues, Rode et al. 2001). According to the current model, mRNP export is initiated by recruiting the THO complex to the mRNP and thereby bringing the Uap56 helicase in close proximity, so that it can ‘sandwich’ the mRNA (Puhringer, Hohmann et al. 2020). The interaction with export adaptor AlyRef is mediated via the Tho-Uap56 complex (Figure 9), so that Tap-p15 can be loaded on the mRNA via AlyRef (Strasser and Hurt 2001, Kohler and Hurt 2007, Hautbergue, Hung et al. 2008).

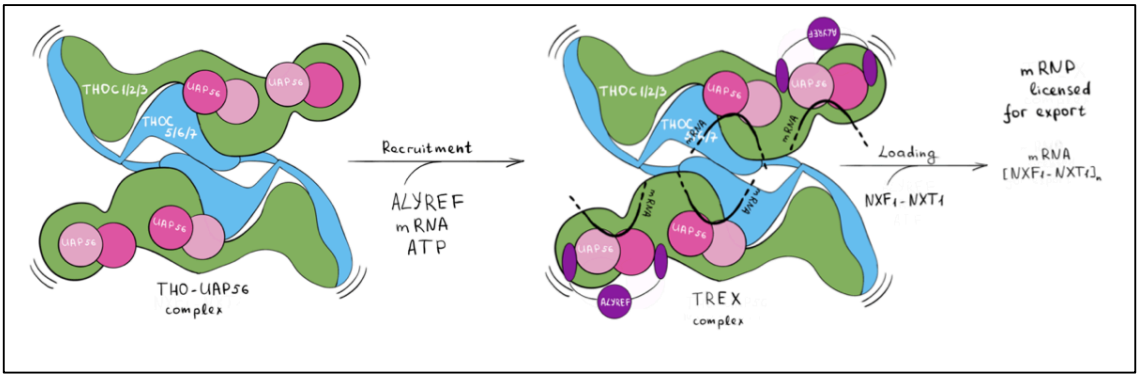


Figure 9: Model of TREX-dependent RNA export mediated by loading of Tap-p15 (here NXF1-NXT1) via AlyRef (Puhringer, Hohmann et al. 2020). Interactions between Tho-Uap56 complex are necessary to recruit

mRNA export factor AlyRef in presence of ATP, leading to RNA binding and loading of export adaptors NXF1-NXT1 (Tap-p15)

According to its name, TREX is involved in both, transcription and mRNA export, and is therefore a factor coupling interfaces of both processes (Jensen, Dower et al. 2003, Aguilera 2005, Reed and Cheng 2005, Katahira and Yoneda 2009, Rondon, Jimeno et al. 2010). The TREX complex is recruited to transcribed pre-mRNAs already during transcription and travels along with RNA pol II. Coupled to the process of transcription, the TREX complex interacts with different protein co-factors to facilitate loading of required adaptor proteins to form export-competent mRNPs (Strasser, Masuda et al. 2002, Zenklusen, Vinciguerra et al. 2002, Jensen, Boulay et al. 2004).

1.1.4 mRNA decay

mRNA decay is a very important posttranscriptional regulatory process in eukaryotic gene expression. The importance of this process is underlined by the presence of several mRNA decay pathways in the cytoplasm. In general, lifetime and fate of a mRNA is determined by its innate features and nucleotide sequence and is strongly related to the function of the encoded protein. Matured mRNAs, that are produced in the nucleus, are protected against non-specific exonucleases by the presence of a 5′-cap and a 3′-poly(A) tail, which can be directly linked to regulation of translation efficiency and mRNA stability (Wahle and Winkler 2013). Cytoplasmic mRNA degradation is generally initiated by deprotection of the mRNA transcript in either way, removing the 5′-cap or the 3′-poly(A) tail and thereby determining the degradation pathway (Figure 10). Deadenylated mRNAs can be degraded by the exosome in 3′ to 5′ direction or after additional decapping, Xrn1 degrades unprotected mRNAs in 5′ to 3′ direction (Meyer, Temme et al. 2004, Parker and Song 2004, Houseley and Tollervey 2009). Besides the major cytoplasmic mRNA decay pathways, quality control pathways exist as well, which are essential to remove aberrant mRNA transcripts (see paragraph 1.1.4.3).

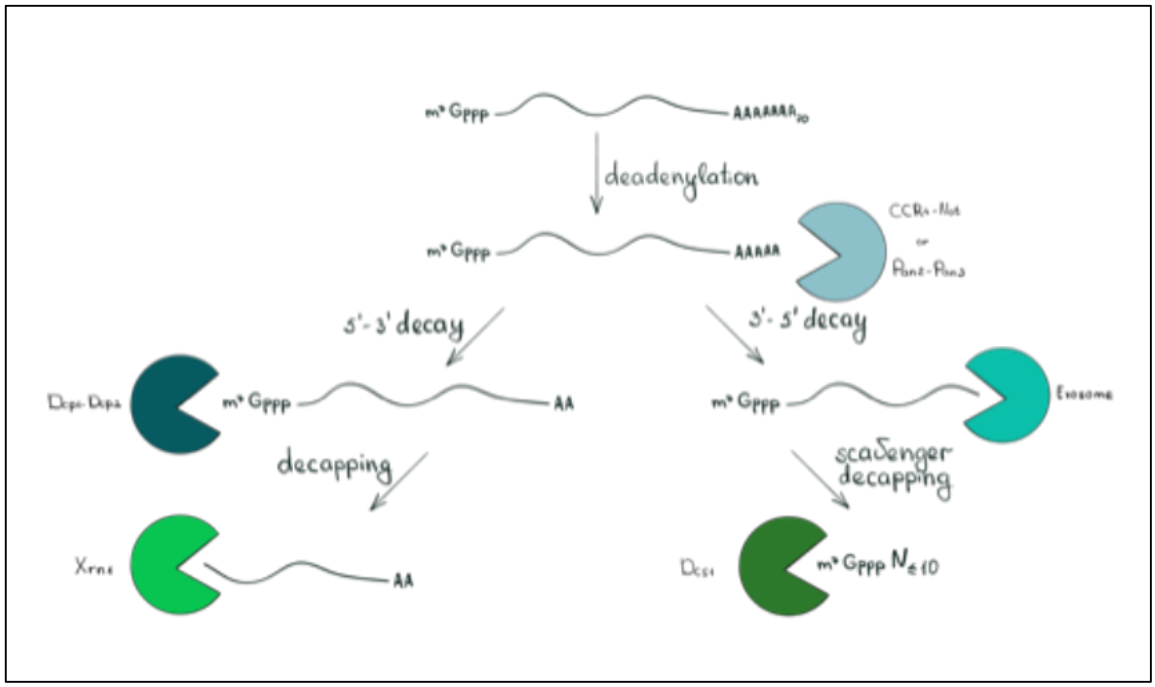


Figure 10. Schematic representation of the two mRNA degradation pathways in the cytoplasm (adapted from (Braun and Young 2014)). Both pathways start with deadenylation of the mRNA by the CCR4-Not and/or Pan2-Pan3 deadenylases. After decapping by the Dcp1/2 decapping complex, the RNA is degraded by Xrn1 in 5' to 3' direction (left panel). The exosome directly degrades the deadenylated RNA in 3' to 5' direction (right panel) and the remaining 5' cap structure is removed by the DcpS protein.

1.1.4.1 Deadenylation-dependent 3′to 5′mRNA decay

When mRNAs enter the cytoplasm, their poly(A) tails are either stabilized by the binding of PABPs for translation or shortened by exonucleases. Shortening of the poly(A) tail, so-called deadenylation, is a first step triggering mRNA degradation. This process is performed by two different 3′ to 5′ exonucleases, Pan2-Pan3 and CCR4-NOT (Wahle and Winkler 2013). However, there are several enzymes that are capable of trimming the poly(A) tail at different stages in the pathway, which shows the dynamism of poly(A) tail length control in regulation of mRNA stability (Goldstrohm and Wickens 2008). By being the first step in the mRNA degradation pathway, deadenylation can be the bottleneck for mRNA decay speed. Therefore, mRNA degradation in 3′ to 5′ direction occurs in a deadenylation-dependent manner. The human CCR4-NOT deadenylase consists of 10 subunits (Lau, Kolkman et al. 2009), of which two subunits are associated with catalytic activity (CCR4: CNOT6 and Caf1: CNOT7/8; (Doidge, Mittal et al. 2012). Enzymatically active Pan2 associates with two Pan3 subunits to form an heterotrimer (Jonas, Christie et al. 2014, Schafer, Rode et al. 2014, Wolf, Valkov et al. 2014, Schafer, Yamashita et al. 2019). According to the current model, deadenylation of mammalian mRNAs is initiated by Pan2-Pan3 in a slow distributive manner until the poly(A) tail is shortened to around 110 nucleotides. When the CCR4-NOT complex takes over,

deadenylation progresses until the poly(A) tail reaches a length of around 10 adenines (Yamashita, Chang et al. 2005, Chen and Shyu 2011).
Directly after deadenylation, attached PABPs are released from the mRNA, which can then be attacked at the 3′-end by the major eukaryotic exoribonucleases, the RNA exosome complex (Braun and Young 2014, Siwaszek, Ukleja et al. 2014). The exosome is a multi-subunit complex consisting of 9 subunits forming the exosome core (Exo9) associating with a ribonuclease subunit (Exo10). The catalytic subunit is called hRrp44 and has two orthologues depending on the subcellular localization, DIS3 and DIS3L (Tomecki et al., 2010). The exosome is responsible for 3′ to 5′ degradation of mRNAs and processing and quality control of almost all RNA species in the nucleus and cytoplasm (Januszyk and Lima, 2014). *In vivo* activity and substrate specificity of the exosome relies on presence of different co-factors like the Ski-complex (Araki, Takahashi et al. 2001, Halbach, Reichelt et al. 2013).
The cap of the remaining RNA oligonucleotide is removed in a so-called salvage pathway by the scavenger enzyme DcpS (Figure 10). DcpS is a pyrophosphatase specific for the m^7G cap structure and can directly hydrolyze capped mRNA substrates within a length of 10 nucleotides (Milac, Bojarska et al. 2014, Labno, Tomecki et al. 2016). Besides that, it can be involved in maintaining cap structure concentrations in the process of mRNA splicing (Shen, Liu et al. 2008).

1.1.4.2 Xrn1-mediated 5′ to 3′ mRNA decay

The 5′ to 3′ mRNA degradation pathway is a multistep process and plays important roles in mRNA quality control and cell growth (Chen, Xu et al. 1995, Andersen, Jensen et al. 2013, Lykke-Andersen and Jensen 2015). This degradation pathway is initiated by deadenylation of the 3′-poly(A) tail, directly followed by removal of the 5′-cap (m^7G cap), a process which is called decapping (Moore 2005, Parker 2012). Decapping is performed by very the conserved Dcp2 decapping enzyme, belonging to the Nudix hydrolase family of proteins (Dunckley and Parker 1999, Piccirillo, Khanna et al. 2003, Li and Kiledjian 2010). Dcp2 is bridged to its direct activator Dcp1 by the Edc4 protein (Gavin, Bosche et al. 2002, Chang, Bercovich et al. 2014). According to previous studies, different decapping co-activators (e.g. Edc3) interact in a mutually exclusive manner, suggesting regulation of different mRNAs (Badis, Saveanu et al. 2004, He, Li et al. 2014, He, Celik et al. 2018). Based on this complex protein-protein interaction network, it is suggested that formation of the decapping machinery is initiated by recruiting decapping factors directly by the 3′-deadenylation complex (Mugridge, Coller et al. 2018). After removal of the 5′-cap, the 5′-monophosphorylated RNA can be attacked by the conserved exonuclease Xrn1 (Nagarajan, Jones et al. 2013) in a processive manner (Figure 10). By interacting directly with the decapping complex via Dcp1 and Edc4, Xrn1 directly

connects the decapping process to 5′ to 3′ degradation (Nissan, Rajyaguru et al. 2010, Braun, Truffault et al. 2012, Jonas and Izaurralde 2013).

1.1.4.3 mRNA surveillance pathways

Besides canonical mRNA degradation pathways mentioned in the text above, several so-called mRNA surveillance pathways exist, to maintain the quality of mRNA transcripts in eukaryotic cells. Without quality control mechanisms, translation of aberrant mRNA transcripts would lead to synthesis of potentially deleterious or toxic proteins. There are three co-translationally occurring mRNA surveillance pathways evolved in eukaryotic cells, to deal with aberrantly transcribed mRNAs (Shoemaker and Green 2012, Simms, Thomas et al. 2017): non-stop decay (NSD), no-go decay (NGD) and non-sense-mediated decay (NMD). NMD is a very well-studied pathway of mRNA quality control, dealing with mRNA transcripts containing pre-mature stop codons (PTC) (Wittkopp, Huntzinger et al. 2009). Thereby, NMD inhibits translation of PTC-containing mRNAs into C-terminally truncated proteins by recognizing PTCs on mRNAs via the Exon Junction Complex (EJC) and recruitment of various proteins to initiate decay of the mRNA (Nagy and Maquat 1998, Palacios, Gatfield et al. 2004). The second mRNA surveillance pathway, NGD, acts on mRNA transcripts harboring pro-longed ribosome stalling during translation elongation stalling elements (Clement and Lykke-Andersen 2006, Doma and Parker 2006). The NGD pathway targets these mRNAs for endonucleolytic cleavage followed by RNA degradation via the exosome and Xrn1 (Doma and Parker 2006, Passos, Doma et al. 2009). The last mRNA surveillance pathway, NSG, deals with mRNA transcripts lacking a stop codon (Frischmeyer, van Hoof et al. 2002, Karamyshev and Karamysheva 2018). Ribosomes are stalled at the 3′-end of RNAs lacking stop-codons, which triggers endonucleolytic cleavage of the mRNA and subsequent degradation (Tsuboi, Kuroha et al. 2012, Karamyshev and Karamysheva 2018).

1.2 The human cleavage and polyadenylation machinery

More than 80 proteins could be co-purified with the human core 3′-end processing machinery in previous studies (Zhao, Hyman et al. 1999, Mandel, Bai et al. 2008, Shi, Di Giammartino et al. 2009, Shi and Manley 2015), forming a huge protein complex required for regulation of 3′-end formation and proper definition of the cleavage site (Proudfoot and O'Sullivan 2002, Lutz 2008, Millevoi and Vagner 2010). The core of the human 3′-end processing machinery was initially thought to consist of five major protein complexes, which were identified in early biochemical studies (Christofori and Keller 1988, Gilmartin and Nevins 1989, Takagaki, Ryner et al. 1989). In later studies, the list of core components was extended, so that the human core 3′-end processing machinery (see table 1) now includes RNA polymerase II (RNA pol II), poly(A) polymerase (PAP), poly(A) binding proteins (PABPs) and four protein complexes consisting of several subunits (Figure 11): Cleavage and Polyadenylation Specificity Factor (CPSF), Cleavage Stimulation Factor (CstF) and Cleavage Factors I_m and II_m (CF I_m and CF II_m) (Colgan and Manley 1997, Mandel, Bai et al. 2008). Those factors are needed to deliver endonuclease, poly(A) polymerase and protein phosphatase activities to cover the main steps in 3′-end processing, including recognition of the cleavage site, the cleavage step itself, addition of the poly(A) tail and the connection to transcription termination. Based on enzymatic functions and on recent structures of parts of the yeast and human 3′-end processing machinery (Casanal, Kumar et al. 2017, Clerici, Faini et al. 2017, Clerici, Faini et al. 2018, Sun, Zhang et al. 2018, Hill, Boreikaite et al. 2019, Sun, Zhang et al. 2020, Zhang, Sun et al. 2020), human CPSF can be divided in two modules (Table 1): polymerase module (CPSF1, WDR33, hFip1, CPSF4) and nuclease module (CPSF2, CPSF3, Symplekin). The third module, the so-called phosphatase module, contains two additional proteins, PP1A and SSU72, which are not part of any of the big complexes (Mandel, Kaneko et al. 2006, Sullivan, Steiniger et al. 2009, Schonemann, Kuhn et al. 2014, Kumar, Clerici et al. 2019). In contrast to yeast poly(A) polymerase Pap1, human PAP is not stably associated to the polymerase module.

Table 1: Overview of the human and yeast 3'-end processing machinery (adopted from Kumar et al., 2019).

Human					Yeast		
complex	protein	alt. name	subcomplex	module	complex	protein	module
CPSF	CPSF1	CPSF160	mPSF	polymerase	*CPF*	Cft1	polymerase
	WDR33		mPSF			Pfs2	polymerase
	CPSF4	CPSF30	mPSF			Yth1	polymerase
	hFip1		mPSF			Fip1	polymerase
	PAP		poly(A) polymerase			Pap1	poly(A) polymerase
CPSF	CPSF2	CPSF100	mCF	nuclease		Cft2	nuclease
	CPSF3	CPSF73	mCF; endonuclease			Ysh1	nuclease; endonuclease
	Symplekin		mCF			Pta1	phosphatase
	RBBP6					Mpe1	nuclease
	PP1A			phosphatase		Glc7	phosphatase
	Ssu72					Ssu72	
CstF	CstF1	CstF50			*CFIA*	/	
	CstF2	CstF64				Rna15	polymerase
	CstF3	CstF77				Rna14	
CF II_m	hPcf11					Pcf11	
	hClp1					Clp1	
CF I_m	CFI25	CPSF5			*CFIB*	Hrp1	
	CFI59	CPSF7					
	CFI68	CPSF6					

CPSF, PAP and PABP are required for polyadenylation, whereas CPSF, CstF, CF I_m and II_m and PAP are required for an effective cleavage reaction (Wahle and Ruegsegger 1999, Zhao, Hyman et al. 1999). Although there are differences in the poly(A) signal sequences on the RNA transcript among different species, most mammalian pre-mRNA 3′-end processing factors have homologues in other species, which is an indication for the conservation of the 3′-end processing steps (Chan, Choi et al. 2011).

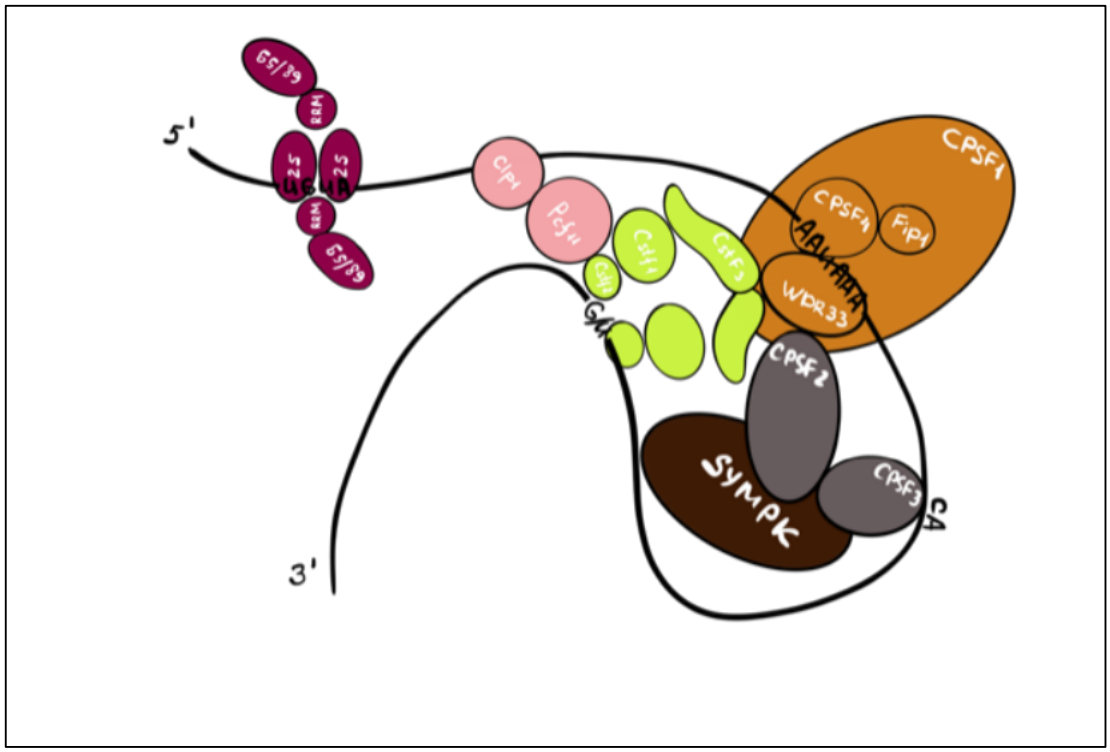


Figure 11. Cartoon of protein complexes of the human 3′-end cleavage and polyadenylation machinery. Proteins of the human 3'-end processing machinery are assembled on a pre-mRNA target containing three sequence elements (polyA signals; PAS). UGUA: upstream element – AAUAAA: conserved hexamer – G/U: GU-rich downstream element. The cleavage site is depicted by the conserved CA dinucleotide. Purple: CF I_m – rose: CF II_m – orange: mPSF – grey: mCF – brown: Symplekin – green: CstF

1.2.1 Sequence elements of mRNA involved in cleavage site definition

Assembly of the human 3′-end processing machinery and definition of the cleavage site relies on multiple protein – RNA interactions of components of the 3′-end processing machinery with distinct sequence elements on the pre-mRNA. This so called *cis*-elements or poly(A) signals (PAS) are either located upstream or downstream of the cleavage site (poly(A) site) (Colgan and Manley 1997, Zhao, Hyman et al. 1999, Shi, Di Giammartino et al. 2009, Millevoi and Vagner 2010). The human cleavage and polyadenylation machinery has to be accurately positioned on the mRNA to define the cleavage site, which has no consensus sequence but is often characterized by a CA dinucleotide (Sheets, Ogg et al. 1990). Correct positioning of the 3′-end processing factors is mediated via a tripartite mechanism, where three consensus *cis*-elements on the pre-mRNA are bound by different complexes of the human 3′-end processing machinery. These three *cis*-elements (Zhao, Hyman et al. 1999) include the very conserved hexameric AAUAAA, which is located 10-30 nucleotides (nt) upstream of the cleavage site (poly(A) site) and a less conserved G/U-rich sequence element 15-30 nt downstream of the cleavage site, which are together sufficient to determine the location of the cleavage site (Chen, MacDonald et al. 1995). Additionally, multiple UGUA motifs located 40-100 nt upstream

of the cleavage site (upstream element; USE), and various auxiliary downstream elements (auxDSE) have been identified to fine tune the 3′-processing machinery by providing binding sites for regulatory factors (Hu, Lutz et al. 2005).

Almost 70% of human poly(A) signals contain the hexameric AAUAAA motif or a very close variant, AUUAAA (15%), which was shown by bioinformatically investigating around 14 000 human genes (Tian, Hu et al. 2005). Those PAS are called canonical, whereas non-canonical PAS are lacking the hexameric AAUAAA motif. It is not clear yet, if non-canonical PAS are bound by the same 3′-end processing factors as canonical PAS (Colgan and Manley 1997, Zhao, Hyman et al. 1999, Zarudnaya, Kolomiets et al. 2003, Tian, Hu et al. 2005, Dickson and Wilusz 2010, Millevoi and Vagner 2010). As shown by mutational studies, the AAUAAA hexamer is essential for both, cleavage reaction and polyadenylation, (Wells and Kedes 1985, Connelly and Manley 1988, Wahle and Keller 1992) and is one of the most conserved sequence motifs known so far (Proudfoot 1991). The distance (10-30 nucleotides) between the hexamer and the cleavage site is critical (Fitzgerald and Shenk 1981, Chen, MacDonald et al. 1995, Beaudoing, Freier et al. 2000, Hu, Lutz et al. 2005), since the AAUAAA hexamer is bound by CPSF4 and WDR33 (Schonemann, Kuhn et al. 2014) and at the same time, the cleavage site has to be contacted by endonuclease CPSF3 (Mandel, Kaneko et al. 2006). Generally, the AAUAAA hexamer is located 13 nucleotides upstream of the cleavage site (Chen, MacDonald et al. 1995).

The second important *cis*-element is located downstream of the cleavage site and therefore is called downstream element (DSE). It is less conserved than the AAUAAA hexamer and characterized by G/U- or U-rich sequence elements, bound by the CstF complex (MacDonald, Wilusz et al. 1994, Beyer, Dandekar et al. 1997). Downstream elements were proposed to consist of spaced sequence elements, a proximal G/U-rich sequence and distal U-rich sequence element (McDevitt, Hart et al. 1986, Gil and Proudfoot 1987, Zarudnaya, Kolomiets et al. 2003, Salisbury, Hutchison et al. 2006). However, no consensus sequence is known yet, but proximity of the DSE within 30 nucleotides to the cleavage site impacts cleavage site selection and also cleavage efficiency (Mason, Elkington et al. 1986, McDevitt, Hart et al. 1986, Gil and Proudfoot 1987, MacDonald, Wilusz et al. 1994, Takagaki and Manley 1997).

The last *cis*-element is located 40-100 nucleotides upstream the cleavage site (USE) and is generally U-rich (Hu, Lutz et al. 2005). Cleavage Factor I_m was shown to bind two UGUA sequence elements within one mRNA simultaneously, which can impact poly(A) site selection and 3′-end processing efficiency as shown by mutational analysis (Venkataraman, Brown et al. 2005, Yang, Gilmartin et al. 2010). In case of non-canonical PAS, UGUA sequence elements can play a role in AAUAAA - independent 3′-end processing (Venkataraman, Brown et al. 2005, Yang, Gilmartin et al. 2010).

The poly(A) site itself is determined by location of the hexameric AAUAAA motif and the DSE (Chen, MacDonald et al. 1995). Although the surrounding sequence is not conserved, the cleavage site is characterized by a CA dinucleotide in 60 % of the genes analyzed (Sheets, Ogg et al. 1990). Protein complexes, which are necessary for cleavage site definition by binding to different poly(A) signals, are discussed in following sections (paragraph 1.2.2-1.2.8).

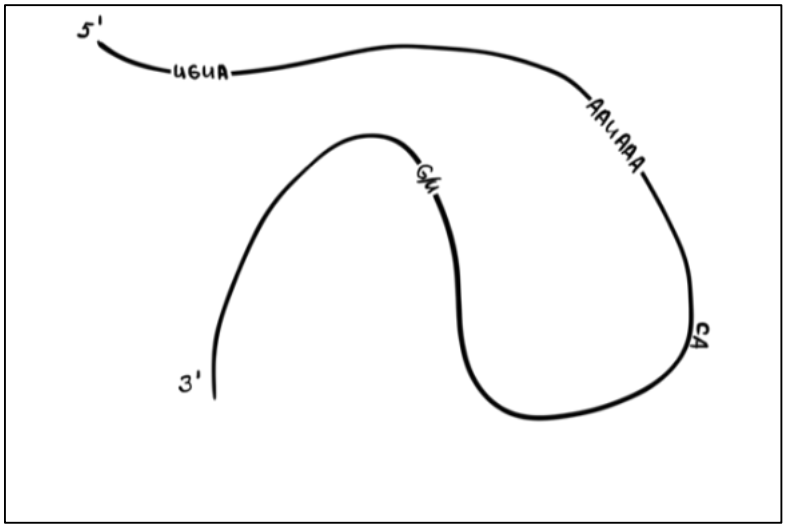


Figure 12. Poly(A) signals on pre-mRNA defining the cleavage site. A set of three distinct *cis*-elements on the pre-mRNA, also called poly(A) signals (PAS) define location of the cleavage site (poly(A) site) CA. UGUA: Upstream element (USE) located 40-100 nucleotides upstream the cleavage site. AAUAAA: conserved hexameric PAS located within 10-30 nucleotides upstream the cleavage site. G/U: G/U-rich downstream sequence element (DSE) located within 30 nucleotides downstream the cleavage site.

1.2.2 The Cleavage and Polyadenylation Specificity Factor CPSF

CPSF is a multi-subunit protein complex, that is very important for poly(A) site definition by binding to the AAUAAA hexameric poly(A) signal (Bienroth, Wahle et al. 1991, Chan, Huppertz et al. 2014, Schonemann, Kuhn et al. 2014). The CPSF complex is required for both, endonucleolytic cleavage (Ryan, Calvo et al. 2004, Mandel, Kaneko et al. 2006) and addition of the poly(A) tail to the mRNA. Besides that, it provides an anchor for other 3′-end processing components (Barabino, Hubner et al. 1997, Colgan and Manley 1997, Zhao, Hyman et al. 1999, Mandel, Bai et al. 2008, Shi, Di Giammartino et al. 2009, Sullivan, Steiniger et al. 2009). First, it has been assumed that purified CPSF consists of CPSF1 (CPSF160), CPSF2 (CPSF100), CPSF3 (CPSF73) and CPSF4 (CPSF30) (Bienroth, Wahle et al. 1991, Murthy and Manley 1992), but in later preparation WDR33, hFip1 and Symplekin were identified as components of the CPSF complex as well (Takagaki and Manley 2000, Kaufmann, Martin et al. 2004, Shi, Di Giammartino et al. 2009). Within the last years, it was shown that CPSF forms two subcomplexes, mammalian Polyadenylation Specificity Factor (mPSF) and mammalian

Cleavage Factor (mCF) (Bienroth, Wahle et al. 1991, Chan, Huppertz et al. 2014, Schonemann, Kuhn et al. 2014). mPSF consists of CPSF1, WDR33, CPSF4 and hFip1 and mediates binding to the AAUAAA poly(A) signal via subunits WDR33 and CPSF4 (Bienroth, Wahle et al. 1991, Chan, Huppertz et al. 2014, Schonemann, Kuhn et al. 2014). In parallel, it is necessary for polyadenylation, because hFip1 is involved in recruiting PAP to start addition of the poly(A) tail (Kaufmann, Martin et al. 2004, Meinke, Ezeokonkwo et al. 2008). mCF containing CPSF2, CPSF3 and Symplekin (SYMPK), catalyzes cleavage of the RNA, because CPSF3 was identified to be the endonuclease responsible for the cleavage reaction (Mandel, Kaneko et al. 2006).

hCPSF1 (CPSF160) is the largest subunit of the CPSF complex and shows a high conservation among eukaryotes. It is a 1443 amino acid long protein and consists of tandem WD40 repeats, which are organized in three major propellers BPA, BPB and BPC (Neuwald and Poleksic 2000, Clerici, Faini et al. 2017, Clerici, Faini et al. 2018, Sun, Zhang et al. 2018). It functions as a large scaffold protein to position WDR33 and CPSF4, which are the RNA binding components in the CPSF complex (Clerici, Faini et al. 2018, Sun, Zhang et al. 2018). The last residues (1352-1443) are organized into a C-terminal domain (CTD). The common WD40 domain often serves as scaffold (Stirnimann, Petsalaki et al.), but can also interact with nucleic acid (Scrima, Konickova et al. 2008). This fits to the fact that CPSF1 is involved in both: protein-protein interactions and also protein-RNA interactions. Based on several lines of evidence, it was believed that CPSF1 was the major subunit interacting with the AAUAAA sequence element (Moore, Chen et al. 1988, Gilmartin and Nevins 1989, Keller, Bienroth et al. 1991, Murthy and Manley 1995). However, studies within the last years showed, that the CPSF complex lacking WDR33 was not able to recognize the AAUAAA element. WDR33 was identified to be an, in early studies detected, 160kDa band UV-cross-linked to AAUAAA containing RNA (Moore, Chen et al. 1988, Chan, Huppertz et al. 2014, Schonemann, Kuhn et al. 2014). Since the CPSF1 subunit recognizes sequence elements close to the AAUAAA hexamer (Martin, Gruber et al. , Bilger, Fox et al. 1994, Gilmartin, Fleming et al. 1995), it is assumed, that CPSF1 can be involved in upstream interactions of the CPSF complex with mRNA (Schonemann, Kuhn et al. 2014).

hCPSF3 (CPSF73) gained attention when it was identified to be the endonuclease performing the cleavage reaction of pre-mRNAs (Mandel, Kaneko et al. 2006). Several evidences were collected over years, indicating that CPSF3 possesses nuclease activity (Ryan, Calvo et al. 2004). Since it contains a β-CASP domain inserted into the N-terminal metallo-β-lactamase (MBL) domain (Figure 13 B), it is a member of the β-CASP subfamily and the MBL superfamily of proteins, which are mostly metal-dependent nucleases (Callebaut, Moshous et al. 2002, Dominski 2007). Besides that, mutational studies in yeast showed, that yeast cells carrying mutations in the zinc binding region of the CPSF3 homologue Ysh1 are

lethal (Ryan, Calvo et al. 2004). By solving the crystal structure of the N-terminus of CPSF3, evidence for its endonuclease activity has been proved (Ryan, Calvo et al. 2004, Mandel, Kaneko et al. 2006). The MBL domain (residues 1-460) is the catalytical domain for the zinc-dependent endonucleolytic cleavage reaction (Mandel, Kaneko et al. 2006, Bebrone 2007) and is split into two parts by the central β-CASP domain (residues 209-394), which is organized in a $\alpha/\beta/\alpha$ domain structure (Figure 13 B). This characteristic structure was observed in other members of the β-CASP family like CPSF2 and RNaseJ as well (Dominski 2007, Li de la Sierra-Gallay, Zig et al. 2008, Mandel, Bai et al. 2008). The catalytic MBL domain is organized in a four-layer sandwich $\alpha/\beta/\beta/\alpha$ and provides the active site for zinc binding, which is required for nuclease activity of CPSF3 (Mandel, Kaneko et al. 2006, Bebrone 2007). The active site is sandwiched between the MBL domain and the β-CASP domain and contains two zinc ions in the crystal structure (Mandel, Kaneko et al. 2006, Bebrone 2007). Biochemical studies of bacterial expressed CPSF3 N-terminal domain (NTD) show weak ribonuclease activity without other members of the 3′-end processing machinery (Mandel, Kaneko et al. 2006, Bebrone 2007). Endonucleolytic activity of CPSF3 is not very sequence specific, demonstrated by the fact that CPSF2 (CPSF100) and CstF2 (CstF64) help to find the exact cleavage site on pre-mRNA substrates (Mandel, Bai et al. 2008).

Since it is also a member of the β-CASP family, hCPSF2 (CPSF100) has a high conservation towards CPSF3, but is lacking one out of the six residues building the zinc binding motif (Mandel, Kaneko et al. 2006, Kolev, Yario et al. 2008). Consequently, CPSF2 is expected to bind none or only one zinc ion (Kolev, Yario et al. 2008) and therefore possesses none or only weak catalytical activity (Aravind 1999, Callebaut, Moshous et al. 2002). It was shown that members of the MBL protein family remained weakly active, when only one of the two required zinc ions is bound (Bebrone 2007). The exact function of CPSF2 still remains unclear, but it shows a tight association with CPSF3 and both subunits are present in various 3′-end processing machineries (Kyburz, Sadowski et al. 2003, Xu, Zhao et al. 2006, Sullivan, Steiniger et al. 2009). Their strong association is mediated specifically through their CTDs (Dominski, Yang et al. 2005), making this heterodimer comparable to other β-CASP proteins. Additionally, this tight interaction provides a possible mechanism, which requires CPSF2-CPSF3 dimerization for catalysis (Dominski 2007). As already mentioned, the NTD of CPSF3 alone has only weak nuclease activity *in vitro* (Mandel, Kaneko et al. 2006), indicating that the C-termini of CPSF3 and CPSF2 are important for CPSF3 exonuclease activity in 3′-end processing of histone pre-mRNA (Yang, Sullivan et al. 2009). Recent studies identified a short segment in CPSF2 mediating interaction to the mPSF complex and therefore being called PSF interaction motif (PIM). Mutational studies revealed, that this motif is necessary for CPSF formation (Zhang, Sun et al. 2020).

hCPSF4 (CPSF30) is the smallest subunit of CPSF and is required for cleavage and polyadenylation (Barabino, Hubner et al. 1997, Barabino, Ohnacker et al. 2000). It has five CCCH zinc finger (ZF) motifs and one CCHC zinc knuckle motif at the C-terminus, which is missing in the yeast homolog Yth1 (Barabino, Hubner et al. 1997). These structures often function in RNA recognition (D'Souza and Summers 2004, Hudson, Martinez-Yamout et al. 2004), suggesting that CPSF4 might bind RNA. This is supported by the fact that CPSF4 can be UV cross-linked to polyU-rich RNA stretches (Barabino, Hubner et al. 1997), which are often located near PAS (Barabino, Ohnacker et al. 2000, Hu, Lutz et al. 2005). Yeast homolog Yth1 also binds mRNA close to the poly(A) site via its zinc fingers, because upon deletion of the zinc fingers, RNA binding and 3′-end processing is decreased (Tacahashi, Helmling et al. 2003). Recent studies identified that RNA binding is directly mediated via ZF2 and ZF3 of CPSF4 (Clerici, Faini et al. 2018, Sun, Zhang et al. 2018). Via its zinc fingers, CPSF4 also binds other proteins in cleavage and polyadenylation, e.g. hFip1, CPSF1 and PAPB (Barabino, Hubner et al. 1997, Chen, Li et al. 1999, Barabino, Ohnacker et al. 2000, Tacahashi, Helmling et al. 2003). CPSF4 was also identified to be involved in poly(A)-dependent transcription pausing, by interacting with RNA pol II (Nag, Narsinh et al. 2007).

hFip1 (Factor interacting with Pap1p) was identified later than other CPSF subunits by sequence analysis of a HeLa cell cDNA library (Kaufmann, Martin et al. 2004). Its yeast homologue was identified earlier in a yeast two-hybrid screen screening binding partners of yeast Poly(A) Polymerase (Pap1) (Preker, Lingner et al. 1995). In the crystal structure of a peptide of yeast Fip1 interacting with Pap1, it was found that amino acids 80-105 of Fip1 are required for interaction with the CTD of Pap1. This interaction impacts structural arrangement of Pap1, but not polymerase activity (Meinke, Ezeokonkwo et al. 2008). hFip1 is required for the cleavage and polyadenylation step and shows high flexibility (Kaufmann, Martin et al. 2004). It contains an acidic segment near the N-terminus (Figure 13 A), which is responsible for binding to human PAP, followed by a highly conserved 70 residue long part, which mediates binding to CPSF4, and a proline-rich region (Kaufmann, Martin et al. 2004). Besides binding of CPSF1, the C-terminal RD- and R-rich domains might also interact with U-rich RNA (Kaufmann, Martin et al. 2004).

WDR33 is a 146 kDa protein and was identified to belong to the 3′-end processing components by co-eluting with CPSF during gelfiltration. It seems to be necessary for the cleavage reaction *in vitro* (Shi, Di Giammartino et al. 2009). WDR33 consists of an N-terminal WD40 domain, a collagen-like domain in the middle and a C-terminal GPR (glycine-proline-arginine) domain with unknown function (Ito, Sakai et al. 2001). Together with hFip1, CPSF1 and CPSF4, it forms the minimum core complex mPSF, active in both, recognition of the AAUAAA sequence element and polyadenylation (Schonemann, Kuhn et al. 2014). In this

complex, WDR33 turned out to be the component binding to the AAUAAA polyadenylation signal (Schonemann, Kuhn et al. 2014).

Symplekin (SYMPK) was identified in early studies to be a tight junction plaque protein (Keon, Schafer et al. 1996), but showed sequence similarity to the yeast Pta1 protein (Preker, Ohnacker et al. 1997, Zhao, Hyman et al. 1999, Takagaki and Manley 2000, Ghazy, He et al. 2009). Pta1 is part of the yeast polyadenylation machinery by serving as a scaffold for various protein – protein interactions (He, Khan et al. 2003, Kyburz, Sadowski et al. 2003, Zhelkovsky, Tacahashi et al. 2006, Ghazy, He et al. 2009). The N-terminal domain of Symplekin is folded into α-helices, which are arranged pairwise in an antiparallel manner (Kennedy, Frazier et al. 2009, Xiang, Nagaike et al. 2010). This fold is characteristic for HEAT repeats, that are usually involved in protein-protein interactions (Andrade, Petosa et al. 2001). Previous studies showed, that the NTD of Symplekin is interacting with Ssu72, which is a phosphatase essential for transcription termination, and mediates transition between RNA pol II initiation and elongation (Ganem, Devaux et al. 2003, Rosado-Lugo and Hampsey 2014). Therefore, the Symplekin-Ssu72 complex is involved in transcription coupled polyadenylation (Ghazy, He et al. 2009, Xiang, Nagaike et al. 2010). Besides that, Symplekin was also shown to bind to the CstF2 hinge region in a mutually exclusive manner with CstF3 (Takagaki and Manley 2000, Ruepp, Schweingruber et al. 2010, Ruepp, Schweingruber et al. 2011). Mutational analysis of CstF2 could abolish interaction with Symplekin, while CstF3 binding was not affected. The same mutations had direct impact on 3′-end processing of histone pre-mRNAs (Takagaki and Manley 2000, Ruepp, Schweingruber et al. 2010, Ruepp, Schweingruber et al. 2011), indicating that Symplekin might be involved in different 3′-end processing pathways by being part of different subcomplexes. Indications that Symplekin also binds to CPSF2 and CPSF3 (Hofmann, Schnolzer et al. 2002, Zhelkovsky, Tacahashi et al. 2006, Ghazy, He et al. 2009, Sullivan, Steiniger et al. 2009) and thereby is part of the so called mCF complex, were confirmed by recent studies, where the mCF structure was solved by cryo-EM, containing the CTDs of CPSF2, CPSF3 and a part of Symplekin (Zhang, Sun et al. 2020).

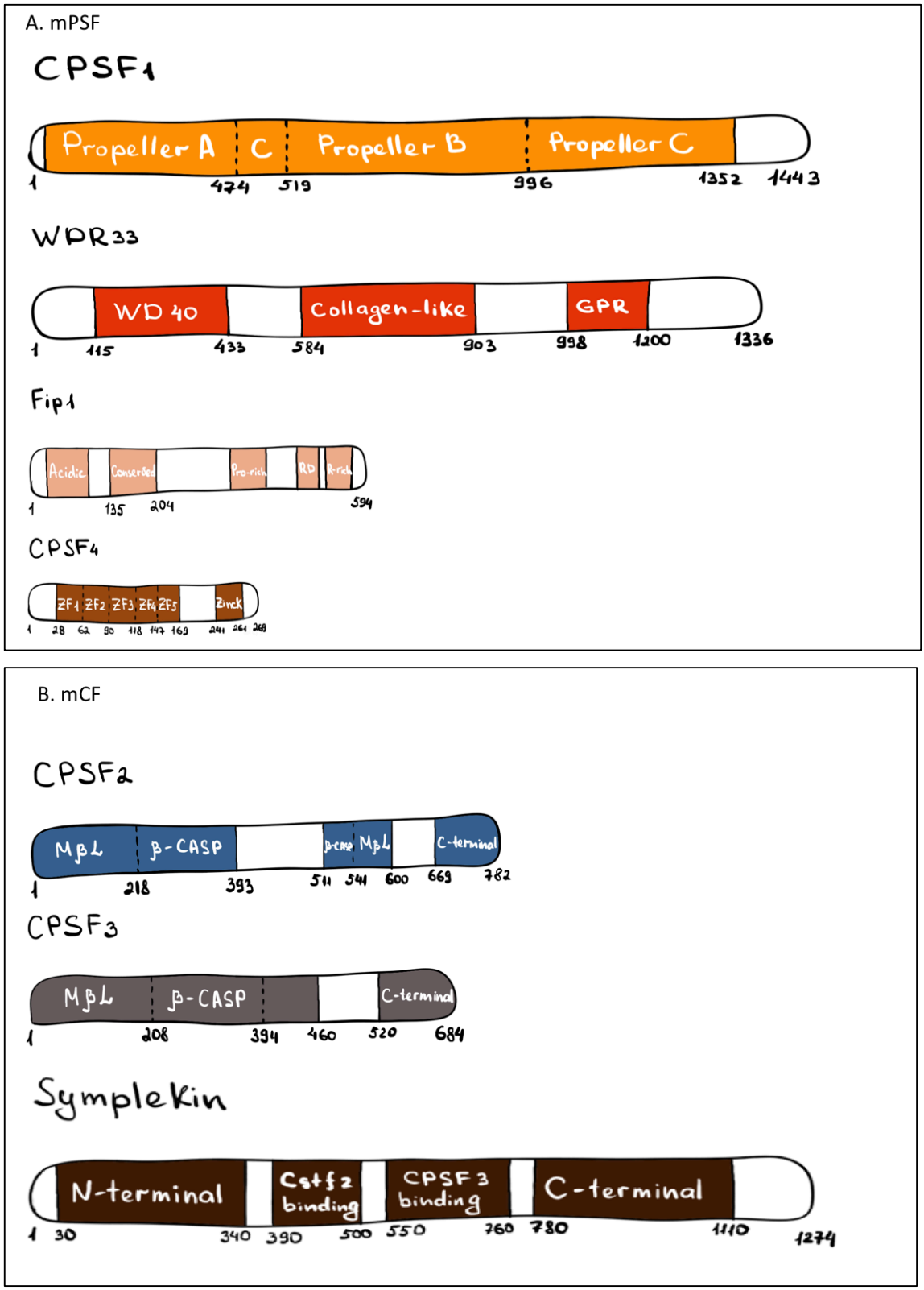


Figure 13. Domain organization of the human CPSF complex consisting of mPSF and mCF subcomplexes. A) The mPSF subcomplex consisting of CPSF1. WDR33, Fip1 and CPSF4. CPSF1 consists of tandem WD40 repeats that are organized in three propellers (A-C). WDR33 consists of an N-terminal WD40 domain, followed by a collagen-like domain and a C-terminal glycine-proline-arginine (GPR) domain. Fip1 has an N-terminal acidic segment, followed by a highly conserved 70 residue long part. It has a proline-rich region in the middle and C-terminal RD- and R-rich domains. CPSF4 has five CCCH zinc finger (ZF) motifs and one C-terminal CCHC zinc

knuckle. B) CPSF3 is the endonuclease and contains a β-CASP domain inserted into the N-terminal metallo-β-lactamase (MBL) domain. CPSF2 is highly conserved to CPSF3.The C-termini of both proteins are fold into a CTD. Symplekin has a N-terminal domain folded into α-helices and a C-terminal domain separated by the binding sites for CstF2 and CstF3.

1.2.3 The Cleavage Stimulation Factor CstF

Cleavage Stimulation Factor (CstF) was identified in early experiments as a factor necessary for cleavage of pre-mRNAs and recognition of *cis*-acting sequence elements on mRNAs (Takagaki, Ryner et al. 1989). The protein complex specifically binds to G/U-rich sequence elements on pre-mRNAs located within 30 nt downstream of the cleavage site (Takagaki and Manley 1997). Interaction with the mRNA occurs in a cooperative manner with the CPSF complex binding to the poly(A) signal AAUAAA (Wilusz, Shenk et al. 1990) et al., 1990), to define the correct location of the cleavage site (Gilmartin and Nevins 1991, MacDonald, Wilusz et al. 1994, Takagaki and Manley 1994, Chen, MacDonald et al. 1995). Although it has an important role in the cleavage reaction, the CstF complex is not required for addition of the poly(A) tail (Takagaki, Manley et al. 1990, Wahle, Lustig et al. 1993).

The CstF complex is also a multi-protein complex consisting of three subunits – CstF1 (CstF50), CstF2 (CstF64) and CstF3 (CstF77). According to previous studies, it is assumed that CstF might function as a heterodimeric complex, assembling two copies of each subunit. (Bai, Auperin et al. 2007, Legrand, Pinaud et al. 2007, Yang, Hsu et al. 2018)

CstF3 (CstF77) is the largest subunit (717 amino acids) of the CstF complex and consists of an N-terminal HAT (half a tetratricopeptide repeat (TPR); residues 1-550, Figure 14) domain with 12 repeat elements (Preker and Keller 1998), which is mostly involved in protein-protein interactions (Lamb, Tugendreich et al. 1995). The HAT domain can be divided into two subdomains, HAT-N and HAT-C (Bai, Auperin et al. 2007, Bai, Auperin et al. 2007, Legrand, Pinaud et al. 2007). The crystal structure of murine CstF3 HAT domain shows, that it strongly homodimerizes in a tail-to-tail manner via the six C-terminal HAT repeats (HAT-C) (Bai, Auperin et al. 2007, Bai, Auperin et al. 2007, Legrand, Pinaud et al. 2007). According to self-association of CstF3, the whole CstF complex might exist in a 2:2:2 stoichiometry (Yang, Hsu et al. 2018). The bow-shaped dimerization interface of the HAT domain is potentially involved in protein – protein interactions: A yeast two-hybrid screen showed, that CstF3 interacts with CPSF1 via the HAT-C dimer (Bai, Auperin et al. 2007, Bai, Auperin et al. 2007), which was confirmed by a recently published cryo-EM structure of CPSF interacting with CstF3 (Zhang, Sun et al. 2020). In this structure, the HAT domain of CstF3 not only contacted CPSF1, but also WDR33. Surprisingly, the interaction mode does not fit to the dimeric structure of the CstF3 HAT domain. The first two loops of one monomer, connecting helices of the HAT repeats 6 to 10, are binding to WDR33, whereas the remaining three loops contact the first propeller

(BPA) of CPSF1. The second HAT monomer binds to the BPA propeller of CPSF1 via the last two HAT repeats (repeats 11 and 12) (Zhang, Sun et al. 2020). The five helices of the HAT-N domain are not involved in dimerization and provide an anchor platform for further protein – protein interactions (Legrand, Pinaud et al. 2007). The C-terminal proline-rich region following the HAT domain (Figure 14) forms a bridge between CstF2 and CstF1, which actually do not interact with each other (Takagaki and Manley 1994, Takagaki and Manley 2000). Therefore, CstF3 is a key element for CstF assembly, since it interacts with both: the hinge region of CstF2 and the WD40 domain of CstF1 (Hatton, Eloranta et al. 2000, Takagaki and Manley 2000, Bai, Auperin et al. 2007, Hockert, Yeh et al. 2010, Yang, Hsu et al. 2018). This interaction is conserved for yeast homologues Rna14 (CstF3 homologue) and Rna15 (CstF2 homologue). The crystal structure of the Rna14 CTD (called monkey tail) and Rna15 hinge domain showed that they interact in the same regions as CstF2 and CstF3, forming a locked conformation. Although the two domains are tightly associated, their position with respect to the HAT domain is highly flexible due to the long linker by which they are connected to the HAT dimer (Legrand, Pinaud et al. 2007, Moreno-Morcillo, Minvielle-Sebastia et al. 2011, Paulson and Tong 2012).

CstF2 (CstF64, 577 amino acids) was one of the first proteins identified as a member of the 3′-end processing machinery, because it could be UV-cross-linked to AAUAAA containing mRNAs (Wilusz and Shenk 1988). Later, it was shown that CstF2 itself binds to G/U-rich sequence elements downstream of the poly(A) site (MacDonald, Wilusz et al. 1994, Beyer, Dandekar et al. 1997, Takagaki and Manley 1997, Deka, Rajan et al. 2005) and that the earlier reported UV-cross linking to AAUAAA containing RNAs was mediated by interaction with the CPSF complex (Wilusz, Shenk et al. 1990, Gilmartin and Nevins 1991). RNA binding of CstF2 is mediated via its conserved RNA recognition motif (RRM) at the N-terminus, which is on its own sufficient for RNA binding and U-rich sequence selection (Takagaki, MacDonald et al. 1992, MacDonald, Wilusz et al. 1994, Beyer, Dandekar et al. 1997, Takagaki and Manley 1997, Perez Canadillas and Varani 2003, Deka, Rajan et al. 2005, Pancevac, Goldstone et al. 2010). The NMR structure of the CstF2 RRM domain identified the surface of the central β-sheet as RNA binding site, which gains its selectivity for G/U-rich sequences by variable contacts outside the active binding site (Perez-Canadillas and Varani, 2003). However, the exact binding mechanism is still unclear. The highly conserved hinge region (residues 111-199) following the RRM (Figure 14), mediates protein-protein interactions with CstF3 and Symplekin in a mutually exclusive manner (Hatton, Eloranta et al. 2000, Takagaki and Manley 2000, Hockert, Yeh et al. 2010, Ruepp, Schweingruber et al. 2010, Ruepp, Schumperli et al. 2011). A long proline-glycine rich region follows the hinge domain, which is interrupted by a helical region consisting of repeated motifs of the pentapeptide MEARA/G. The function of this part of the protein is unknown, since it is not present in the yeast homologue Rna15 (Takagaki, MacDonald et al. 1992, Richardson, McMahon et al. 1999). The last 50 residues (529-577)

build up the highly conserved C-terminal domain (CTD). By forming a 3-helical bundle, this domain has a conformation similar to other proteins involved in protein-protein interactions like Dia1, a cytoskeletal protein, or cyclin-dependent kinase inhibitor p21 (Rose, Weyand et al. 2005, Ye and Patel 2005). The structure of this domain exposes a set of conserved residues, which are essential in the yeast homologue Rna15 for binding to a subunit of yeast CFIA, Pcf11p. Human Pcf11 is part of the human Cleavage Factor II_m (Gross and Moore 2001, Qu, Perez-Canadillas et al. 2007). Based on information about its yeast counterpart, the CTD of human CstF2 is also expected to interact with hPcf11 and transcription co-activator PC4 (Calvo and Manley 2001, Qu, Perez-Canadillas et al. 2007).

CstF1 (CstF50) is the smallest CstF subunit (431 amino acids) and structurally dominated by seven WD40 repeats beginning 80 residues from the N-terminus (Mandel, Bai et al. 2008). WD40 propellers are conserved structure motifs and usually serve as platforms for protein complex formation (Smith, Gaitatzes et al. 1999, Li and Roberts 2001). The CstF1 WD40 repeats (residues 80-431), that fold into a seven-blade-β-propeller, mediate interaction with CstF3 (Takagaki and Manley, 2000, Yang, Hsu et al. 2018). The structure of the N-terminal homodimerization domain of CstF1 was solved in previous studies and is very important for self-association of the protein (Moreno-Morcillo, Minvielle-Sebastia et al. 2011, Takagaki and Manley, 2000). This also promotes heterodimeric arrangement of the CstF complex. The homodimerization domain of a CstF1 monomer is characterized by three helices, which are interacting with the N-terminal dimerization domain of the second monomer via a hydrophobic core (Moreno-Morcillo, Minvielle-Sebastia et al. 2011). Besides homodimerization, the NTD is also crucial for interaction of CstF1 with the C-terminal domain of RNA pol II, thereby linking transcription and 3′-end processing (McCracken, Fong et al. 1997, Proudfoot and O'Sullivan 2002).

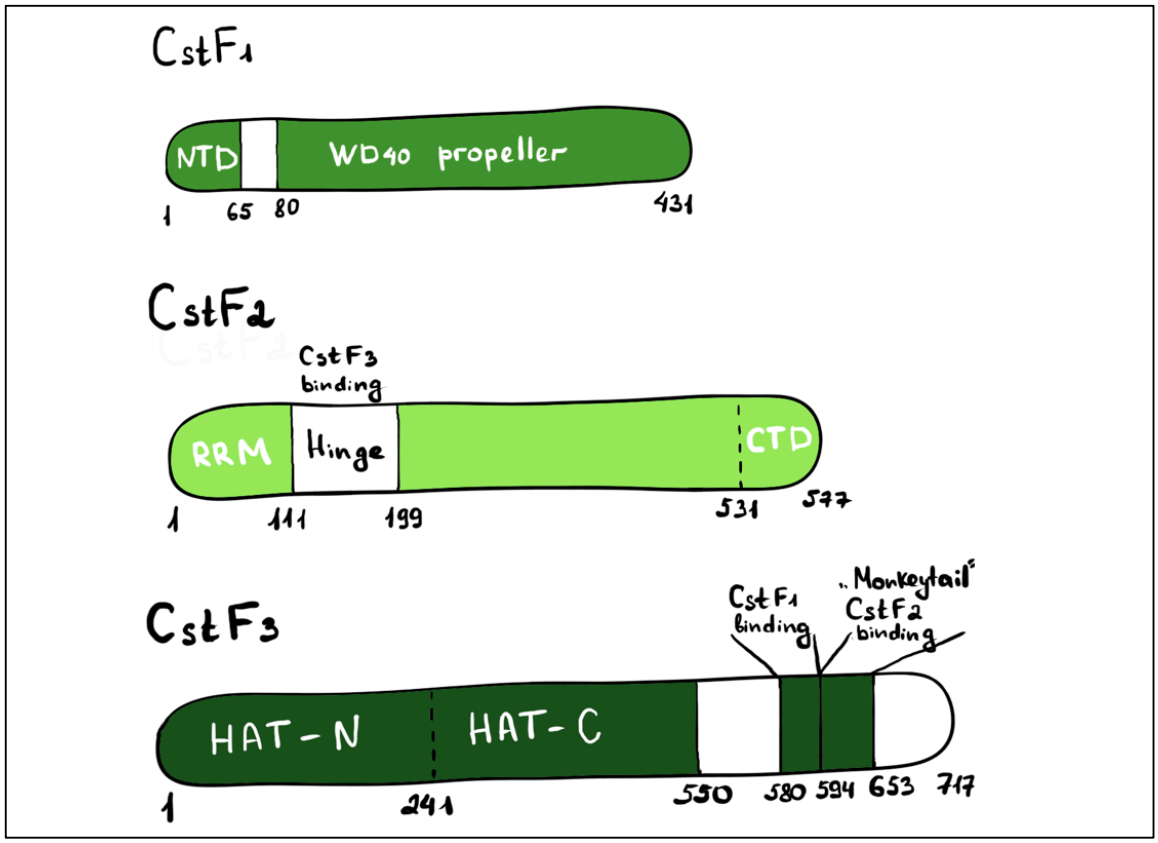


Figure 14. Cartoon of CstF subunits. Depiction of CstF subunits and their domain organization. Top row: CstF1 has a N-terminal homodimerization domain (NTD) followed by a WD40 propeller. Middle row: CstF2 comprises of a N-terminal RNA Recognition Motif (RRM), followed by the hinge domain, which binds to CstF3 and a C-terminal domain (CTD). Bottom row: CstF3 contains a Half a TPR (HAT) domain, divided in the N-terminal (HAT-N) and C-terminal (HAT-C) part, followed by binding region for CstF1 and the monkeytail, which is binding to CstF2.

1.2.4 The Cleavage Factor I CF I_m

Cleavage Factor I_m (CF I_m) was identified early in HeLa nuclear extracts, by co-purifying of a small 25 kDa, a 59 kDa, 68 kDa and a 72 kDa subunit (Ruegsegger, Beyer et al. 1996). Along with CstF and CPSF, CF I_m was shown to be necessary for poly(A) site recognition (Ruegsegger, Beyer et al. 1996, Venkataraman, Brown et al. 2005) and is required only for the cleavage reaction (Colgan and Manley 1997, Zhao, Hyman et al. 1999, Mandel, Bai et al. 2008). Later, it turned out that CF I_m exists as a heterodimer (Coseno, Martin et al. 2008, Yang, Gilmartin et al. 2010), by assembling two copies of the small 25 kDa (CFI25) subunit and two copies of either the 59 kDa (CFI59) or the 68 kDa (CFI68) subunit or a combination of both. CFI59 and CFI68 are encoded by different genes, whereas CFI72 turned out to be a product of alternative splicing of the CFI68 gene (Ruepp, Schumperli et al. 2011). Recombinantly purified CFI25 and CFI68 can be reconstituted to a complex with similar activity in cleavage assays as CF I_m purified from HeLa nuclear extract (Takagaki, Ryner et al. 1989, Ruegsegger, Blank et al. 1998). All CF I_m subunits can be cross-linked to RNA (Ruegsegger, Beyer et al.

1996). Early SELEX experiments showed that CF I_m prefers binding to UGUA motifs, which are located upstream of the cleavage site and were thought to belong to the tripartite *cis*-elements necessary for poly(A) site definition (Brown and Gilmartin 2003, Hu, Lutz et al. 2005, Venkataraman, Brown et al. 2005). CF I_m – RNA interactions via UGUA motif were shown to be mediated by the CFI25 subunit (Brown and Gilmartin 2003, Yang, Gilmartin et al. 2010) and are essential for fine tuning the cleavage reaction (Brown and Gilmartin 2003, Venkataraman, Brown et al. 2005). However, recent studies showed that the UGUA sequence element is not an essential *cis*-element, but rather functions as an enhancer motif for 3'-end processing stimulation (Zhu, Wang et al. 2018). It was observed, that the UGUA sequence element is not required for cleavage and polyadenylation *in vitro,* but that 3′-end processing efficiency is increased by the presence of one or two UGUA elements upstream of the hexameric AAUAAA PAS in a position dependent manner (Zhu, Wang et al. 2018).

CFI25, also known as CPSF5, belongs to the Nudix phosphohydrolase superfamily of proteins (Coseno et al, 2008) and consists of a Nudix (Nucleoside diphosphate linked to some other moiety, x) domain (residues 77-202). The Nudix domain adopts a $\alpha/\beta/\alpha$ fold without any metal ions present, suggesting that CFI25 has no hydrolase activity (Dettwiler, Aringhieri et al. 2004, Coseno, Martin et al. 2008, Tresaugues, Stenmark et al. 2008). Instead, the Nudix domain of CFI25 is modulated to obtain RNA binding capability by two special features: First, the CFI25 Nudix domain lacks two conserved catalytic glutamate residues, so that it can only bind and not hydrolyze dinucleotide substrates (Mildvan, Xia et al. 2005, Coseno, Martin et al. 2008). Second, an alternative binding pocket is formed, which is used to specifically select for UGUA sequence elements on the pre-mRNA (Yang, Gilmartin et al. 2010). The dimeric architecture of CF I_m allows binding of two UGUA sequence elements on one pre-mRNA by providing two copies of CFI25. Besides RNA binding, CFI25 also helps assembling other 3′-end processing factors, since it was shown to interact with PAP (Kim and Lee 2001) and PAPBN1 (Dettwiler, Aringhieri et al. 2004, Mandel, Bai et al. 2008).

CFI68/CFI59 contains a N-terminal RNA recognition motif (RRM), which mediates binding to the 25 kDa subunit (Dettwiler, Aringhieri et al. 2004). The RRM is followed by a proline-rich region and a C-terminal arginine/serine-rich region (RS-domain, Figure 15) (Li, Tong et al. 2011). The crystal structure of CFI68 RRM shows that it adopts a four stranded antiparallel β-sheet, sandwiched between two α-helices (Yang, Coseno et al. 2011). A third α-helix covers the β-sheet, which is the canonical RNA binding surface (Yang, Coseno et al. 2011), thereby explaining why CFI68 only shows weak interaction with RNA (Dettwiler, Aringhieri et al. 2004). RNA binding can be increased by simultaneously interacting with a CFI25 dimer, but has no effect on the specificity for two UGUA sequence motifs (Li, Tong et al. 2011, Yang, Coseno et al. 2011). A crystal structure of the CFI25-CF68RRM-RNA complex, revealed a non-canonical heterotetrameric organization where two CFI68 RRM monomers are

flanking the CFI25 homodimer via a novel identified RRM-protein interaction (Yang, Coseno et al. 2011). Recent studies identified by pulldown assays direct interaction between the RS-domain of CFI68/CFI59 and the C-terminal RE/D region of Fip1 in a phosphorylation dependent manner. Additionally, it was shown that the CFI68 subunit is important for interaction between CF I_m and CPSF (Zhu, Wang et al. 2018).

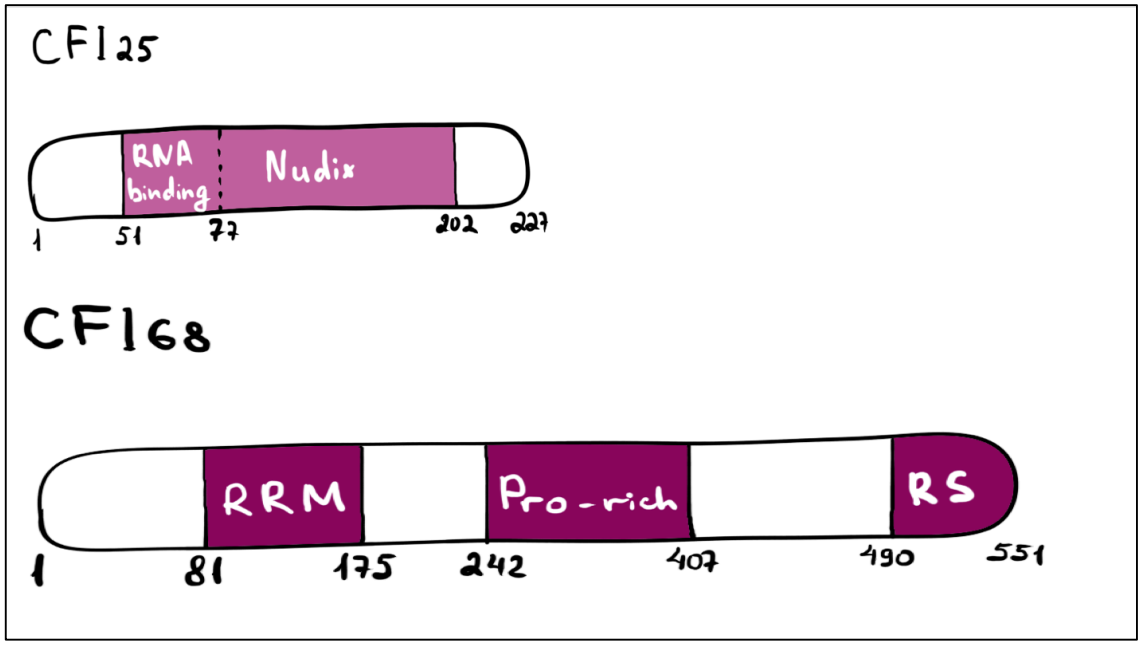


Figure 15. Cartoon of CF I_m subunits and their domain organization. Top row: CFI25 is mainly characterized by a Nudix domain, which is capable of RNA binding. Bottom row: CFI68 contains a N-terminal RNA recognition motif (RRM), followed by a proline-rich region in the middle of the protein and a C-terminal arginine-serine rich domain (RS).

1.2.5 The Cleavage Factor CF II_m

Cleavage Factor II (CF II_m) is the least characterized complex in the human 3′-end processing machinery. Although it is known that CF II_m is required for the cleavage reaction (Colgan and Manley 1997, Zhao, Hyman et al. 1999, Mandel, Bai et al. 2008), the exact function of CF II_m is still not known. Initial studies suggested, that CF II_m consists of two subunits, hPcf11 and hClp1, which are highly conserved from yeast to human (de Vries, Ruegsegger et al. 2000). Yeast homologues of hPcf11 and hClp1 stably associate in a complex with yeast homologues of CstF2 (Rna15) and CstF3 (Rna14), to form the so-called Cleavage Factor I A (CF IA) (Amrani, Minet et al. 1997, Minvielle-Sebastia, Preker et al. 1997, Gross and Moore 2001, Gordon, Shikov et al. 2011, Stojko, Dupin et al. 2017).

hPcf11 is almost twice as long as its yeast counterpart, so that both proteins only share homologous parts at the N-terminus (de Vries, Ruegsegger et al. 2000). hPcf1 has a N-terminal Pol II interacting domain (CID), which interacts with the RNA pol II C-terminal domain

(CTD), preferably with the Serine 2-phosphorylated form of the CTD (Barilla, Lee et al. 2001, Sadowski, Dichtl et al. 2003, Proudfoot 2004, Zhang and Gilmour 2006, Hsin and Manley 2012). Although Pcf11 proteins do not share much sequence homology, their function seems to be evolutionary conserved, because interaction with the RNA pol II CTD suggests a role of Pcf11 in transcription termination (Zhang, Fu et al. 2005). This was shown in previous studies, where Pcf11 knockdown in HeLa cells had direct impact on cleavage efficiency and transcription termination (West and Proudfoot 2008). Pcf11 therefore links 3′-end processing to transcription termination (Sadowski, Dichtl et al. 2003, Luo, Johnson et al. 2006, Zhang and Gilmour 2006, Porrua and Libri 2015) and transcript export (Johnson, Cubberley et al. 2009, Johnson, Kim et al. 2011). Mutations in Pcf11 cause retention of mRNA transcripts in the nucleus, because interaction between Pcf11 and the export factor AlyRef (Yra1 in yeast) is lost (Johnson, Cubberley et al. 2009, Johnson, Kim et al. 2011). Consequently, a correct assembly of the Transcription Export (TREX) complex is not possible (Chi, Wang et al. 2013). In human Pcf11, the domain following the N-terminal CID (residues 14-142) adopts a helical fold (residues 183-297) with unknown function (Xu, Perebaskine et al. 2015), followed by a region with a high content in charged amino acids (residues 295-565) (Schafer, Tuting et al. 2018). A proline-glycine rich region (Figure 16) is built of 13 repeated amino acids (short: FEGP repeats; resiudes 770-1123), which shows high conservation among vertebrate Pcf11 proteins (Schafer, Tuting et al. 2018) and demethylation of arginines within the 13 repeated residues (Guo, Gu et al. 2014, Schafer, Tuting et al. 2018). The two C-terminal zinc fingers (Barilla, Lee et al. 2001, Sadowski, Dichtl et al. 2003) are separated by the Clp1 binding region (Noble, Beuth et al. 2007, Schafer, Tuting et al. 2018).

hClp1 is highly conserved among eukaryotes and interacts with CF I_m and CPSF (de Vries, Ruegsegger et al. 2000, Gross and Moore 2001). The crystal structure of yeast Clp1 revealed that it has a central ATPase domain with an ATP molecule bound, but it showed no ATPase activity (Noble, Beuth et al. 2007). Mutational analysis indicated that this region is important for binding to Pcf11 and directly affects 3′-end processing and transcription termination (Ghazy, Gordon et al. 2012, Haddad, Maurice et al. 2012). In contrast to its yeast counterpart, which has no active kinase activity, hClp1 was identified to be an active RNA 5′-OH kinase (Weitzer and Martinez 2007). The question, if the RNA kinase activity is essential in 3′-end processing, is not solved yet, but genetic screens in mice indicated that kinase-dead Clp1 mice show no defect in 3′-end processing (Hanada, Weitzer et al. 2013). Apart from 3′-end processing, hClp1 is also implied in t-RNA splicing, where the RNA 5′-OH kinase activity is needed to phosphorylate the 5′-end of the 3′-exon for the ligation step (Weitzer and Martinez 2007).

In recent studies, human CF II_m was reconstituted consisting of hPcf11 and hClp1 in a heterodimeric association (Schafer, Tuting et al. 2018), which was active in AAUAAA

dependent cleavage assays in contrast to hClp1 alone. Initial indications, that hClp1 kinase activity is not essential for 3′-end processing was proven by mutational analysis of the active kinase site and the impact on 3′-end processing efficiency (Schafer, Tuting et al. 2018). It was identified, that RNA binding of CF II_m is mediated by hPcf11 and that its two C-terminal zinc fingers are the RNA binding domains with preference for G-rich RNA sequences (Schafer, Tuting et al. 2018).

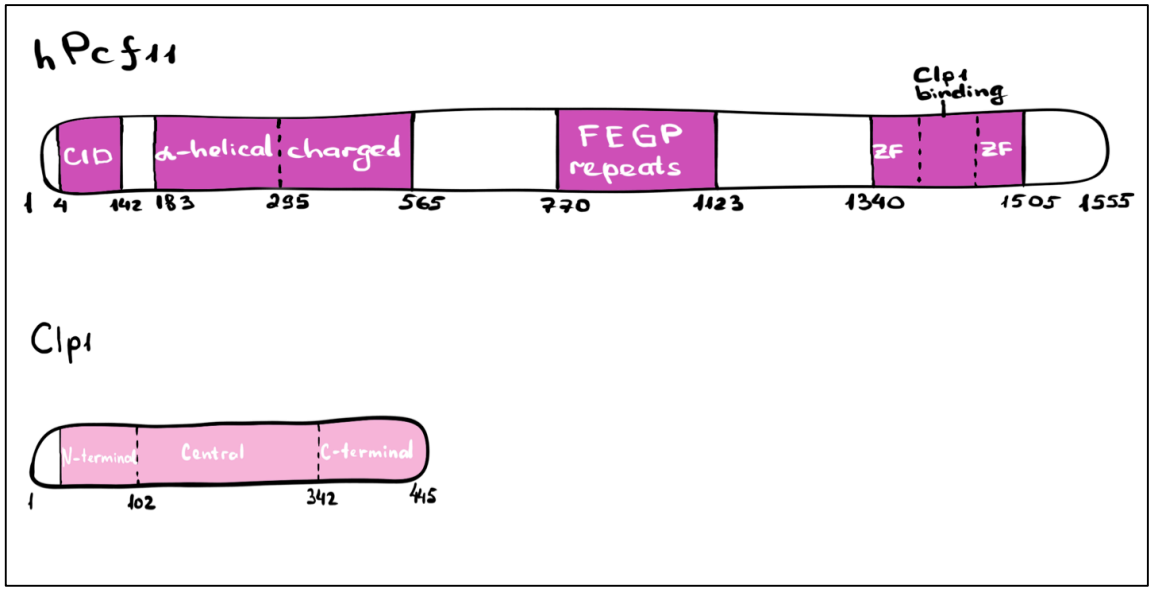


Figure 16. Cartoon of human CF II_m subunits and their domain organization. Top row: Mammalian Pcf11 has a N-terminal RNA polymerase II interacting domain (CID), followed by a helical region. A highly charged region (amino acids 295-565) with a high serine content is followed by a proline-glycine-rich region with repeated FEGP motifs (amino acids 770-1123). The C-terminal zinc fingers (ZF) are separated by the Clp1 interacting region (amino acids 1340-1505). Bottom row: Clp1 has a central ATPase domain (central) flanked by two domains at the N-terminus and the C-terminus.

1.2.6 The RNA polymerase II RNA pol II

There are three different RNA polymerases in eukaryotes (RNAP I-III), of which RNA pol II is responsible for generation of all mRNA transcripts (Hsin and Manley 2012). RNA pol II consists of 12 subunits and is very conserved among eukaryotes (Khatter, Vorlander et al. 2017, Engel, Neyer et al. 2018). The largest subunit, Rpb1, forms the catalytic center separated from the very conserved C-terminal domain (Cramer, Bushnell et al. 2001), which is characterized by several repeated heptapeptides with a conserved sequence (YSPTSPS). The number of repeats is different from yeast to vertebrates (Bartkowiak, Mackellar et al. 2011, Egloff, Dienstbier et al. 2012, Hsin and Manley 2012, Zhang, Rodriguez-Molina et al. 2012, Heidemann, Hintermair et al. 2013). Previous studies showed, that the CTD is required for polyadenylation *in vitro* and *in vivo* (McCracken, Fong et al. 1997, Hirose and Manley 1998). Additionally, the CTD is target of various PTMs, especially phosphorylation, which creates a

diversity platform to interact with many different factors of the 3′-end processing machinery (Bartkowiak, Mackellar et al. 2011, Heidemann, Hintermair et al. 2013, Jasnovidova and Stefl 2013). Besides its role in transcription, RNA pol II is involved in co-transcriptional mRNA processing steps by promoting a stable binding platform for interactions with various 3′-end processing factors via the CTD (McCracken, Fong et al. 1997, Hirose and Manley 2000, Barilla, Lee et al. 2001, Buratowski 2003, Kyburz, Sadowski et al. 2003, Sadowski, Dichtl et al. 2003, Meinhart and Cramer 2004, Bentley 2005, Kyburz, Friedlein et al. 2006). Thereby it links polyadenylation to the transcription process.

1.2.7 The Poly(A) Polymerase PAP

Polyadenylation in context of 3′-end processing of pre-mRNAs is mediated by an enzyme called poly(A) polymerase (PAP) in complex with other factors of the huge 3′-end processing machinery (Shi, Di Giammartino et al. 2009). Canonical PAP is one of the best characterized proteins of the 3′-end processing machinery (Colgan and Manley 1997, Zhao, Hyman et al. 1999, Mandel, Bai et al. 2008) and exists in several isoforms: PAPα, TPAP (PAPβ) and neo-PAP (PAPγ) (Raabe, Bollum et al. 1991, Kashiwabara, Zhuang et al. 2000, Lee, Lee et al. 2000, Kyriakopoulou, Nordvarg et al. 2001, Le, Kim et al. 2001, Topalian, Kaneko et al. 2001). Initially, canonical PAPs were reported to be the only polymerases to perform nuclear polyadenylation of pre-mRNA transcripts. Besides canonical PAPs, also different nuclear non-canonical PAPs (NcPAPs) were identified. One of them is Star-PAP (Speckle Targeted PIPKα Regulated Poly(A) Polymerase), which exists in the nucleus and cytoplasm (Lee, Lee et al. 2000, Chan, Choi et al. 2011). Star-PAP was reported to also take part in polyadenylation of pre-mRNAs, but does not share the same domain architecture as the canonical PAP (Mellman, Gonzales et al. 2008, Li, Laishram et al. 2012).

PAPα is very conserved from yeast to humans and belongs to the family of DNA polymerases β (Edmonds and Abrams 1960). The N-terminus of PAP is comprised by a highly conserved catalytic nucleotidyl transferase domain (NTD), spanning over the first 500 residues (Martin and Keller 1996, Martin, Keller et al. 2000). PAP is recruited to the 3′-end processing machinery acting on pre-mRNAs via interactions with CF I_m and CPSF (Colgan and Manley 1997, Zhao, Hyman et al. 1999, Mandel, Bai et al. 2008, Meinke, Ezeokonkwo et al. 2008). PAP alone binds to the pre-mRNA in a non-specific manner via the RNA binding domain located in the middle part of the protein (Figure 17). Two nuclear localization signals (NLS) are located in the C-terminal part of the RNA binding domain. The C-terminus of PAP contains a serine and threonine rich sequence part (Raabe, Murthy et al. 1994, Martin and Keller 1996, Martin, Keller et al. 2000), which is a target for post translational modifications (PTMs). These PTMs play an important role in regulation of PAP activity and localization (Colgan, Murthy et

al. 1996, Kim, Lee et al. 2003, Shimazu, Horinouchi et al. 2007). Besides that, the C-terminus of PAPα can interact with U1A and U2AF65 splicing factors (Colgan, Murthy et al. 1996, Colgan, Murthy et al. 1998, Zhao and Manley 1998), thereby delivering another level of regulation. The crystal structure of mammalian PAPα showed, that it adopts a globular shape with a central cleft forming the active site (Bard, Zhelkovsky et al. 2000, Martin, Keller et al. 2000). The active site can be closed by direct interaction of the PAPα NTD and CTD in a so-called induced-fit mechanism (Martin, Moglich et al. 2004, Balbo and Bohm 2007, Balbo, Toth et al. 2007). Besides that, the active site is characterized by an aspartic triad, which is necessary for ATP hydrolysis by coordinating metal ions (Davies, Almassy et al. 1994, Martin, Jeno et al. 1999, Bard, Zhelkovsky et al. 2000, Martin, Keller et al. 2000).

TPAP (PAPβ) is the smallest form of canonical PAPs and is mainly expressed in testis from a different, intronless gene than PAPα. It is localized in the nucleus and cytoplasm and seems to be required for processing of pre-mRNA transcripts during spermatogenesis (Kashiwabara, Zhuang et al. 2000, Lee, Lee et al. 2000, Le, Kim et al. 2001). It is assumed, that all canonical PAP forms have their origin in a common gene, which was duplicated into different forms PAPOLA, PAPOLB and PAPOLG (Kashiwabara, Zhuang et al. 2000, Lee, Lee et al. 2000, Le, Kim et al. 2001). Therefore, TPAP shares domain similarity with PAPα.

Neo-PAP (PAPγ) is the third form of canonical PAPs and has the same domain organization as PAPα. Besides that, its function in 3′-end processing of pre-mRNA transcripts seems to be similar to that of PAPα, because previous studies showed that it has polyadenylation activity *in vitro* (Kyriakopoulou, Nordvarg et al. 2001, Topalian, Kaneko et al. 2001). In addition to normal polyadenylation activity, also monoadenylation of small RNAs was observed *in vitro* (Perumal, Sinha et al. 2001). *In vivo* functions remain poorly characterized, but neo-PAP activity was identified in tumorigenesis (Kyriakopoulou, Nordvarg et al. 2001, Topalian, Kaneko et al. 2001).

Star-PAP is a non-canonical PAP located in the nucleus and was identified interacting with phosphatidyl inositol phosphate kinase Iα (PIPKIα) (Mellman, Gonzales et al. 2008). Its domain architecture differs from the one of the canonical PAP, because it has a N-terminal zinc finger domain followed by an RNA recognition motif (Mellman, Gonzales et al. 2008, Laishram 2014). The catalytic domain (NTP) is separated by a 200-residue long proline-rich region (PRR) followed by a PAP-associated domain. The C-terminus is characterized by a RS-domain and the NLS (Mellman, Gonzales et al. 2008). Star-PAP functions in 3′-end processing of selected pre-mRNA transcripts in complex with PIPKIα and CPSF subunits (Mellman et al., 2008; Laishram et al., 2006) and is directly regulated by phosphatidyl inositol 4,5-bisphosphate ($Pi4,5P_2$), a lipid messenger (Doughman, Firestone et al. 2003, Bunce, Bergendahl et al. 2006, Barlow, Laishram et al. 2010).

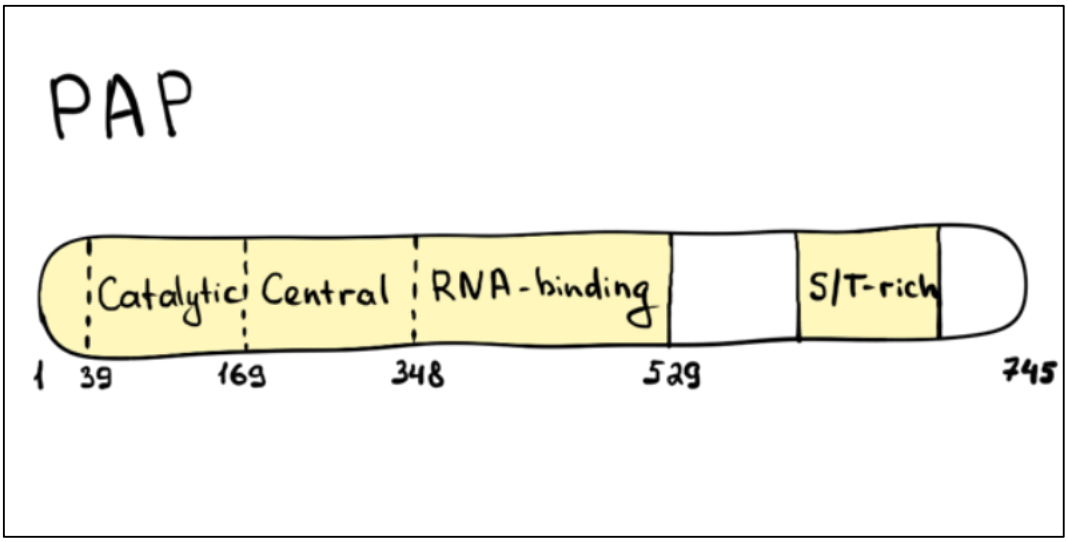


Figure 17. Domain organization of poly(A) polymerase PAP. The N-terminus of PAP contains a conserved catalytic nucleotidyl transferase domain and an RNA binding domain in the middle. The C-terminus is rich in serine and threonine (S/T-rich) and prone to posttranslational modifications.

1.2.8 Poly(A) binding proteins

In humans, there are several poly(A) binding proteins (PABPs). One of them exists in the nucleus (PABPN1) and was identified later than its four cytoplasmic counterparts (PABPC1, 3, 4, 5) (Blobel 1973, Wahle 1991). Nuclear and cytoplasmic PABPs don't share a similar domain architecture, because PABPN1 has a very acidic glutamate-rich (E-rich) N-terminus, preventing unwanted interactions with PAP (Kerwitz, Kuhn et al. 2003). A coiled-coil region following the N-terminus (Figure 18) is involved in stimulation of PAP for processive poly(A) tail elongation (Kerwitz, Kuhn et al. 2003). The crystal structure of the more C-terminally located RRM showed that it dimerizes in solution (Ge, Zhou et al. 2008), which is in line with previous studies that predicted a self-dimerizing capability of the C-terminal domain (CTD) of PABPN1 (Kuhn, Nemeth et al. 2003). Both domains, RRM and CTD, are necessary for binding to around 10 nt long poly(A) RNA stretches (Nemeth, Krause et al. 1995, Ge, Zhou et al. 2008). Besides stimulating PAP (Wahle 1991), PABPN1 was shown to control poly(A) tail length (Bienroth, Keller et al. 1993). By being recruited to the slowly emerging poly(A) tail produced in presence of only CPSF, PABPN1 stabilized the polyadenylation machinery by coating the poly(A) tail in a way, that a spherical shape is produced (Keller, Kuhn et al. 2000). This brings the CPSF complex and the elongating PAP in close proximity to maintain their interaction (Kuhn, Gundel et al. 2009). Consequently, processivity is switched to a rapid mode, resulting in fast poly(A) tail elongation until a length of around 250 nt is reached (Bienroth, Keller et al. 1993, Wahle, Lustig et al. 1993).

Cytoplasmic PABPs bind to poly(A) tails in the cytoplasm (Baer and Kornberg 1983) and therefore take part in translation initiation by serving as scaffolds to bridge the 5′-cap of

mRNAs to the 3′-poly(A) tail (‘closed loop model’) (Jacobson and Peltz 1996, Tarun and Sachs 1996, Borman, Michel et al. 2000, Mangus, Evans et al. 2003, Kuhn and Wahle 2004). PABC proteins are very conserved and share a common domain architecture. They consist of four RRMs, which are essential for binding to poly(A) RNA (Burd, Matunis et al. 1991, Kuhn and Pieler 1996). A proline-rich region (PRR) connects the RRMs to a C-terminal mademoiselle (MLLE) domain (Passmore and Coller 2022), which recognizes a short, so-called poly(A)-interacting motif 2 (PAM2), present on various eukaryotic proteins (Xie, Kozlov et al. 2014). RRM1 and RRM2 mediate binding to a poly(A) stretch of around 12 nt with high affinity (Burd, Matunis et al. 1991, Kuhn and Pieler 1996), but the whole protein covers a length of around 30 nt, so that the longer the poly(A) tail, the more PABPCs can be bound (Schafer, Yamashita et al. 2019).

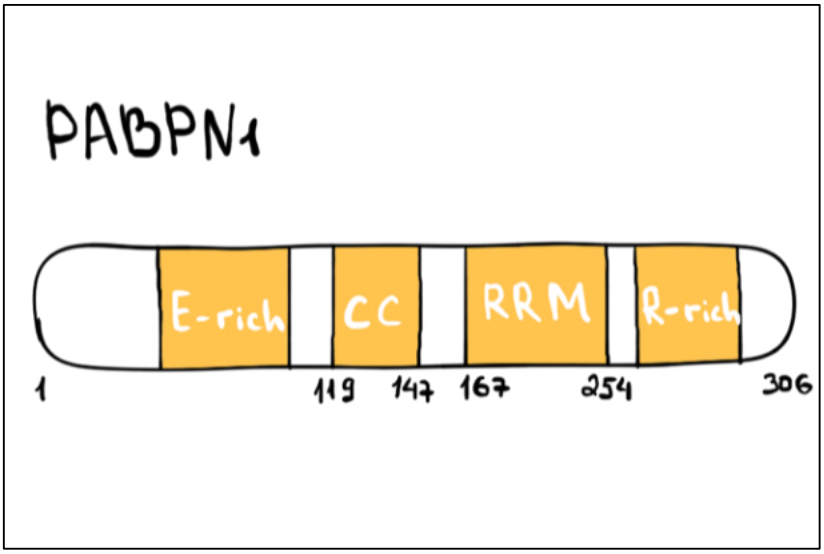


Figure 18. Domain organization of PABPN1. PABPN1 has an acidic N-terminus, which is rich in glutamates (E-rich). The coiled coil (cc) domain in the middle is involved in stimulation of PAP. The following RRM and arginine-rich (R-rich) CTD can self-associate (Kuhn, Nemeth et al. 2003, Ge, Zhou et al. 2008).

1.3 Molecular interactions of the human 3′-end processing machinery with pre-mRNA

1.3.1 Recognition of the poly(A) signal AAUAAA by CPSF complex

As already mentioned in the text above, human CPSF1, WDR33, hFIp1 and CPSF4 form a stable subcomplex, the so-called mPSF, which is the equivalent to the yeast polymerase module. This complex is required for recognition of the very conserved hexameric AAUAAA poly(A) signal via subunits WDR33 and CPSF4 (Schonemann, Kuhn et al. 2014). So far, the molecular mechanism of the specific AAUAAA recognition was not unveiled, until two independent groups solved the structure of human mPSF complex bound to an AAUAAA containing RNA ligand (Clerici, Faini et al. 2018, Sun, Zhang et al. 2018). Studies of both groups show small differences in construct design used for their cryo-EM studies. However, both groups reconstituted a CPSF1-WDR33-CPSF4 complex bound to AAUAAA containing RNA stretches at high resolution, which allowed determination of the molecular mechanism of recognition of the AAUAAA poly(A) signal. The AAUAAA RNA, which adopts an S-shaped fold, is located in the interface of the WD40 propeller of WDR33 and zinc finger (ZF) 2 and 3 of CPSF4. The S-shape mediates, that A_1 and A_2 of the poly(A) signal point in a perpendicular direction away from A_4 and A_5. U_3 – A_6 bases are forming a Hoogsteen pair pointing in the opposite direction of A_4 and A_5. The first bases A_1 and A_2 are bound by ZF2 of CPSF4, A_4 and A_5 interact with ZF3 of CPSF4 and the Hoogsteen base pair U_3 – A_6 is recognized by WDR33. Both structures show, that the N-terminal amino acids 41-55 of WDR33 cover the RNA, directly interacting with U_3, A_5 and A_6 and thereby stabilizing the kinked backbone shape of the poly(A) signal. The N-terminal residues Lys46, Arg47, Arg49 and Arg54 of WDR33 directly interact with the backbone of the RNA (Clerici, Faini et al. 2018). There is no base-specific recognition of the Hoogsteen base pair, but it is flanked by two phenylalanine residues of WDR33 (Phe43 and Phe153) and consequently stabilized by π- π interactions. Besides that, Lys117 and Ile156 are covering the other site of the Hoogsteen base pair. Bases A_1, A_2, A_4 and A_5, which are bound by CPSF4, are located in pockets forming π- stacking interactions with conserved residues. The side chain of Phe84 forms π- stacking interactions with A_1, which further interacts with residues in ZF2 of CPSF4 (Lys69 and Val67). The second base A_2 is only contacted by one hydrogen bond via Lys77, which together with Lys78 surrounds the π- stack formed by His70. Bases A_4 and A_5 are bound by sequence-specific hydrogen bonds of residues within ZF3 of CPSF4. The π- stacking interactions formed between Phe112 and A_4 are further stabilized by interaction with the main-chain amide of Tyr97 and the carbonyl of Glu95. Besides π-stacking with Phe98, A_5 is recognized by Ser106 and Asn107. The complicated and specific interaction network between proteins and the AAUAAA hexamer is consistent with the high conservation of the hexameric motif (Hu, Lutz et al. 2005, Derti, Garrett-Engele et al. 2012,

Gruber, Schmidt et al. 2016). However, less specific recognition of the A_2 base goes in line with the variance in hexameric poly(A) signals especially in position 2 (Sheets, Ogg et al. 1990, Hu, Lutz et al. 2005, Derti, Garrett-Engele et al. 2012, Gruber, Schmidt et al. 2016). Substitution of single bases can lead to strong reduction of the affinity of the poly(A) signal towards the CPSF complex and therefore impact efficiency of 3′-end processing, which was shown to happen in human diseases like α- and β-thalassemia (Higgs, Goodbourn et al. 1983, Orkin, Cheng et al. 1985). Although there is no base-specific recognition of the U_3 – A_6 base pair, its conservation can be explained by the binding pocket of WDR33, which is not compatible with other base combination and would form non-ideal hydrogen bonds upon substitution of bases.

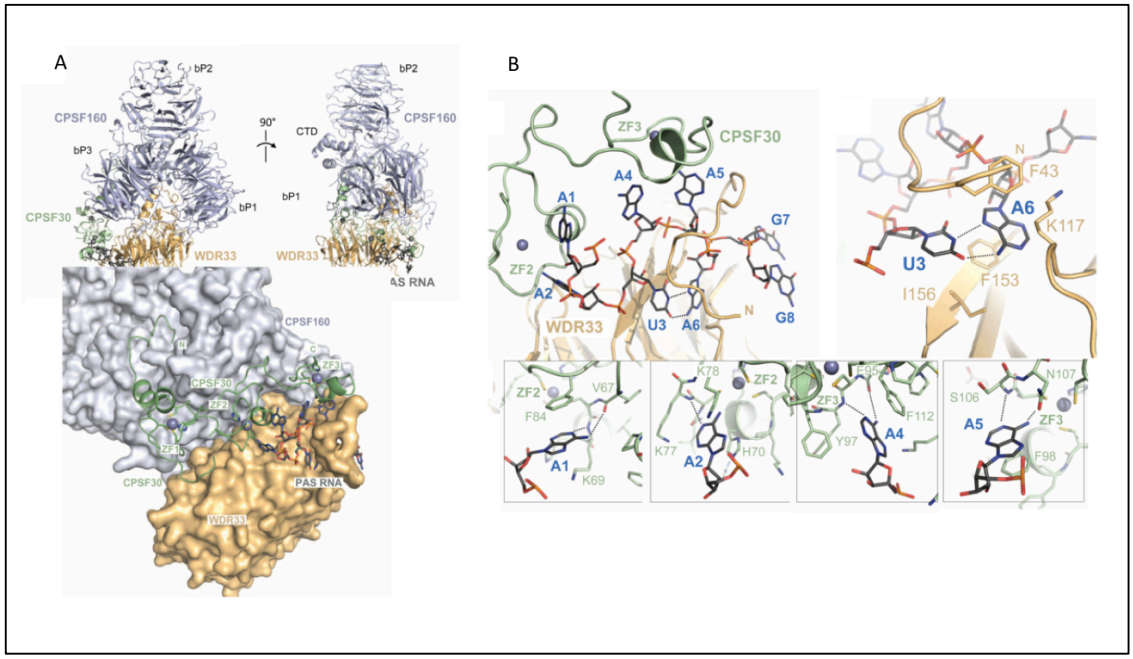


Figure 19. Molecular mechanism of AAUAAA PAS recognition by the CPSF complex (Clerici, Faini et al. 2018). A) upper panel left: front view of the CPSF1-WDR33-CPSF4-AAUAAA structure. Proteins are shown in cartoon and RNA nucleotides as sticks. Upper panel right: sideview of the CPSF1-WDR33-CPSF4-AAUAAA structure. Proteins are shown in cartoon and RNA nucleotides as sticks. Bottom panel: Zoomed view of the CPSF4 binding cleft formed between CPSF1 and WDR33. CPSF4 (green) is shown as cartoon and the AAUAAA hexamer as sticks. WDR33 and CPSF1 are shown in surface format. B) Upper panel left: Structural insights in the binding of the AAUAAA PAS to zinc fingers (ZF) 2 and 3 of CPSF4 and WDR33. Proteins are shown as cartoons and RNA as sticks. Dotted lines show Hoogsteen base-pair formation of U3 and A6. Upper panel right: Zoomed view of the U3-A6 Hoogsteen base-pair flanked by WDR33 F43 and F153/I156. Bottom panel: Molecular mechanism of the recognition of the nucleotides A1, A2, A4 and A5 by ZF2 and ZF3 of CPSF4.

1.3.2 Recognition of G/U-rich downstream elements by CstF2 RRM

The CstF complex consisting of CstF1, CstF2 and CstF3 is involved in binding of G/U-rich DSEs on the pre-mRNA via the RRM domain of CstF2. DSEs are essential poly(A) signals (PAS) on the pre-mRNA, that help defining the exact location of the cleavage site (Chen, MacDonald et al. 1995, Beyer, Dandekar et al. 1997, Takagaki and Manley 1997). In contrast

to the very conserved hexameric AAUAAA sequence motif, G/U-rich DSEs show high variability in mammalian PAS. Therefore, it is not known yet, how CstF can distinguish between various sequences to select for G/U-rich RNA. Since the CstF2 protein is binding to G/U-rich downstream elements without any strong consensus sequence, the RRM of CstF2 must recognize certain elements on the RNA without directly binding to a particular sequence. Besides that, it also has to discriminate against A/C-rich motifs. Previous studies showed, that presence of two consecutive Uracils has a huge effect on RNA binding by CstF2 (Perez Canadillas and Varani 2003). The group solved the structure of the CstF2 RRM by NMR (Perez Canadillas and Varani 2003) and modelled it with a UU-dinucleotide based on its similarity to the HuD-*cfos* complex (Wang and Tanaka Hall 2001). Chemical shifts in the NMR spectra indicated unfolding of the C-terminal helix (helix C; residues 94-105) and local motion in the main β-sheet, representing the RNA binding interface, upon presence of two Uracils in the RNA stretch. Consequently, they hypothesized that the RNA binding specificity is mediated by the dynamic behavior of the RRM, especially by creating a binding pocket for two Uracils upon unfolding of helix C. RNA nucleotides neighboring the UU-dinucleotide are supposed to fine-tune interaction with the RRM, thereby providing different binding affinities for different G/U-rich RNA elements. Flexibility of the C-terminal helix of the RRM is crucial to expose the RNA binding site: the loop β_1/α_1 and the central β-sheet of the RRM, which are hidden by helix C in absence of RNA (Perez Canadillas and Varani 2003, Pancevac, Goldstone et al. 2010). In a non-RNA bound state, hydrophobic residues of the C-terminal helix (E100, L101 and L104) interact with aromatic residues (F19 and F61) in the two RNP consensus sequences (RNP1 and RNP2). Further stabilization of helix C is obtained by hydrogen bonds between N91 and N97.

Based on similarity to the HuD-*cfos* complex (Wang and Tanaka Hall 2001) and data available, the following molecular mechanism for recognition of two uracils has been proposed by Perez-Canadillas and Varani, 2003 (Figure 20). The 5′-uracil of the RNA is recognized specifically by H-bonds between the O4 carbonyl and side chain of S17 and the O2 carbonyl and R46 side chain. These two hydrogen bonds are able to discriminate against G and A for the first position due to their different size. Since the H-bond between N91 and N97 side chains is lost upon RNA binding due to unfolding of helix C, N91 can recognize the O4 carbonyl of the second uracil U_2. Besides that, another hydrogen bond is formed between the NH group of U_2 and the protein main chain. These interactions would not be possible, if C or A were in the second position of the RNA. Additionally, interactions formed by aromatic side chains from RNP1 and RNP2 (F19 and F61), which are keeping helix C in its correct position in the apo conformation, are lost upon unfolding of helix C, what allows the aromatic residues to form intermolecular stacking interactions with the two uracils instead. Taken together, the RNA binding specificity

of the CstF2 RRM for G/U-rich sequences is obtained by formation of a tight binding pocket in combination with a network of base specific interactions, that discriminate against A and U.

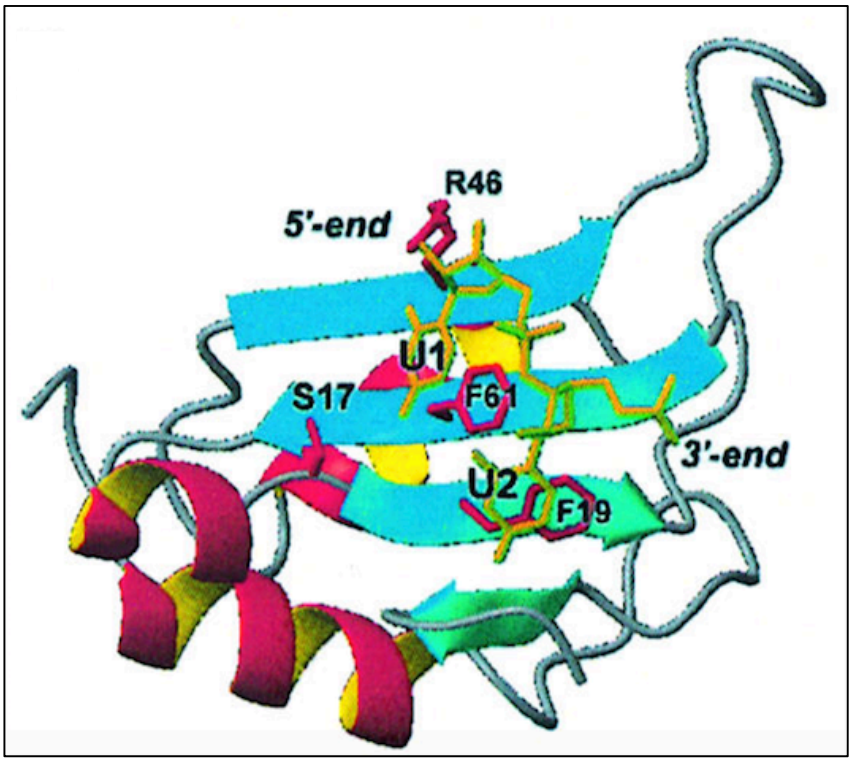


Figure 20. Model of UU-dinucleotide recognition by the CstF2 RRM domain (Perez Canadillas and Varani 2003). Cartoon structure of the CstF2 RRM modelled with the UU-dinucleotide. Side chains of the residues involved in the binding of the RNA (yellow) are shown in pink.

1.3.3 Recognition of the UGUAN USE by CF I_m

The so-called upstream sequence element on the pre-mRNA, that is one of the tripartite *cis*-elements involved in cleavage site definition, is bound by Cleavage Factor I_m via the Nudix domain of the 25 kDa subunit CFI25 (Yang, Gilmartin et al. 2010). However, it was not clear how Nudix proteins bind with sequence specificity to mRNA ligands. The UGUAN motif was identified by different studies to be a consensus sequence for CF I_m - pre-mRNA interactions (Brown and Gilmartin 2003). The crystal structure of the CFI25 homodimer reconstituted with UGUAAA and UUGUAU RNA species (Yang, Gilmartin et al. 2010) solved the molecular mechanism of sequence specific CFI25 – UGUAN interaction. Besides that, crystal structures of a heterotetrameric CF I_m complex consisting of the CFI25 homodimer flanked by two CFI68 RRM molecules, identified the binding mode by which CF I_m recognizes two UGUAN binding sites on the pre-mRNA (Li, Tong et al. 2011, Yang, Coseno et al. 2011).

The structure of homodimeric CFI25 reconstituted with UGUA-containing RNA ligands adopts a similar conformation as observed for the apo complex (Coseno, Martin et al. 2008, Tresaugues, Stenmark et al. 2008). As already described in paragraph 1.2.4, the CFI25 Nudix domain follows the consensus $\alpha/\beta/\alpha$ fold (Mildvan, Xia et al. 2005). The big differences between the apo structure and RNA-bound CFI25 is, that the N-terminal residues 21-29 are not flipped towards the second monomer in the RNA-bound structure and that the connection between α/β is incorporated into the active RNA binding site (Yang, Gilmartin et al. 2010). Most nucleotides of one hexameric RNA ligand ($UU_1G_2U_3A_4U$ and $U_1G_2U_3A_4AA$) are bound by one molecule of the CFI25 dimer and only part of the nucleotides by the second monomer. Both RNA sequences are arranged and twisted in a way, that specifically the UGUA part of the RNA sequence is bound by the CFI25 dimer (Yang, Gilmartin et al. 2010). Sequence-specific recognition of the $U_1G_2U_3A_4$ tetrameric RNA core is initiated by formation of hydrogen bonds from U_1 to the main chain amide and carbonyl groups of Phe104 and stabilized by complex interactions with Glu81, Leu106, Thr102, Tyr208 and Gly209 (Figure 21 A) (Brown and Gilmartin 2003, Auweter, Oberstrass et al. 2006, Coseno, Martin et al. 2008). All interactions together define U_1 as the first nucleotide of the $U_1G_2U_3A_4$ sequence. The second nucleotide G_2 not only forms direct interactions with the side chain of Glu55 and indirect interactions via a water molecule, but also binds intramolecularly with the fourth RNA nucleotide A_4 (Figure 21B). An incorporated water molecule bridges between A_4 and Thr102 and Phr103, thereby specifying the fourth position of the RNA sequence being a purine base (Figure 21 B). Specificity of the second position G_2 of the RNA tetramer is gained by interaction with Glu55 and is further fixed by stacking interactions with Phe103, thereby indirectly determining the fourth base A_4 (Figure 21 D). Sequence specific G_2 - A_4 – CFI25 interactions are further extended by van der Waals contacts between A_4 and Leu99. The RNA base U_3 is also positioned by a strong interaction network, thereby delivering specificity for the third position of the RNA sequence. U_3 directly interacts with the guanidium group of Arg63 and forms intramolecular interactions to A_4 mediated by a water molecule. A second water molecule bridges U_3 to Glu55 and Asp57 (Figure 21 C). Yang, Gilmartin et al. 2010 also tested simultaneously binding to two UGUA sequence elements by EMSA, but structural evidence for this hypothesis was delivered in later studies (Li, Tong et al. 2011, Yang, Coseno et al. 2011), when the structure of a CFI25-CFI68RRM-RNA complex was solved.

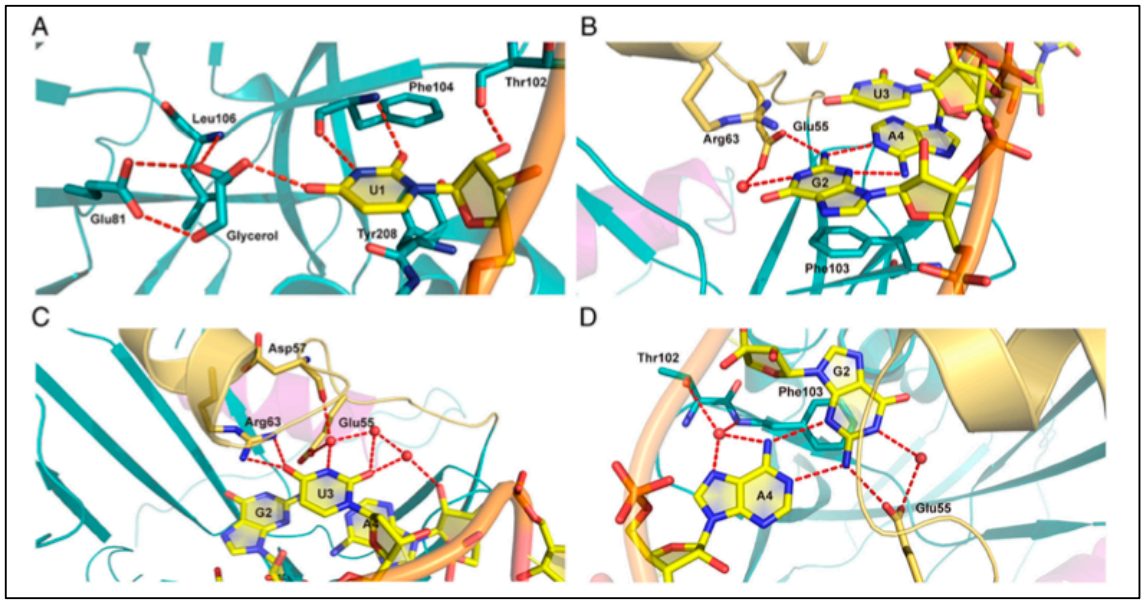


Figure 21. Crystal structure shows the molecular mechanism of the CFI25-UGUA interaction (Yang, Gilmartin et al. 2010). RNA backbone is shown in orange and hydrogen (H) bonds are shown as red dashed lines. Detailed view of the CFI25 subunit interacting with all bases of the UGUA USE. A) U1 B) G2 C) U3 D) A4.

In the text above, the molecular mechanism of specific binding of UGUA RNA sequence elements by the CFI25 Nudix domain is described. As a consequence of the dimeric association of CF I_m, a bipartite binding mode to two UGUA elements is suggested, as was shown by Li, Tong et al. 2011 and Yang, Coseno et al 2011. They solved the crystal structure of a CF I_m heterotetramer bound to two UGUA sequence elements and thereby explained the molecular mechanism of the poly(A) site determination by upstream RNA sequence elements. Additionally, they investigated the role of the CFI68 subunit and especially its RRM on RNA binding. Besides that, CF I_m was shown to be involved in RNA looping, indicating a potential role in the selection of alternative poly(A) sites (Yang, Coseno et al. 2011). The crystal structure, consisting of a CFI25 dimer flanked by two CFI68 RRM domains, was shown to adopt a unique way of interaction, because two CFI68 RRMs do not interact with individual CFI25 subunits each, but are flanking the CFI25 dimer in a way that each of the CFI68 RRMs can contact both CFI25 monomers. This structural arrangement was also shown for a heterotetramer consisting of a CFI25 dimer, sandwiched between two CFI59 RRM domains (PDB: 3N9U, Treasaugues et al., to be published). In the crystal structure solved by Yang, Coseno et al. 2011, the CFI68 RRM folds into the typical RRM fold ($\beta_1/\alpha_1/\beta_2/\beta_3/\alpha_2/\beta_4$), forming an antiparallel four stranded β-sheet, covered by the C-terminal α_3-helix. Presence of a C-terminal α-helix on top of the β-sheet was already reported for several RRMs (Perez Canadillas and Varani 2003, Dominguez, Fisette et al. 2010). Two CFI68 RRMs were shown to have impact on RNA binding affinity of the CFI25 subunit, that specifically recognizes UGUA RNA sequence elements (Yang, Gilmartin et al. 2010). Presence of two CFI68 subunits

increased the binding affinity towards a UGUA-containing USE on poly(A) polymerase α pre-mRNA (PAPOLA; GGGUGUAAAACAGAUGAUGUAU). In contrast to the CFI25 homodimer, which binds to only one UGUA-containing RNA stretch, two RNA molecules were bound in the crystal structure of the CFI25-CFI68 tetramer. EMSA experiments with wild type PAPOLA USE RNA and with different linkers inserted between two UGUA sequence motifs confirmed, that the CFI25/CFI68 complex binds to two UGUA sequence elements with different spacers on the pre-mRNA. Due to the anti-parallel arrangement of the CFI25 subunits, it was assumed that simultaneous binding to two UGUA sequence elements requires loop formation of the RNA to fit to the anti-parallel binding mode. RNA looping by CF I_m is consistent with its ability to bind to two UGUA upstream elements with different spacers (Venkataraman, Brown et al. 2005). In contrast to the CFI25 homodimer (Yang, Gilmartin et al. 2010), longer spacer length between two UGUA sequence elements of PAPOLA USE increased binding affinity for the CFI25/CFI68 heterotetramer, suggesting that the CFI68 subunit contributes to loop formation of RNA between two UGUA binding sites. Although the RRM domain is usually directly involved in RNA binding via residues in its RNP1 and RNP2 consensus binding motifs (Dominguez, Fisette et al. 2010), mutational analysis showed that these motifs are not required for RNA binding of the CFI25/CFI68 complex (Yang, Coseno et al. 2011). Among RRMs, not only RNP1 and RNP2 containing β-sheets are involved in RNA binding, but also loops connecting β_1/α_1 and β_2/β_3 of the RRM can be involved in RNA interactions (Clery, Blatter et al. 2008, Dominguez, Fisette et al. 2010). The CFI25/CFI68 crystal structure showed, that these loops are located in clefts between both subunits (Yang, Coseno et al. 2011). Mutational analysis of the loops showed, that both clefts seem to participate in RNA binding and location of the mutated residues gives information about the position of the RNA loop connecting both UGUA binding motifs. By further introducing different spacer length between two UGUA sequence elements, it was assumed that the RNA loop wraps around the CFI68 RRM (Yang, Coseno et al. 2011). A model was purposed (Figure 22), where two clefts function as entry and exit channel. An RNA ligand containing two UGUA motifs is guided to its correct position at the binding sites of the CFI25 monomers by looping out nucleotides connecting the two UGUA sequences.

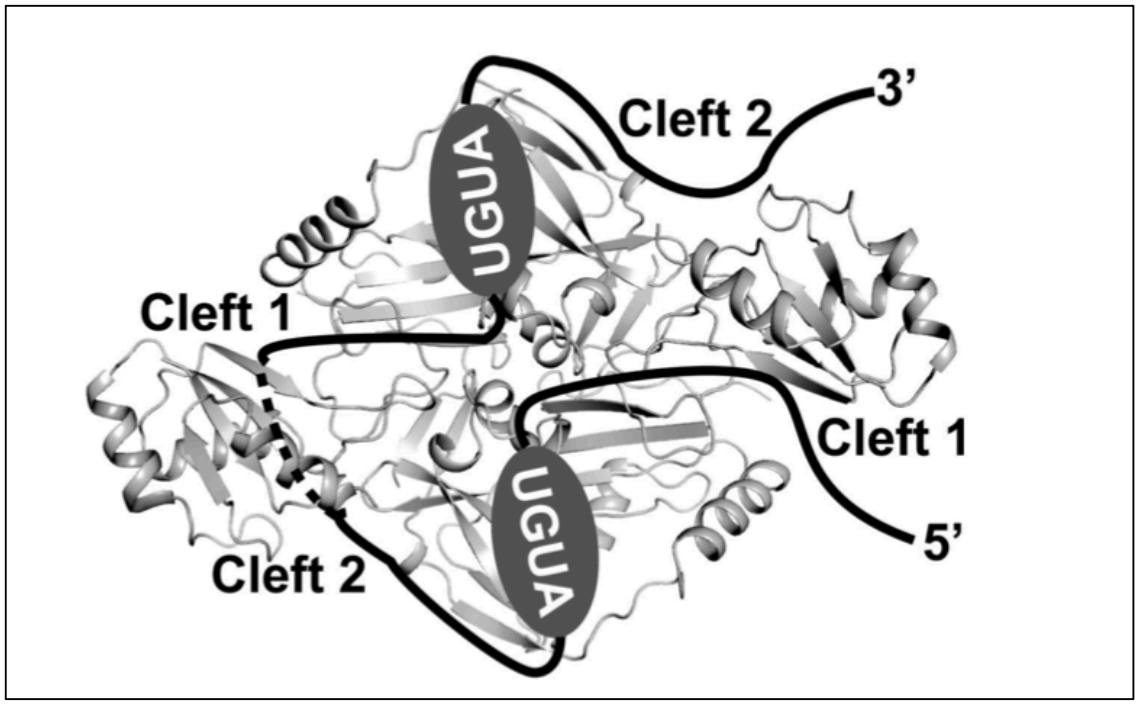


Figure 22. Model for RNA looping mediated by the CF I_m subunit CFI68 (Yang, Coseno et al., 2011). RNA is shown as continuous line with two UGUA upstream motifs highlighted by ovals. Dashed line indicates, that RNA is running below the RRM of CFI68.

Based on their data, Yang, Coseno et al. 2011 proposed a model for the involvement of CF I_m in alternative polyadenylation, as reported by previous studies (Kubo, Wada et al. 2006, Sartini, Wang et al. 2008). By looping out parts of the pre-mRNA, CF I_m can on the one hand combine different UGUA USEs and thereby influence selection of alternative poly(A) sites (Figure 23 A), or on the other hand directly loop out the whole poly(A) site, leading to selection of a downstream poly(A) site (Figure 23 B).

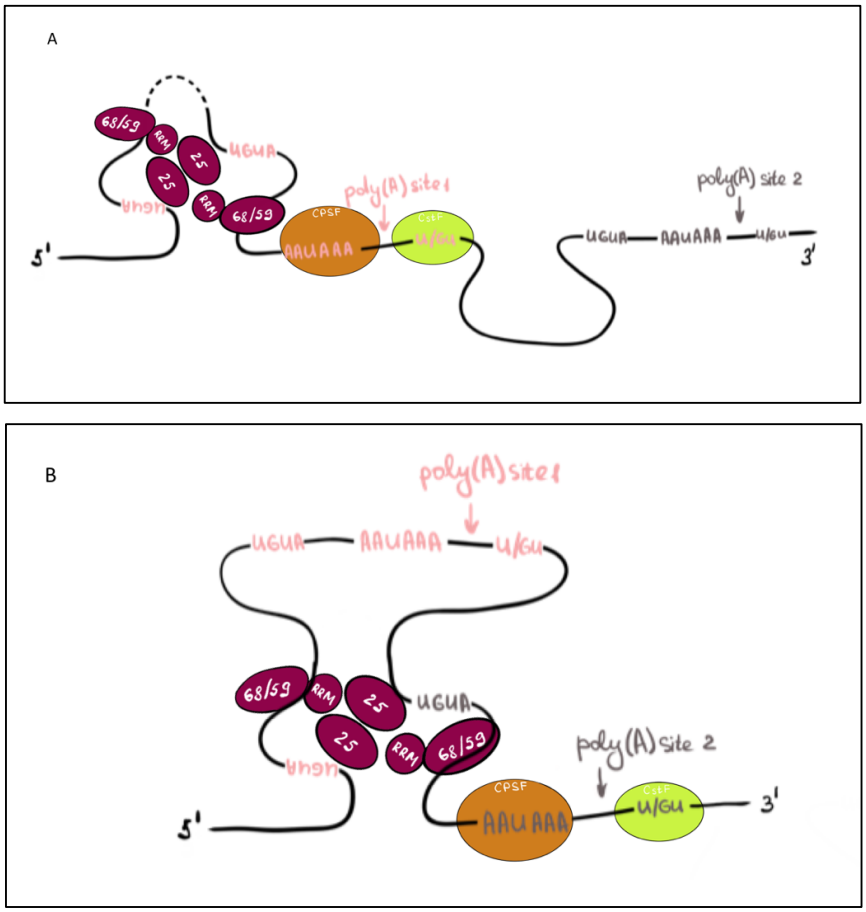


Figure 23: Model of CF I_m interaction with two UGUA upstream elements on pre-mRNA (adopted from Yang, Coseno et al., 2011). A) CF I_m binds to canonical UGUA sequence elements and pre-mRNA cleavage occurs at the normal cleavage site (poly(A) site 1). B) CF I_m loops out the whole poly(A) site 1 with its *cis*-elements and binds to a downstream UGUA binding site. Cleavage of pre-mRNA occurs at an alternative cleavage site (poly(A) site 2).

2 Results

2.1 Expression and purification of recombinant human CstF complex and its subcomplexes in insect cell expression system

In order to get more information about the structural assembly of the CstF complex, arrangement of its subunits and its biochemical properties, high amounts of pure protein and clean reconstituted complex was needed. The complex was not only characterized structurally by cryo-Electron Microscopy (cryo-EM), but also biochemical experiments were used to examine RNA binding of CstF and binding affinities were determined by Fluorescence Anisotropy (FA) or Isothermal Titration Calorimetry (ITC) analysis.

First, I will give an overview, how different protein samples were purified for specific experimental purposes. The purification paragraph in this thesis will start with purification of the full-length CstF complex, then continue with several subcomplexes and finally describe purification of single domains. Full-length subunits (CstF1, CstF2 and CstF3) of the CstF complex were purified as recombinantly tagged proteins, expressed in insect cells due to low expression levels in bacterial expression systems. At the beginning of the project, coding sequences for full-length CstF1, CstF2 and CstF3 were cloned into a single co-expression vector for protein expression in insect cells. One of the subunits, CstF2, proved to be especially difficult to express. On the one hand, it was expressed in sub stoichiometric amounts, and on the other hand it was not properly detectable in initial pull downs from expression tests, because it was not recombinantly tagged on either N- or C-terminus. Consequently, I started optimizing constructs by fusing N-terminal Strep tags to all subunits. A TwinStrep tag (Schmidt, Batz et al. 2013) was cloned to the N-terminus of CstF2 and Strep II (Voss and Skerra 1997) tags to the N-termini of CstF1 and CstF3. Additionally, a His_8 tag was fused to the C-terminus of CstF3, because it was prone to C-terminal degradation. Besides that, I re-cloned all three genes coding for the subunits into individual expression vector each, because subunits were expressed in very different levels from a single expression vector, making it difficult to obtain a stoichiometric complex in purifications. In the end, I was able to obtain high yields of pure recombinant CstF complex in an amount suitable for initial structural studies by negative stain Electron Microscopy (EM).

Once, purification results could be validated by the presence of clearly visible single particles in negative stain EM, the purification protocol was used to prepare full-length CstF for biochemical and biophysical studies. Sample preparation for single particle cryo-EM studies had to be further optimized, since the CstF complex turned out to be unstable under cryogenic conditions and was prone to dissociation during the plunging procedure. The results of

processing of different screening datasets showed density for CstF1 and CstF3 in various conformations, but no density could be clearly assigned to the CstF2 subunit. Consequently, I removed this subunit from the CstF1-CstF2-CstF3 complex to obtain a more minimal tetramer consisting of two copies of CstF1 and CstF3. CstF1-CstF3 was more stable in biochemical experiments and in cryo-EM screening datasets. Therefore, purification of the CstF1-CstF3 subcomplex was further optimized to allow collection of high-resolution data for single particle analysis. Final cryo-EM datasets were collected using a sample purified by a combination of density-gradient-ultracentrifugation and in-batch cross-linking.

Besides full-length CstF, different subcomplexes were needed for biochemical and biophysical studies, either as controls or to perform further experiments. Various subcomplexes could be stably assembled by co-lysis, but their purification procedures had to be optimized individually, depending on the complex composition and in regard to the experiment they were needed for. For RNA binding studies, I designed various mutants that contain single or double mutations in the CstF2 RNA binding domain (RBD), which corresponds to amino acids 1-111 of CstF2 (Figure 24). CstF2 mutants were stably bound by CstF1 and CstF3, so that purification and buffers were optimized to get pure, stable complex for optimal results in RNA binding studies. To have a closer look, if RNA binding behavior changed with changing complex composition and size, CstF2 alone was used to dissect its RNA binding properties. Unfortunately, CstF2 turned out to be very unstable and was co-purifying with some stably associated contaminants. In regard of this challenges, I designed and optimized a purification protocol for CstF2, where I could remove most contaminants and obtain single CstF2 in reasonable amounts, so that it could be used for biochemical studies. Going to the protein domain level, single RNA recognition motif (RRM) of CstF2 and a fusion of two RRMs to mimic dimeric association, were cloned for bacterial expression. All mentioned constructs of CstF components are visually summarized in Figure 24. Following paragraphs will describe more details of protein complex purification.

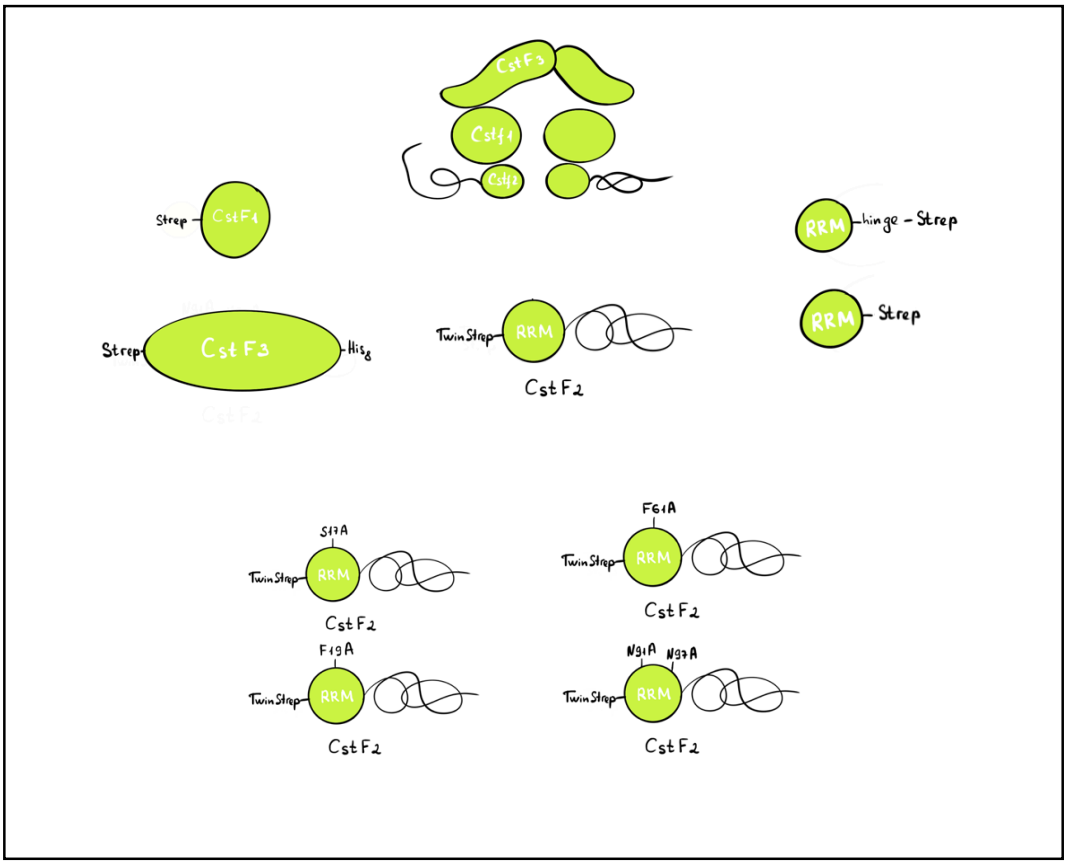


Figure 24. CstF subunit and construct scheme. Depiction of CstF subunits and constructs generated and used in this study to purify the hexameric CstF complex, distinct subcomplexes and the single RRM domain. Full-length CstF3 is N-terminally tagged with a Strep II tag and carries a C-terminal His_8 tag. Full-length CstF2 carries a N-terminal TwinStrep tag, as well as all CstF2 mutants (S17A, F19A, F61A and N91A-N97A). RRM-Hinge domain containing construct and single RRM construct carry a C-terminal Strep II tag. A N-terminal Strep II tag is fused to full-length CstF1 subunit.

2.1.1 High yield purification of full-length CstF complex for biochemical studies using a combination of affinity tag purification and Size Exclusion Chromatography

Human CstF complex forms a heterodimer consisting of three subunits: the smallest 49 kDa subunit CstF1, which is 431 amino acids long (also called CstF50; UniProt: Q05048), the 61 kDa subunit CstF2, consisting of 577 amino acids (also called CstF64; UniProt: P33240) and the largest subunit, CstF3 (also called CstF77, UniProt: Q12996), which has a molecular weight of 83 kDa and is 717 amino acids long.

Coding sequences of all three subunits were cloned into individual MuliBac^{TM} acceptor vectors (Berger, Fitzgerald et al. 2004) for baculoviral mediated expression in insect cells. A N-terminal Strep-tag II was fused in frame to the CstF1 and CstF3 subunits, although CstF3 had to be re-cloned a second time, because it showed C-terminal degradation in first purifications trials. The degradation product was identified by in-gel mass-spectrometry as being CstF3 spanning from amino acids 1-657. Consequently, an additional C-terminal His_8-tag was added in frame

after the coding sequence of CstF3. This C-terminal His_8-tag should protect CstF3 from C-terminal degradation and allowed purification of only non-degraded CstF3 by pulling on its C-terminus. The C-terminal degradation was negligible in purifications of the full-length CstF complex, but became stronger when CstF2 was missing in purification of the CstF1-CstF3 subcomplex (Paragraph 2.1.2). Human CstF2 turned out to be the least stoichiometrically expressed and purified subunit and was therefore cloned with a N-terminal TwinStrep tag, so that it would bind preferentially to StrepTactin resin compared to CstF1 and CstF3 carrying a simple Strep-tag II.

Complex formation was done by co-lysis of three cell pellets, each obtained from 1.5 L insect cell expression cultures expressing one subunit of the CstF complex. Cleared lysate containing all CstF subunits was applied to a StrepTrap column and then further purified via Heparin column, where the CstF1-CstF2-CstF3 complex eluted towards the end of a salt gradient at a concentration of around 600 mM NaCl, demonstrating CstF complex stability even in higher salt concentrations. Complex containing fractions were pooled, concentrated and injected on a Superose6i 10/300 column to perform a final Size Exclusion Chromatography (SEC) step. Full-length CstF showed a symmetric single peak in the elution profile from the SEC column, containing high amounts of pure complex as shown on the SDS PAGE in figure 25.

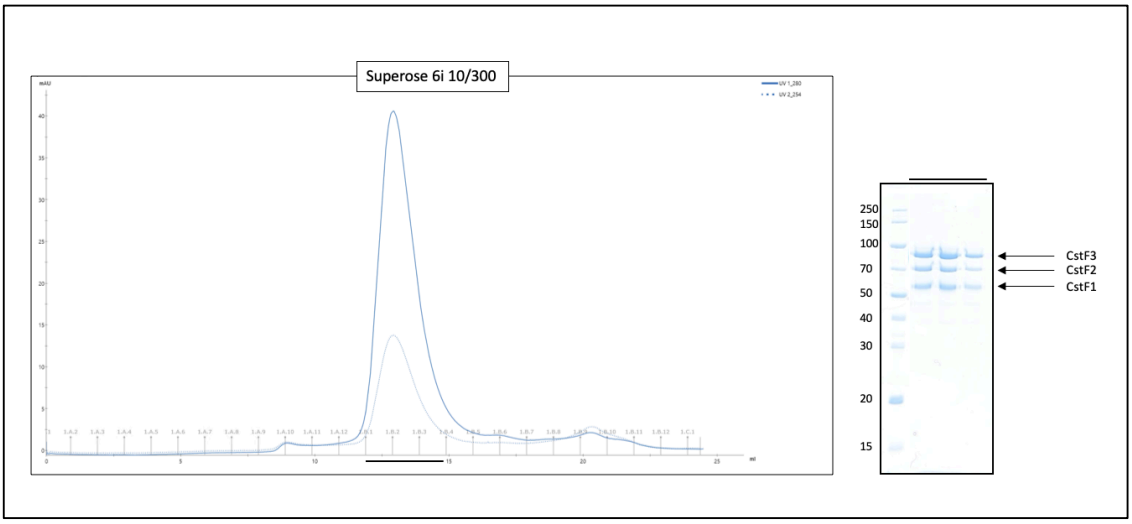


Figure 25. Purification of human CstF complex. SEC profile with A_{280} (blue line) and A_{260} (dotted line). Full-length CstF complex elutes at a retention volume of 13.5 ml in a symmetric peak (black line). Peak fractions (back line) are loaded on the SDS PAGE on the right side, containing three bands at 50 kDa (CstF1), 70 kDa (CstF2) and 85 kDa (CstF3). Lane 1: Molecular weight marker.

Furthermore, the ratio of A_{260}/A_{280} = 0.38 confirmed that the complex was free from nucleic acid contamination, which was very important for further biochemical or biophysical RNA binding experiments. After SEC, all protein containing fractions were pooled, concentrated to a high stock concentration of around 10 mg/ml and stored at -80°C for further studies. This

standard purification usually led to a final yield of 5-10 mg of pure CstF complex. The capability of full-length CstF to stably bind G/U-rich RNA species was examined briefly with analytical gel filtration. In the chromatogram depicted in figure 38 of paragraph 2.3.1, increase in the A_{260}/A_{280} ratio confirms that recombinantly purified CstF complex can bind to G/U-rich RNAs.

2.1.2 Purification of the CstF1-CstF3 subcomplex and CstF containing a C-terminal truncated version of CstF2 using a combination of His- and Strep-tag affinity purification

In full-length CstF, the 64kDa subunit CstF2 is the protein known to mediate RNA binding to G/U-rich sequence elements on the mRNA located downstream of the cleavage site (Takagaki and Manley 1997). Binding of CstF to mRNA is proposed to help definition of the cleavage site for mRNA cleavage by endonuclease CPSF3 (Chen, MacDonald et al. 1995, Mandel, Kaneko et al. 2006).

In previous studies, there was no RNA binding activity observed for CstF1 and CstF3 (Yang, Hsu et al. 2018), so that I used the CstF1-CstF3 subcomplex as control in Fluorescence Anisotropy (FA) experiments to examine the RNA binding mechanism of CstF2 in context of the full-length complex and on a single protein level. Based on available structures (pdb 2OOE, 2XZ2), CstF1-CstF3 assembly is expected to occur in a dimeric manner. First, two copies of CstF3 homodimerize via their HAT domain to the very stable HAT dimer (Bai, Auperin et al. 2007, Legrand, Pinaud et al. 2007) and can therefore assemble two copies of CstF1, since CstF3 is directly binding to CstF1. Second, the potential two CstF1 subunits can in turn homodimerize by themselves through their N-terminal homodimerization domains (Moreno-Morcillo, Minvielle-Sebastia et al 2011). Due to lack of CstF2, which was the bottleneck in terms of protein stability, the CstF1-CstF3 subcomplex showed a more stable behavior during purifications and gave more homogeneous data in Electron Microscopy (EM) studies, as described later in this thesis (Paragraph 2.5).

Purification of the CstF1-CstF3 subcomplex

For purification of CstF1-CstF3, constructs already cloned for purification of full-length CstF were used. As already described, both subunits were carrying a N-terminal Strep-tag II and a C-terminal His_8-tag was fused to CstF3 to only purify intact CstF3, because C-terminally degraded CstF3 does not contain the His-tag. When CstF2 was not present in the CstF1-CstF3 subcomplex, a clear degradation product of CstF3 was observed appearing as a third protein band in SDS PAGE analysis, running directly below CstF3 at a molecular weight of around 70 kDa (Figure 26). In-gel mass spectrometry analysis of the degradation product identified it as human CstF3 degraded from the C-terminus until amino acid 657**.** This was in line with

secondary structure prediction using the PredictProtein web server (Technical University of Berlin, Germany). This analysis showed structured parts in CstF3 ending at amino acids 645 (Arg645). Additionally, the structure of CstF3 predicted by AlphaFold (Jumper, Evans et al. 2021) depicted a highly unstructured part spanning from amino acid 661 (Gly661) to the C-terminal end of CstF3. Based on literature (Ruepp, Schweingruber et al. 2011, Yang, Hsu et al. 2018) and a structure from yeast homologs (Moreno-Morcillo, Minvielle-Sebastia et al. 2011), human CstF2 is hypothesized to bind to amino acids 595-653 of CstF3 (see paragraph 2.6.1). Since CstF2 is missing in the CstF1-CstF3 complex, the unstructured C-terminus of CstF3 after the binding region for CstF2 (residues 595-653 of CstF3) is completely unprotected and therefore easily accessible for degradation (Moreno-Morcillo, Minvielle-Sebastia et al. 2011, Yang, Hsu et al. 2018).

A detailed description of the optimized CstF1-CstF3 purification procedure is summarized in paragraph 4.2.7.2 of Material and Methods. Briefly, the subcomplex was formed by co-lysis of two pellets and a His-tag affinity step was performed after first purification step, a Strep column, to get rid of the CstF3 degradation product. Contaminants and the most of the CstF3 degradation product, were washed off with low imidazole concentrations and the target complex was eluted from the HisTrap. Fractions were pooled, concentrated and loaded on a Superose6i 10/300 column to re-buffer the sample during final SEC. The CstF1-CstF3 subcomplex eluted in high amounts in a single peak from the SEC column and was concentrated to a high stock concentration and stored at -80°C for further studies. Minor amounts of C-terminally degraded CstF3 were still present in the final SDS-PAGE (Figure 26). From a standard large-scale purification of the CstF1-CstF3 subcomplex, I obtained around 5-10 mg of pure target complex.

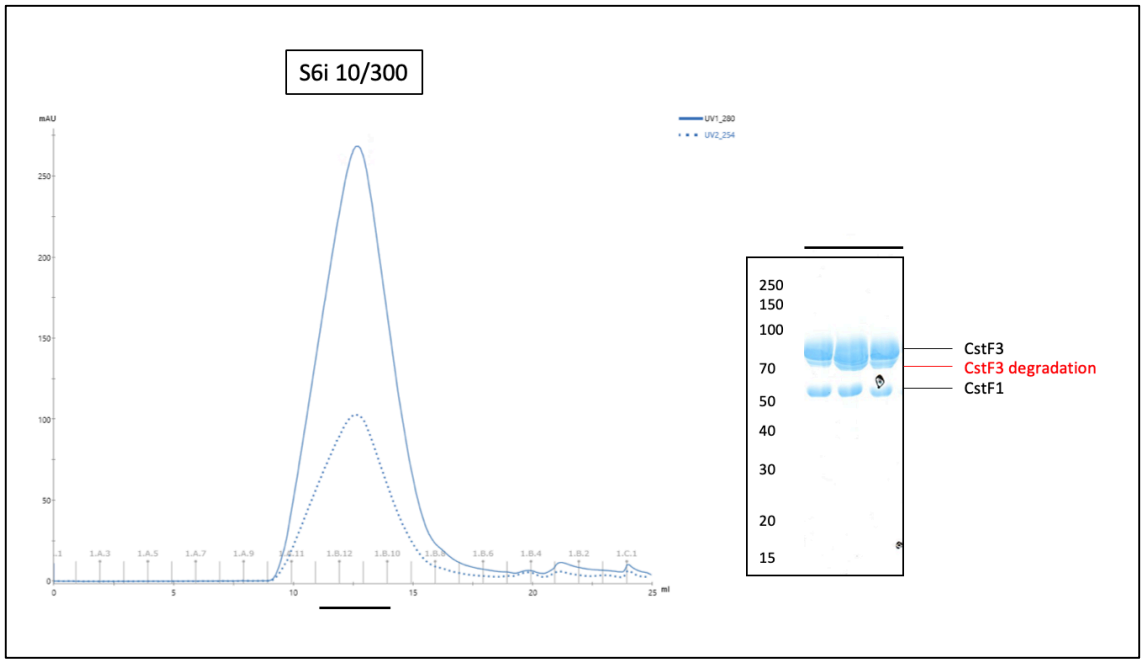


Figure 26. Purification of the human CstF1-CstF3 complex. SEC profile with A280 (blue line) and A260 (dotted line). Full-length CstF1-CstF3 complex elutes at a retention volume of 12.5 ml in a symmetric peak (black line). Peak fractions (back line) are loaded on the SDS PAGE on the right side, containing two bands at 50 kDa (CstF1) and 85 kDa (CstF3). A minor band of C-terminally degraded CstF3 was visible (red line). Lane 1: Molecular weight marker

Purification of the minimal CstF1-CstF2$^{1-204}$-CstF3 complex

To obtain a minimal CstF1-CstF2$^{1-204}$-CstF3 complex, the so-called hinge domain of CstF2 has to be present, because it binds to CstF3 (Ruepp, Schweingruber et al. 2010, Moreno-Morcillo, Minvielle-Sebastia et al. 2011, Yang, Hsu et al. 2018) and therefore indirectly mediates connection to CstF1. Since the disordered C-terminal part of CstF2 was never visible in cryo-EM studies in this thesis (Paragraph 2.4), I designed a construct of CstF2 spanning over the N-terminal RRM and hinge domain (CstF2-RH; residues 1-204). Consequently, CstF2-RH should still be able to bind RNA and form a complex with CstF1 and CstF3. This CstF1-CstF2$^{1-204}$-CstF3 complex is termed minimal CstF (short CstFdC) in this context. CstF2-RH was highly expressed in insect cells and clearly visible in pull downs. Therefore, a protocol for large-scale purification of minimal CstF was optimized in order to get a more stable complex for cryo-EM studies. CstFdC was also used for RNA binding assays, to observe if the truncated residues 205-577 of CstF2 have an indirect impact on binding capability of CstF to G/U-rich RNA species (Paragraph 2.3.4). Purification procedure followed exactly the strategy developed for the CstF1-CstF3 subcomplex. After initial Strep affinity chromatography, the sample was loaded on a HisTrap to further purify the truncated CstFdC complex. As clearly visible on the elution profile of final SEC, the target complex eluted in a single peak with a slight shoulder containing CstFdC with all three subunits present (Figure 27). The major peak contained excess of CstF1 and CstF3, which were in general over represented. To sum up, it

was possible to form the truncated CstFdC complex, but it was not as homogeneous as the full-length CstF complex, indicating that full-length proteins are necessary for complex stability and homogeneity.

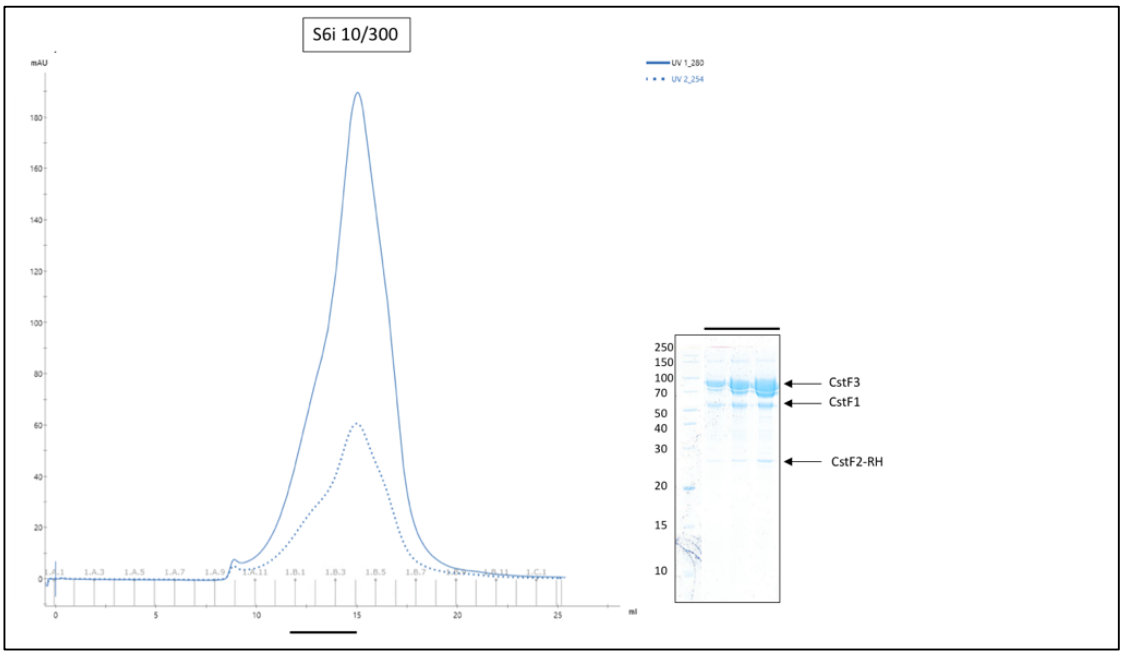


Figure 27. Purification of the minimal CstF complex. SEC profile with A_{280} (blue line) and A_{260} (dotted line). Minimal CstFdC complex elutes at a retention volume of 13 ml in a small shoulder of the main peak (black line). Peak fractions (back line) are loaded on the SDS PAGE on the right side, containing three bands at 25 kDa (CstF2-RH), 50 kDa (CstF1) and 85 kDa (CstF3). Lane 1: Molecular weight marker

2.1.3 Purification of human CstF2 derivatives and CstF2-CstF3 subcomplex using a combination of TwinStrep-tag and Heparin column

As mentioned in paragraph 2.1.2, RNA binding ability of the CstF complex derives from the 64 kDa subunit CstF2, which carries a N-terminal RRM (Takagaki, MacDonald et al. 1992). Initial information about this subunit was already given in paragraph 2.1, where the CstF2 RRM mutants were introduced. Besides binding affinities for the single RRM and a truncated CstF1-CstF2$^{1\text{-}200}$-CstF3$^{241\text{-}717}$ (Yang, Hsu et al. 2018), there is still information missing to draw a whole picture of the RNA binding behavior of full-length CstF and its binding site selection on the mRNA target.

In order to shed light on that, I purified full-length CstF2 alone and the CstF2-CstF3 subcomplex for biochemical and biophysical characterization of their RNA binding capability. Not only wild type CstF2 was purified, but also derivatives containing single or double mutations in the RNA binding domain. As already mentioned, CstF2 alone was quite difficult to handle, both during protein expression and at the stage of purification. Depending on the batch of cells used for protein expression in insect cells, CstF2 showed weaker expression

levels than the other subunits, especially in expression volumes larger than 250 ml. Initially, it was preferably expressed in Sf21 cells, but after changing virus production by using the Fugene transfection reagent (Promega, Walldorf, Germany) instead of Polyethylenimine (PEI), it was possible to express this subunit in Hi5 cells, which then delivered better expression yields. The CstF2 virus generated with PEI based transfection was either not infecting Hi5 cells or it was too weak.

Purification of TwinStrep-tagged CstF2

For purification of TwinStrep-tagged CstF2 alone, the cell pellet from a large-scale expression culture (3 L) was used and cell lysis was performed as described in Material and Methods, section 4.2.7.2. As for the other complexes, the first purification step was an affinity step based on the N-terminal TwinStrep tag of CstF2. After being bound to the stationary phase, the column was washed and the target protein was eluted. Unfortunately, CstF2 co-eluted with a tightly bound contamination band, which was visible on SDS PAGEs in an almost stoichiometric ratio and was also resistant towards high salt washing steps.

To exclude that this additional band was the C-terminal degradation product of CstF3 natively pulled down with CstF2, I performed pull down assays, where I co-lysed CstF1 and CstF2 separately and then co-lysed all three subunits to obtain full-length CstF. As clearly visible in the SDS PAGE below (Figure 28), the contaminant indicated by the red arrow, was already visible in the CstF1-CstF2 pull down and presence of CstF3 had no impact.

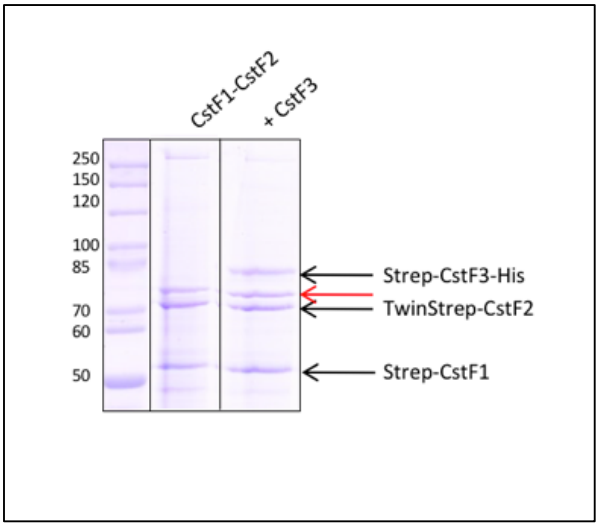


Figure 28. Pulldown of the human CstF complex. SDS PAGE of a pull down of co-lysed CstF1-CstF2 (lane 2) and co-lysed CstF1, CstF2 and CstF3 (lane 3). The contamination band running at 75 kDa (indicated by red arrow) is present in lane 2 and lane 3, therefore co-purifying with CstF2. Lane 1: molecular weight marker

The stoichiometric contamination band was identified by in-gel mass spectrometry (MS) protein identification as a protein expressed from Spodoptera frugiperda Baculovirus (SfAV) open reading frame 046 (ORF046). Since for RNA binding assays, there was need for clean and pure single CstF2, I developed a Heparin column-based strategy to separate the baculoviral protein from CstF2. Before loading the Strep elution on a cation exchange column (HiTrap Heparin 5 ml or MonoS 5ml), salt concentration of the sample was decreased to around 75 mM. Under this condition, SfAV ORF046 protein remained bound to the cation exchange column and CstF2 alone was detected in the flow through. CstF2 was then concentrated and re-buffered to working conditions for RNA binding assays. The SDS PAGE in figure 29 shows the flow through and wash fraction of the HiTrap Heparin containing high amounts of CstF2. The contaminant in complex with minor amounts of CstF2 was eluted from the Heparin column (Figure 29).

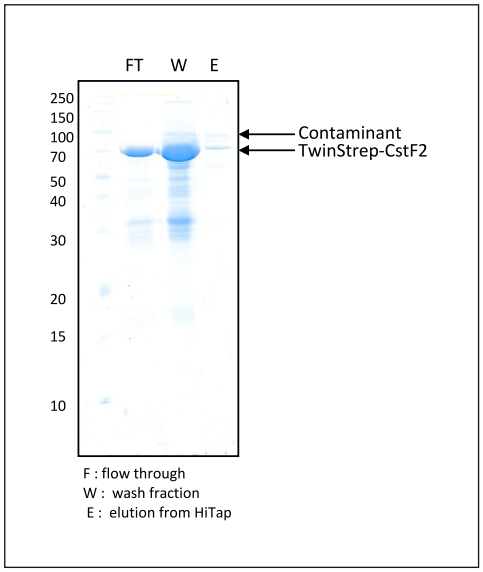


Figure 29. Purification of human CstF2. SDS PAGE of the HiTrap Heparin column shows a band at 70 kDa corresponding to human CstF2 in the flow through (FT, lane 2) and wash fraction (W, lane 3). CstF2 co-purifying with the contaminant visible in the elution (E, lane 4). lane 1: molecular weight marker

Purification of the CstF2-CstF3 subcomplex

Besides the heterodimeric CstF1-CstF3 subcomplex (see paragraph 2.1.2), another dimer consisting of CstF2 and CstF3 could be formed, since the largest 77 kDa subunit CstF3 interacts with both, CstF1 and CstF2 (Yang, Hsu et al. 2018). CstF3 binds the so-called hinge region (amino acids 112-199) of CstF2 with a stretch of roughly 60 amino acids (amino acids 594-653), which is called monkey tail (Moreno-Morcillo, Minvielle-Sebastia et al. 2011, Yang, Hsu et al. 2018). CstF2-CstF3 is theoretically expected to exist as a tetramer with two copies

of each protein, since CstF3 is homodimerizing via its HAT domain and would bridge two copies of CstF2.

Complex formation by co-lysis and the detailed purification protocol is described in section 4.2.7.2 of the Material and Methods part of this thesis. The optimized purification strategy was similar to the purification procedure of CstF2, as described in the text above. CstF2-CstF3 was also loaded on a HiTrap Heparin column after elution from the Strep Trap with a sodium chloride (NaCl) concentration of 100 mM. In contrast to CstF2 alone, CstF2-CstF3 did bind to the Heparin column, where it was eluted with a salt gradient (Figure 30). CstF2 was not as stable as in the full-length CstF complex, because it degraded after several freezing and thawing cycles of the stock solution. Therefore, proteins were directly used after purification for further tests and experiments.

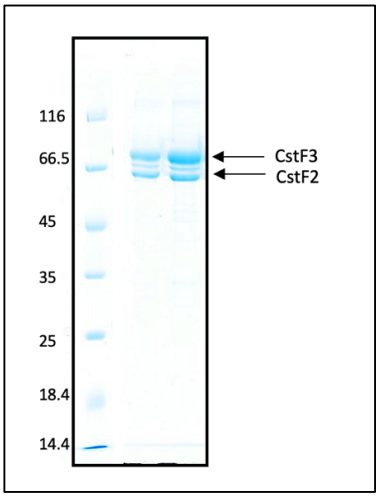


Figure 30. SDS PAGE of purified CstF2-CstF3 complex. SDS PAGE of the elution fractions from the HiTrap Heparin column show a band at 61 kDa corresponding to TwinStrep-tagged CstF and at 70 kDa corresponding to CstF3. Lane 1: molecular weight marker.

2.1.4 Optimizing purification of CstF complex for cryo-EM studies by reconstituting it with G/U-rich RNA in combination with Gradient Fixation (GraFix) and analytical Size Exclusion Chromatography

In paragraph 2.1.1, it was already described how CstF was produced to high purity and stoichiometry by an affinity purification step followed by a final SEC. Besides biochemical studies, the complex was subjected to cryo-EM data collection for single particle analysis (SPA). Initial sample screening sessions in negative stain EM and cryo-EM (Paragraph 2.4.1 and 2.4.2) showed, that samples obtained from this purification protocol were not homogeneous and stable under cryogenic conditions, which are necessary to collect cryo-EM data for structural reconstruction. Therefore, the purification protocol of full-length CstF was changed and optimized to address challenges that were faced during sample preparation and

EM studies, to obtain a more stable complex where ideally all subunits are present and visible in cryo-EM.

Screening of cross-linking conditions for GraFix of full-length CstF

CstF complex without RNA always looked disassembled on EM grids after different purification steps, including SEC. As a consequence, cross-linking (x-linking) experiments were performed to stabilize the complex for cryo-EM preparation. Initially, cross-linking reactions were performed in-batch to screen for optimal reaction conditions and different cross-linkers. Table 2 shows different cross-linking reagents as well as different conditions, that were tested on the full-length complex.

Table 2. Cross-linking screening. Screening of different cross-linkers and cross-linking conditions using 1.8-bismaleimido-diethyleneglycol (Bm(PEG)), glutaraldehyde (GA) and bis(sulfosuccinimidyl)suberate (BS3). Cross-linker concentrations are increased in steps (column two) at distinct conditions (Temperature, column three) and quenched with different reagents (column four). CstF complex is used in 1x PBS buffer.

Cross-linker	Coss-linker concentration							Temperature	Quenching
µM Bm(PEG)	1	5	10	20	40	60	80	5min @ 30°C	10mM DTT
µM Bm(PEG)	40	60	80	100	120	150	/	60min, RT	10mM DTT
% GA	0.001	0.005	0.01	0.05	0.1	0.2	0.3	5min @ 30°C	25mM Tris
% GA	0.01	0.02	0.03	0.04	0.05	/	/	20min @ RT	25mM Tris
µM Bm(PEG) + mM BS3	20 2	40 2	60 2	80 2	100 2	120 2	/ /	**60min @ RT**	10mM DTT
mM BS3	0.02	0.1	1	1.5	2	2.5	3	30min @ 26°C	25mM Tris
CstF @ 0.5µM	1x PBS buffer								

BS3 and Glutaraldehyde showed the best results in in-batch cross-linking (see figure 31 A and B) and were used for all further sample preparations. Unfortunately, in-batch cross-linking of the CstF complex purified via SEC did not result in any improvement in sample stability (see section 2.4.2). Therefore, I decided to replace the gelfiltration with a density gradient ultracentrifugation step. This technique can be combined with cross-linking approaches (Gradient-Fixation), as shown by Stark et al., 2010. For the Gradient-Fixation (GraFix) step, I initially used glutaraldehyde, because this cross-linker was used in the original method developed by Stark et al., 2010.

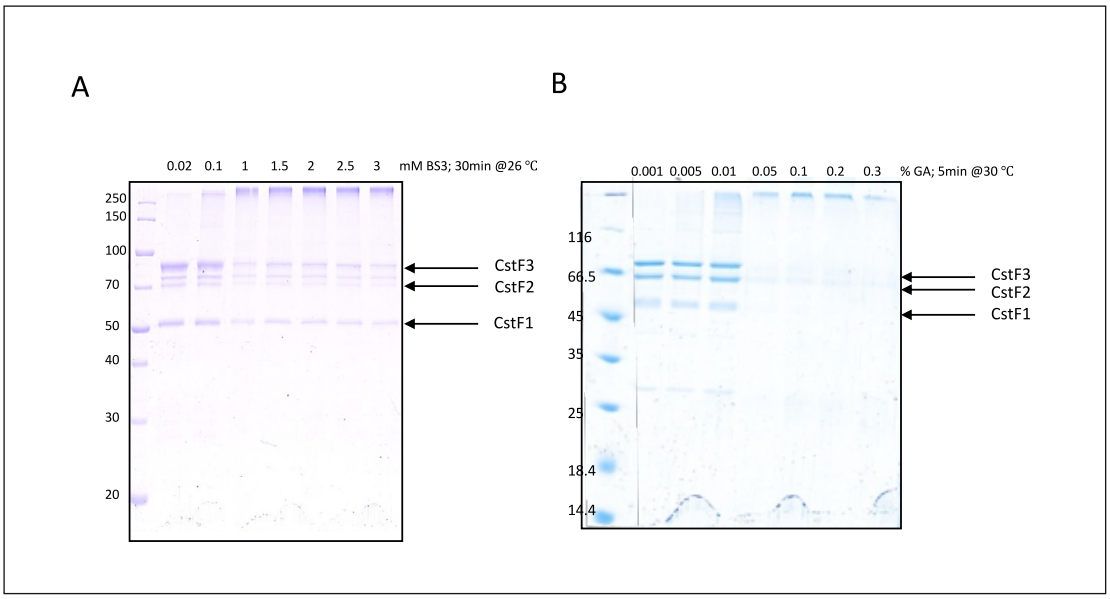


Figure 31: BS3 and GA cross-linker screening. SDS PAGES of screening BS3 and GA in different concentrations. A). Full-length CstF complex is cross-linked with BS3 in concentrations from 0.02 to 3 mM for 30 min at 26 ℃. The SDS-PAGE shows three bands at 50 kDa (CstF1), 70 kDa (CstF2) and 85 kDa (CstF3) and a fourth band shifted to higher molecular weight corresponding to the cross-linked complex. Lane 1: molecular weight marker. B) Full length CstF complex is cross-linked with GA in concentrations from 0.0001 to 0.3 % for 5 min at 30 ℃. The SDS-PAGE shows three bands at 50 kDa (CstF1), 70 kDa (CstF2) and 85 kDa (CstF3) and a fourth band shifted to higher molecular weight corresponding to the cross-linked complex. Lane 1: molecular weight marker

As observed on the SDS PAGE in figure 31 B, cross-linking of the CstF complex was initiated at a GA concentration of 0.01% in-batch, so I decided to use the same concentration for the GraFix procedure. Sucrose density gradient tubes were prepared as described in section 4.2.7.3 and two tubes were used for every ultracentrifugation. One of them contained the cross-linking reagent applied to the 25% sucrose solution and the other one was a classical 5%-25% sucrose gradient tube.

Purification of CstF via GraFix

First steps of the purification protocol remained the same as already described for full-length CstF in paragraph 2.1.1. After elution from the StrepTrap, the complex was now concentrated and roughly 200 µl of protein at a concentration of 10 – 15 mg/ml were carefully loaded on each of the sucrose tubes. After over-night ultracentrifugation, gradients were fractionated and fractions were analyzed by SDS PAGE. Cross-linked sample and non-crosslinked complex eluted in almost the same fractions, indicating that the overall complex composition was similar. Unfortunately, the elution from sucrose gradient could not directly be used for cryo-EM studies, because the sucrose content in the sample was too high. Sucrose concentrations of more than 3% in a protein sample would result in higher background noise on cryo-EM images. Therefore, I introduced an analytical SEC step after the GraFix and density gradient to

exchange buffer and thereby remove sucrose from the sample. Both, the sample from the gradient and the one from GraFix eluted in a single peak from analytical S6i. However, the elution profile of the GraFix sample showed a slight shoulder (Figure 32 B), which could correspond to over cross-linked complex. The SDS PAGE of non-cross-linked CstF showed all three subunits present in stoichiometric amounts as depicted in figure 32 A. The SDS PAGE from analytical SEC of the GraFix sample (Figure 32 B) showed a band shifted towards higher molecular weight, corresponding to cross-linked complex.

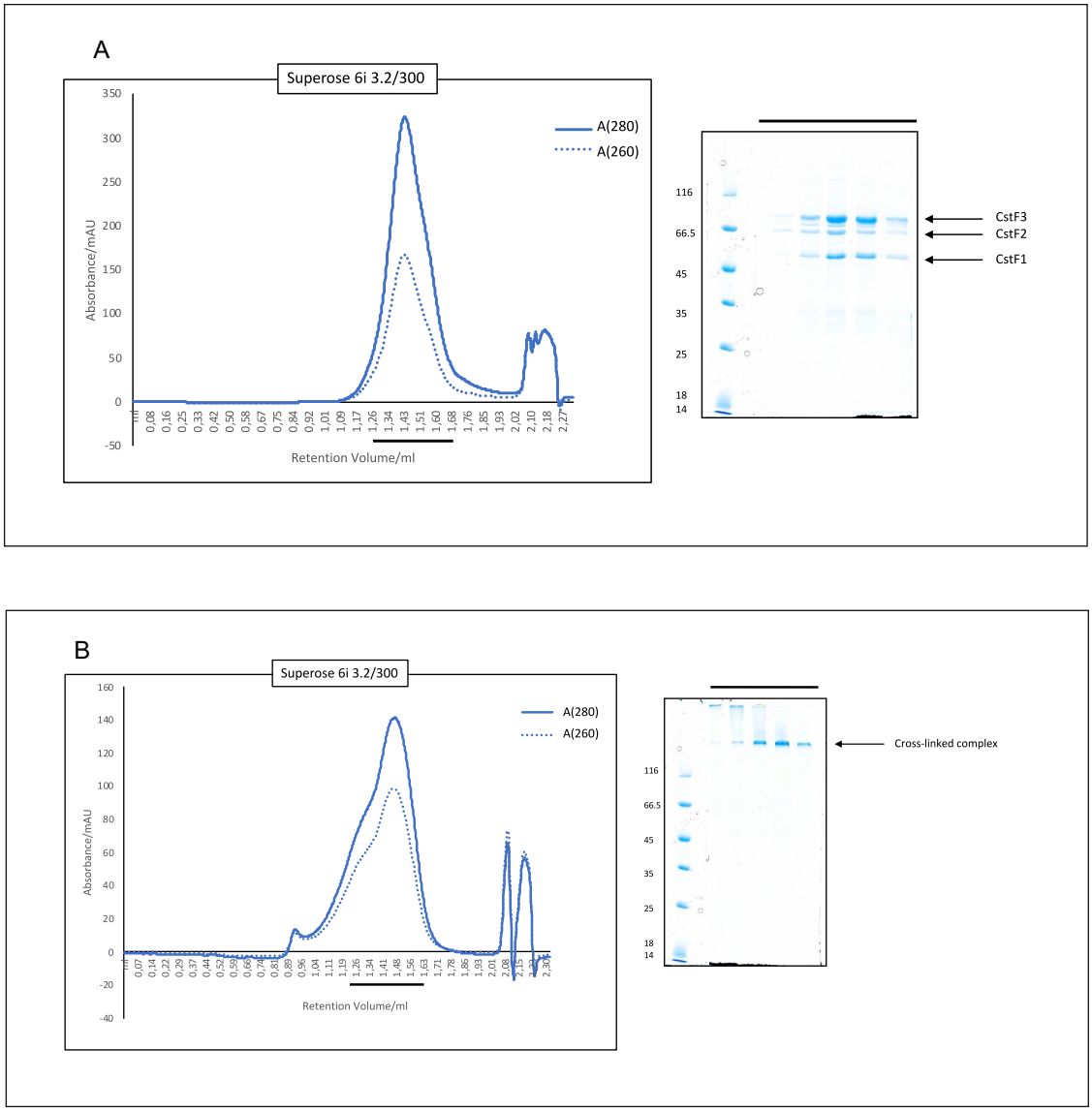
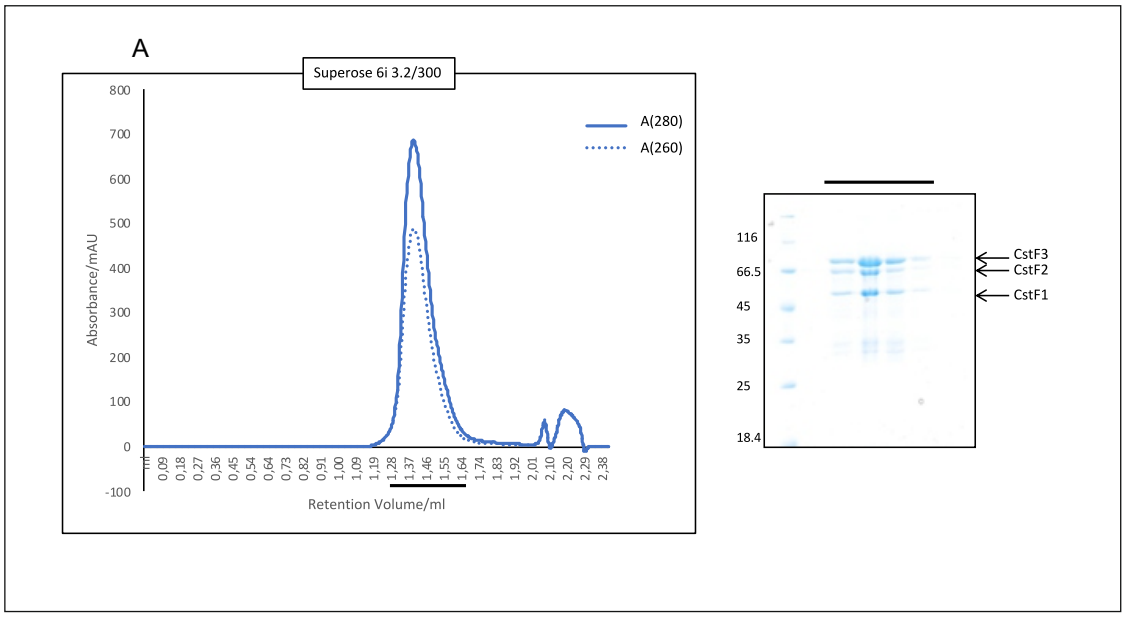


Figure 32. Purification of human CstF complex via sucrose density gradient and GraFix. Elution profile of the analytical SEC and corresponding SDS PAGE on the right of the non-cross-linked CstF (A) and the GraFix CstF (B). A) SEC profile with A280 (blue line) and A260 (dotted line). CstF complex elutes at a retention volume of 1.45 ml in a symmetric sharp peak (black line). Peak fractions (back line) are loaded on the SDS PAGE on the right side, containing three bands at 50 kDa (CstF1), 70 kDa (CstF2) and 85 kDa (CstF3). Lane 1: Molecular weight marker. B) SEC profile with A280 (blue line) and A260 (dotted line). Cross-linked CstF complex elutes at a retention volume of 1.45 ml in the main peak (black line) with a small shoulder before the main peak. Peak fractions (back line) are loaded on the SDS PAGE on the right side, containing one band shifted to high molecular weight corresponding to the cross-linked complex. Lane 1: Molecular weight marker

Purification of RNA-bound CstF via GraFix

Since GraFix alone did not deliver the desired results in cryo-EM screening data collection (described in detail in paragraph 2.4.3), I introduced a second step in order to improve compositional and conformational homogeneity of the sample. As already mentioned in paragraph 2.1.2, CstF2 binds to G/U-rich DSEs downstream of the cleavage site (MacDonald, Wilusz et al. 1994, Beyer, Dandekar et al. 1997, Perez Canadillas and Varani 2003, Pancevac, Goldstone et al. 2010). Thus, reconstitution of CstF with G/U-rich RNA could help to stabilize the sample for grid preparation. In section 2.3.1, I showed that intact CstF was bound with high affinity to the so-called *CstF01* RNA and I wanted to test, whether this RNA would serve as a binding platform to stabilize the CstF2 subunits in a certain conformation for cryo-EM studies.

To purify the CstF complex with RNA, the initial Strep-tag based affinity purification step was performed analogously to the description in 2.1.1. After elution from the Strep column, the sample was concentrated to high stock concentration of around 10 mg/ml. 400 µl of the concentrated complex were incubated on ice with a two-fold excess of *CstF01* RNA and the reconstituted complex was loaded afterwards on two sucrose density gradient tubes, one containing the cross-linking reagent as described above. RNA-containing CstF, cross-linked and not cross-linked, was collected in the same fractions of the gradient as wild type CstF. Peak fractions were pooled and concentrated to be loaded on the analytical gelfiltration column to remove sucrose from the sample. Both samples eluted in a single peak, with an increase in the A_{260}/A_{280} ratio compared to apo CstF, indicating that RNA was bound to the complex (Figure 33 A). Again, the elution profile of the GraFix sample had a slight shoulder before the main peak, which might be cross-linking artefact (Figure 33 B).



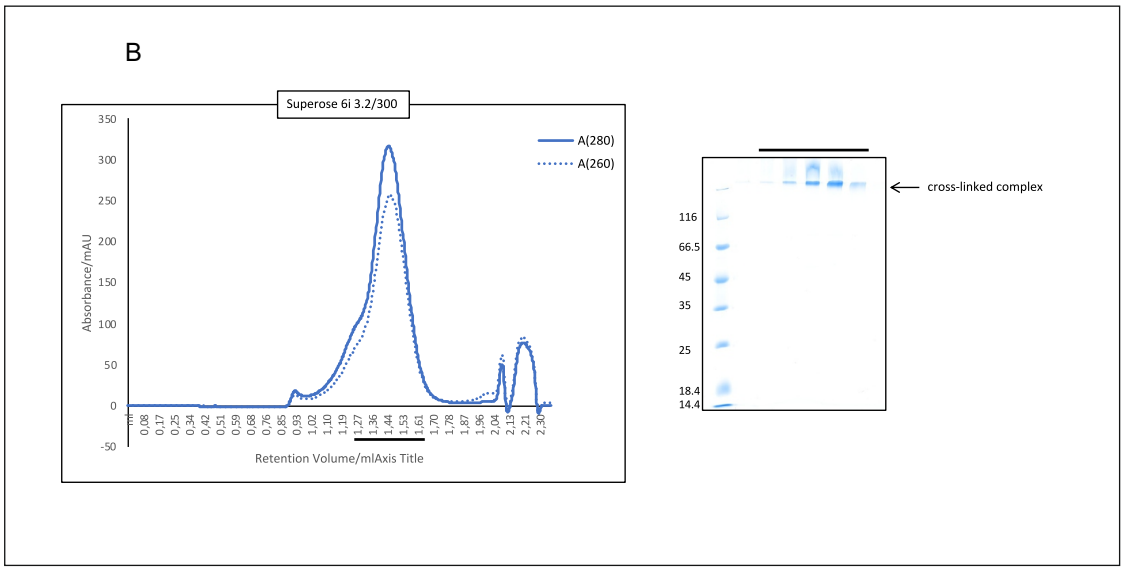


Figure 33. Purification of RNA bound CstF complex via sucrose density gradient and GraFix. Elution profile of the analytical SEC and corresponding SDS PAGE on the right of non-cross-linked CstF (A) and the GraFix CstF (B) both bound to RNA. A) SEC profile with A_{280} (blue line) and A_{260} (dotted line). Hexameric CstF complex elutes at a retention volume of 1.45 ml in a symmetric sharp peak (black line) with increased A_{260} (dotted line), indicating that RNA is bound. Peak fractions (back line) are loaded on the SDS PAGE on the right side, containing three bands at 50 kDa (CstF1), 70 kDa (CstF2) and 85 kDa (CstF3). Lane 1: Molecular weight marker. B) SEC profile with A280 (blue line) and A260 (dotted line). Cross-linked CstF complex elutes at a retention volume of 1.45 ml in the main peak (black line) with a small shoulder before the main peak and increased A_{260} (dotted line), indicating that RNA is bound. Peak fractions (back line) are loaded on the SDS PAGE on the right side, containing one band shifted to high molecular weight corresponding to the cross-linked complex. Lane 1: Molecular weight marker

2.1.5 High-yield purification of the CstF1-CstF3 subcomplex for cryo-EM high-resolution data collection using an optimized density-gradient-ultracentrifugation based cross-linking protocol

So far, there are only structures of parts or single domains of the CstF complex available, but the overall structure of full-length CstF is still not solved. Due to its molecular weight of 385 kDa, the complex is a suitable target for single particle analysis by cryo-EM.

In this thesis, sample preparation for EM studies was accompanied by regular sample screenings by negative stain EM after all steps of purification and all different purification trials. Once the sample looked reasonable in negative stain EM, first screening sessions in Cryo-EM were performed. The protein complex behaved different under cryogenic conditions, because it was completely dissociating for conditions optimized in negative stain screening (see paragraph 2.4.1 and 2.4.2). Additionally, full-length CstF1-CstF2-CstF3 turned out to be too heterogeneous in cryo-EM studies (see paragraph 2.4). Instead, the more stable CstF1-CstF3 subcomplex was used for further sample optimization and for final data collection (see

paragraph 2.5). The final purification protocol (see Material and Methods 4.2.7.1) for the CstF1-CstF3 complex for single particle analysis will now be briefly discussed.

Similar to purifications described above, CstF1-CstF3 was formed by co-lysing two separately expressed cell pellets for CstF1 and CstF3. After elution from the initial StrepTrap, there was no need to include a Heparin column, because the CstF1-CstF3 subcomplex did not show any nucleic acid contamination during the whole purification procedure. Concentrated eluate from the first purification step was now cross-linked in-batch with BS3, before being loaded on a sucrose gradient tube for further sample preparation by density gradient ultracentrifugation. After centrifugation overnight in a swingout rotor, partially cross-linked CstF1-CstF3 subcomplex eluted in a single peak from the gradient. Non-cross-linked subcomplex was loaded on a second tube as a control, migrating similar to its cross-linked derivate in the density gradient. A final analytical SEC step was introduced in the purification procedure to exchange the sucrose buffer to the optimized sample buffer for cryo-EM grid preparation. The concentration of CstF1-CstF3 to be loaded on analytical gel filtration column, was chosen in such a way, that the eluted fractions could directly be used for grid preparation without further concentration or dilution steps (Figure 34).

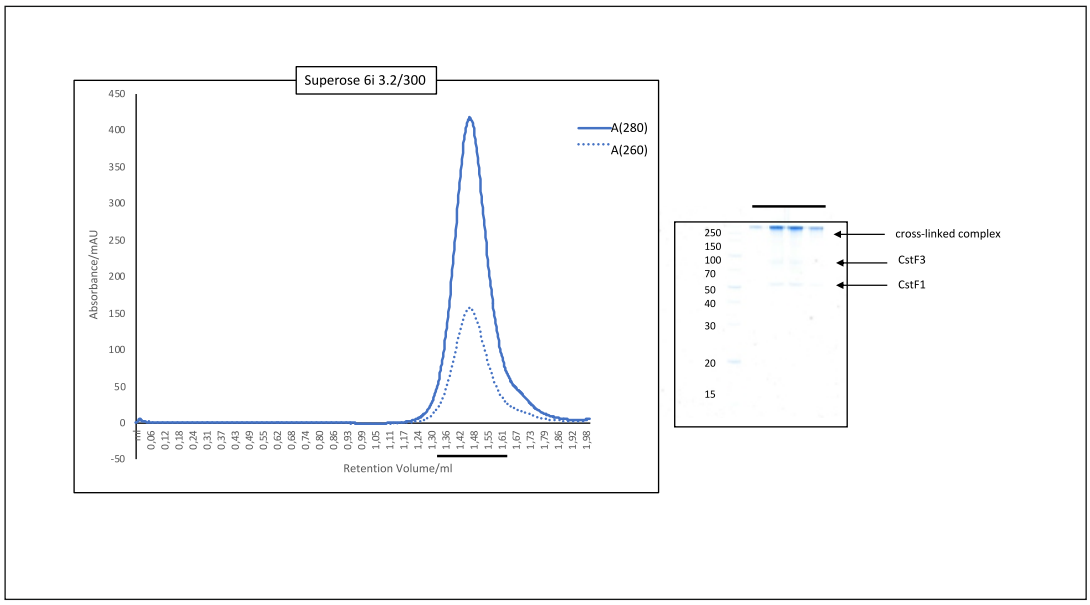


Figure 34. Purification of the CstF1-CstF3 subcomplex via a combination of in-batch BS3 cross-linking and sucrose density gradient ultracentrifugation. SEC profile with A_{280} (blue line) and A_{260} (dotted line). Cross-linked CstF1-CstF3 complex elutes at a retention volume of 1.42 ml in a symmetric sharp peak (black line). Peak fractions (back line) are loaded on the SDS PAGE on the right side, containing three bands at 50 kDa (CstF1), 85 kDa (CstF3) and at high molecular weight (cross-linked complex). Lane 1: Molecular weight marker.

2.2 Generation and purification of recombinant human CstF2 RNA recognition motifs from bacterial expression system

As already mentioned in paragraph 2.1.2, the 64kDa subunit of CstF, CstF2, contains a N-terminal RRM, which binds to U-/GU-rich downstream element on mRNAs to take part in cleavage site definition (Chen, MacDonald et al. 1995).

To get additional insights in the RNA binding mechanism of the CstF2 RRM domain, I designed various mutations of amino acids located in the RNA-binding interface of the RRM, which will be described in detail in paragraph 2.3.6. Mutations are named by the original amino acids, position and the mutated amino acid name and are summarized in figure 35.

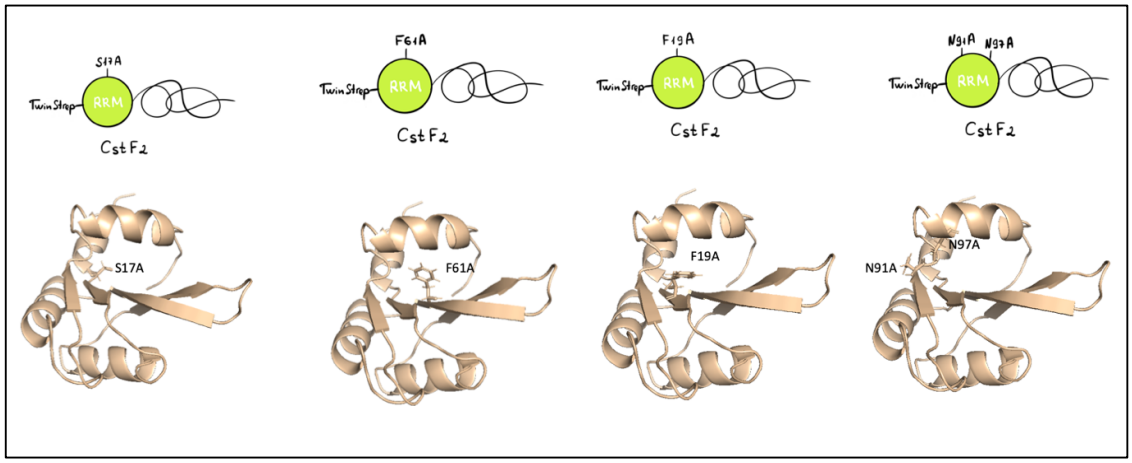


Figure 35: CstF2 RRM mutants. Top row: Four different mutants of RRM domain of full-length CstF2 were generated. Mutants are named with name of the original amino acid, position and name of the mutated amino acid in one letter code. Bottom row: NMR structure (pdb 1P1T) of the CstF2 RRM domai carrying corresponding mutations (Perez-Canadillas and Varani, 2003).

Point mutations in the CstF2 coding sequence were obtained by site directed mutagenesis using the TwinStrep-tagged CstF2 construct for expression in insect cells as template. Expression levels of CstF2 mutants were similar to wild type proteins.

2.2.1 Purification of CstF carrying CstF2 RRM mutations using a combination of Strep- and His-tag affinity purification

CstF containing mutated CstF2 ($CstF2^{mut}$) was prepared the same way as the wild type complex by co-lysing three pellets from around 1.5 L of insect cells, each expressing one subunit of the complex. The detailed purification protocol is described in section 4.2.7.2, but in general it followed the procedure optimized for the CstF1-CstF3 subcomplex with minor exceptions or changes. Usually, CstF1-$CstF2^{mut}$-CstF3 (CstFmut) was already very clean and stoichiometric after the Strep column, so that CstFmut was directly concentrated to high stock concentrations and re-buffered to the desired working buffer for downstream experiments. If

strong contaminants were present on the SDS PAGE, the complex was loaded on a HisTrap column directly after elution from the Strep affinity column. The HisTrap column was then extensively washed with low imidazole concentrations to remove contaminants and the CstF3 degradation product, which sometimes co-migrated with the complex over the Strep column. Similar to wild type CstF, mutated CstF was eluted from the HisTrap with imidazole as a stable complex (see elution profile and SDS PAGE for CstF-F19A in figure 36), containing all three subunits.

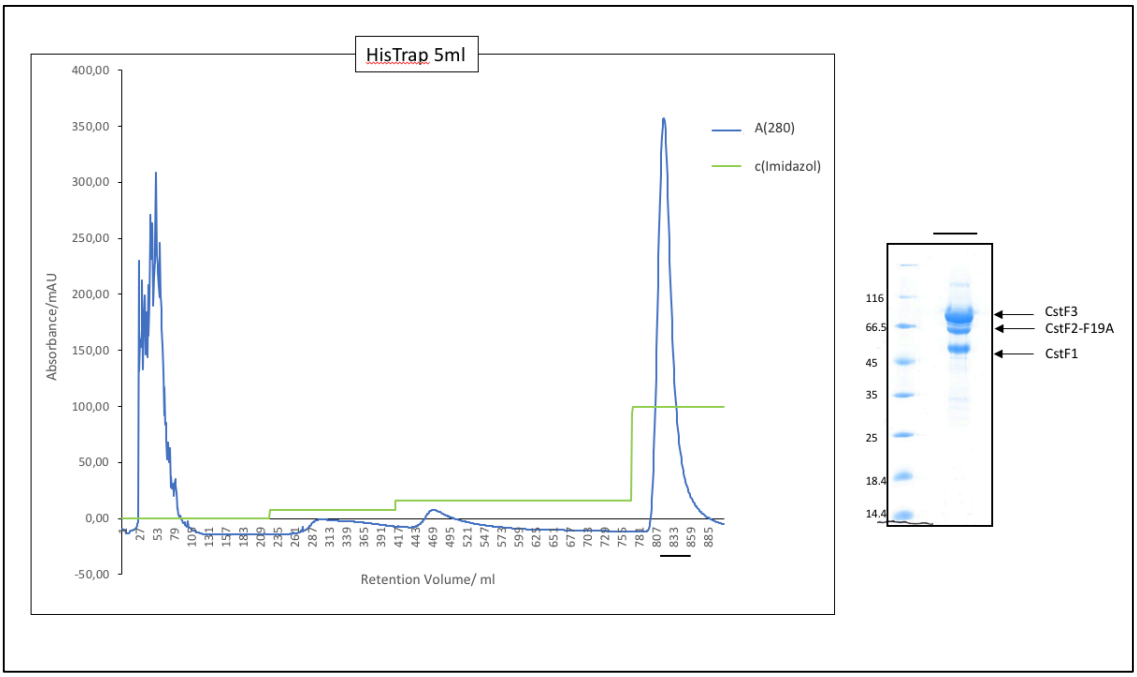


Figure 36. Purification of the human CstF complex carrying the F19A mutation in its CstF2 subunit. HisTrap elution profile with A_{280} (blue line) and concentration of imidazole (green line). Imidazole concentration is increased in steps from 0 mM, to 8mM to 16mM and finally 250mM. Mutated CstF complex elutes at a retention volume of 83.3 ml in a symmetric sharp peak (black line). Peak fraction (black line) is depicted on the SDS PAGE on the right side, containing three bands at 50 kDa (CstF1), 70 kDa (CstF2-F19A) and 85 kDa (CstF3). Lane 1: Molecular weight marker.

Mutations in the CstF2 RRM did not seem to influence stability of the CstFmut complex, since it could be concentrated to high concentrations. Re-buffered stocks were stored at -80 °C for RNA binding studies. Final SDS gels of all purified mutated CstF complexes are depicted in figure 37.

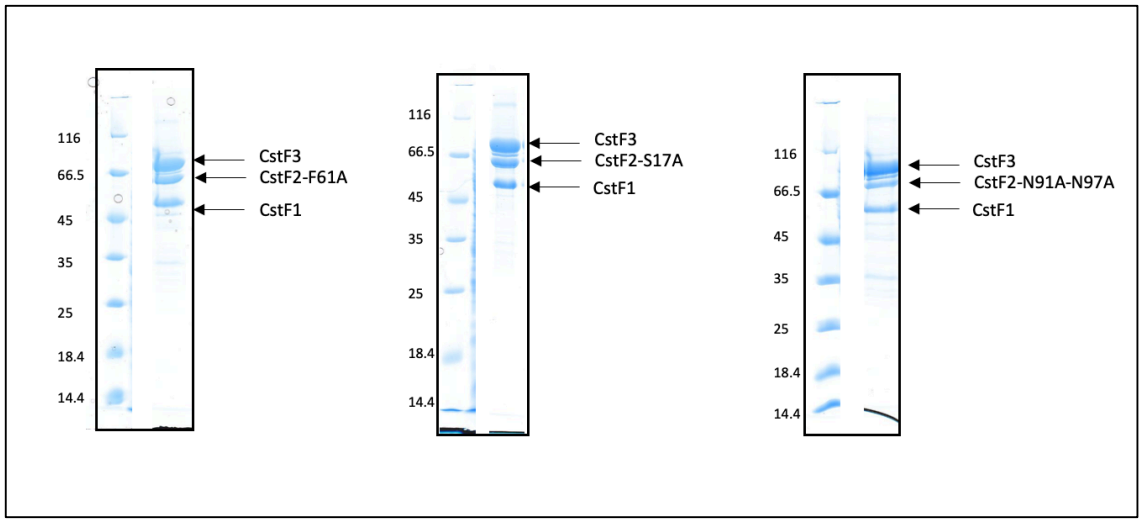


Figure 37. Purification of all mutated CstF complexes. Peak fraction of each HisTrap column is shown for each mutated complex on the corresponding SDS PAGE. Each SDS PAGE contains three bands at 50 kDa (CstF1), 70 kDa (CstF2mut) and 85 kDa (CstF3). Lane 1: Molecular weight marker.

2.2.2 Generation and purification of recombinant human CstF2-RNA binding motifs from bacterial expression system

In parallel to studying the RNA binding behavior of the full-length CstF complex or its subcomplexes, containing full-length CstF2 (wildtype or mutant), I also cloned the RRM of CstF2 alone for studies of the single protein domain. The N-terminal RRM of CstF2 has a molecular weight (MW) of around 12.5 kDa and adopts a typical fold for RRMs (Nagai, Oubridge et al. 1990, Varani and Nagai 1998, Nagaike and Manley 2011). So far, there is still biochemical data missing about the RNA binding mechanism and sequence preference of the CstF2 RRM domain.

Purification of the single RRM domain of CstF2

For RNA binding studies using CstF2 RRM, I designed a construct based on its structure solved by NMR (Perez Canadillas and Varani 2003) and cloned it with a C-terminal Strep-tag into pET-vectors for expression in *E.coli.* The single RRM domain showed high expression levels in bacterial expression systems and high amounts of cell pellet could be generated from one large-scale expression (3L) of *E.coli* culture. In contrast to all purification strategies so far, HEPES-based buffers were used for cell lysis and all purification steps to purify CstF2 RRM. Under this condition, single RRM was stably bound to the Strep column and was eluted in very high amounts and purity. In final SEC, few contaminants which were still present after the Strep column, were separated from the target protein (Figure 38). Additionally, the buffer was exchanged during SEC to the working buffer for downstream experiments. The RRM domain eluted in a symmetric peak from SEC. Peak fractions were pooled and concentrated to generate protein stocks with concentrations from 10 – 20 mg/ml.

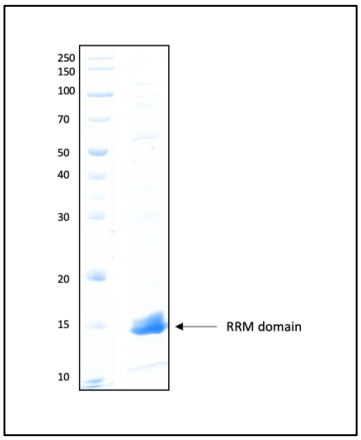


Figure 38. Purification of the CstF2 RRM domain. Peak fraction of the final SEC is shown on the SDS PAGE, containing a thick band at around 15 kDa corresponding to the CstF RRM domain (MW of single RRM 12.5 kDa). Lane 1: Molecular weight marker.

Purification of an RRM fusion construct to mimic dimeric complex assembly

As already described, the CstF complex is supposed to exist as a heterodimer (Moreno-Morcillo, Minvielle-Sebastia et al. , Bai, Auperin et al. 2007). This means, it is arranged in a 2:2:2 stoichiometry and assembles two copies of each subunit, thereby potentially influencing CstF RNA binding behavior and affinity. If two copies of CstF2 were present in the complex, this would lead to the presence of two RRMs as well, which is thought to increase RNA binding affinity (Yang, Hsu et al. 2018). In previous studies, only RNA binding affinity of one RRM domain binding to a set of different G/U-rich RNA templates was measured (Perez Canadillas and Varani 2003, Pancevac, Goldstone et al. 2010, Yang, Hsu et al. 2018). One difficulty of the experimental setup using only one RRM to determine binding affinities is, to ensure that one, not more RRMs, assembles on the RNA. In order to exclude, that one single RRM does not ‘slide’ on the GU-stretch, but is stably bound in one position, I tried to mimic the dimeric arrangement of RRMs by designing a fusion construct of two single RRMs separated by a short Glycine-Serine-linker (GS-linker). Both, two wildtype RRMs and RRMs carrying the S17A mutation were fused together. Similar to constructs designed for expression and purification of the single RRM, an in-frame Strep-tag was added to the C-terminus of the second RRM. Additionally, a 3C cleavage site was cloned in between the coding sequence of the last RRM and Strep-tag coding sequence. Two RRMs fused together including GS linker and C-terminal Strep-tag, have a molecular weight of roughly 25 kDa. In total, there were four fusion constructs generated, which are summarized in table 3 and visualized in figure 39 for better understanding.

Table 3. Fusion constructs of either wild type CstF2 RRM or RRM carrying the S17A mutation. Two RRMs are fused together by a GS-linker. Four fusion constructs were generated containing either two wild type RRMs, two mutated RRMs (S17A mutation, called single mutant) or either the first or the second RRM mutated (S17A, called hybrid)

Construct name	RRM1	RRM2	comment
RRM-GS-RRM	Wild type	Wild type	Wild type
RRM-GS-RRM(S17A)	Wild type	S17A	Hybrid
RRM(S17A)-GS-RRM	S17A	Wild type	Hybrid
RRM(S17A)-GS-RRM(S17A)	S17A	S17A	single mutant

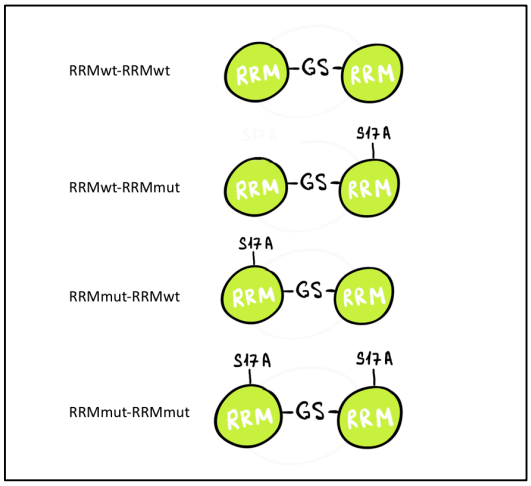
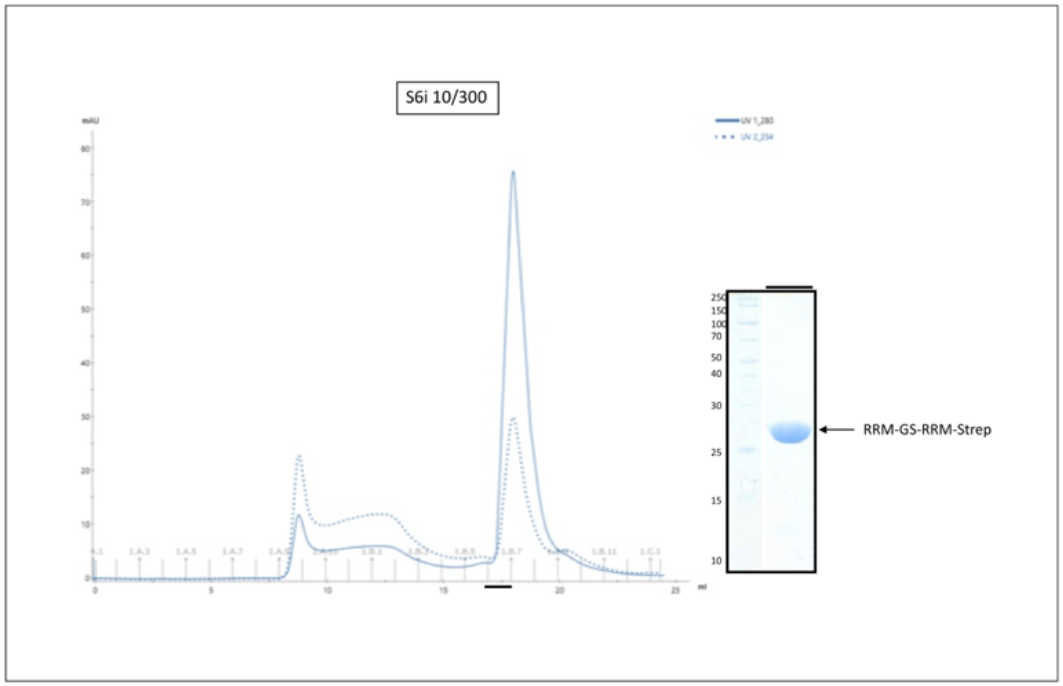


Figure 39: CstF2 RRM fusion constructs. Four different fusion constructs of the RRMs were generated by fusing either two wild type RRMs (row 1), one wild type and one mutated RRM (row 2 and row 3) or two mutated RRMs (row 4) together. For simplification, a RRM domain carrying the S17A mutation is called RRMmut.

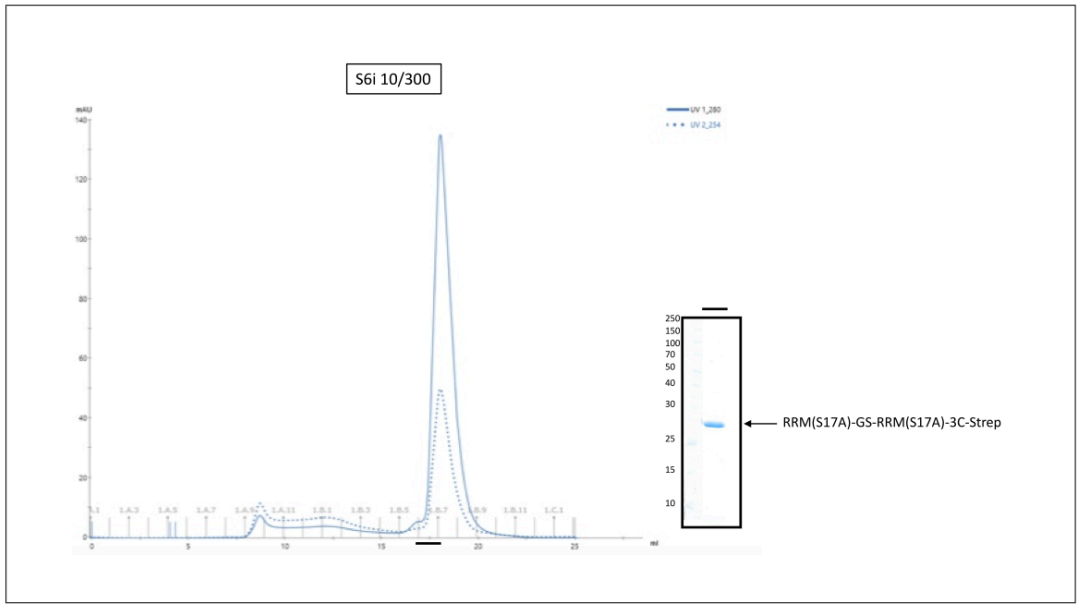
Fusion constructs of wildtype and mutated RRMs showed high expression levels in bacterial expression system. From one 3 L expression culture, I could obtain enough protein for biophysical assays and biochemical characterization.

The purification protocol of wildtype RRMs was very similar to the one developed for the single RRM construct. However, different buffers were used for cell lysis and purification (Paragraph 4.2.7.2). As already mentioned, fusions of CstF2 RRMs were also expressed in *E.coli*, so cell lysis and following steps until the Strep-tag based affinity purification step were performed in a manner similar to the single RRM. However, the washing protocol of the StrepTrap was extended with one high salt wash step (1 M NaCl) to remove any contaminants. RRM fusion constructs showed a high ratio of absorbance at 260 nm to absorbance at 280 nm (A_{260}/A_{280}) after elution from the Strep column. This indicated nucleic acid contamination, which was removed in final SEC. Contaminants were clearly separated from pure CstF2 RRMs, eluting in a single peak from the column with a reasonable ratio of A_{260}/A_{280} (see figures 40 A-D).

A


S6i 10/300
RRM-GS-RRM-Strep

B


S6i 10/300
RRM(S17A)-GS-RRM(S17A)-3C-Strep

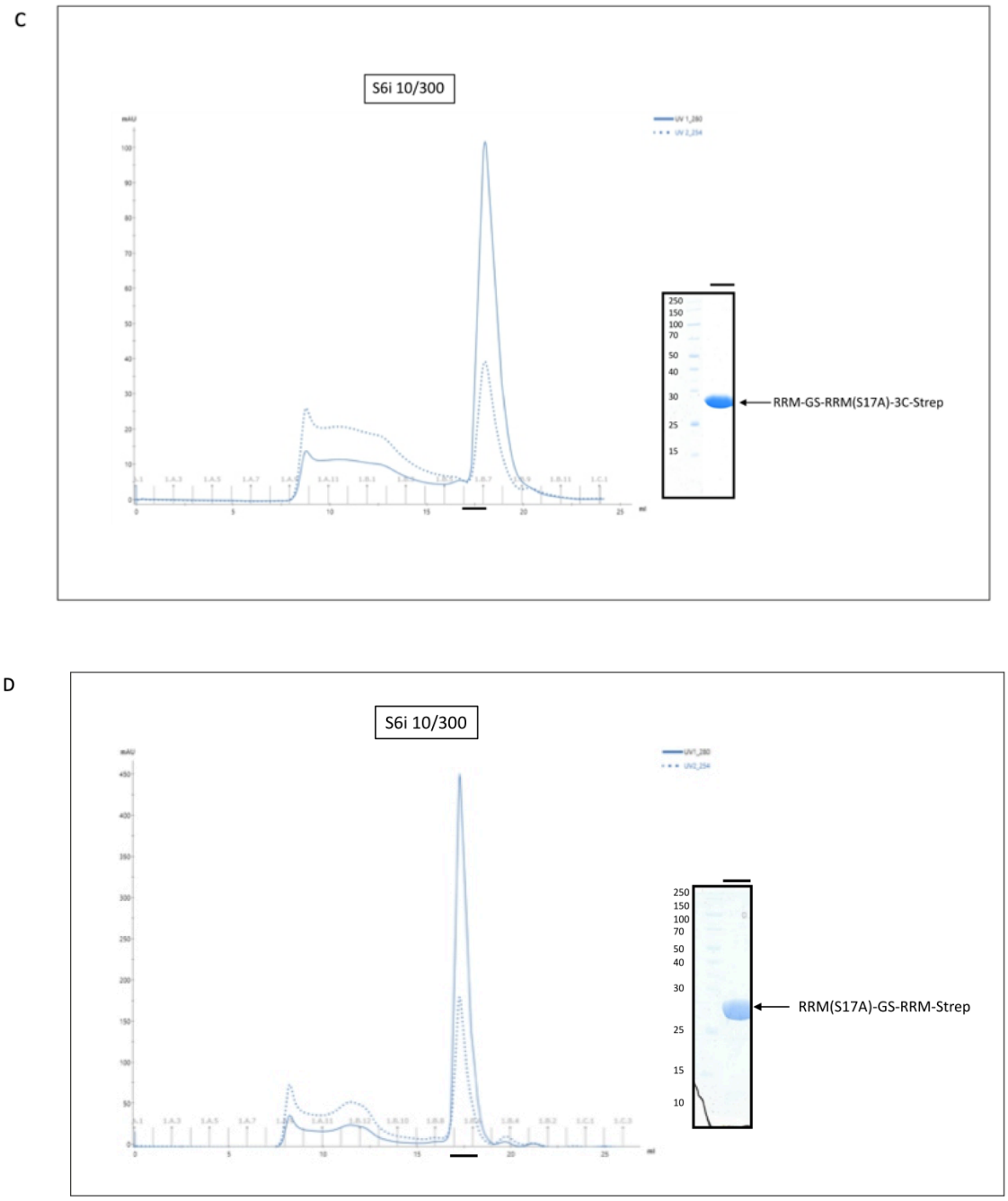


Figure 40. Purification of CstF2 RRM fusion constructs. Elution profile of SEC and corresponding SDS PAGE on the right of RRM-RRM (A), RRM(S17A)-RRM(S17A) (B), RRM-RRM(S17A) (C) and RRM(S17A)-RRM fusion (D). A) SEC profile with A_{280} (blue line) and A_{260} (dotted line). RRM-RRM construct elutes at a retention volume of 17.5 ml in a sharp peak (black line). Peak fraction (back line) is loaded on the SDS PAGE on the right side, containing a thick band at 27.5 kDa (RRM fusion). Lane 1: Molecular weight marker. B) SEC profile with A_{280} (blue line) and A_{260} (dotted line). RRM(S17A)-RRM(S17A) construct elutes at a retention volume of 17.5 ml in a sharp peak (black line). Peak fraction (back line) is loaded on the SDS PAGE on the right side, containing a thick band at 27.5 kDa (RRM fusion). Lane 1: Molecular weight marker. C) SEC profile with A_{280} (blue line) and A_{260} (dotted line). RRM-RRM(S17A) construct elutes at a retention volume of 17.5 ml in a sharp peak (black line). Peak fraction (back line) is loaded on the SDS PAGE on the right side, containing a thick band at 27.5 kDa (RRM fusion). Lane 1: Molecular weight marker. D) SEC profile with A_{280} (blue line) and A_{260} (dotted line). RRM(S17A)-RRM construct elutes at a retention volume of 17.5 ml in a sharp peak (black line). Peak fraction (back line) is loaded on the SDS PAGE on the right side, containing a thick band at 27.5 kDa (RRM fusion). Lane 1: Molecular weight marker.

Peak fractions containing a fusion of two RRMs were pooled and concentrated to generate protein stocks with very high concentrations of 20 mg/ml, to be stored at -80°C. Since wild type RRMs and mutants expressed quite well, I could easily obtain 5-10 mg protein out of one purification.

2.3 Biochemical analysis of RNA binding behavior of CstF complex

CstF is involved in definition of the cleavage site by binding to sequence elements on the mRNA, which are located 10-30 nucleotides (nt) downstream of the cleavage site (poly(A) site). DSEs have a high content in U/GU nucleotides (MacDonald, Wilusz et al. 1994, Beyer, Dandekar et al. 1997, Takagaki and Manley 1997). Binding of CstF to mRNA is mediated via the N-terminal RRM domain of CstF2, which can bind to U/GU-rich sequence elements with low μM affinity. G/U specificity is mediated by the ability to discriminate between G/U and A/C nucleotides (Perez Canadillas and Varani 2003, Deka, Rajan et al. 2005, Pancevac, Goldstone et al. 2010, Yang, Hsu et al. 2018).

In order to check, if protein samples, which were prepared for structural studies by Electron Microscopy (EM), were capable of binding to G/U-rich RNA species, I designed various RNA oligos based on early SELEX experiments (Beyer, Dandekar et al. 1997). In this study, CstF purified from calf thymus cell extract and HeLa nuclear extract was used to select for RNA ligands. It was shown, that G/U-rich sequence elements at a certain length (15 to 16 nt) and with spacers (varying up to four nt) separating CstF specific binding elements, were selected by CstF in several rounds of SELEX. Based on these results, different RNA oligos were designed in this thesis (Table 4). The aim of testing different RNA species was to find an RNA ligand, which is strongly bound by CstF in order to form a stable CstF-RNA complex for cryo-EM studies. The working hypothesis was to obtain a stable and homogeneous complex, where potentially flexible or loosely attached subunits are fixed on the RNA.

2.3.1 Recombinantly purified full-length CstF complex is capable of binding to a G/U-rich RNA oligo with high affinity

Binding experiments were initially performed as quality control to verify that recombinant complexes purified in this thesis were capable of binding specific RNAs, and to identify an optimal RNA oligo for structural studies using cryo-EM.

Binding of CstF to RNA was one potential way to stabilize or conformationally fix the CstF2 subunits for structural studies. Detailed information about sample preparation, screening and optimization for cryo-EM studies is mentioned later in this thesis (paragraph 2.4**)**. For this, CstF complex, purified as described in 2.1.1 with all three subunits present in stoichiometric amounts, was used in RNA reconstitution experiments using analytical SEC and Electrophoretic Mobility Shift Assay (EMSA). The first RNA ligand used for reconstitution experiments was designed based on SELEX studies of Beyer et al., 1997 (Table 4). In their study, they identified three major RNA elements (element 1: AUGCGUUCCUCGUCC, element 2a: YGUGUYUUYAYUGYGU and element 2b: UUGYUAUUUACU(U/G)YCU, where Y=U/C).

Results

I designed a 16-nucleotide long RNA oligo (Table 4, referred to as *CstF01* RNA) based on the consensus sequence of element 2a (YGUGUYUUYAYUGYGU, where Y=U/C).

Table 4. RNA oligonucleotides for biochemical and biophysical assays

Oligo name	*characteristics*	*Sequence (5'-3')*
CstF01	SELEX RNA	UGU GUU UUU A UUG UGU

The sequence of the resulting *CstF01* RNA (UGUGUUUUUAUUGUGU) was almost similar to one of two most abundant RNA species in element 2a (CGUGUUUUUAUUGUGU, the RNA named A-2 in Beyer et al., 1997). Further studies identified several hexameric sequence elements to act as DSE (UGUUUU, UGUGUU and UUUUUU) based on genomic alignment information obtained from *D.melanogaster* (Graber, Cantor et al. 1999). The second hexameric sequence (UGUGUU) of this study is, if inverted, identical to the last 6 nucleotides of the A-2 RNA (CGUGUUUUUAUUGUGU) in SELEX studies of Beyer et al., 1997. Subsequently, I replaced the first cytosine in A-2 RNA (**C**GUGUUUUUAUUGUGU) sequence by uracil, so that both potential binding sites for two CstF2 RRMs were identical to each other (UGUGUUUUUAUUGUGU). Additionally, they were identical to the second hexamer (UGUGUU) identified by Graber, Cantor et al., 1999. In all following experiments this RNA oligo is referred to as *CstF01*.

Apo CstF complex (10 μM) eluted in a symmetric single peak from analytical SEC with a ratio of A_{260}/A_{280} = 0.4. Successful reconstitution with RNA would lead to an increase in the A_{260}/A_{280} ratio in the elution profile. For RNA reconstitution, protein complex of the same concentration (10 μM) was incubated with 2 μM *CstF01* RNA on ice, because at this ratio no free RNA was left (see EMSA, figure 41). CstF with RNA eluted at approximately the same volume as the apo complex. Figure 41 shows the overlaid elution profiles for the CstF and CstF-*CstF01* complexes. Dashed lines correspond to the A_{260} of each sample, which shows in comparison a clear increase for the CstF-*CstF01* sample. This indicates that recombinantly purified CstF complex is capable of binding to G/U-rich RNA species. RNA binding was further confirmed by EMSA. For this, *CstF01* RNA was maintained at a constant concentration and recombinantly purified full-length CstF complex was added in increasing concentrations. At a protein to RNA ratio of 2:1, almost all unbound RNA was shifted to the top of the gel, indicating RNA binding by the protein complex (Figure 41).

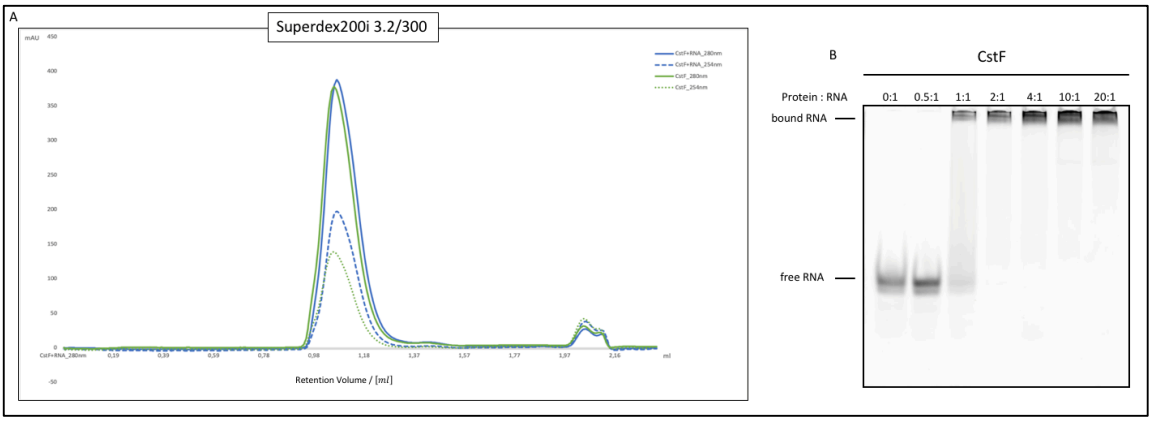


Figure 41. Binding of the CstF complex to G/U-rich RNA species. Analytical SEC of apo-CstF and CstF bound to *CstF01* RNA (A) and EMSA of CstF binding to CstF01 RNA (B). A) SEC profile with A_{280} (continuous line) and A_{260} (dotted line). Apo-CstF complex (green line) and RNA bound complex (blue line) elute at a retention volume of 1.05 ml in a symmetric sharp peak. Increased A_{260} (dotted blue line) of the RNA bound CstF indicates presence of *CstF01* RNA. B) TBE gel of an EMSA of CstF and *CstF01* RNA. At a protein to RNA ratio of 2:1, all RNA is shifted towards the top of the gel.

2.3.2 Full-length CstF complex shows selectivity towards G/U-rich RNA species in Fluorescence Anisotropy experiments

In studies published, binding affinity of full-length CstF complex to RNA has not yet been determined in a quantitative way. There are some data available obtained from CstF1-CstF2$^{1-199}$-CstF3$^{241-717}$ binding to $(GU)_n$ stretches of different lengths, spanning from $(GU)_6$ to $(GU)_{14}$. Binding affinities were determined for single CstF2 RRM, CstF1-CstF2$^{1-199}$-CstF3$^{241-717}$ complex and CstF2$^{1-199}$-CstF3$^{241-717}$complex (Yang, Hsu et al. 2018). In the same study, it was verified, that CstF can discriminate between U/G and A/C stretches. Second showed no detectable binding in ITC experiments to protein constructs that were used. The open question remaining is, if the presence of full-length proteins in the complex influences RNA binding behavior in any way. Therefore, I used Fluorescence Anisotropy (FA) to precisely measure binding affinities of the CstF complex to different RNA species. FA measurements allowed to determine the dissociation constant K_D of protein-RNA interactions. The K_D is defined as the concentration of protein at which 50% of labelled RNA is incorporated into the complex.

Experiments were started with *CstF01* RNA, for which binding to CstF complex was confirmed by analytical gelfiltration (see paragraph 2.3.1). Figure 42 shows three FA measurements for full-length CstF and G/U-rich *CstF01* 16-mer plotted against the logarithmic protein concentration. The dissociation constant K_D = 52.1 ± 10.3 nM was calculated as an average from these curves. This affinity is slightly higher than the one determined by Yang and co-workers for truncated CstF1-CstF2$^{1-199}$-CstF3$^{241-717}$ binding to $(GU)_{14}$ RNA, which showed the highest affinity K_D = 120 ± 10 nM in their hands (Yang, Hsu et al., 2018). In order to shed more

light on this and better understand contribution of RNA sequence on binding affinities, I went on to test several different oligo sequences.

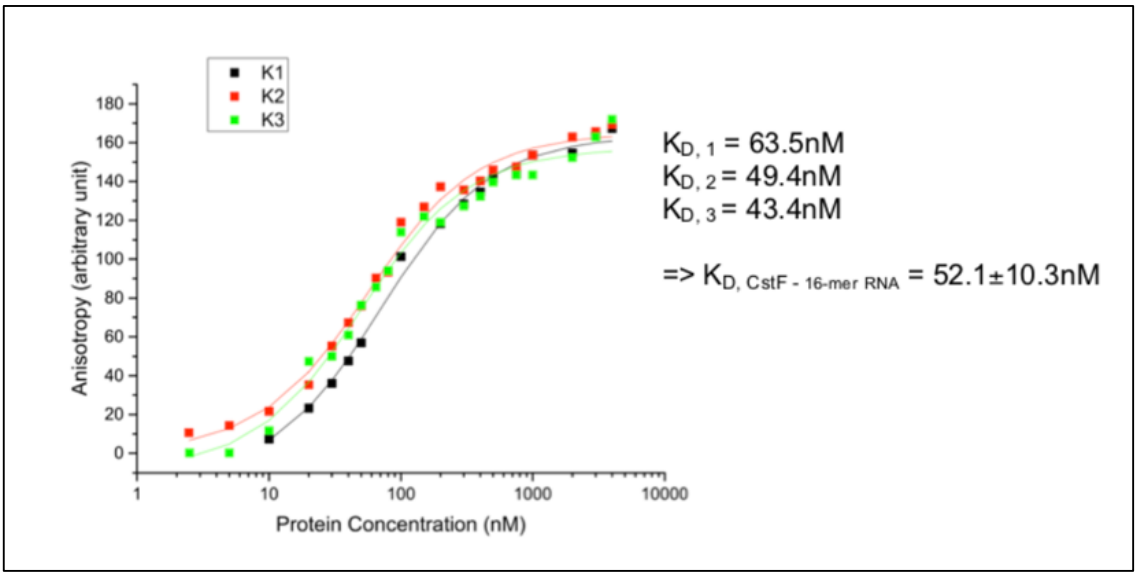


Figure 42: Determination of K_D of human CstF complex binding to G/U-rich CstF01 RNA by FA measurements. *CstF01* RNA (here named CstF) was designed based on previous SELEX experiments (Beyer et al., 1997). The graph shows the anisotropy plotted in dependency of the logarithmic protein concentration (X-axis). Measurements were repeated three times resulting in a K_D = 52.1 ± 10.3 nM.

I selected three more RNA oligos lacking the optimized G/U-rich elements (UGUGUU) of *CstF01*, listed in Table 5.

Table 5: RNA oligos used in FA experiments. A 6-Carboxyfluorescin (6-FAM) is fused to the 5' end of all RNAs. The G/U-rich binding motifs of *CstF01* RNA are underlined. 15U: 15 uracil stretch; 15A: 15 adenine stretch; ARE: random RNA containing U-rich elements

Oligo name	*characteristics*	*Sequence (5'-3')*
6-FAM-15U	polyU	5'Fl-UUU UUU UUU UUU UUU
6-FAM-15A	polyA	5'Fl-AAA AAA AAA AAA AAA
6-FAM-ARE		5'Fl-UUU CUA UUU AUU UUG
6-FAM-CstF01	SELEX-RNA	5'Fl-<u>UGU GUU</u> UUU A <u>UUG UGU</u>

Figure 43 shows that there was a clear difference in binding to various RNA species. As expected, *CstF01* 16-mer was bound with highest affinity (52.1 ± 10.3 nM) followed by *ARE* RNA (931 ± 202 nM), which also contains U-rich sequence motifs. Moderate binding to this RNA as well as to polyU (1.82 ± 0.38 μM) stretch was expected, since CstF2 RRM is known to bind U-rich RNAs (Gil and Proudfoot 1987, MacDonald, Wilusz et al. 1994). The last RNA ligand tested was a 15-nucleotide long polyA RNA. Supposedly, CstF2 should be able to discriminate adenine nucleotides and consequently has very weak to no binding (Perez Canadillas and Varani 2003, Pancevac, Goldstone et al. 2010). Corresponding to this finding, data for the polyA 15-mer confirmed, that full-length CstF binds with very low affinity to a polyA

RNA oligo in FA measurements. In summary, *CstF01* RNA was bound with highest affinity as expected, whereas polyU or U-enriched (ARE) sequences show moderate binding. Almost no binding occurs in case of polyA stretches.

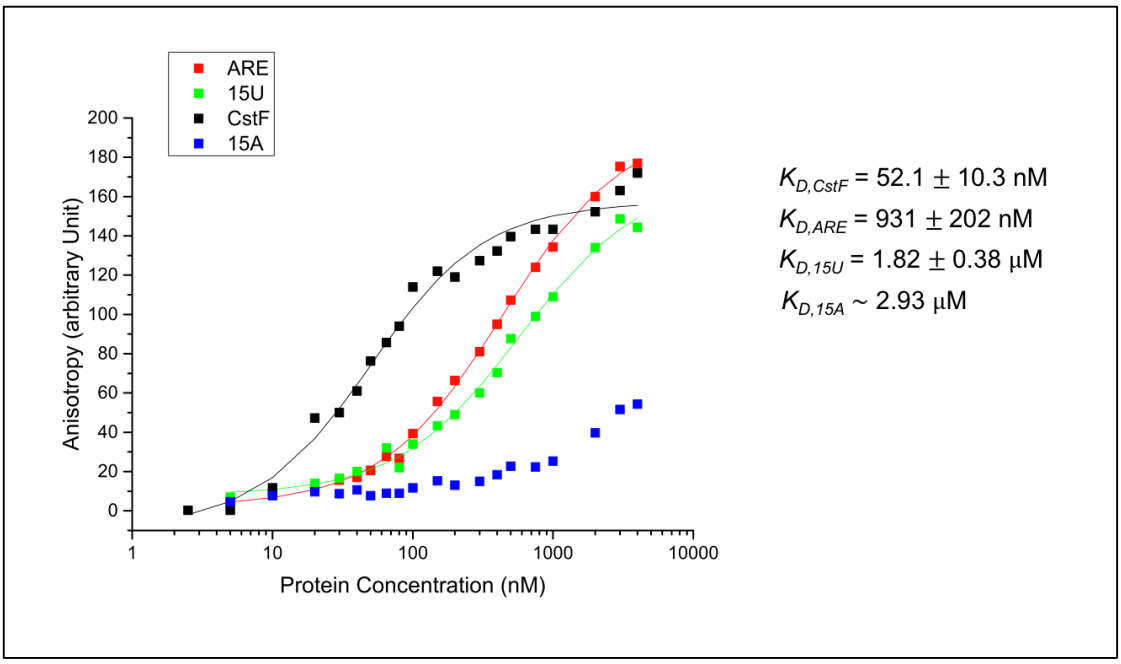


Figure 43: Determination of binding specificity of human CstF complex to G/U-rich RNA by FA measurements. The graph shows the anisotropy plotted in dependency of the logarithmic protein concentration (X-axis). *CstF01* RNA (named CstF) was designed based on previous SELEX experiments (Beyer et al., 1997). A polyU (15U) and polyA (15A) RNA were used as control. ARE is a random selected RNA containing U-rich sequence elements.

2.3.3 Full-length CstF complex recognizes bipartite G/U-rich DSEs with high affinity

Because of conformational and compositional heterogeneity of the full-length CstF complex in cryo-EM screening datasets, I tested further RNA ligands which might potentially bind with stronger affinity than *CstF01* RNA to stably fix the CstF2 subunits on the RNA. As a consequence of the dimeric structure of the CstF complex, there can be two RRMs present in the complex, which can bind to G/U-rich RNA motifs. The minimum length of an RNA oligo to allow binding of two RRMs of CstF simultaneously, was identified to be 10 nucleotides (Yang, Hsu et al., 2018). However, in previous studies, metazoan DSEs were shown to exist in a bipartite manner (McDevitt, Hart et al. 1986, Zarudnaya, Kolomiets et al. 2003, Salisbury, Hutchison et al. 2006), meaning that a proximal G/U-rich sequence followed by a distal polyU element are acting together to form the DSE (Figure 45, Gil and Proudfoot, 1987).

The existence of a U-rich distal binding pattern was examined by two independent SELEX studies (Beyer, Dandekar et al. 1997, Takagaki and Manley 1997), which identified different consensus motifs for CstF and the CstF2 RRM (see paragraph 2.3.1). For those reasons, I did not modify the second G/U-rich binding motif of *CstF01* RNA (UGUGUU UUU A UUGUGU) to a polyU sequence.

As described in section 2.4.2**,** the CstF complex showed high flexibility and heterogeneity in cryo-EM studies, raising the question if the distance between the G/U-rich binding motifs on *CstF01* RNA is sufficient to lock the complex in a homogeneous conformation. In order to test this, I increased the linker length between CstF binding elements by inserting a stretch of $(AC)_n$ nucleotides in multiple repeats (see table in figure 44). In contrast to previous studies, where a polyA stretch was used as spacer between GU-repeats (Yang, Hsu et al. 2018), I decided not to use an A_n linker because secondary structure formation could not be excluded, if the polyA stretch exceeds a certain length. RNAs used for determination of binding affinities by FA measurements are listed in the table in figure 44. The plotted graphs in figure 44 represent the results of the linker screen obtained from FA measurements with full-length CstF complex. *CstF01* RNA was measured for every new experimental setup as reference to judge if determined K_D values were in a reasonable range. Binding affinities obtained from triplicate FA measurements for the CstF complex are listed in the table in figure 44.

RNA	linker / [nt]	K_D / [µM]
CstF01	4	0.025 ± 0.009
CstF12	/	0.102 ± 0.035
CstF13	control	2.16 ± 0.41
CstF14	2	0.100 ± 0.68
CstF15	4	0.151 ± 0.052
CstF16	6	0.153 ± 0.61
CstF17	8	0.256 ± 0.102
CstF18	16	1.45 ± 0.71

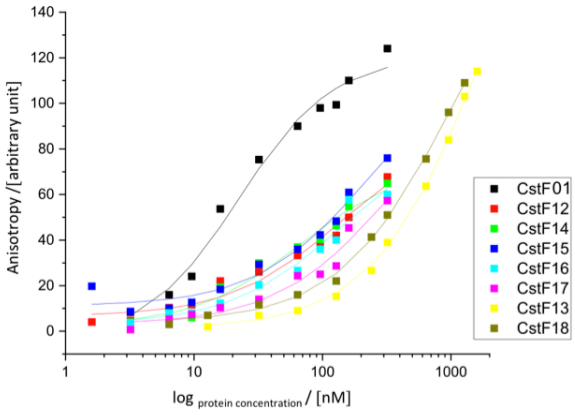


Figure 44. Determination of linker preference between G/U-rich downstream elements of the human CstF complex by FA measurements. RNAs were designed based on the G/U-rich binding motifs of *CstF01* RNA with a repetitive AC-linker in between spanning from two to 16 nucleotides. Measurements were repeated three times each and are depicted in the graph on the right. The graph shows the anisotropy Y-axis) plotted in dependency of the logarithmic protein concentration (X-axis). Exact determined of K_D values was not possible for curves that did not reach saturation, so that K_D values listed in the table on the left are considered to vary within a nanomolar range for RNAs *CstF12* to *CstF17*. However, overall tendency of CstF binding preferably to RNA species with shorter linker length between the G/U-rich elements can be observed in this experiment.

These results indicate, that full-length CstF complex is able to recognize symmetric G/U-rich sequence pattern spaced by a distance of two to eight nucleotides with high affinity. RNA without linker between G/U-rich sequence elements (*CstF12* RNA) showed decreased binding compared to *CstF01* RNA. For spacer lengths of 16 nt and longer between both G/U rich elements, the K_D drastically increased, suggesting that if both G/U-rich sequences are bound simultaneously, the optimal distance is around 4-6 nucleotides. Results did not confirm previous studies (Yang, Hsu et al., 2018), where no spacer length dependency of binding affinities was observed for a truncated CstF1-CstF2$^{1\text{-}199}$-CstF3$^{241\text{-}717}$ complex. Within the next paragraph, further experiments on the impact of CstF subunits on RNA binding mediated by the CstF2 RRMs will be described.

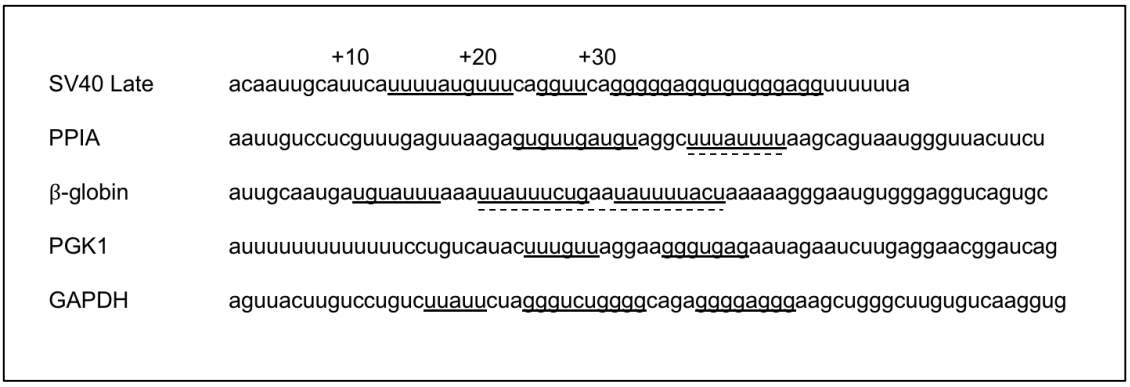


Figure 45. pre-mRNA sequences downstream of pre-mRNA poly(A) sites. Bipartite G/U-rich DSEs consisting of a proximal and a distal sequence element (underlined) are located within 30 nucleotides after the cleavage site. Distal U-rich sequence elements are indicated by dashed lines. SV40: Simian virus 40, PPIA: Protein phosphatase 1; PGK1: Phosphoglycerate Kinase 1; GAPDH: Glyceraldehyde-3 phosphate dehydrogenase

Additionally, this experiment suggested questioning of the polyU stretch forming the distal DSE (Figure 45). When using the *CstF01* RNA oligo with identical G/U-rich (UGUGUU) sequence elements separated by a four-nucleotide linker, measured affinity was almost six-fold higher compared to that of an RNA with the same spacer length, where the distal sequence element was replaced by U_5 (Yang, Hsu et al. 2018). To sum up, although the CstF complex is able to recognize various DSEs in pre-mRNAs, an RNA with identical G/U-rich binding sites separated by 4-6 nucleotides was preferably bound by full-length CstF in FA experiments.

2.3.4 CstF1 and CstF3 have a stimulatory effect on RNA binding of CstF2

Paragraphs 2.3.1 through 2.3.3 indicated that RNA oligo length, RNA sequence and distance between G/U-rich binding motifs are important factors for RNA binding by CstF. Besides that, CstF complex composition and subunit length may influence RNA binding mediated by the CstF2 RRMs. Also, previous studies have shown, that CstF1-CstF2$^{1\text{-}199}$-CstF3$^{241\text{-}717}$complex

binds to $(GU)_n$ RNAs stronger than the CstF2-RRM domain alone or a truncated $Cst2^{1-199}$-$CstF3^{242-717}$ subcomplex, missing the predicted unstructured C-terminal part of CstF2 and the N-terminal HAT domain of CstF3 (Takagaki and Manley, 1997; Yang, Hsu et al. 2018). So far, there is no data available on the RNA binding of all CstF full-length components.

In order to dissect the contribution of individual CstF subunits, I purified single components and subcomplexes of CstF as described in section 2.1 and performed FA measurements using *CstF01* RNA with the highest binding affinity determined in this thesis so far. Binding affinities were determined for just CstF2, CstF2-3, full CstF1-CstF2-CstF3 and CstF1-CstF3. CstF1-CstF3 was missing the RRMs of CstF2 and therefore should have low affinity for G/U-rich RNA (Figure 46).

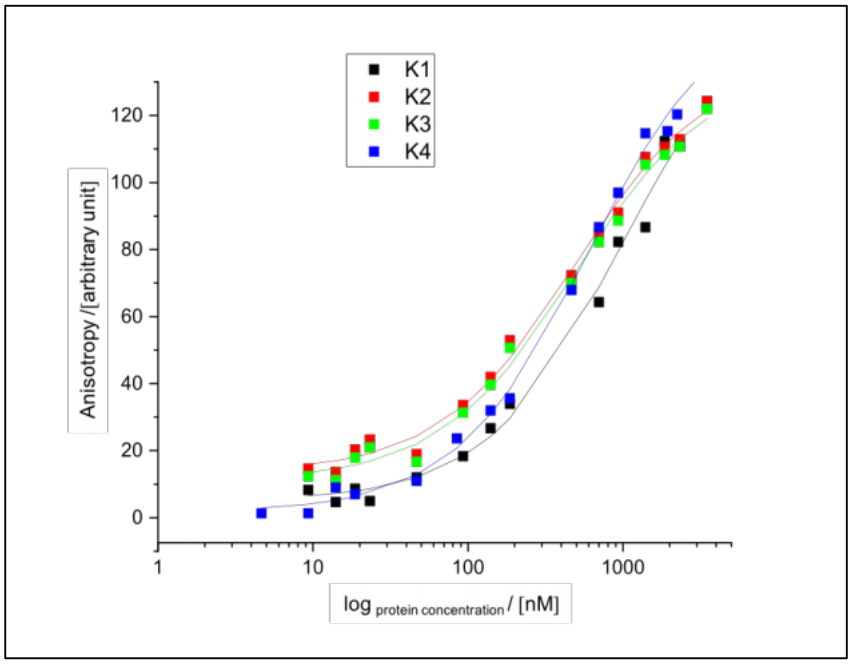
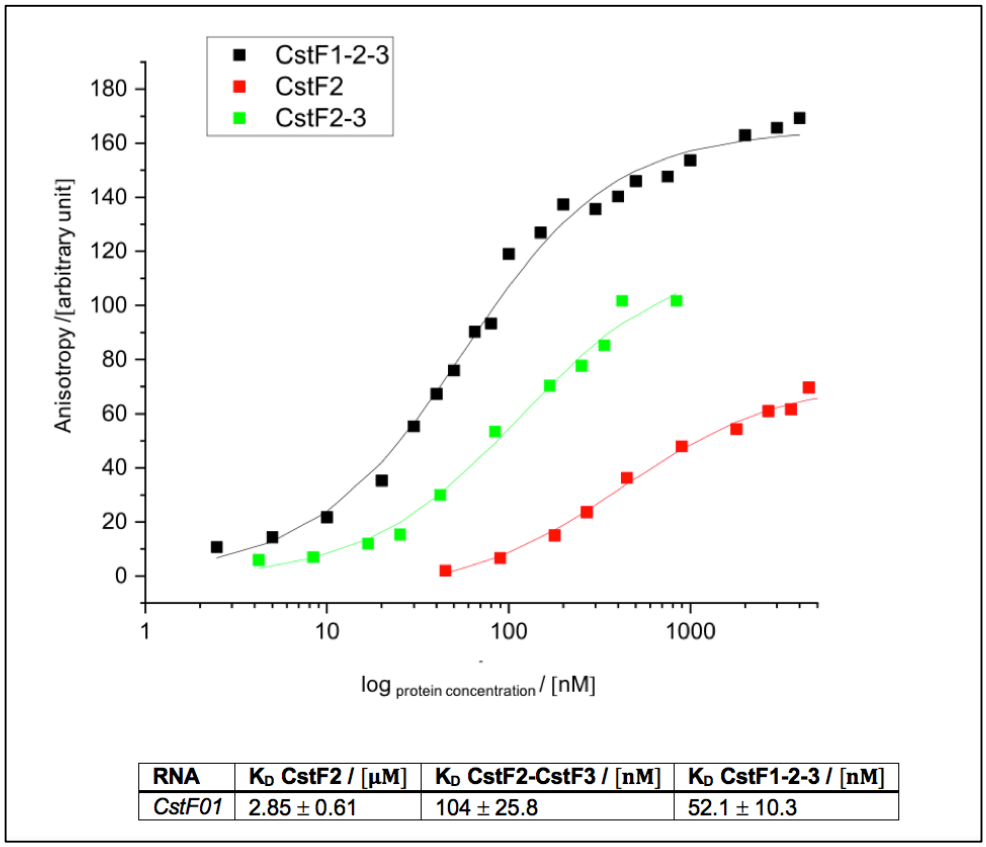


Figure 46. RNA binding affinity of CstF1-CstF3 subcomplex determined by FA. *CstF01* RNA was designed based on previous SELEX experiments (Beyer et al., 1997). Measurements were repeated four times each as depicted in the graph and resulting in a $K_D = 0.5 \pm \mathbf{0.066}\ \boldsymbol{\mu M}$. The graph shows the anisotropy plotted in dependency of the logarithmic protein concentration.

The plot below (Figure 47) shows the curves obtained for different CstF samples binding to *CstF01* RNA and respective K_D values are listed in the table below.



RNA	K_D CstF2 / [µM]	K_D CstF2-CstF3 / [nM]	K_D CstF1-2-3 / [nM]
CstF01	2.85 ± 0.61	104 ± 25.8	52.1 ± 10.3

Figure 47. Stimulatory effect of CstF subunits on binding to G/U-rich RNA sequences. *CstF01* RNA was designed based on previous SELEX experiments (Beyer et al., 1997). Measurements were repeated three times each resulting in K_D values listed in the table and depicted in the graph above. The graph shows the anisotropy plotted in dependency of the logarithmic protein concentration.

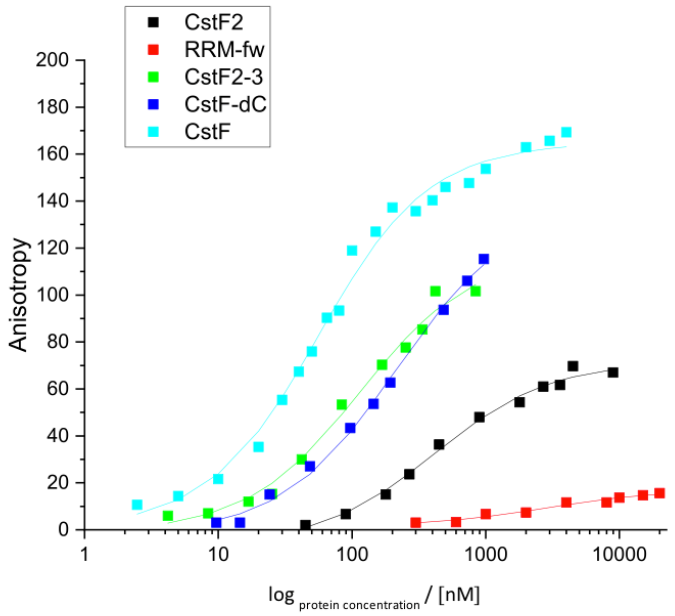
I started with full-length CstF2 (containing the RRM and the C-terminal part, which function is not yet known) and looked at its RNA binding capability as a single protein and in complex with other CstF subunits. As clearly visible in the graph (Figure 47), CstF2 alone has a rather low affinity of $K_D = 2.85 \pm 0.61\ \mu M$ for *CstF01* RNA. In presence of either CstF3 or CstF1 and CstF3, affinity increases by a factor of 20 or 40, respectively. This might indicate, that the presence of one or both other subunits is helping to form an RNA binding platform for DSEs, that is more compatible with a bipartite binding motif. Presence of CstF3 resulted in a drastic decrease of $K_D = 104 \pm 25.8$ nM for CstF2-CstF3, as compared to $K_D = 2.85 \pm 0.61\ \mu M$ for CstF2 alone. Most likely, presence of CstF3 has a huge effect on complex composition, since it self-dimerizes via its HAT domain. Given it binds two copies of CstF2, results in a pre-orientation of two RRMs with restricted conformational space, leading to an increased RNA binding affinity. Although there is no data available about the role of CstF1 in RNA binding, it had only small effect in the complex, since upon its presence $K_D = 52.1 \pm 10.3$ nM decreased to a low nanomolar range. Results obtained from the dissection of CstF into its of full-length components resembled what Yang, Hsu et al., 2018 had shown for a truncated CstF1-CstF2$^{1\text{-}199}$-CstF3$^{242\text{-}717}$ complex. They observed a minor effect of CstF1 on binding affinities for RNAs with GU-repetitive sequence elements.

Also, in their hands, binding affinities of CstF1-CstF2$^{1-199}$-CstF3$^{242-717}$ and Cst2$^{1-199}$-CstF3$^{242-717}$ were in the same range for different spacers (1-19 nt polyA linker) between GU/U-rich binding elements. I performed FA measurements, repeating the linker screening experiment from paragraph 2.3.3, but using the CstF2-CstF3 subcomplex formed by full-length proteins. In this experiment, I observed the same linker length dependency as for CstF1-CstF2-CstF3, but overall K_D values were five- to six-fold increased for CstF2-CstF3 (Appendix figure 102).

Consequently, I wanted to test if residues 200-577 of CstF2, which were deleted in experiments Yang, Hsu et al., 2018 performed, have an unknown impact on the RNA binding capability of the N-terminal RRM. To investigate this, I performed FA measurements using recombinantly purified minimal CstF1-Cst2$^{1-204}$-CstF3 (referred to as CstFdC) complex, consisting of a truncated version of CstF2 only containing RRM and hinge region (CstF2-RH). The CstF2-RH construct was designed based on following sequence alignment (Figure 48) including the conserved amino acid asparagine 204 (N 204).

Figure 48: Sequence alignment of the CstF2 RRM and hinge domain. Sequences were aligned using ClustalW (Madeira, Park et al. 2019) and visualized with Jalview (Waterhouse, Procter et al. 2009).

Measurements were performed with *CstF01* RNA to directly compare the K_D of CstFdC to full-length CstF and the CstF2-CstF3 subcomplex. If the C-terminal part of CstF2 has an indirect impact on RNA binding affinity of the RRM, this should be true in context of the full-length CstF complex and also for CstF2 alone. To test this effect, I included single CstF2 and its RRM domain alone in this study. In this case, the hinge domain was not present in the RRM construct, because it mediates complex formation by interacting with CstF3. Consequently, it was difficult to purify CstF2-RH (residues 1-204) alone without natively co-purifying CstF1 and CstF3. All measurements performed with *CstF01* RNA are summarized in figure 49 and corresponding K_D values are listed in the table below.



RNA	K_D RRM / [µM]	K_D CstF2 / [µM]	K_D CstF2-CstF3 / [nM]	K_D CstF(dC) / [nM]	K_D CstF / [nM]
CstF01	4.1 ± 0.8	2.85 ± 0.61	104 ± 25.8	196.6 ± 64	52.1 ± 10.3

Figure 49: Influence of the C-terminal part of CstF2 on RNA binding affinity determined by FA. *CstF01* RNA was designed based on previous SELEX experiments (Beyer et al., 1997). Measurements were repeated three times each resulting in K_D values listed in the table and depicted in the graph above. The graph shows anisotropy plotted in dependency of the logarithmic protein concentration. RRM: single RRM domain CstF(dC): Minimal CstF complex, with CstF2 only containing the RRM and hinge domain

When comparing the output of the FA measurements, it is clear that there is a difference if only the RRM domain or full-length CstF2 is present. For both of them, the affinity is in low micromolar range ($K_{D,\,RRM}$ = 4.1 ± 0.8 µM, $K_{D,\,CstF2}$ = 2.86 ± 0.61 µM) similar to previous studies (Perez Canadillas and Varani 2003, Pancevac, Goldstone et al. 2010, Yang, Hsu et al. 2018). Full-length CstF2 has a 1.5-fold higher affinity than the RRM domain alone. For CstFdC, affinity towards *CstF01* RNA is decreased about four-fold compared to full-length CstF, indicating that the C-terminal part (residues 205-577) of CstF2 might be involved in RNA binding also in context of the full complex.

Compositional setup of the CstFdC complex was similar to that in studies done by Yang, Hsu et al., 2018. The overall affinities they determined for GU_{12}/GU_{14} RNA by ITC, and K_D values measured in this thesis for CstFdC binding to *CstF01* RNA were in a comparable range of 120 – 220 nM. Since Yang and co-workers did not observe decreasing binding affinities upon increasing linker length between G/U-rich binding elements, I performed the same experiment as described in paragraph 2.3.3 for the CstFdC complex with RNA species listed in figure 50. A clear decrease in binding affinities with increasing linker length was observed, same as for full-length CstF and the CstF2-CstF3 subcomplex (Figure 44 and appendix figure 102).

RNA	linker / [nt]	K_D / [µM]
CstF01	4	0.196 ± 0.064
CstF12	/	0.410 ± 0.036
CstF13	control	2.62 ± 0.4
CstF14	2	0.334 ± 0.020
CstF15	4	0.355 ± 0.001
CstF16	6	0.705 ± 0.013
CstF17	8	1.14 ± 0.4
CstF18	16	1.40 ± 1

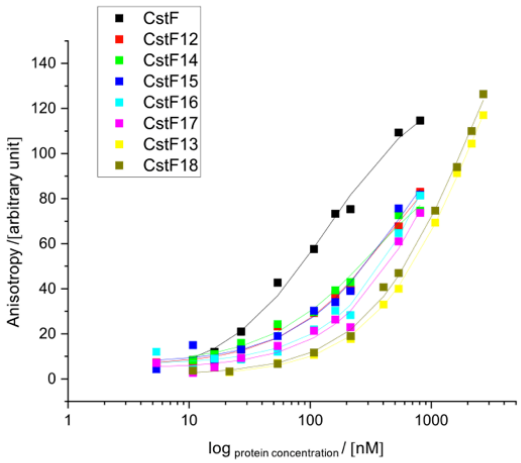


Figure 50: Determination of linker preference between G/U-rich downstream elements of human CstFdC complex by FA measurements. RNAs were designed based on the G/U-rich binding motifs of the *CstF01* RNA with a repetitive AC-linker in between spanning from two to 16 nucleotides. Measurements were repeated three times each resulting in K_D values listed in the table and depicted in the graph on the right. The graph shows the anisotropy plotted in dependency of the logarithmic protein concentration.

2.3.5 Proximity of two CstF2 RRM domains shows increased RNA binding

As shown by previous studies (Yang, Hsu et al. 2018) and in experiments described in the paragraph above, CstF3 significantly boosts RNA binding mediated by CstF2. This positive impact probably results from self-dimerization of the CstF3 protein via its HAT domain, resulting in assembly of two RRMs in unknown proximity to each other. Based on the bipartite sequence architecture of DSEs (McDevitt, Hart et al. 1986, Gil and Proudfoot 1987, Zarudnaya, Kolomiets et al. 2003, Salisbury, Hutchison et al. 2006) and on data observed in paragraph 2.3.4, where increased spacing between proximal and distal G/U-rich sequence element reduced the binding affinity, two RRMs are expected to be pre-orientated in restricted conformational space.

To mimic CstF dimerization with two RRMs in close proximity, I cloned and purified a fusion construct consisting of two single RRMs connected by a short linker. If dimerization was one of the key factors, driving increase in binding affinity of the CstF complex, the RRM fusion would have a clearly increased affinity towards G/U-rich RNAs compared to the single RRM or CstF2 alone. First confirmation for this assumption is indicated in EMSAs performed with the single RRM domain and the RRM fusion depicted in figure 51. To allow proper comparison, *CstF01* RNA was used in this experiment as well.

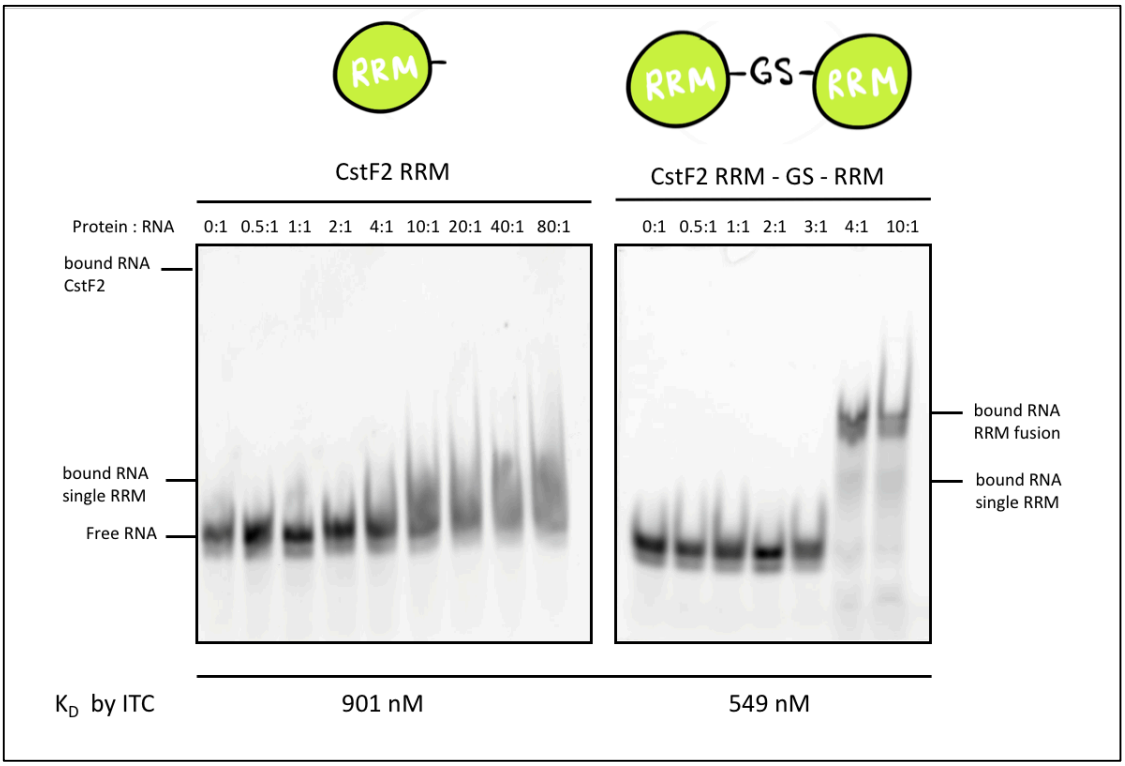


Figure 51: Increased RNA binding by the presence of two RRMs in close proximity observed in EMSA. Left: TBE gel of an EMSA of the single RRM domain (left) binding in increasing steps to a constant amount of *CstF01* RNA. Upon binding at a protein to RNA ratio of 10: 1, RNA is shifted towards higher molecular weight indicating RNA-protein complex formation. Right: TBE gel of an EMSA of the RRM fusion construct (right) binding in increasing steps to a constant amount of *CstF01* RNA. Upon binding at a protein to RNA ratio of 4: 1, RNA is shifted towards higher molecular weight indicating RNA-protein complex formation. K_D values for binding affinities of single RRM and RRM fusion were determined by ITC and are listed below the gels.

It was suggested from EMSA as well as from K_D values determined by ITC (data shown below), that presence of two RRMs in limited space had a huge effect on the RNA binding capability. For the RRM fusion construct, almost all RNA is bound at a protein to RNA ratio of 4:1. By contrast, for the single RRM, the RNA band is slowly starting to shift upwards at a protein to RNA ratio of 10:1. In case of the single RRM, there is no clear band shift between free RNA and bound RNA, which is why the outcome of this EMSA had to be verified by further experiments. In the gel for the RRM fusion construct, there is a second shifted band with lower intensity visible, running roughly at the same height as bound RNA for the single RRM. Lanes of free RNA show a second band with less intensity below the main free RNA band, which could be degradation or impurity of RNA. The second band could therefore correspond to the impurity band being shifted.

In order to determine the K_D to directly compare the observed effect of two RRMs in close proximity enhancing binding to G/U-rich RNA elements, Isothermal Titration Calorimetry (ITC) experiments were performed on the same samples used for EMSA and the *CstF01* RNA oligo (Figure 52).

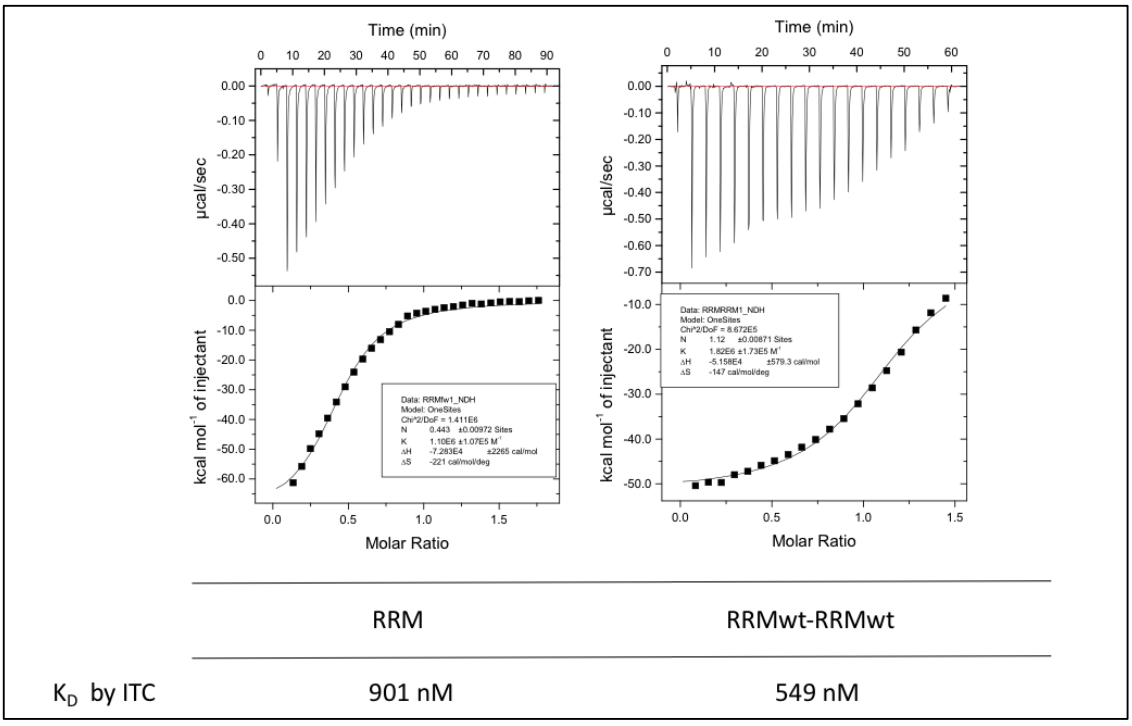


Figure 52: Binding affinities of CstF2 RRM and of two RRMs in close proximity to CstF01 RNA determined by ITC. ITC measurements were performed on the single RRM (left) and the RRM fusion construct (right) using *CstF01* RNA, which was designed based on previous SELEX experiments (Beyer et al., 1997). The upper panel of the graphs describes the release of heat (µcal/sec) over the time (min), with each peak corresponding to one injection. The lower panel of the graph depicts the binding isotherm, meaning the integrated peak injections resulting in the thermal energy ($\Delta\mathbf{H}$ in kcal/ml) plotted against the molar ratio of RNA/protein. Stoichiometry (n), change in enthalpy ($\Delta\mathbf{H}$) and the dissociation constant K_D can be derived from the curves. K_D values of each measurement are listed below the corresponding graphs.

Protein concentrations for both measurements were in the same range (25-30 µM). This could lead to overestimation of the K_D for the single RRM due to too high protein concentration and underestimation of the K_D for the RRM fusion. Consequently, decrease and increase of protein concentration for both measurements could lead to slightly differing K_D values. The output of ITC measurements (Figure 52) of the single RRM domain and the RRM fusion construct, however, supported the results of the EMSA experiment (Figure 51), that presence of two RRMs with restricted conformational space might positively influence the RNA binding to *CstF01* RNA. K_D values determined by ITC demonstrate an increase in RNA binding affinity for the RRM fusion ($K_{D,RRMfusion}$ = 549 nM) compared to the single RRM domain ($K_{D,RRM}$ = 901 nM). Stoichiometry for binding of the RRM fusion to RNA was almost 1:1, as expected, whereas stoichiometry of RNA to the single RRM domain was 1:2, indicating that two single RRMs could be bound to one RNA.

2.3.6 Identification of CstF2 residues important for RNA binding to G/U-rich RNA

The structure of CstF2 RRM was solved by NMR and modelled with an UU-dinucleotide ligand based on its homology to the HuD-*cfos* complex, identifying side chains involved in UU-RNA recognition (Perez Canadillas and Varani 2003) (Figure 53).

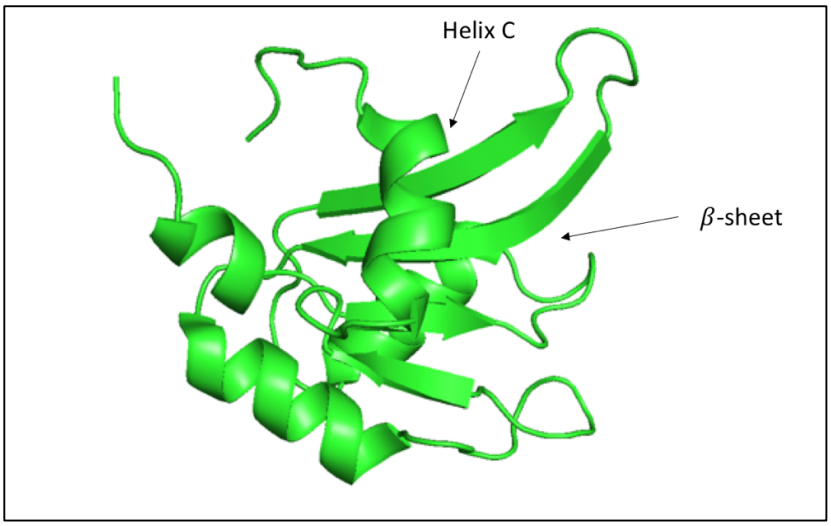


Figure 53. NMR structure of the RRM domain of CstF2 (Perez-Canadillas and Varani, 2003). CstF2 RRM domain (amino acids 94-105) adopts a classical RRM fold with a central four-stranded β-sheet (arrow) mediating the RNA binding surface and the C-terminal helix (labelled Helix C) lying perpendicular across the β-sheet. PDB: 1P1T.

Comparing different RNAs shows, that the protein achieves its specificity profile through a binding pocket for two uracils (Perez Canadillas and Varani 2003). In general, the structure follows features of a classical RRM domain (Nagai, Oubridge et al. 1990, Varani and Nagai 1998). A C-terminal helix, called helix C (amino acids 94-105), covers the β-sheet, which represents the RNA binding platform (Perez Canadillas and Varani 2003). Presence of a long C-terminal α-helix is also known for other RRM domains, like the one of U1A (Avis, Allain et al. 1996) or HuD (Wang and Tanaka Hall 2001). However, the C-terminal helix is not lying parallel on top of the β-sheet in any of them, making this conformation a unique structural feature of the CstF2 RRM. Residues involved in stabilization of helix C on the β-sheet are very conserved in vertebrates (Figure 54), but so far none of these residues has been addressed by site-directed mutagenesis in context of their impact in RNA binding.

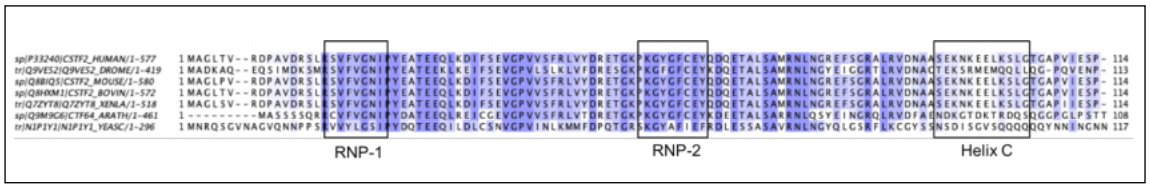


Figure 54: Sequence alignment of the human CstF2 RRM (amino acids 1-111). RNA binding motif 1 (RNP-1), RNA binding motif 2 (RNP-2) and the C-terminal helix (Helix C) are highlighted in black rectangles. Sequences were aligned using ClustalW (Madeira et al., 2019) and visualized with Jalview (Waterhouse et al., 2009).

Sequence alignment in figure 54 shows conservation of the CstF2 RRM among different species. The C-terminal helix and two RNP motifs (RNP-1 and RNP-2), which are conserved sequence motifs found in RNA-binding proteins (Landsman, 1992) and seem to be involved in RNA binding, are highlighted. Based on the sequence alignment (Figure 54) and the model purposed by Perez-Canadillas and Varani, 2003 (see paragraph 1.3.2), I designed various mutants of the conserved amino acids in the CstF2 RRM domain listed in figure 35.

CstF was purified as described in paragraph 2.1.2, carrying desired mutations in the RRM domain of CstF2. Initial EMSA experiments performed with mutated CstF and fluorescently labeled *CstF01 RNA* (Figure 55) already showed a difference in the RNA binding ability between mutants and wild type complex. For the wild type complex, almost all free RNA was shifted upwards when a two-fold protein excess was used (Fig. 52, Gel 1, lane 4). In comparison to that, for all mutants, a ratio of 4:1 (protein:RNA) was needed to observe a clear shift towards the upper part of the gel (Figure 55, Gel 2-4).

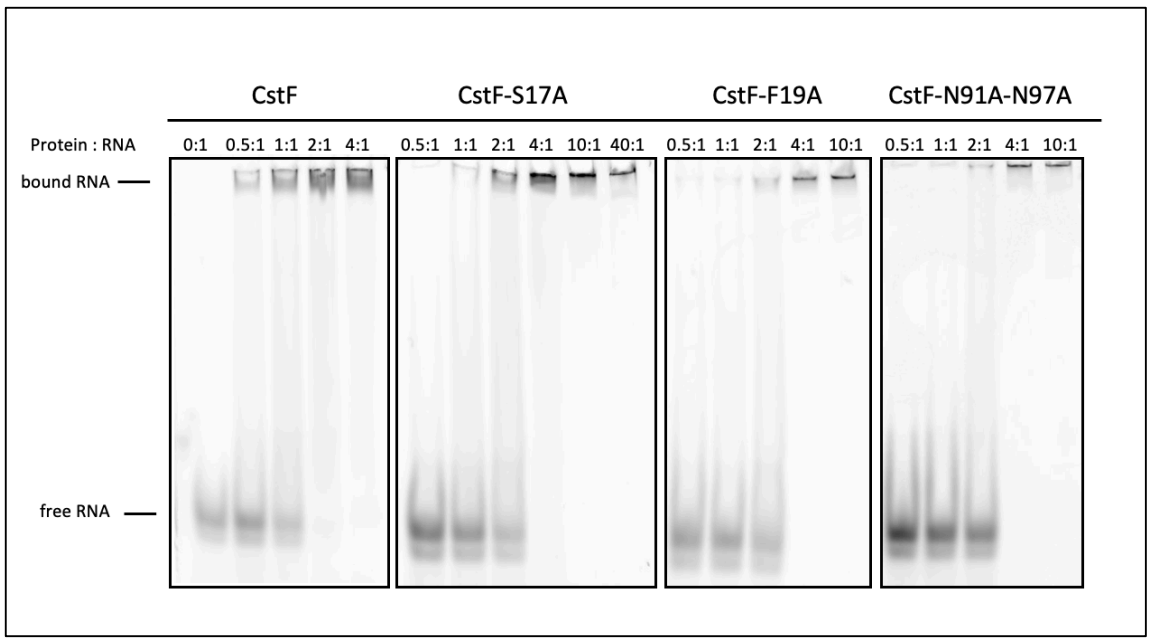
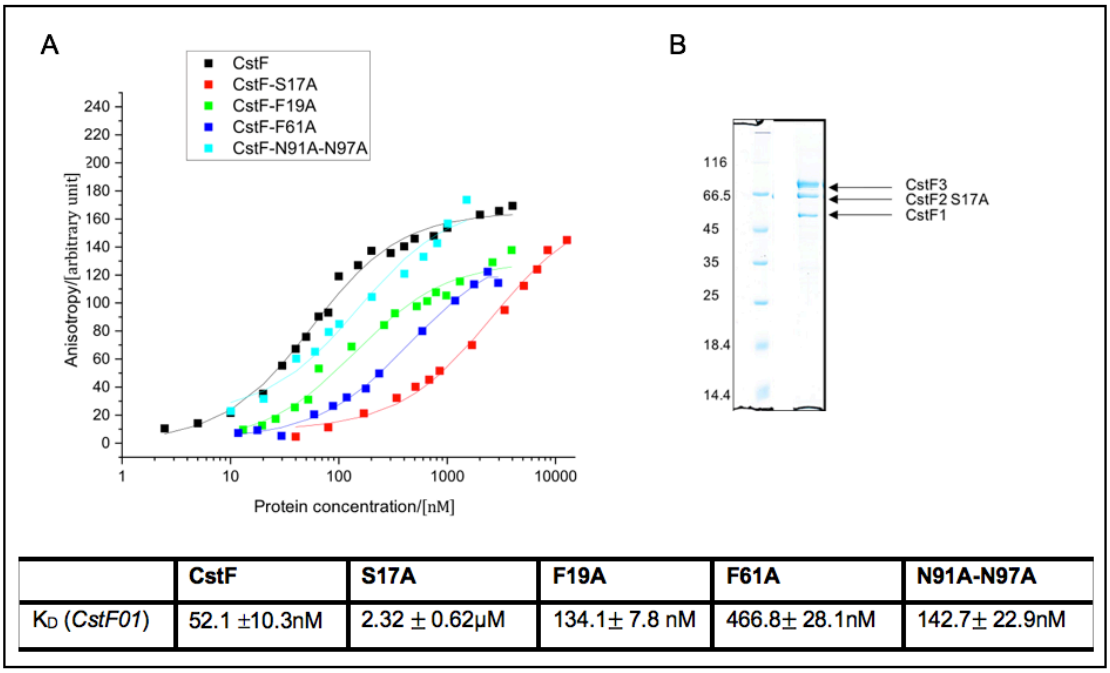


Figure 55. Mutations in the CstF2 RRM show decreased RNA binding of full length CstF in EMSA. Left panel: TBE gel of an EMSA of the wild type CstF (left) binding in increasing steps to a constant amount of *CstF01* RNA. Upon binding at a protein to RNA ratio of 0.5:1, RNA is shifted towards higher molecular weight indicating RNA-

protein complex formation. Second panel: TBE gel of an EMSA of CstF carrying S17A mutation in CstF2 subunit binding in increasing steps to a constant amount of *CstF01* RNA. Upon binding at a protein to RNA ratio of 4:1, RNA is shifted towards higher molecular weight indicating RNA-protein complex formation. Third panel: TBE gel of an EMSA of CstF carrying F19A mutation in CstF2 subunit binding in increasing steps to a constant amount of *CstF01* RNA. Upon binding at a protein to RNA ratio of 4:1, RNA is shifted towards higher molecular weight indicating RNA-protein complex formation. Right panel: TBE gel of an EMSA of t CstF carrying N91A-N97A mutations in t CstF2 subunit binding in increasing steps to a constant amount of *CstF01* RNA. Upon binding at a protein to RNA ratio of 4:1, RNA is shifted towards higher molecular weight indicating RNA-protein complex formation.

Following this first insights gained from EMSA experiments, I performed FA measurements to determine the K_D for all mutants. Since the binding affinity for wild type CstF to *CstF01* RNA was already known, one could directly compare the output and see if any of the point mutations affected RNA binding. The plot in figure 56 A shows the results of FA measurements plotted over the logarithmic protein concentration.



	CstF	S17A	F19A	F61A	N91A-N97A
K_D (*CstF01*)	52.1 ±10.3nM	2.32 ± 0.62µM	134.1± 7.8 nM	466.8± 28.1nM	142.7± 22.9nM

Figure 56: Mutations in the CstF2 RRM show decreased binding affinities to G/U-rich CstF01 RNA determined by FA. A) FA measurements show decreased K_D values of CstF complex carrying mutations in the RRM to *CstF01* RNA compared to wild type CstF (CstF, black curve). *CstF01* RNA was designed based on previous SELEX experiments (Beyer et al., 1997). Measurements were repeated three times each resulting in K_D values listed in the table below. The graph shows anisotropy plotted in dependency of the logarithmic protein concentration. B). SDS PAGE of CstF complex carrying S17A mutation in CstF2 RRM domain shows three bands at bands at 50 kDa (CstF1), 70 kDa (CstF2-S17A) and 85 kDa (CstF3). Lane 1: Molecular weight marker.

The results from initial EMSA experiments were verified by this method, since K_D values for all mutants are significantly lower than for wild type CstF. The majority of mutants (F19A, F61A, N9A-N97A) shows a moderate drop in binding affinity since their K_D values are still in nanomolar range ($K_{D,\ F19A}$= 134.1 ± 7.8 nM; $K_{D,\ F61A}$ = 466.8 ± 28.1 nM; $K_{D,\ N91A\text{-}N97A}$ = 142.7 ± 22.9 nM). However, the S17A mutation seems to drastically decreased the binding affinity since its K_D is about 55-fold lower compared to the wild type complex (Figure 56 A). To exclude

that this was an artefact of bad sample quality, degradation of one of the subunits or complex instability, determination of T_m was performed with a thermal shift assay for wild type CstF and CstF(S17A). T_m curves were comparable for both complexes (Figure 57) and also all three subunits of the mutated complex were present in stoichiometric amounts as clearly visible on the SDS PAGE in figure 56 B.

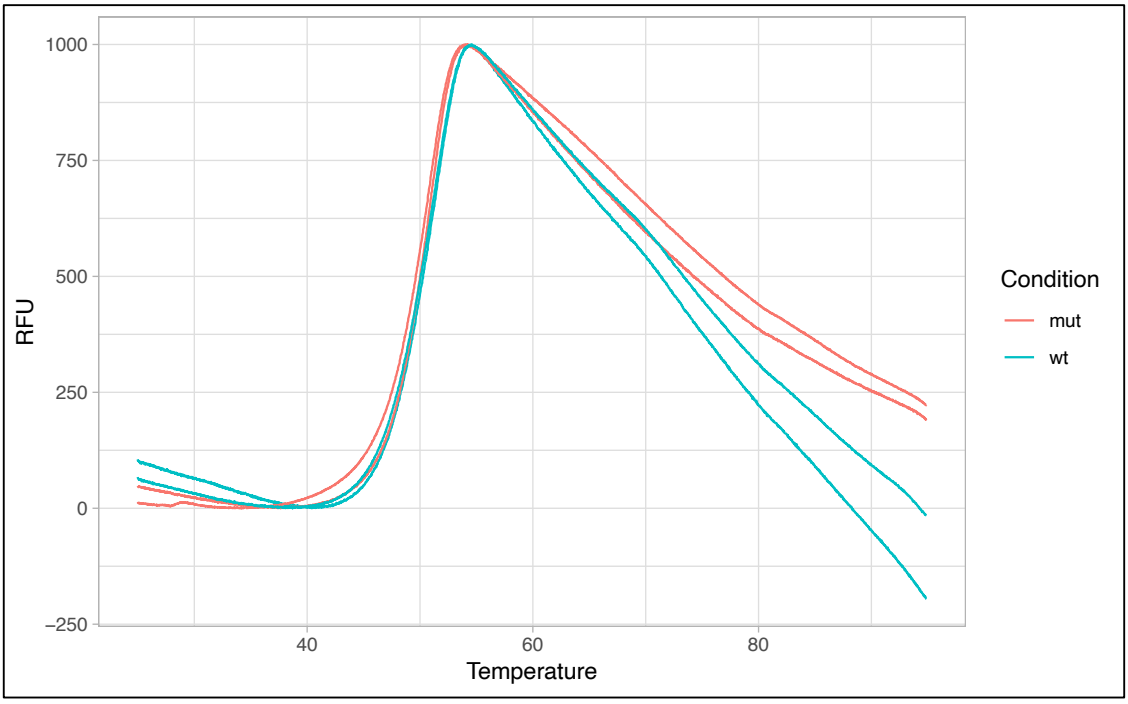
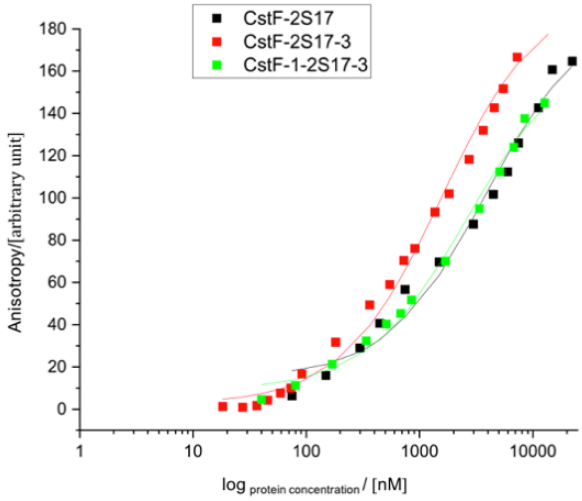


Figure 57. Determination of the T_m of wild type CstF and CstF(S17A) with a thermal shift assay. Relative fluorescence unit (RFU) is plotted against the temperature (℃). Two independent measurements were performed for each, wild type CstF (blue) and CstF-S17A (red) using the same sample as for FA measurements.

As shown by Yang, Hsu et al., 2018 for truncated CstF1- Cst2$^{1\text{-}199}$-CstF3$^{242\text{-}717}$, as well as in the previous paragraph 2.3.4 of this thesis for full-length CstF, presence of subunits CstF1 and CstF3 has a positive impact on the RNA binding behavior of CstF2. After identification of the S17A mutation in CstF2, which heavily decreased the RNA binding affinity, I wanted to check if the stimulatory effect of CstF1 and CstF3 was also true for CstF(S17A). Therefore, FA experiments were performed using CstF2(S17A), CstF2(S17A)-CstF3 and full-length CstF(S17A). Figure 58 indicates, that in case of the S17A mutant, there was no stimulatory effect of CstF1 and CstF3 on the RNA binding capability of mutated CstF2.



	CstF2-S17A	CstF2S17A-3	CstF1-2S17A-3
K_D (*CstF01*)	3.66 ± 0.76 µM	1.44 ± 0.29 µM	2.32 ± 0.62µM

Figure 58. CstF subunits have no stimulatory effect on RNA binding of CstF2 carrying S17A mutation. FA measurement of CstF2, CstF2-CstF3 subcomplex and CstF binding to *CstF01* RNA, all carrying S17A mutation in CstF2 RRM. *CstF01* RNA was designed based on previous SELEX experiments (Beyer et al., 1997). Measurements were repeated three times each resulting in K_D values listed in the table below. The graph shows anisotropy plotted in dependency of logarithmic protein concentration.

Overall binding affinities measured as K_D, were for in low micromolar range for all three samples. To sum up, the S17A mutation in the RRM has the largest effect of all tested mutants on the RNA binding affinity of CstF. On the single protein level, there is almost no difference between CstF2(S17A) (K_D = 3.66 ± 0.76 µM) and wild type CstF2 (K_D = 2.85 ± 0.61 µM). This raised the question, if this was also the case on the single domain level, when only the dimeric RRM fusion is used for mutational studies. Therefore, the RRM domain mutated in S17A was cloned for expression in bacterial systems. So far, nothing is known if there is any directionality in RNA binding, when two RRMs are present. Consequently, different combinations of mutated RRMs were generated as depicted in figure 59.

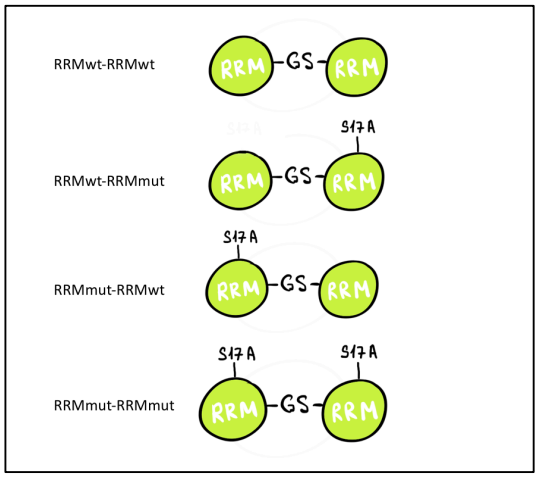


Figure 59: RRM fusion constructs containing wild type and mutated RRMs. Four different fusion constructs of the RRMs were generated by fusing either two wild type RRMs (row 1), one wild type and one mutated RRM (row 2 and row 3) or two mutated RRMs (row 4) together.

Two fused wild type RRMs were compared to constructs where either one RRM was mutated or both of them carried the S17A mutation (Figure 59). Due to a too low difference in molecular mass and size, FA measurement could not be applied and samples were subjected to Isothermal Titration Calorimetry (ITC) instead. All different samples were purified the same way and in the same buffer. To directly compare binding affinities to all data generated so far, *CstF01* RNA was used for this experiment.

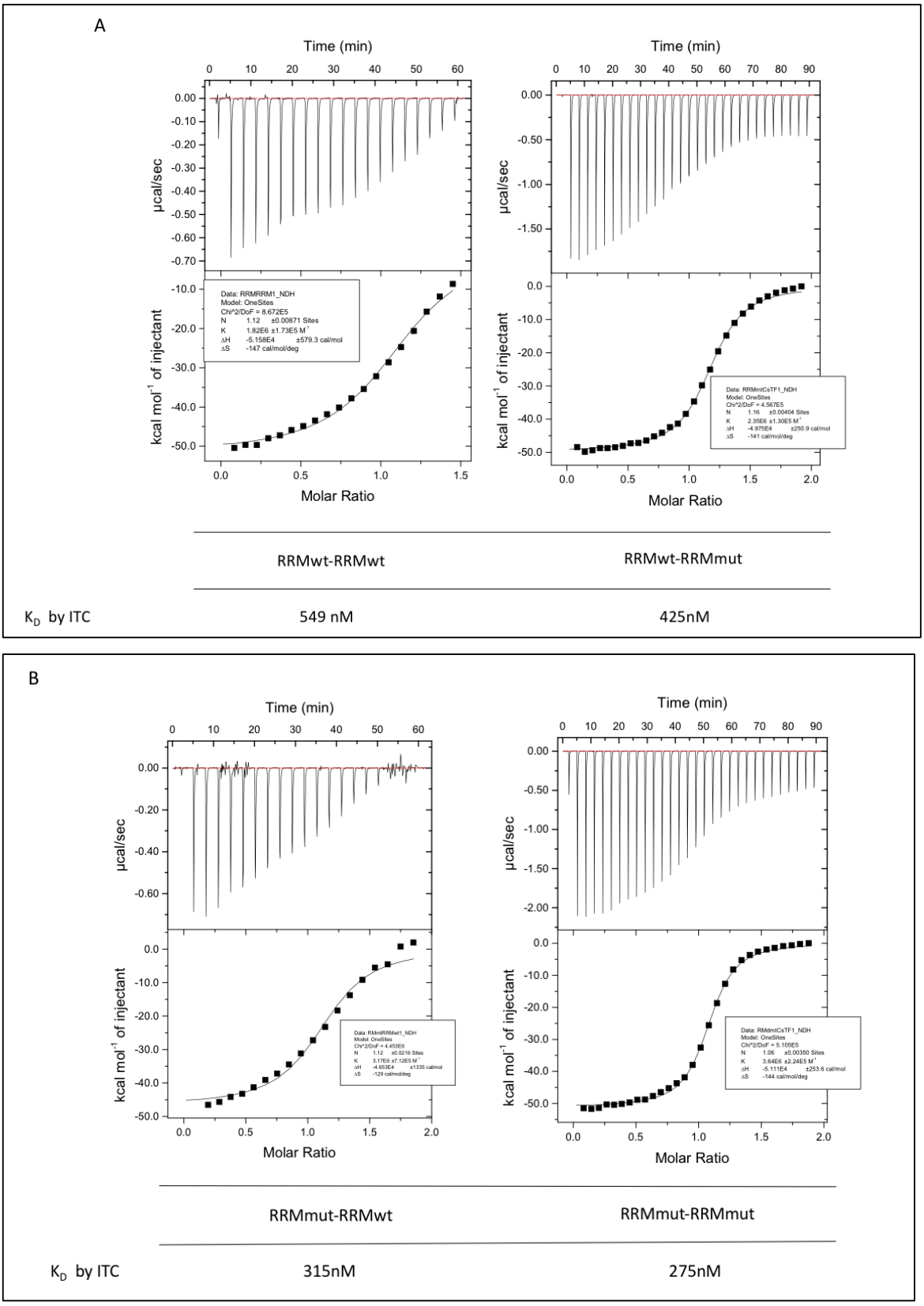


Figure 60. ITC measurements of RRM fusion constructs show stimulatory effect of the S17A mutation on RNA binding. ITC measurements were performed on the fusion constructs of two wild type RRMs (panel 1 A), one wild type and one mutated RRM (panel 2 A and panel 3 B) or two mutated RRMs (panel 4 B) using *CstF01* RNA, which was designed based on previous SELEX experiments (Beyer et al., 1997). The upper panel of the graphs describes the release of heat (µcal/sec) over the time (min), with each peak corresponding to one injection. The

lower panel of the graph depicts the binding isotherm, meaning the integrated peak injections resulting in the thermal energy ($\Delta\mathbf{H}$ in kcal/ml) plotted against the molar ratio of RNA/protein. Stoichiometry (n), change in enthalpy ($\Delta\mathbf{H}$) and the dissociation constant K_D can be derived from the curves. K_D values of each measurement are listed below the corresponding graphs. Stoichiometry for all reactions was 1:1.

Different ITC curves depicted in figure 60 show clear RNA binding for all samples. Remarkably, the fusion construct containing two wildtype RRMs has the lowest affinity (K_D = 549 nM) towards G/U-rich RNA, when compared to the mutants (Figure 60 A, left panel). Furthermore, comparison of the two fusion constructs, where either the first or the second RRM was mutated (Fig. 60 A, right panel and Fig. 60 B, left panel), shows a clear difference. Mutation of the first RRM increases RNA binding affinity (K_D = 315 nM)slightly more than mutation of the second RRM (K_D = 425 nM). This could be an indication, that there might be a directionality in RNA recognition between both RRMs. When the S17A mutation is present in both RRMs, RNA binding affinity (K_D = 275 nM) is increased about two-fold compared to wildtype RRMs (K_D = 549 nM). This is exactly the opposite effect to the one that S17A has in context of the full-length complex, indicating that presence of the full-length CstF subunits is important for RNA binding. Structural information is needed to explain why the S17A mutation in the single RRM stimulates RNA binding affinity, while for CstF containing all three full-length subunits, affinity is drastically decreased by this single mutation.

2.4 Structural analysis of the CstF full-length complex using cryo-EM

Another aim of this work was to study the structure of the CstF complex and get information about the overall arrangement of its three subunits CstF1, CstF2, and CstF3. For this purpose, I tried x-ray crystallography and single particle transmission electron microscopy. Unfortunately, trying to crystallize the complex was not successful. With a predicted molecular weight of around 385 kDa, the CstF complex is a good sample for single particle analysis (SPA) using EM. Initial screening and optimization were performed by negative stain EM as described in this paragraph. Optimized samples that looked promising in negative stain, were then subjected to cryo grid preparation.

2.4.1 CstF complex disassembles in initial negative stain EM grid preparations without RNA reconstitution and cross-linking

Once, a purification protocol of the CstF complex was established, negative stain grids were prepared after every purification step to monitor the sample quality and how the protein behaves on the grid. Already after the first Strep-tag affinity purification step, particles were clearly visible on the negative stain grid. It was very clear, that homogeneity of the sample was improving with every purification step. However, even after SEC, particles were not homogeneous in size, indicating that the complex might have disassembled or aggregated during grid preparation procedure. To address particle heterogeneity, one idea was to conformationally constrain the complex by binding an RNA (*CstF01* RNA). This could keep the flexible subunits like CstF2 in a more fixed conformation. Unfortunately, this did not help to overcome the second problem: disassembly of the complex (see figure 58). Thus, the second strategy to improve complex stability was chemical cross-linking of the sample, directly before it was applied to the negative stain grid. For this, I screened two different cross-linkers and two different cross-linking methods (see Material and Methods, 4.2.7.3). First, SEC purified CstF complex was cross-linked in solution using with 2 mM BS3 for 5 minutes at 30 ℃ and directly used for negative stain screening. Second, the CstF complex was subjected to GraFix, as described in paragraph 2.4.1, where 0.01% of GA was added to the sucrose gradient solution. Before negative stain grid preparation, sucrose was removed from the cross-linked GraFix sample by analytical SEC. Figure 61 summarizes the outcome of negative stain sample screening.

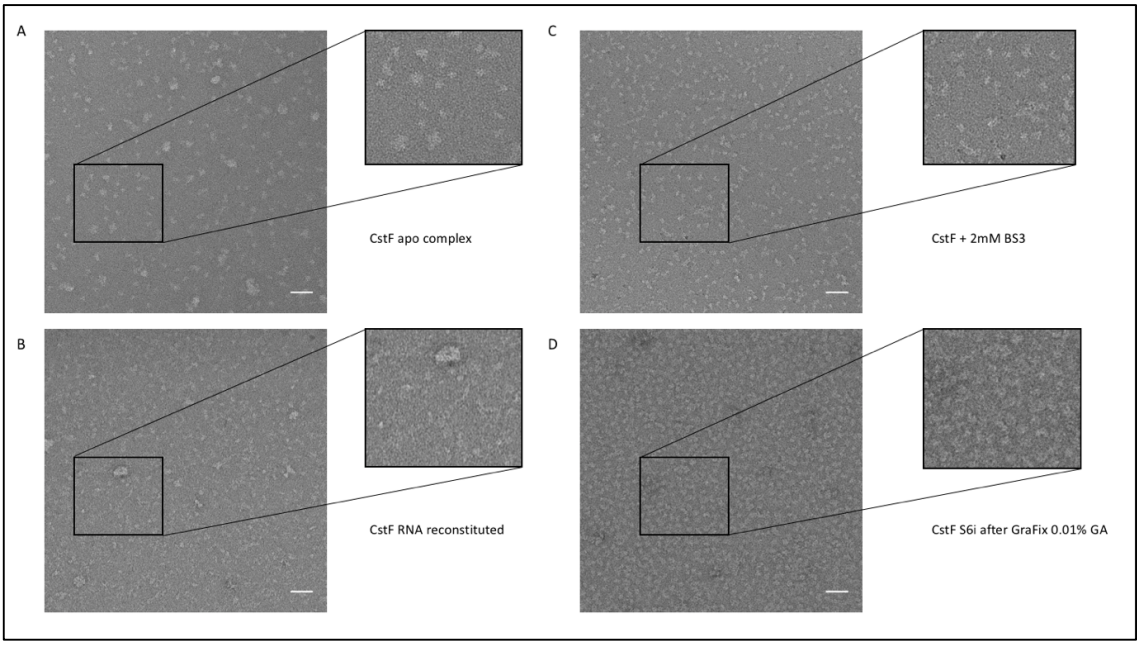


Figure 61. Negative stain micrographs of CstF with and without cross-linker and RNA. Examples of negative stain EM screening of the human CstF complex. A) Micrograph of CstF complex without cross-linking and RNA shows single particles of different sizes as well as aggregation. B) Micrograph of CstF complex bound to *CstF01* RNA shows a mixture of full complex, small disassembled pieces and aggregation (big white dots) in too low concentration. C) CstF complex cross-linked in batch with 2 mM BS3 shows improved particle distribution and more homogenous size of particles (white) in different shapes. D) CstF cross-linked with GA by GraFix and eluted from an analytical SEC shows good particle distribution and homogenous particle size (white dots) without aggregation. Particles are visible as white dots. The scale bar corresponds to 60 nm.

Comparison of figure 61 A and B clearly shows that reconstituting CstF with RNA did not deliver the desired effect of stabilizing particles, since disassembled particles of different sizes are still visible on the micrographs. On the other hand, both cross-linking procedures improved the overall homogeneity of particle size on negative stain grids, as well as particle distribution. In both micrographs (Figure 61 C and D), particles look more comparable in size. In addition to that, it looked like CstF tends to form less aggregates when cross-linking was applied before grid preparation (Figure 61 C and D). Even though no high-resolution structure can be obtained from negative staining EM due to resolution limitation because of stain grain size, it was an efficient method to validate different ways of preparation of the CstF complex for EM studies. Although the sample behavior and conditions optimized for negative stain EM could not be exactly transferred to cryo-EM, it was useful to know that there was need to stabilize CstF1-CstF2-CstF3 for EM studies

2.4.2 Cryo-EM screening of full-length, native CstF1-CstF2-CstF3 showed signs of shows complex disassembly without using cross-linking

In the beginning of my cryo-EM studies, nothing was known about sample preparation and plunging conditions suitable for the CstF complex. Therefore, I started to systematically optimize the purification protocol regarding complex assembly, buffer composition, carbon support grids, protein concentration and conditions for plunging to obtain in the end a sample suitable for high-resolution data collection. To directly track, how changes that were made in the purification protocol and plunging procedure, affected the protein sample in cryogenic conditions, screening session were performed on a Talos Arctica Transmission Electron Microscope (TEM) after every optimization step. The Talos Arctica TEM was equipped with a Falcon 3 camera. For suitable samples, a small dataset was collected for better sample quality assessment. Screening data were generally collected over night, resulting in around 600 to 1200 movies. The data processing procedure followed the same pipeline for all datasets collected on the Talos Arctica TEM. Initially, collected movies were corrected for beam-induced sample motion and integrated to a motion-corrected single frame micrograph using MotionCor2 (Zheng, Palovcak et al. 2017). As a next step, protein particles were picked using a script for template-free picking with Gautomatch (www.mrc-lmb.cam.ac.uk/kzhang). Then, micrographs were imported to either CryoSparc (Punjani, Rubinstein et al. 2017) or Relion (Scheres 2012), where contrast transfer function (CTF) was estimated using CTFFIND4 (Rohou and Grigorieff 2015). After particle extraction with a box size of 256-384 pix, particles were four-times binned and sorted and aligned in several rounds of 2D classification. One 3D initial model was calculated when processing was performed in Relion, or several 3D initial models were calculated in CryoSparc and used as input for 3D classification. Output maps from 3D classification delivered enough information, based on which quality of the sample and the dataset was judged. Initial screening sessions to test different buffers and concentrations revealed, that, no matter what buffer was used, the complex completely disassembled or aggregated when plunged without any cross-linking or RNA reconstitution (Figure 62 A and B). Besides instability, the CstF complex showed a clear tendency to stick to the carbon. Apart from aggregated or disassembled parts, almost no particles were observed in the ice. It seemed that without stabilization, the CstF complex does not survive the conditions it is exposed to during plunging, such as contact with the air-water interface, and is not distributed evenly on the cryo-EM grid.

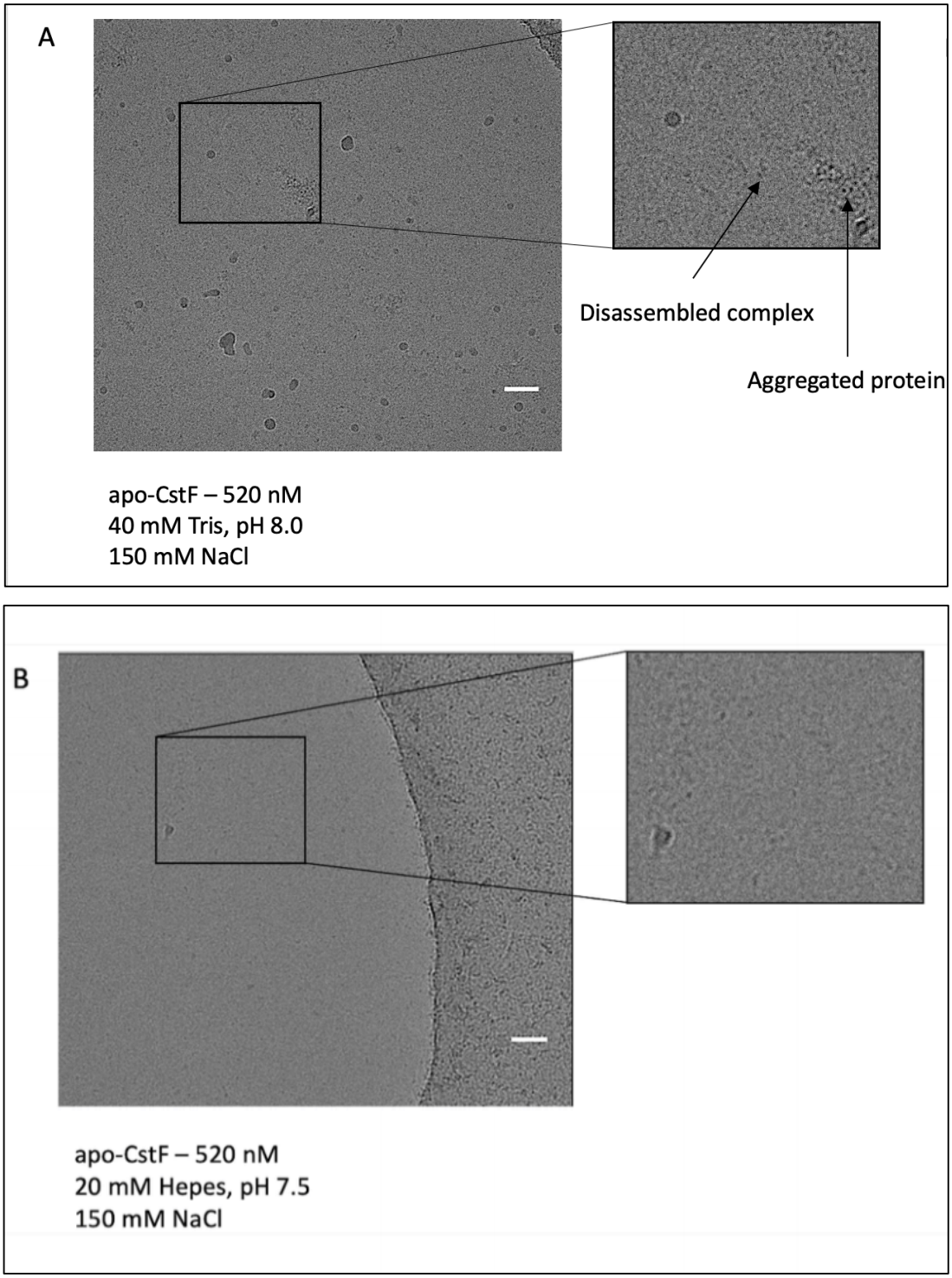
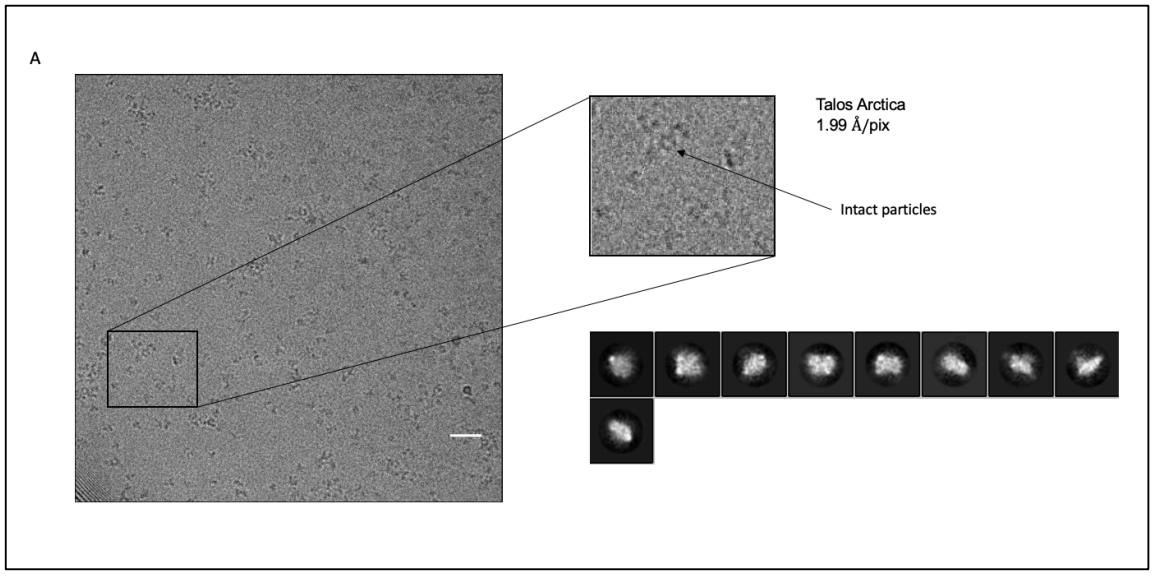


Figure 62: Cryo-EM micrographs of human CstF without cross-linking in different buffers. A) Micrograph of human CstF complex without cross-linker and RNA in 40 mM Tris, pH 8.0; 150 mM NaCl at 520 nM. Dissociated particles are visible as small dark dots and aggregated complex as chain-like arrangements (arrows). B) Micrograph of human CstF complex without cross-linker and RNA in 20 mM Hepes, pH 7.5; 150 mM NaCl at 520 nM. Dissociated particles are visible as small dark dots. Scale bar correspond to 60 nm.

The same observations of sample instability were described in the section above (2.4.1) for negative stain screening of the CstF complex. Improvement in the sample stability for negative staining was obtained by cross-linking the protein complex before preparing grids. Therefore, both cross-linking procedures described in paragraph 2.4.1 were tested in cryogenic conditions as well, but this did not lead to the desired effect of intact particles being visible in the. Most particles were still sticking to the surrounding carbon and holes remained almost empty. Usually, a carbon support layer on the EM grid helps to get a better dispersion of particles within the holes of a grid, but this was not the case for the CstF complex. Although carbon support was combined with one of the cross-linking procedures, successfully used in negative staining, intact particles were still not clearly visible in the holes. The only way to get the CstF complex into the ice, was to decrease the glow discharging time from 30 seconds to 10 seconds and to incubate the sample on the grid before blotting. These two steps in combination with cross-linking the protein complex with GA via GraFix led to clearly visible particles in the grid holes (Figure 63 A), so that a screening dataset was collected on this sample.



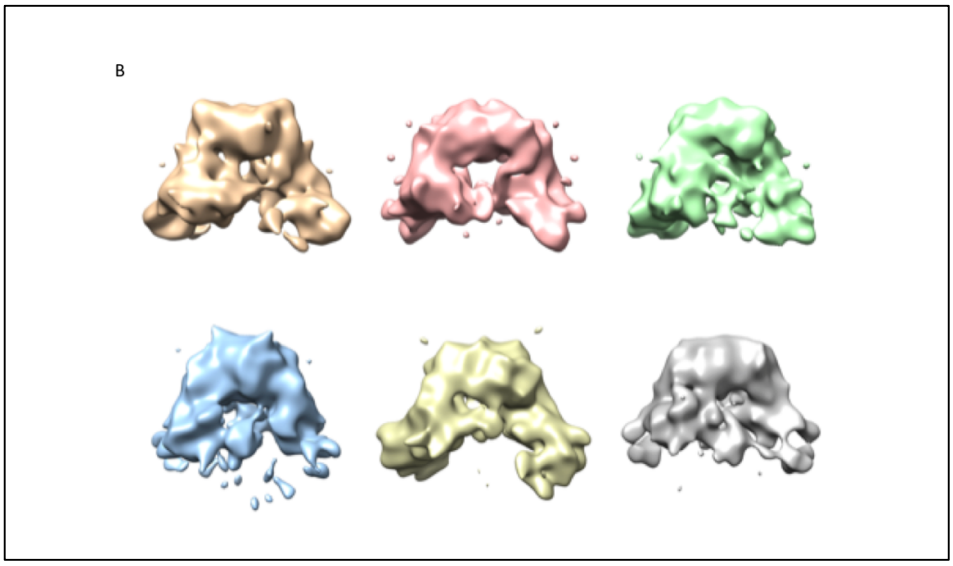


Figure 63: Cryo-EM screening dataset of human CstF complex. Representative micrograph and 2D classes of a screening dataset collected on a Talos Arctica TEM with a Falcon 3 camera show the human CstF complex cross-linked with GA by GraFix. A) Left: Micrograph (scale bar corresponds to 60 nm) showing CstF particles in dark with good particle distribution and contrast. Right: Final 2D classes obtained in Relion after particle cleaning show different particle shapes, but no high-resolution features. B) 3D classes obtained in Relion adopt an overall expected shape for the CstF complex.

The micrograph in figure 63 A clearly shows particles present in the hole, which looked mostly intact and resemble the overall shape expected for the CstF complex. After several rounds of 2D classification, different views of particles were observed in final 2D classes (Figure 63 A). 2D classification and 3D classification did not show any high-resolution features, likely due to the low number of micrographs collected, sample flexibility and thereby heterogeneity (Figure 63 A and B).

This sample preparation protocol including protein cross-linking by GraFix, the glow discharging time of 10 seconds and sample incubation on the grid before blotting was reproducible, and the sample quality was sufficient enough, so that it was decided to collect a dataset on the Titan Krios TEM.

2.4.3 Cross-linking of CstF1-CstF2-CstF3 with or without RNA is able to stabilize the complex but at the cost of high resolution

After optimizing the sample preparation protocol for cryo-EM as described above, a bigger dataset for high resolution reconstruction was collected on the CstF1-CstF2-CstF3 complex, which was prepared as described in Material and Methods, paragraph 4.2.9.1. Data were collected on a Titan Krios TEM and detailed parameters, as well as the processing workflow are described in Material and Methods, paragraph 4.2.10.3. A representative micrograph and final 2D classes, corresponding to 125000 particles, are depicted in figure 64. 2D classes show different views and possibly several conformations of the CstF complex. The most prominent

classes contained only the CstF3 HAT dimer in different orientations, but a few had extra fuzzy density below the HAT (Figure 64 B, enlarged 2D classes). Less represented classes showed one or two spherical densities, which might correspond to the WD40 propellers of CstF1 (Figure 64 B, enlarged 2D classes). In general, most of the 2D classes appeared very fuzzy and secondary structure features were visible only for a few classes containing the HAT dimer. There was no view in 2D, where all subunits seemed to be present and CstF2 was not visible at all. The GraFix cross-linking procedure stabilized the CstF complex to an extent, that single subunits and some subcomplexes survived the plunging method. However, it was unclear if the fuzzy classes were an artefact of GraFix or were due to the limited number of particles and therefore not enough signal for high resolution features. Furthermore, the flexibility of subunits relative to each other can create fuzzy densities. Given the described challenges, I continued optimizing the sample preparation for collection of high-resolution data instead of collecting more data on this sample, which has 2D classes of limited quality. Since no reasonable 3D map was obtained from this dataset, processing was abandoned after 2D classification.

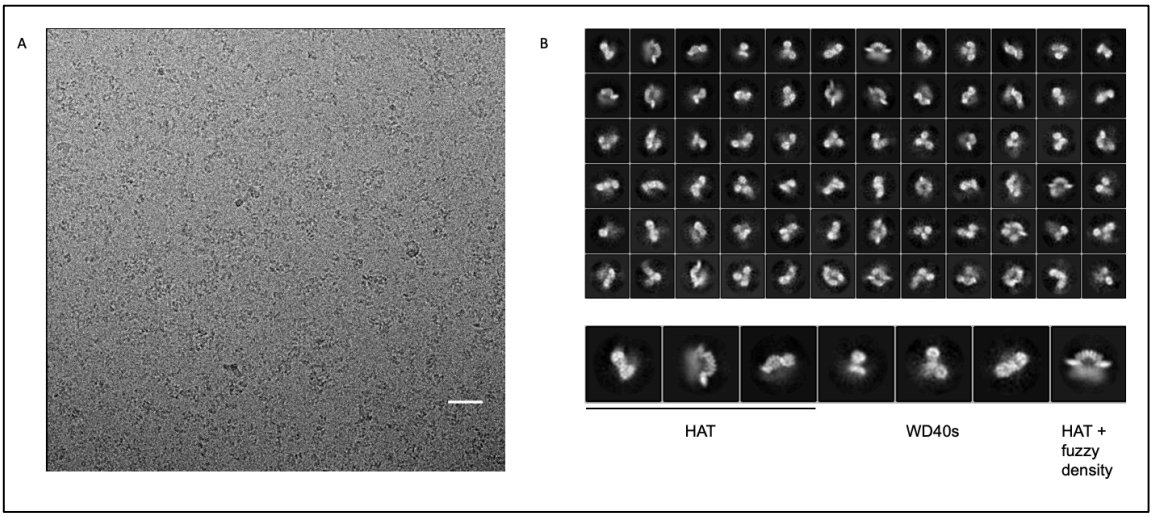


Figure 64: Cryo-EM data collection of human CstF1-CstF2-CstF3 cross-linked with GA by GraFix. Representative micrograph (A) and 2D classes (B) of a dataset collected on a Titan Krios with a K2 camera on a GA-cross-linked CstF obtained from GraFix. A) Micrograph (scale bar corresponds to 60 nm) shows cross-linked CstF particles in dark on a bright background with good particle distribution and contrast. B) Final 2D classes containing 125 k particles obtained in Relion after particle cleaning. Different subunits and domains are clearly visible in 2D classes, as well as secondary structure features for the HAT dimer.

2.4.4 CstF1-CstF2-CstF3 particles obtained from BS3 cross-linked samples in the presence of RNA resulted in improved 2D classes

As described in the paragraph above, GraFix was able to stabilize the CstF complex so far that subunits or subcomplexes were visible in resulting 2D classes, but not the full-length complex. Nevertheless, I wanted to exclude that the limited resolution in 2D classification is due to the use of GA as cross-linking reagent. Previous negative stain and cryo-EM screening

sessions showed, that without cross-linking, it was almost impossible to see intact particles. Instead of using the strong cross-linker GA, I wanted to test in-batch cross-linking with the 'milder' (i.e. less promiscuous) cross-linker BS3 in combination with RNA reconstitution of CstF. In-batch BS3 cross-linking was already shown to stabilize the sample for negative stain (paragraph 2.4.1), although in cryo-EM studies (paragraph 2.4.2) particles were still falling apart. After decreasing the glow discharging time and optimizing the plunging procedure (see Material and Methods 4.2.10.2), CstF1-CstF2-CstF3 purified via sucrose density gradient without cross-linking and an analytical S6i afterwards, was reconstituted with *CstF01* RNA, then cross-linked with BS3 in solution and plunged (Material and Methods 4.2.9.2). The SDS PAGE in figure 65 A shows the CstF complex being shifted upon cross-linking with BS3 to a single band at the top of the gel. Cryo-EM data were collected on a Titan Krios TEM operating at 300 kV. For detailed description of the data collection parameters and the processing workflow, refer to Material and Methods 4.2.10.4. A representative micrograph is shown in figure 655 B. 2D classes obtained in Relion of the final particle stack are depicted in figure 65 C. Unfortunately, during 2D classification in Relion, particles were collapsed in only very few classes, so that some classes were highly populated while others had too few particles per class to allow detection of secondary structure features. Therefore, the same dataset was processed in parallel in CryoSparc, which led to evenly distributed particles in 2D classification (Figure 65 D). Unfortunately, 3D classification did not yield to any high-resolution map resembling the intact CstF complex (Figure 65 E). Only low-resolution reconstructions without any secondary structure features of the HAT dimer with two blobby densities corresponding to the CstF1 WD40 propellers were obtained. No density for the CstF2 subunit was observed.

Although GraFix was not applied to this sample, particles were clearly visible on the micrograph (Figure 65 B). This confirms that the key to observe particles in the holes, was the combination of optimizing the plunging procedure together with different cross-linking strategies. In contrast to the dataset described in the section above (2.4.3), 2D classes obtained from BS3 cross-linked CstF (Figure 65 D) show clear secondary structure features, especially for the HAT dimer. There seem to be different conformations of the WD40 propellers present in the dataset, as observed in 2D classes obtained in Relion as well. If the WD40s are attached to the HAT dimer or disassembled is not clear from 2D classification. However, there are a few classes where density of the WD40 propellers is clearly visible below the HAT dimer, as indicated with the red arrows in figure 65 D. Unfortunately, no clear density can be assigned to CstF2 in those classes.

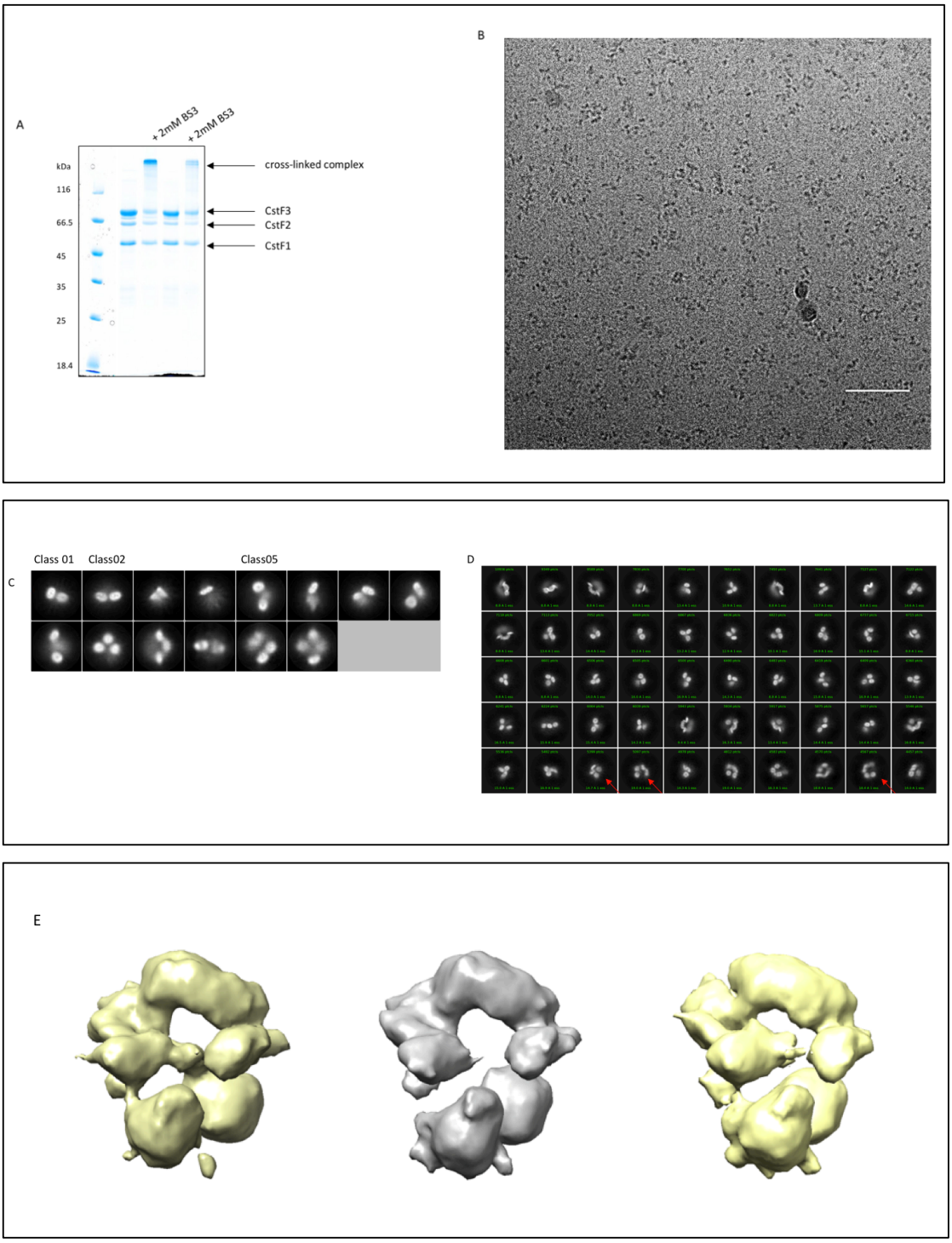


Figure 65: Cryo-EM data of human CstF complex cross-linked in batch with BS3. A) SDS PAGE of in-batch cross-linking with 2 mM BS3, showing four bands at 50 kDa (CstF1), 70 kDa (CstF2), 85 kDa (CstF3) and a band shifted to higher molecular weight corresponding to the cross-linked complex. Lane 1: molecular weight marker. B) Micrograph (scale bar corresponds to 60 nm) shows cross-linked CstF particles in dark on a brighter background in good particle distribution and contrast. C) 2D classes obtained in Relion containing 560 k particles showing group collapse into 14 classes. Aligned classes show the WD40 propellers of CstF1 in different conformations with fuzzy background and surrounding resulting from too many particles being aligned into one class. D) 2D classes containing 1.5 million particles obtained from picking with Topaz implementation in CryoSparc. No group collapse is observed. Secondary structure features are present in several classes, and a few classes contain density for the CstF3 HAT domain and the WD40 propellers of CstF1 (red arrow). E) Final 3D classes, containing in total 280 k particles. Particles are evenly distributed among the classes. Low resolution density for the HAT is visible in all classes, as well as the two WD40 propellers.

Although quality of 2D classes was improved in this dataset and density for the WD40 propellers of CstF1 were observed in 3D classes (Figure 65 E), the complex was either still falling apart or the sample contained a mixture of full complex and extra CstF1. After further 2D classification of classes showing proper density for the WD40 propellers, different conformations of the WD40s became even more visible (Figure 66 A). Although the CstF1 dimer on its own is rather small with a total molecular weight of around 100 kDa, it was possible to classify different conformations of the WD40s also in 3D to a low resolution (Figure 66 B). For high-resolution reconstructions, there were too many different conformations of the smallest subunit of the CstF complex.

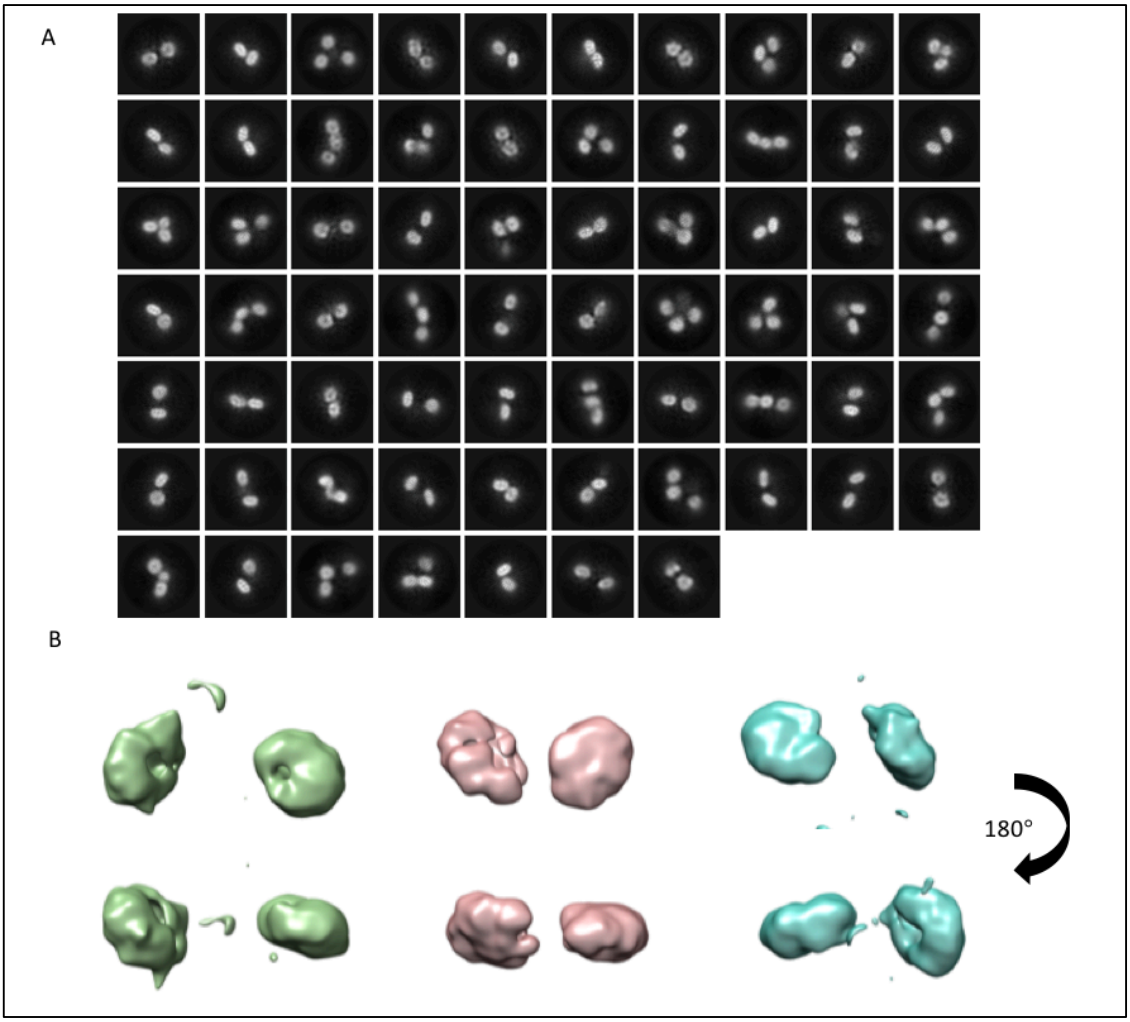


Figure 66. Different conformations of the CstF1 WD40 domains. 2D classes and 3D classes show heterogenous conformations of CstF1 WD40 propellers in a subset of particles from a Krios dataset collected on the CstF complex cross-linked in batch with BS3. This particle subset either corresponds to disassembled CstF1 from the complex or extra CstF1 in the sample. Particles for 2D and 3D classification were picked in CryoSparc using the Topaz implementation. A) Representative 2D classes showing different views of the WD40 propellers. They adopt different conformations towards is each other by rotation and movement of the propellers. B) Conformational heterogeneity is also visible in 3D classification in CryoSparc. 3D classes show the three main conformations that were possible to reconstruct in 3D at low resolution.

To sum up, by using an optimized protocol for sample preparation and plunging, it was possible to stabilize the CstF complex so far, that particles were clearly visible in a good distribution in the holes of cryo-EM grids. Datasets of two different ways of sample cross-linking show the presence of different subdomains and subcomplexes in 2D classification and deliver initial secondary structure features in case of a BS3 cross-linked sample. Besides that, it was shown that the WD40 propellers can adopt different sub conformations, if the CstF1 subunit was dissociated from the complex. It now has to be clarified, if the CstF1 WD40 propellers can adopt this high flexibility also when they are stably incorporated into the CstF complex.
Besides two Krios datasets of the two different cross-linking procedures, that were mentioned in the text above, a combination of both ways of sample preparation was tested. The CstF complex was reconstituted with *CstF01* RNA in-batch, followed by cross-linking with GA by the GraFix method. However, this did not further improve complex stability and homogeneity of particles.

2.5 Structural analysis of the CstF1-CstF3 subcomplex

The two datasets described in the paragraphs above showed that neither harsh cross-linking with GraFix using GA nor in-batch cross-linking with BS3 were sufficient to overcome the compositional sample heterogeneity and complex disassembly.
In order to address this problem, two measures were applied to get a more homogeneous sample. First, the CstF2 subunit was not included in the complex, because so far, no clear density was observed for this protein. Instead, the CstF1-CstF3 subcomplex was readily visible in 2D classes (Figure 65 D, red arrow). Secondly, the sample preparation and cross-linking procedures from the two previous datasets were combined to stabilize the CstF1-CstF3 subcomplex. As described in section 2.1.5, the sample was cross-linked in-batch with BS3 and further purified via a sucrose density gradient and a final analytical SEC. Following the established plunging protocol, cryo-EM grids were prepared and screened on a Talos Arctica. Screening session showed clearly visible particles in a good distribution in the holes.

2.5.1 Cryo-EM data collection of cross linked CstF1-CstF3 shows less sample heterogeneity than full-length CstF complex

Based on the output of the Arctica screening session, a big dataset was collected on the same grid on a Titan Krios. A detailed description of data collection parameters and processing workflow is given in Material and Methods, 4.2.10.5. A representative micrograph is depicted in figure 67.

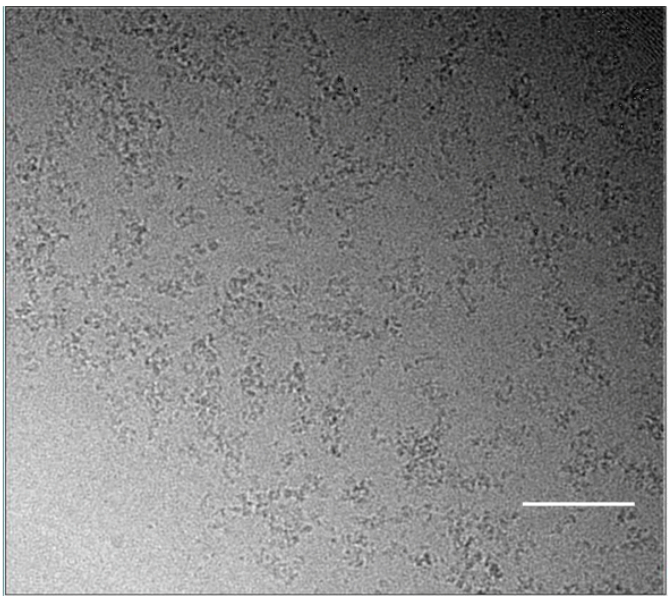


Figure 67: Cryo-EM data collection of the CstF1-CstF3 subcomplex. Data were collected on a Titan Krios TEM with a K3 camera on CstF1-CstF3 subcomplex prepared with a combination of in-batch cross-linking and density gradient ultracentrifugation. Micrograph (scale bar corresponds to 60 nm) shows cross-linked CstF1-CstF3 particles in dark on a brighter background in good particle distribution and contrast. Particles were concentrated at the edge of the holes due to too thin ice.

Resulting 2D classes from data processing in CryoSparc show three major sub populations. The first population containing mostly the HAT dimer, is the most represented one (Figure 68 A, around 54 %) and is characterized by clear secondary structure features for the HAT in different views. In most of the classes, density below the CstF3 HAT domain is only visible as light shadows, indicating that CstF1 is either too flexible to allow alignment in one distinct class by the software or that the CstF1-CstF3 subcomplex was partially disassembling on the grid. The second group of 2D classes shows different views of the CstF1-CstF3 subcomplex consisting of the HAT dimer as well as clear density for the WD40 propellers (Figure 68 B, around 19 %). Different 2D classes from this group show flexibility of the WD40 propellers within the complex, because they can be located in different positions with respect to the HAT dimer and in varying distances to each other. This high flexibility might be a reason, why the density for WD40s is difficult to align in 2D classification and also in 3D reconstructions. The last major population observed in this dataset corresponds mainly to the WD40 propellers in different projections (Figure 68 C). The WD40s are either close together or more separated and in some classes, one of them seems to be rotated relative the other one. It is not possible to judge, if those classes really only contain the CstF1 dimer or if there is some part of the CstF3 subunit attached, which is not visible either due to complex orientation or complex flexibility. The phenomena of high flexibility of the CstF1 WD40 propellers was already observed in a BS3 cross-linked dataset of full-length CstF (Figure 66 C and D, paragraph 2.4.4).

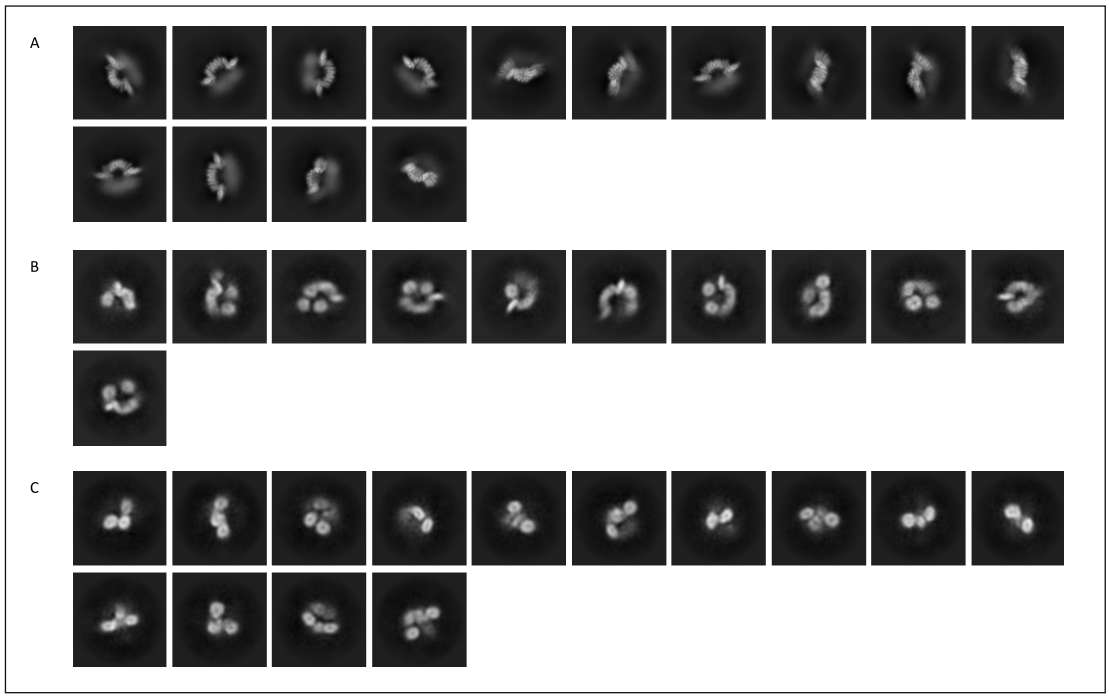


Figure 68: 2D classes of a cryo-EM dataset of the CstF1-CstF3 subcomplex show three major particle populations. 2D classes obtained from a cryo-EM data collection on a Titan Krios TEM with a K3 camera on the CstF1-CstF3 subcomplex prepared with a combination of in-batch cross-linking and density gradient ultracentrifugation. All 2D classes were obtained in CryoSparc using the Topaz implementation to train particle picking. A) 2D classes depicting the CstF3 HAT dimer in different orientations showing secondary structure features and a bright fuzzy density surrounding the HAT domain. B) 2D classes showing the CstF1-CstF3 with the WD40 propellers adopting different positions below the HAT domain. In some classes only one WD40 propeller is visible. C) 2D classes corresponding to the WD40 propellers of CstF1, which adopt different conformations towards each other. In some classes a third density between the WD40 propellers is visible, most likely belonging to the homodimerization domain of CstF1.

In conclusion, compared to data collected on the full-length CstF complex, this dataset of the CstF1-CstF3 subcomplex shows improved homogeneity although there were several sub populations of particles present. In contrast to previous datasets, different conformations were assigned to parts and subdomains of the protein complex.

2.5.2 Reconstruction of the CstF1-CstF3 subcomplex at medium resolution shows flexibility of CstF1 WD40 domains within the complex

Using 2D templates generated in CryoSparc (paragraph 2.5.1), particles were picked from all micrographs of the same dataset with Gautomatch, extracted within Relion, binned and used for several rounds of 2D classification thereby sorting out disassembled or fuzzy particles.

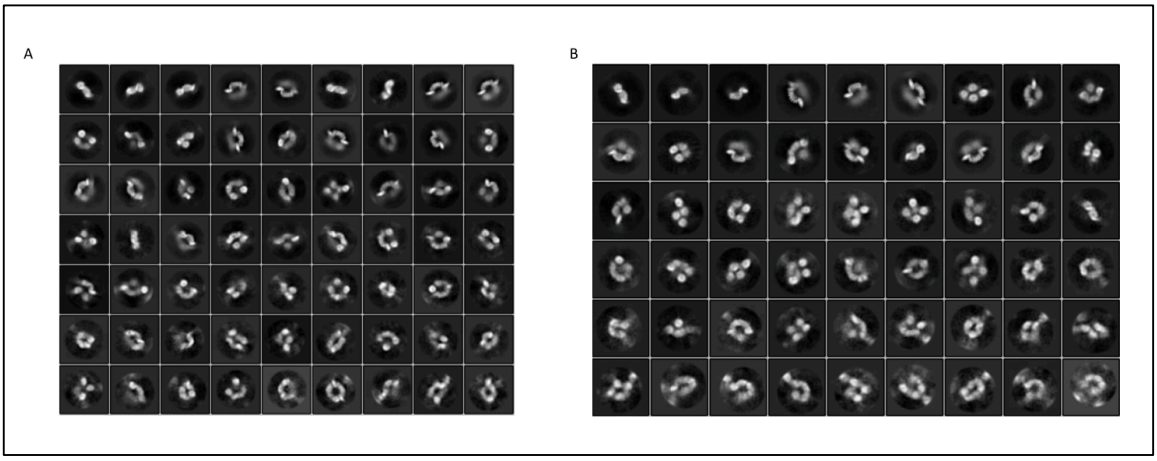


Figure 69: Final 2D classes obtained in Relion. 2D classes were generated from particles picked based on 2D templates from CryoSparc. A) Final classes corresponding to 598 k particles show secondary structure features for the HAT dimer and in some clear density for WD40 propellers below the HAT, indicating that a subset of particles contains the intact CstF1-CstF3 complex. B) 2D classes after clean-up selecting for particles containing density for the HAT and WD40 propellers. Final 2D classes contain 145 k particles.

In the first processing strategy (Material and Methods 4.2.10.5.1), a pre-cleaned particle stack (Figure 69 A) was first used for extensive clean-up in 2D classification by only selecting for visible density of HAT and WD40s. This resulted in final 2D classes shown in picture 69 B. Particles of this final 2D classes were used for 3D classification, which delivered a class showing secondary structure features for the HAT dimer and clearly visible spherical density below the HAT domain. The class contained 21.7 percent of the input particle stack, and was refined and postprocessed to a final resolution of 6.5 Å (Figure 70 A)

To exclude that some conformations of the complex are missed by using only particles generated by focusing 2D classification on the HAT dimer and WD40 propellers, a second processing approach (Material and Methods, 4.2.10.5.1) was used. All pre-cleaned particles (Figure 69 A) were subjected to another round of 3D classification, which delivered a second class of the CstF1-CstF3 subcomplex with slightly different conformation of the spherical density below the HAT dimer (Figure 70 B). After refinement and post processing, the estimated resolution was 5.3 Å. Estimated resolutions were different for both 3D reconstructions, but are expected to be in similar range since both maps show a similar grade of detailed structural information.

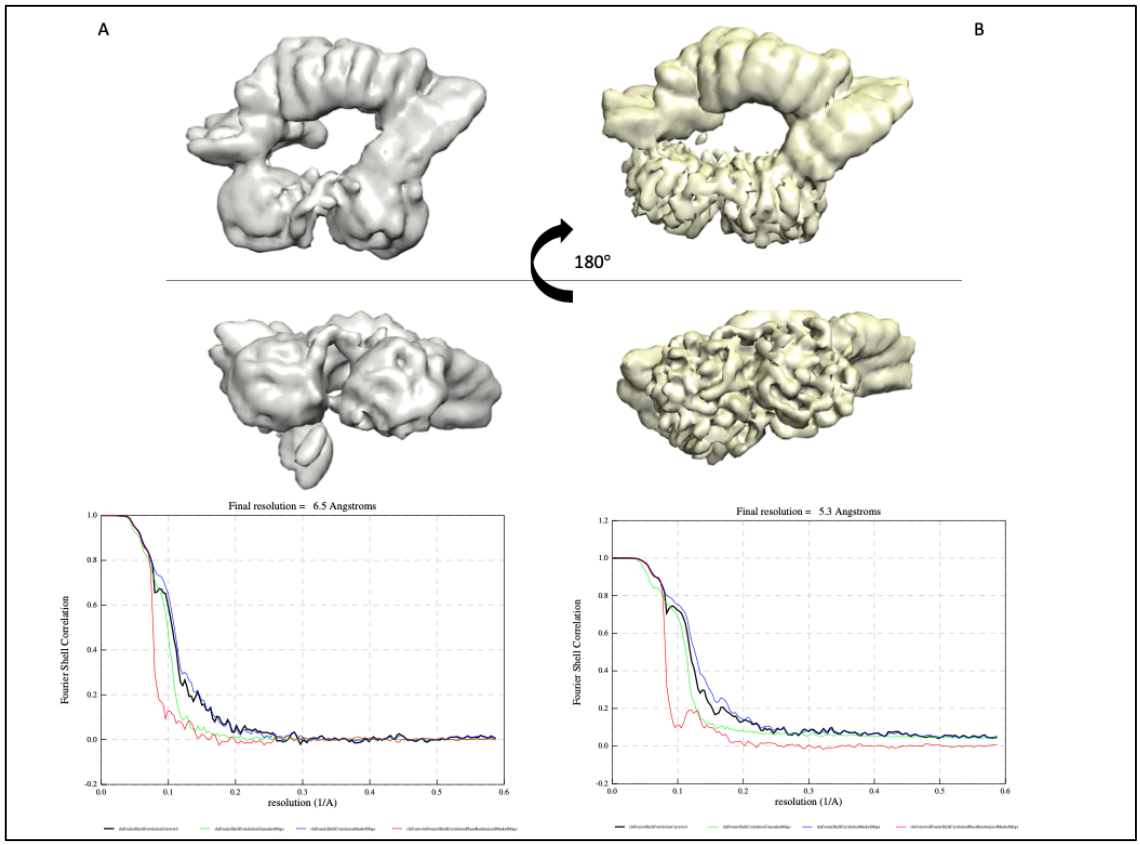


Figure 70: Medium resolution reconstructions of the CstF1-CstF3 subcomplex. Final 3D maps of the main two conformations of CstF1-CstF3 at 6.5 Å and 5.3 Å resolution, respectively obtained from the same cryo-EM dataset. The CstF1-CstF3 complex was prepared using a combination of in-batch cross-linking and density gradient ultracentrifugation. The 3D maps correspond to two major conformations that were possible to be obtained in 3D reconstructions showing flexible behavior of the WD40 propellers of CstF1. A) 3D map of the CstF1-CstF3 subcomplex containing clear density and secondary structures for the CstF3 HAT domain and improved signal for the WD40 propellers. The WD40 are separated from each other in this conformation. In the middle row, there is an additional density visible attached to the HAT. Below: Corresponding Gold standard Fourier shell correlation (GSFSC) indication a final resolution of 6.5 Å. B) 3D map of the CstF1-CstF3 subcomplex containing clear density and secondary structures for the CstF3 HAT domain but only limited signal for the WD40 propellers, which appear spikey after 3D refinement and postprocessing. The WD40 are close to each other in this conformation. Below: Corresponding Gold standard Fourier shell correlation (GSFSC) indicating a final resolution of 5.3 Å.

Maps in figure 70 clearly show, that the dimeric CstF3 HAT domain is resolved to a state where single helices become visible in both classes. Although density of the WD40 propellers of CstF1 is not as highly resolved as the HAT dimer, differences in the conformations are visible. In the map depicted in figure 70 A, the WD40s are more separated from each other than in the other 3D reconstruction (Figure 70 B). To make it more obvious, that depicted densities fit to subunits of the CstF1-CstF3 subcomplex, available crystal structures were fitted into final 3D reconstructions (Figure 71; CstF1: PDB 6P3X, Yang, Hsu et al., 2018; CstF3: 6URO, (Sun, Hamilton et al. 2020).

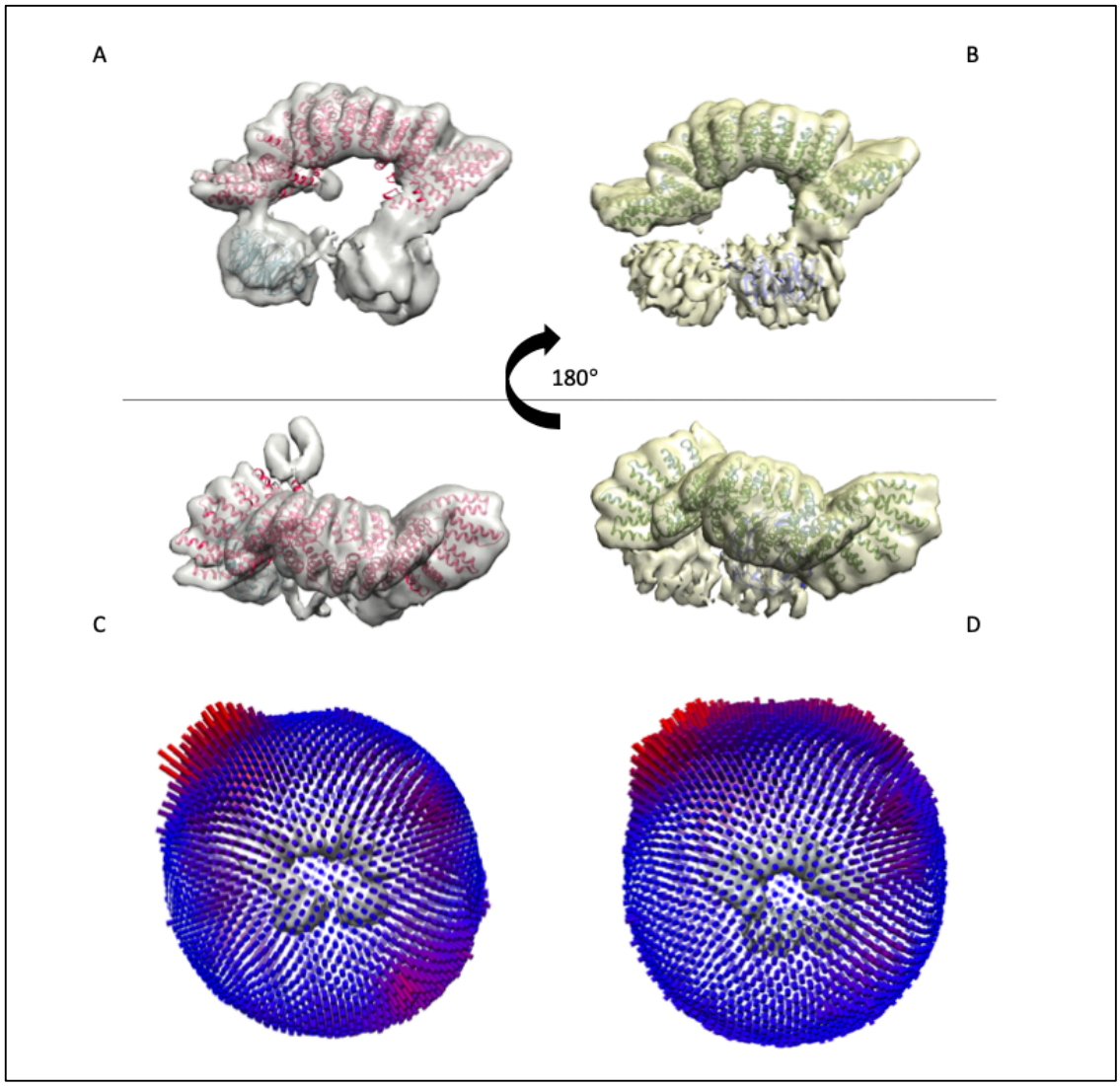


Figure 71: Crystal structures of CstF1 and CstF3 fitted into Cryo-EM reconstructions of CstF1-CstF3 subcomplex. 3D maps at medium resolution obtained in Relion from a cryo-EM data collection on a Titan Krios TEM with a K3 camera on the CstF1-CstF3 subcomplex prepared with a combination of in-batch cross-linking and density gradient ultracentrifugation. The 3D maps correspond to two major conformations that were possible to be obtained in 3D reconstructions showing flexible behavior of the WD40 propellers of CstF1. A) 3D map of the CstF1-CstF3 subcomplex containing clear density and secondary structures for the CstF3 HAT domain but only limited signal for the WD40 propellers, which appear spikey after 3D refinement and postprocessing. The WD40 are close to each other in this conformation. B) 3D map of the CstF1-CstF3 subcomplex containing clear density and secondary structure features for the CstF3 HAT domain and improved signal for the WD40 propellers. The WD40 are separated from each other in this conformation. Fitted structures: CstF1: PDB 6P3X, Yang et al., 2018; CstF3: 6URO, Zhang et al., 2020. C) angular distribution of the 6.5 Å map. D) Angular distribution of the 5.3 Å map.

Final 3D reconstructions of the CstF1-CstF3 subcomplex contained only a small subset of particles (105k particles for 6.5 Å and 31k particles for 5.3 Å reconstruction), which might also be a reason that resolution was limited.

2.5.3 Reconstruction of the CstF3 HAT dimer at high resolution

In order to improve resolution of the CstF1-CstF3 3D reconstructions (Figure 70 A and B), another dataset was collected on a grid which was prepared in same batch as the one used for data collection in the previous section. This means, that exactly the same sample was used for preparation of several grids. A data collection on a Titan Krios TEM was set up for several days. Detailed parameters of the data collection and workflow of data processing (Material and Methods, 4.2.10.5.2) are described in Material and Methods, 4.2.10.5. A representative micrograph is depicted in figure 72 A. A total number of 1 750 900 particles were picked and cleaned by several rounds of 2D classification, selecting only classes with clear density for HAT domain and WD40 propellers. The final particle stack contained around 560 000 particles. 2D classes of final particle stack are shown on figure 72 B.

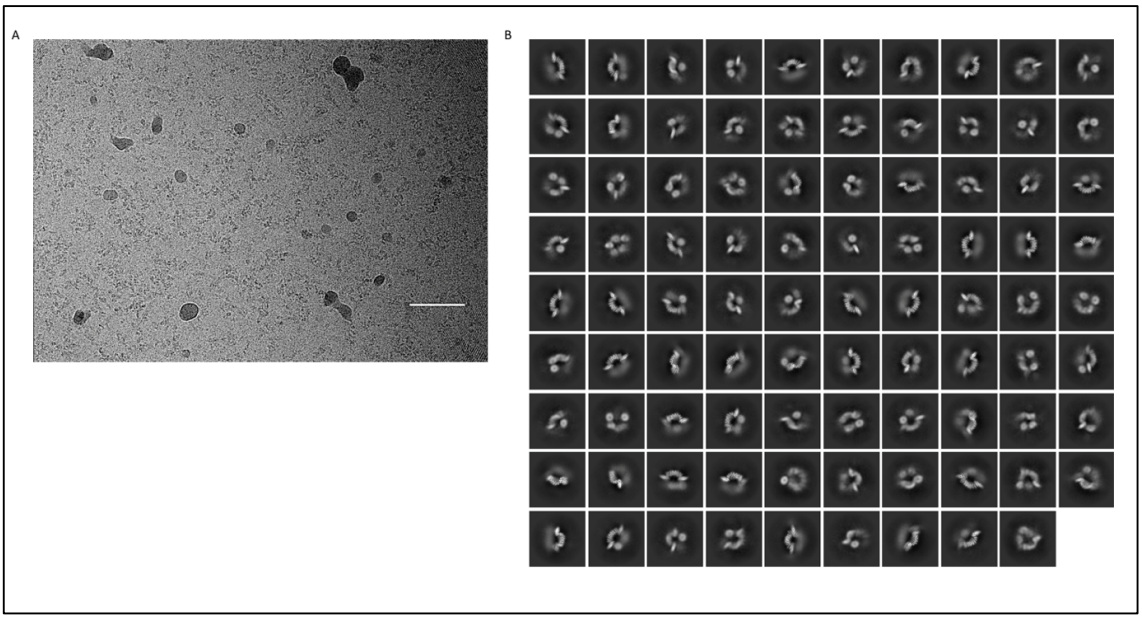


Figure 72: Dataset of the CstF1-CstF3 subcomplex shows complex flexibility mediated by CstF1 WD40 propellers. Micrograph (A) and 2D classes (B) obtained from a cryo-EM data collection on a Titan Krios TEM with a K3 camera on the CstF1-CstF3 subcomplex prepared with a combination of in-batch cross-linking and density gradient ultracentrifugation. All 2D classes were obtained in CryoSparc using the Topaz implementation for trained particle picking. A) Micrograph (scale bar corresponds to 60 nm) shows cross-linked CstF1-CstF3 particles in dark on a brighter background in good particle distribution and contrast. Ethane contamination is visible as big dark dots. B) 2D classes obtained in CryoSparc after using the Topaz implementation for particle picking. The classes contain 560 k particles and show secondary structure features for the HAT dimer and a high content of classes containing density for the WD40 propellers below the HAT, corresponding to the CstF1-CstF3 subcomplex. WD40s are arranged in different conformations and positions below the HAT indicating that the sub complex is highly flexible and can adopt different conformations.

2D classes depicted in figure 69 clearly show density for the CstF1 WD40 propellers visible below the HAT. It was already reported for processing of earlier datasets, that the WD40s adopt several conformations also in complex with CstF3, which can be clearly observed in these 2D classes as well. There are either two WD40 domains present in different

conformations towards each other or only one of the propellers is visible. 3D classification of classes containing densities of CstF1 and CstF3 resulted in different conformations at low-resolution, containing the HAT domain and either one or both WD40 propellers at different positions below the HAT (Figure 73). Even after density subtraction of the HAT domain and focused 3D classification was applied, it was not possible to obtain clear density for both WD40 propellers in 3D reconstructions. The number of particles in different classes was too small to improve resolution by 3D refinement.

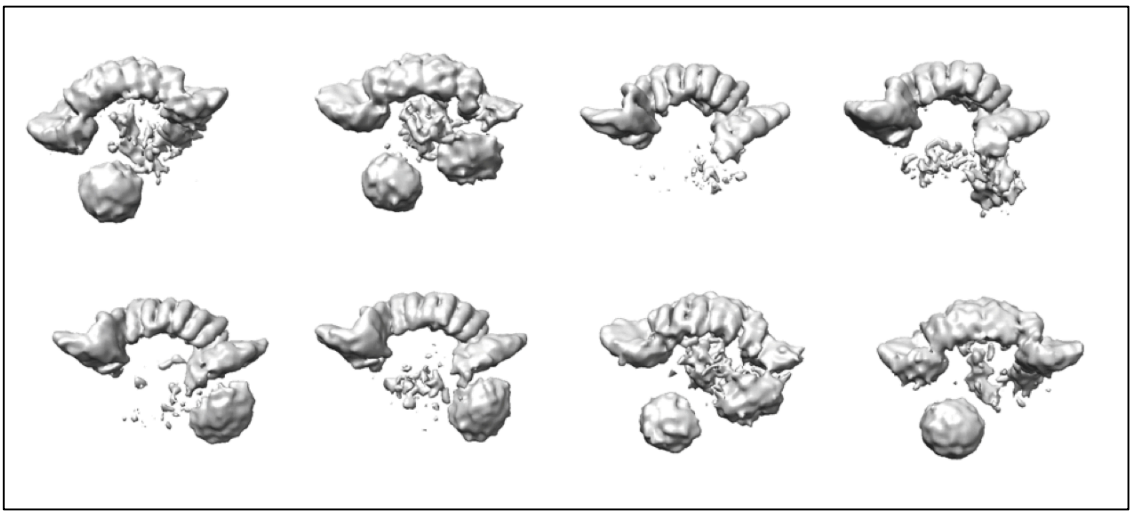


Figure 73: 3D reconstructions of the CstF1-CstF3 subcomplex at low resolution. 3D maps obtained from a cryo-EM data collection on a Titan Krios TEM with a K3 camera on the CstF1-CstF3 subcomplex prepared with a combination of in-batch cross-linking and density gradient ultracentrifugation. All 3D maps were obtained in CryoSparc from 3D classification with 327 k particles. Classes contained between 26 k and 77 k particles. The third class in the top row contained 77 k particles but had no density for the WD40 propellers. Secondary structure features are visible for the HAT dimer in all classes, whereas density for the WD40 propellers remains blobby.

Selection of 3D classes that contained the HAT only, resulted in a final class containing around 90 000 particles, which were refined and post processed to a final resolution of 3.43 Å . The sharpened 3D reconstruction is depicted in figure 74, where the structure of the human CstF3 HAT (pdb: 6URO) domain solved by Zhang and co-workers was fitted in (Zhang, Sun et al. 2020).

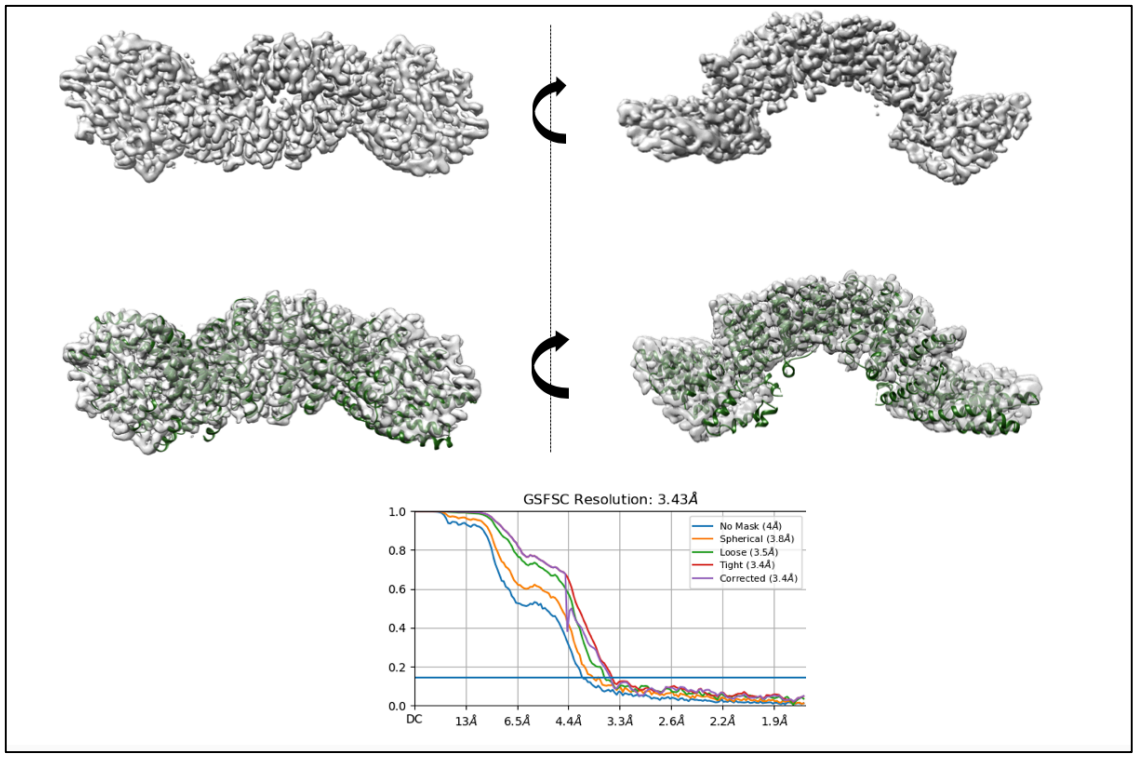


Figure 74: High resolution 3D reconstruction of the CstF3 HAT domain. 3D reconstruction of the human CstF3 HAT dimer at high resolution from a subset of 90 k particles, that did not contain any density for CstF1 WD40 propellers in 3D classification. After 3D refinement and sharpening, a final resolution of 3.43 Å was obtained as indicated by GSFSC. The structure of t human CstF3 HAT dimer obtained by Zhang et al., 2020 was fitted in the EM map.

2.6 Modelling of the CstF complex

2.6.1 Modelling of the CstF2-CstF3 interaction interface using AlphaFold

Previous studies showed that CstF3 binds to the so-called hinge region of CstF2 (Figure 72 A) in a mutually exclusive manner with Symplekin (Takagaki and Manley 2000, Ruepp, Schweingruber et al. 2010). The region of CstF3 binding to CstF2 is called monkeytail based on the crystal structure of the binding interface (Figure 75 B) of yeast homologues Rna14 (CstF3) and Rna15 (CstF2) (Leeper, Qu et al. 2010, Moreno-Morcillo, Minvielle-Sebastia et al. 2011). Yeast proteins Rna14 and Rna15 are part of the cleavage/polyadenylation factor IA (CF IA), which is the closest related factor to the human CstF complex (Minvielle-Sebastia, Preker et al. 1994, Minvielle-Sebastia, Preker et al. 1997, Mandel, Bai et al. 2008). Like the human CstF3 subunit, the Rna14 protein is mainly characterized by a HAT domain, which mediates the homodimeric association of Rna14 (Noble et al., 2004). The C-terminus shares weak sequence similarity to human CstF3. Similar to its human homologue CstF2, the Rna15 protein binds to pre-mRNA in context of CF IA via its N-terminal RRM (Noble, Walker et al. 2004, Leeper, Qu et al. 2010, Pancevac, Goldstone et al. 2010, Paulson and Tong 2012). The region following the RRM is similar to the hinge domain of CstF2 and involved in interaction with monkeytail (residues 593-677) of Rna14 (Legrand, Pinaud et al. 2007, Hockert, Yeh et al. 2010). Therefore, both proteins are required for complex formation. The crystal structure (Figure 75 B) of a minimal complex of Rna14 and Rna15 revealed the molecular interaction between the Rna15 hinge domain and Rna14 monkeytail (Moreno-Morcillo, Minvielle-Sebastia et al. 2011, Paulson and Tong 2012).

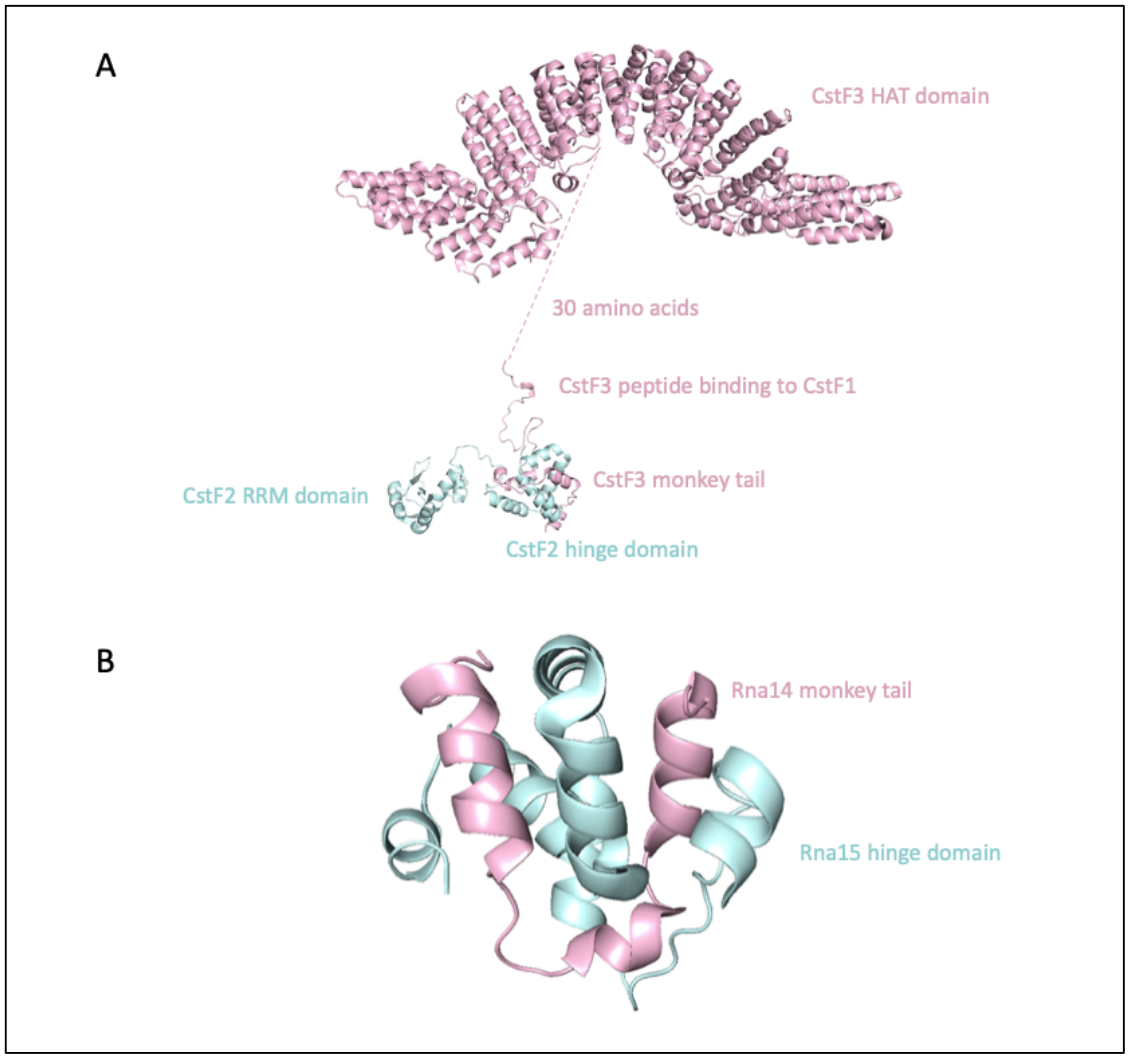


Figure 75. Overview of structures of CstF2 and CstF3 and their yeast homologs. A) Cryo-EM structure of the CstF3 HAT dimer (pdb 6URO) and an AlphaFold model of the interaction between CstF2 and CstF3. The CstF3 construct contains the peptide (residues 580-594), which mediates binding to CstF1, and the monkeytail (residues 595-653), which binds to CstF2. The CstF2 construct contains the hinge domain (residues 112-199) interacting with the CstF3 monkey tail and the CstF2 RRM (residues 1-111). B) Crystal structure of the interaction between yeast Rna14 monkeytail and Rna15 hinge domain (Moreno-Morcillo, Minvielle-Sebastia et al. 2011).

Although the interacting regions are conserved from yeast to human, there is no structure of the human proteins available yet. Therefore, a model of the human CstF2 hinge domain binding to the CstF3 monkeytail (Figure 75 A) was generated using AlphaFold (AF; Jumper et al., 2021). In combination with information generated by sequence alignment between yeast Rna14 and Rna15 and human CstF3 and CstF2 (Figure 76 A and B) and the available structure of yeast proteins, the AF model of the human CstF2-hinge and CstF3-monkeytail interaction was validated and residues important for interaction of both proteins were identified.

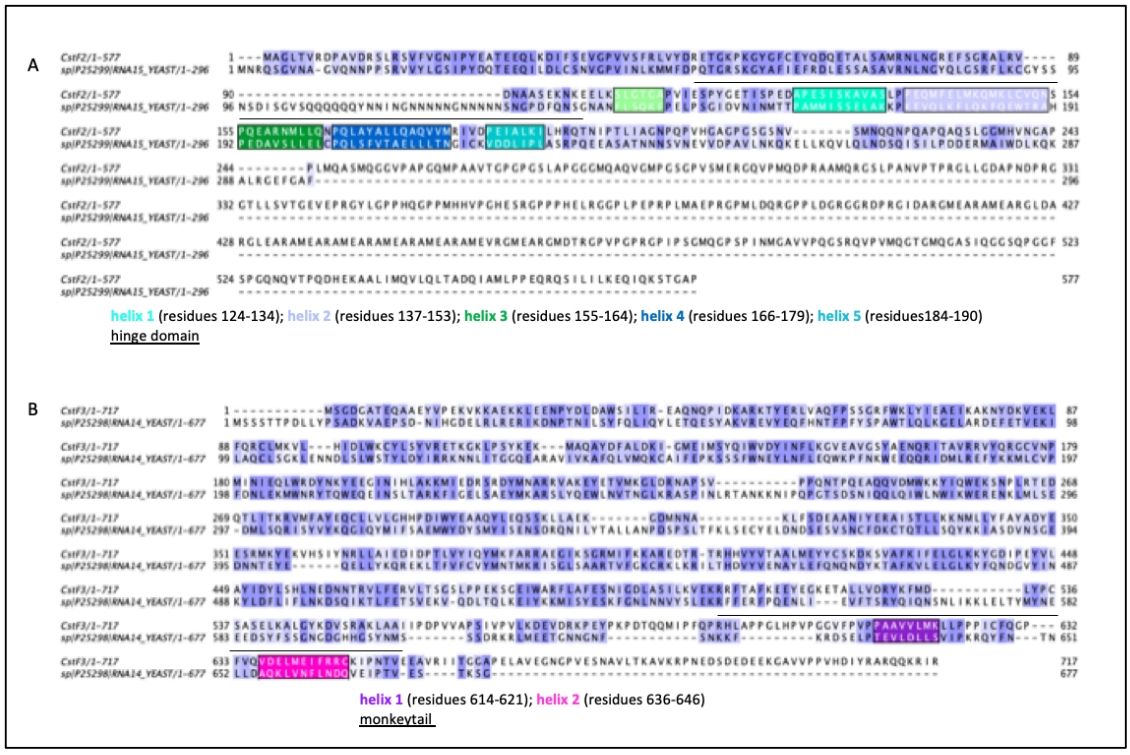


Figure 76. Sequence alignment between human CstF2 and CstF3 and their yeast homologs Rna15 and Rna14. Sequence alignments were edited in Jalview. A) Sequence alignment between human CstF2 and its yeast homolog Rna15 shows high sequence conservation in the N-terminal RRM domain (residues 1-111) and the hinge domain (overlined, residues 112-199). Characteristic helices of the hinge domain are colored based on the yeast structure (Moreno-Morcillo et al., 2011) in different blue and green shades. The first green box depicts the C-terminal helix of the RRM domain. Cyan: helix 1 (residues 124-134); lightblue: helix 2 (residues 137-153); green: helix 3 (residues 155-164); blue: helix 4 (residues 166-179); aquamarine: helix 5 (residues184-190). B) Sequence alignment between human CstF3 and its yeast homolog Rna14 shows high sequence conservation in the throughout the whole protein. Characteristic helices of the monkeytail (overlined, residues 594-653) are colored based on the yeast structure (Moreno-Morcillo, Minvielle-Sebastia et al. 2011) in different pink and purple. Purple: helix 1 (residues 614-621); pink: helix 2 (residues 636-646).

The construct used as AF input for CstF2 contained the N-terminal RRM (residues1-111) followed by the hinge domain (residues 112-199) to avoid any clashes of CstF3 with the RRM of CstF2 during modelling (Yang, Hsu et al. 2018). The input sequence of CstF3 was spanning over residues 580-660 containing the peptide, which is binding to CstF1 and the monkeytail (residues 594-653) (Yang, Hsu et al., 2018). Besides that, structure prediction by AlphaFold also contained the WD40 propeller of CstF1 (residues 80-431) to correctly position CstF3 on CstF1 and avoid later clashing of individual models. Models were calculated using the AF multimer colab version (Jumper, Evans et al. 2021) et al., 2021) with default settings and a number of five models to predict. Resulting models were initially analyzed based on their per-residue confidence plot, called pLDDT, and secondly by aligning them to available pdb structures. The corresponding output of the AlphaFold run containing the pLDDT plot for the final model, consisting of a minimal monomer formed of CstF1, CstF2 and CstF3, is depicted in figure 78.

For structural analysis of the CstF2-CstF3 interaction, the WD40 propeller of CstF1 was hid in following figures to simplify the depicted interacting regions. Residues responsible for interaction in the yeast structure were correlated to human proteins based on the sequence alignment in figure 76 (Figures 76 A and B).

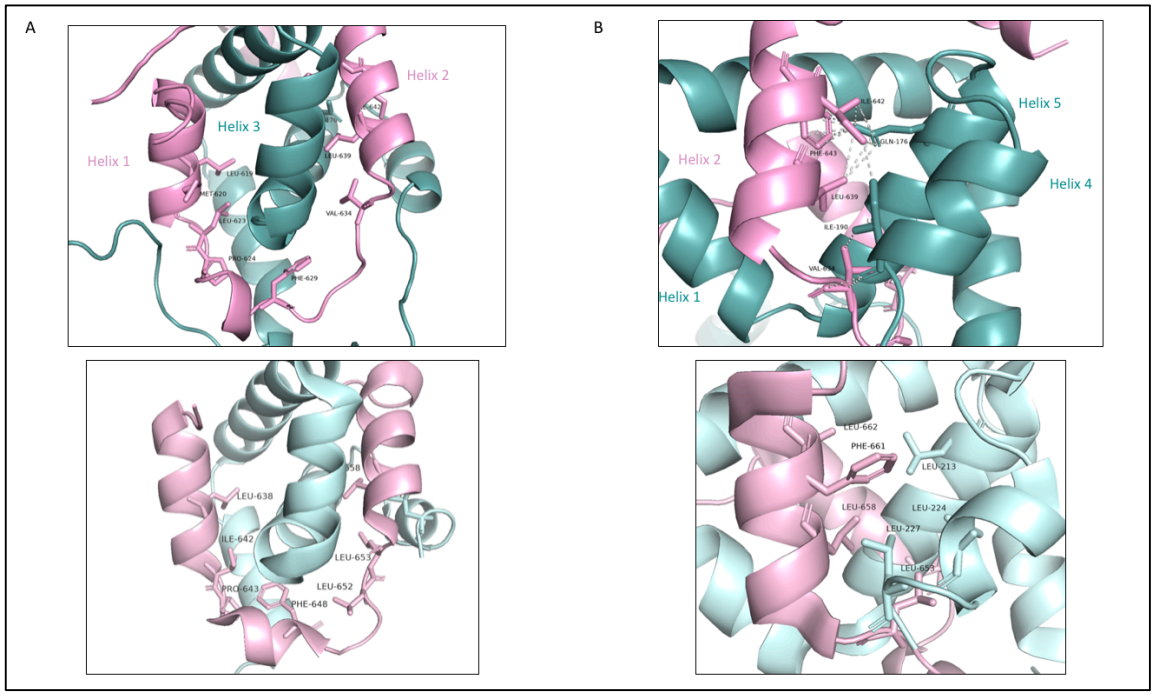


Figure 77. Structure of a heterodimer formed by the human CstF2 hinge domain and CstF3 monkey tail modelled with AlphaFold (Jumper et al., 2021). Top panel A and B: Cartoon representation of the model representing the tight interaction of CstF2 and CstF3. Models were analyzed and edited in PyMol (Schroedinger). CstF2 (Rna15) is depicted in deepteal and CstF3 (Rna14) is shown in pink. Bottom panel A and B: corresponding orientation of the yeast Rna15-Rna14 structure (Moreno-Morcillo, Minvielle-Sebastia et al. 2011) A) Residues of the CstF3 (Rna14) monkey tail contacting helix 3 of the CstF2 (Rna15) hinge domain. B) Side chain contacts formed between helix 2 of CstF3 (Rna14) and helix 4 and helix 5 of CstF2 (Rna15). Hydrogen-bonds stabilizing CstF2 and CstF3 are depicted as grey dashed line.

The structure of the interaction of yeast homologs of CstF2 and CstF3 was described in previous studies as a central bundle of four helices of Rna15 being surrounded by the Rna14 peptide flanking the bundle with helix 1 and helix 2 on each side (Figure 77 A, bottom panel). The same conformation was predicted for human proteins. Sandwiching the central core of CstF2 is mediated by mostly hydrophobic and aromatic side chains in CstF3 forming a hydrophobic pocket for the CstF2 helical core (Figure 77 A, upper panel). Neighboring helices of CstF2 (helix 4 and helix 5) are tightly interacting by hydrophobic intramolecular contacts and Hydrogen-bonds (H-bonds) formed by hydrophobic and polar residues (Ile190, Leu 173, Gln 176), which in turn are closely packed against helix 2 of CstF3. Hydrophobic amino acids of helix 2 are pointed towards the core of the heterodimer, thereby stabilizing the sandwiched structure of both proteins (Figure 77 B, upper panel). To sum up, by modelling the CstF2-CstF3 interaction interface and combining the model with information from homology to the structure of yeast Rna14-Rna15, it was identified that the CstF3 monkey tail tightly wraps around the

helical core of the CstF2 hinge domain by hydrophobic interactions conserved from yeast to human.

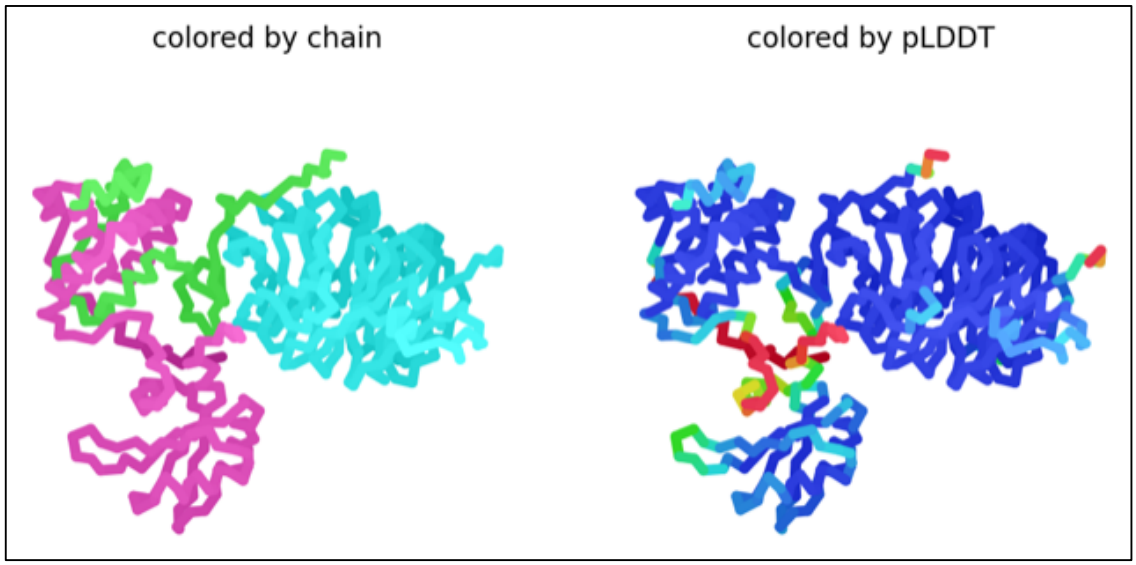


Figure 78. Model of a minimal CstF monomer calculated by AlphaFold (Jumper, Evans et al. 2021). Left panel: calculated model consisting of the WD40 propeller of CstF1 (residues 80-431) depicted in cyan interacting with a peptide of CstF3 (residues 580-594) shown in green, which then continues to form the monkey tail interacting with CstF2 (pink). CstF2 (residues 1-199) contains the N-terminal RRM directly followed by the hinge domain (residues 112-199) forming a locked conformation with CstF3. Left panel: Coloration of the model according to the per-residue confidence. Blue: high confidence, green: medium confidence, red: low confidence.

2.6.2 Modelling of a minimal CstF1-CstF2-CstF3 complex by combining structural information from cryo-EM, XL-MS and AlphaFold

Although no high-resolution reconstruction was obtained from the CstF1-CstF3 subcomplex, structural information from the 5.3 Å reconstruction (Figure 70 and 71) was used to create a model of a minimal Cst1-CstF2-CstF3 complex. Available structures of the HAT domain of CstF3 (pdb: 6URO) and the N-terminal homodimerization domain of CstF1 (pdb: 2XZ2) were directly used for the model. To correctly position the CstF3 monkeytail and CstF2 hinge domain in relation to a WD40 propeller of CstF1, a model of a minimal CstF monomer was calculated using AlphaFold. Following model was calculated with one CstF1 WD40 propeller (residues 101-431) bound to CstF3 (residues 576-595), which is downstream interacting with the CstF2 hinge domain via the monkeytail (residues 596-660). Since the hinge domain directly follows the N-terminal RRM, the CstF2 construct used for model calculation contained residues 1-197. To simplify the model, the long loop of CstF3 connecting CstF-interacting peptide (residues 576-593) and monkeytail (residues 607-660) was deleted as well as unstructured residues 108-119 of CstF2 (Figure 79). XL-MS data were generated as described in Material and Methods by in-batch cross-linking of full CstF complex with 2mM BS3 before loaded on a sucrose density gradient followed by analytical SEC. XL-MS data were analyzed and mapped

using Chimera Xlink analyzer (Kosinski, von Appen et al. 2015). Fitting available and theoretical (AlphaFold) structures into the final 5.3 Å reconstruction of CstF1-CstF3 (see paragraph 2.5.2) was done by first recognizing the most prominent map features and relating them to available structures. Second, XL-MS data obtained in this study were used to find the most likely orientation of domains and overall validate the fit. Fitted structures and models were oriented in a way, that as many cross-links as possible of one monomer were within a reasonable length of 30 Å (Figure 79). First, the HAT domain (pdb: 6URO) was fitted in the corresponding density of the cryo-EM map (Figure 79 A). Due to map resolution, structure of CstF3 HAT dimer could automatically fitted in corresponding density by Chimera. Next, the remaining prominent densities correspond most likely to the doughnut shaped WD40 domain of CstF1, which could be fitted well (Figure 79 B). Due to limited resolution of the map, the rotational orientation of the WD40 domain was not obvious. But with help of XL-MS data, the WD40 propeller was fitted in the most probable orientation (Figure 79 B). The CstF3 peptide was stably anchored on the WD40 propeller based on the crystal structure (pdb: 6B3X). Since there was no density for the CstF1 NTD observed in the cryo-EM reconstruction, positioning of the NTD is a rough estimation based on the few cross-links (Figure 79 C) formed from CstF3 HAT domain.

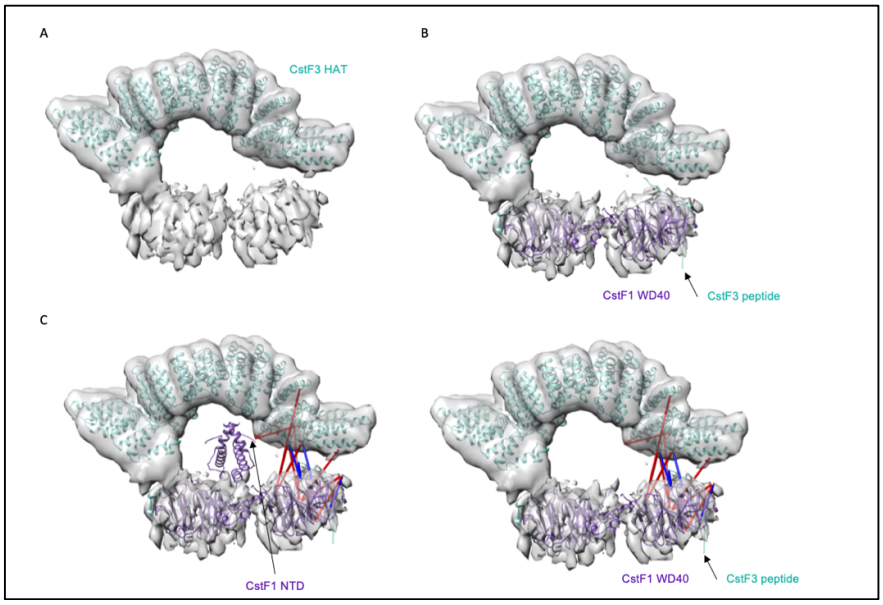


Figure 79. Structures of CstF1 and CstF3 fitted into and a 5.3 A reconstruction of CstF1-CstF3 with help of XL-MS data. EM maps well as structures and models were displayed in Chimera (Pettersen, Goddard et al. 2004). CstF1 (residues 4-65 and 101-431) is shown in dark purple, CstF2 RRM and hinge domain (1-197) are shown in wheat and CstF3 (residues 25-550 and 576-660) is colored in deepteal. CstF1 NTD: CstF1 N-terminal homodimerization domain, CstF3 peptide: residues 576-593 of CstF3 binding to WD40 propeller of CstF1. A) Cryo-EM map of CstF1-CstF3 with fitted HAT structure (pdb 6URO). B) AlphaFold model of CstF1 WD40 propeller containing the bound CstF3 peptide fitted into cryo-EM reconstruction (upper panel). Displayed cross-links used to orient WD40 propeller below the HAT domain. C) CstF1 N-terminal homodimerization domain (NTD) placed without corresponding density based on cross-links formed to CstF3 HAT domain. Cross-links were mapped on each CstF1-CstF2-CstF3 monomer with a minimal cross-link score of 90. Intra cross-links were hidden in the figure. Blue: Cross-links shorter than 30 Å red: cross-links longer than 30 Å.

Afterwards, the locked conformation of CstF3 monkeytail (MT) and CstF2 hinge domain (paragraph 2.6.1) was placed according to displayed cross-links (Figure 80 A). Last, CstF2 RRM was positioned in a cleft formed between CstF1 WD40 and CstF3 HAT (Figure 80 B and 82 D).

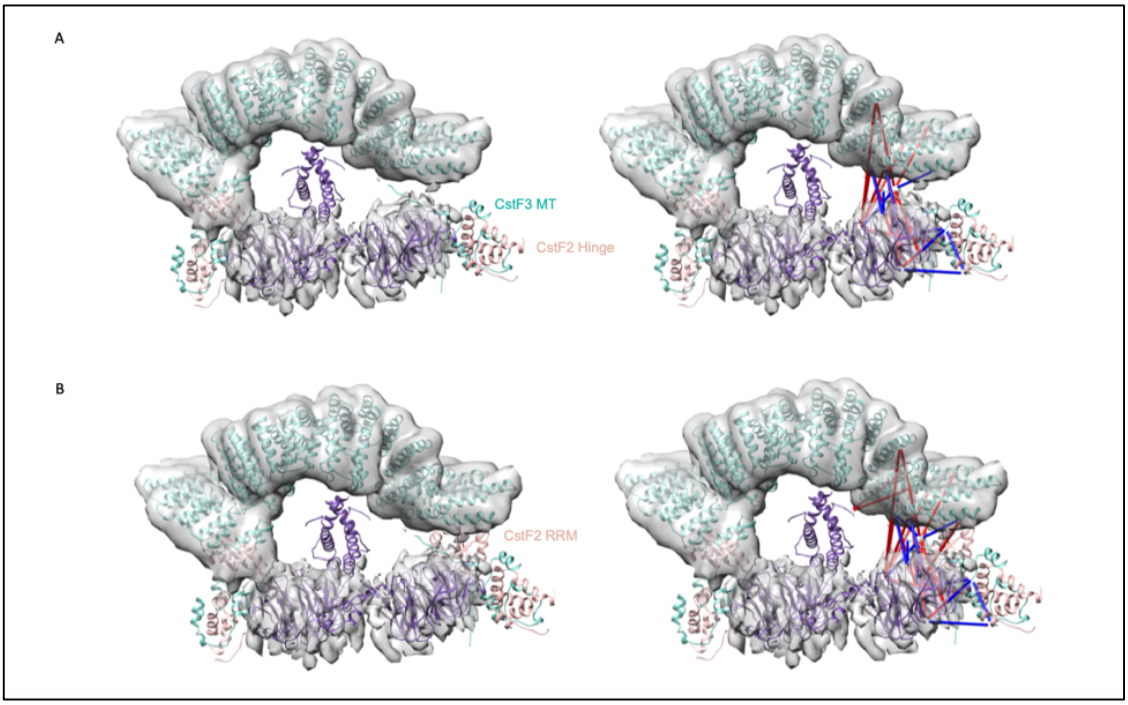
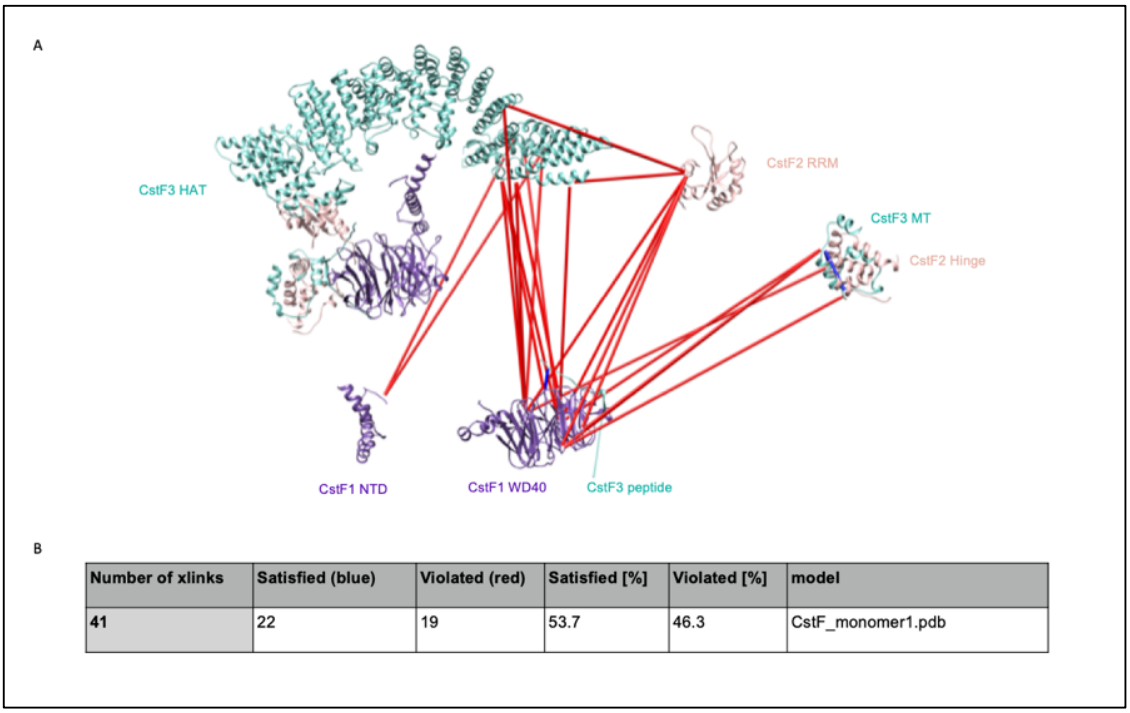


Figure 80. Structure of CstF2-CstF3 monkeytail-hinge conformation and CstF2 RRM placed close to density for CstF1-CstF3 with help of XL-MS. CstF1 (residues 4-65 and 101-431) is shown in dark purple, CstF2 RRM and hinge domain (1-197) are shown in wheat and CstF3 (residues 25-550 and 576-660) is colored in deepteal. A) Possible localization of CstF2-CstF3 monkeytail-hinge structure based on displayed cross-links in one monomer (right panel of A) CstF3 MT: monkeytail of CstF3. B) Possible position of CstF2 RRM domain based on displayed cross-links within one monomer. Only inter-subunit crosslinks with a minimum cross-linking score of 90 are displayed. Blue: Cross-links with a length of 30 Å and shorter. Red: cross-links longer than 30 Å.

Structural arrangement of the CstF subunits fitted into the cryo-EM reconstruction of CstF1-CstF3 (Figure 80 B) is reasonable in a way, that distances of most of the displayed cross-links are within a range of 30 Å. This distance constraint between C_α atoms of cross-linked lysine residues was shown to be appropriate for the BS3 cross-linker (Merkley, Rysavy et al. 2014). The cross-linking interface between all three subunits is visible in an exploded model in figure 81 A and shows different cross-linking clusters for each subunit.



B

Number of xlinks	Satisfied (blue)	Violated (red)	Satisfied [%]	Violated [%]	model
41	22	19	53.7	46.3	CstF_monomer1.pdb

Figure 81. Exploded model of cross-links displayed on CstF subunits. A) In total 41 cross-links with a minimum cross-linking score of 90 are displayed on an exploded CstF monomer. CstF1 (residues 4-65 and 101-431) is shown in dark purple, CstF2 RRM and hinge domain (1-197) are shown in wheat and CstF3 (residues 25-550 and 576-660) is colored in deepteal. CstF1 NTD: CstF1 N-terminal homodimerization domain, CstF3 peptide: residues 576-593 of CstF3 binding to WD40 propeller of CstF1. CstF3 MT: monkey tail. B) Summary of satisfied cross-links in the final CstF model. Satisfied (blue): Cross-links with a length of 30 Å and shorter. Violated (red): cross-links longer than 30 Å.

In the resulting model of the minimal CstF complex (CstF1-CstF2^{RH}-CstF3^{HAT-MT}), which is depicted in figure 82 A, 53.7% of all displayed cross-links (Figure 81 B) showed a distance constraint of 30 Å or less (Figure 81 B). According to the description of the way of fitting single CstF subunits, is was possible to obtain a final model, where each of the subunits or structures contained several cross-links with a distance less than 30 Å (Figure 82 C-F). In combination with the high cross-linking core of over 90, the arrangement of structures in the final fit makes sense (Figure 82). Figure 82 and the following text gives a detailed overview of the cross-linking interface of all structures.

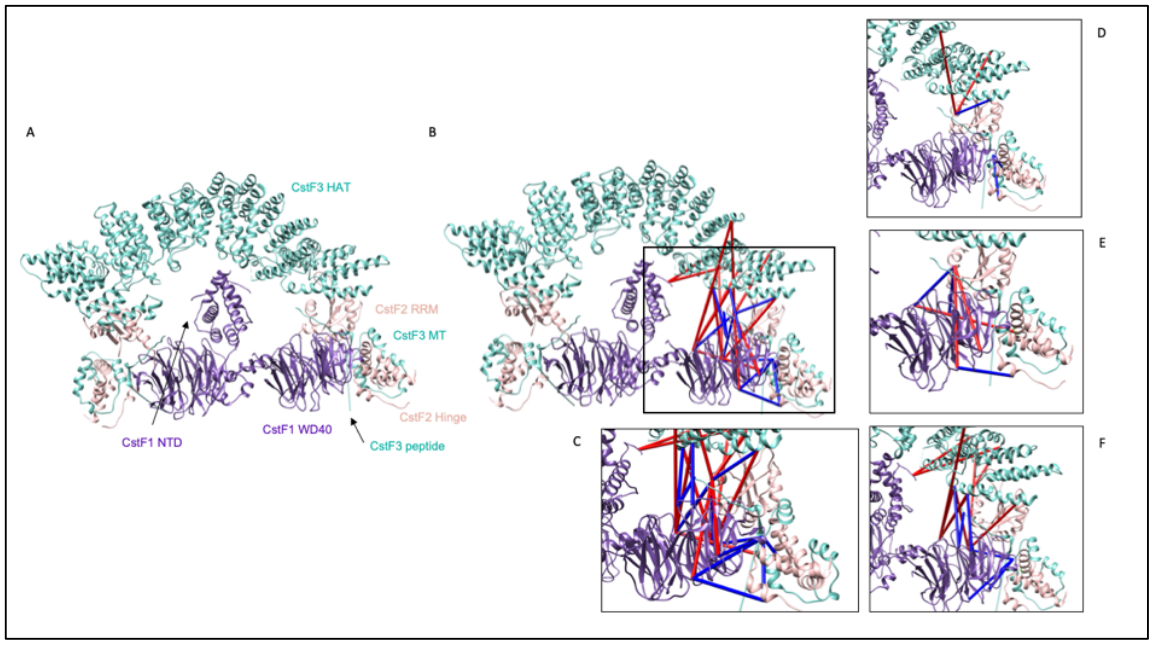


Figure 82. CstF model and detailed view of the cross-linking interface between its subunits. CstF1 (residues 4-65 and 101-431) is shown in dark purple, CstF2 RRM and hinge domain (1-197) are shown in wheat and CstF3 (residues 25-550 and 576-660) is colored in deepteal A) CstF model without cross-links. CstF1 NTD: CstF2 N-terminal homodimerization domain, CstF3 MT: monkeytail of CstF3. CstF3 peptide: residues 576-593 of CstF3 binding to WD40 propeller of CstF1. B and C) Detailed view of the cross-linking interface between CstF1, CstF2 and CstF3. Only inter-subunit crosslinks with a minimum cross-linking score of 90 are displayed. Blue: Cross-links shorter than or equal 30 Å red: cross-links longer than 30 Å. D) Cross-links between CstF2 and CstF3. E) Cross-links between CstF1 and CstF2. F) Cross-links between CstF1 and CstF3.

The CstF1 NTD only showed few cross-links at Lysin 5 pointing towards the HAT domain of CstF3. However, the WD40 propeller of CstF1 has a set of Lysin residues, that were either cross-linking to different residues of the CstF3 HAT or either RRM or hinge domain of CstF2 (Figure 79 E and F). All those Lysines (K204, K212, K302, K316, K319 and K326) were located in loops pointing to the outer surface on one side of the WD40 propeller, thereby allowing determination of the ‘upper side’ of the propeller, which is arranged towards CstF3 HAT-N. This orientation of CstF1 WD40 propellers is supported by additional density in the cryo-EM map, which could correspond to the peptide of CstF3 binding to CstF1 (Figure 83). The N-terminal end of the peptide should be connected to the HAT-C domain of CstF3 via 30 amino acids. Based on this connection to the CstF3 HAT domain and on XL-MS data in this thesis, I excluded the possibility that WD40 propellers are flipped horizontally.

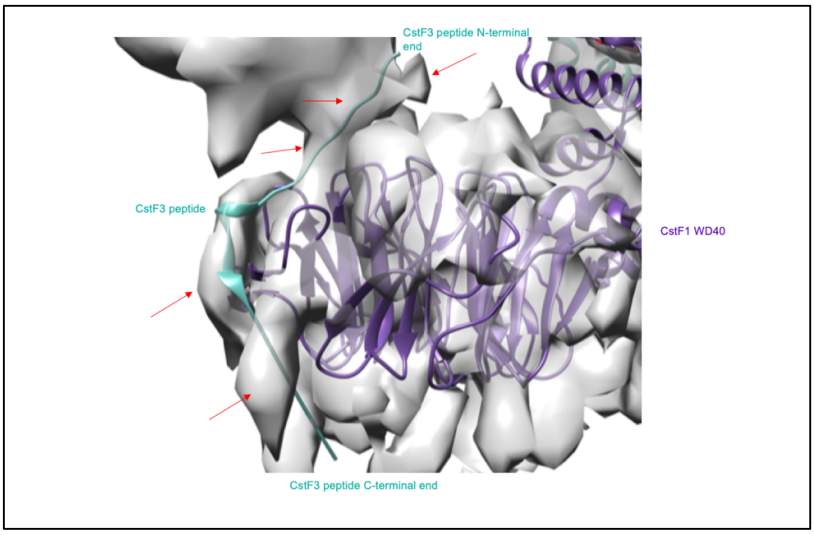


Figure 83. Additional density in the cryo-EM reconstruction of CstF1-CstF3. CstF1 (residues 4-65 and 101-431) is shown in dark purple and CstF3 (residues 576-660) is colored in deepteal. CstF3 peptide bound to WD40 propeller of CstF1 was fitted into EM density based on XL-MS data. Additional density in the cryo-EM map indicated by red arrows, is directly located where CstF3 peptide was fitted.

The CstF2 RRM only had one major cross-linking cluster, K96, which showed a number of cross-links to CstF1 and CstF3 (Figure 82 D). This allowed rough positioning of CstF2 RRM between the cleft formed by one WD40 propeller and the N-terminal HAT domain (Figure 84 B). By pointing into this cleft with the loop containing K96, the β-sheet representing the RNA binding interface is rotated in a way, that it would be accessible for RNA binding. However, exact rotation of the RRM domain could not be determined based on the cross-links to K96. The hinge domain of CstF2 was cross-linking to WD40 propellers of CstF1 and due to their locked conformation to the monkeytail of CstF3 via Lysines 148 and 189 (Figure 82 E). Consequently, the CstF2 hinge domain in complex with CstF3 monkeytail was positioned on the outer side of CstF1 WD40 propellers below HAT-N. CstF3 was forming most cross-links via the N-terminal HAT domain, thereby allowing positioning of the CstF1 WD40 propeller in a certain position below the HAT. The number of cross-links is quantified per residue in figure 84 A. Therefore, all cross-links with a minimum cross-linking score of 80 were counted per monomer.

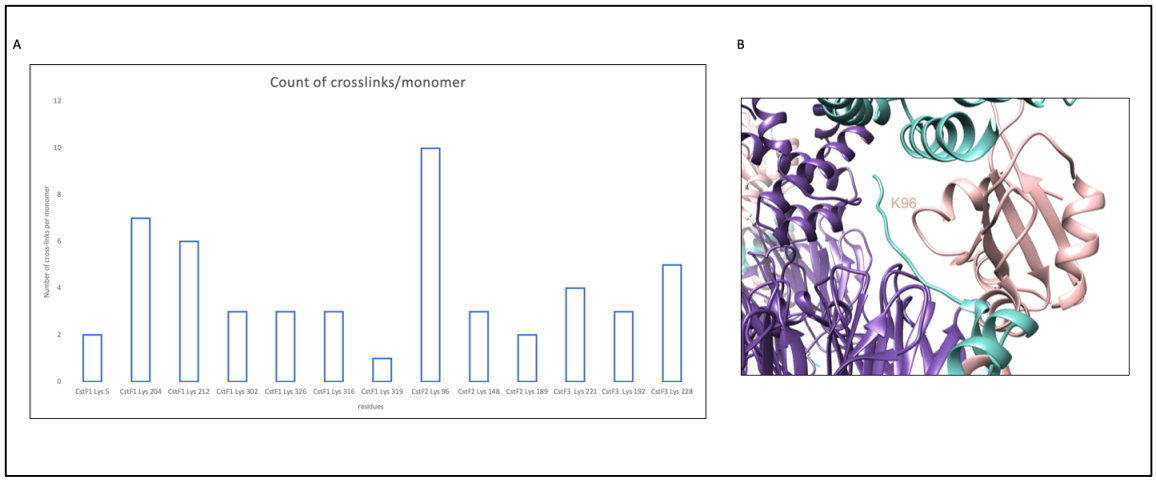


Figure 84. Quantification of cross-links per residue and monomer. A) All Cross-links of one monomer with a minimum cross-linking score of 80 were quantified per residue. B) Lysine 96 of CstF2 RRM pointing into a cleft formed between CstF1 and CstF3.

3 Discussion

Coupled to the process of transcription, a pre-mRNA transcript has to undergo several processing steps until a mature mRNA can be exported into the cytoplasm, where it can be translated into the corresponding protein sequence. 3′-end processing consisting of cleavage of the pre-mRNA at the poly(A) site followed by addition of a 200-250 nt long poly(A) tail is one of the essential steps of pre-mRNA maturation. A huge protein machinery is necessary to perform the crucial actions required for correct positioning of the 3′-end processing machinery on the pre-mRNA and subsequent cleavage and polyadenylation at the poly(A) site. The so-called Cleavage Stimulation Factor CstF is one component of the human 3′-end processing machinery involved in definition of the cleavage site by its ability to bind to G/U-rich sequence elements on the pre-mRNA downstream of the poly(A) site (Cheng et al., 1995; Graber et al., 1999; MacDonald et al., 1994; Takagaki and Manley, 1997). According to its name, the CstF complex was shown to be involved in cleavage of pre-mRNA targets (Takagaki et al., 1989). Human CstF complex consists of three proteins, CstF1, CstF2 and CstF3, which are believed to assemble in a 2:2:2 ratio, by means of the self-dimerizing capabilities of CstF1 and CstF3 (Yang et al., 2017; Bai et al., 2007). Binding to G/U-rich sequence elements on the pre-mRNA is mediated via the N-terminal RRM domain of CstF2 (Takagaki et al., 1992; Perez-Canadillas and Varani, 2003; MacDonald et al., 1994). Previous studies identified a model for recognition of a UU-dinucleotide, which explains discrimination between G/U and A/C nucleotides (Perez-Canadillas and Varani, 2003). In this study, recombinantly expressed and purified CstF complex consisting of its full-length subunits was used to, first, undertake structural studies by cryo-EM and second, shed light on RNA binding to a G/U-rich RNA using full-length or minimal, dimeric RRM fusion constructs.

3.1 A baculoviral protein co-elutes with human CstF2 during purification

To better understand the molecular mechanism of recognition and binding to RNA by the human CstF complex, full-length components had to be obtained in reasonable amounts to be able to reconstitute the full complex or subcomplexes. Besides that, single subunits, in this case CstF2, were used for RNA binding experiments. Purification protocols for the CstF complex, as well as CstF1-CstF3 and CstF2-CstF3 subcomplexes, were successfully established after optimizing baculovirus-mediated expression in insect cells and using combinations of certain purification techniques (paragraph 2.1). According to my results, initial purification trials of the single CstF2 protein revealed a stably associated contaminant (Paragraph 2.1.3). So far, not much information was available in literature about the purification of overexpressed human full-length CstF2, because either a truncated version containing only

the RRM and hinge domain (Yang, Hsu et al. 2018) was used, or the protein was pulled out from HeLa total cell extract (Ruepp, Schweingruber et al. 2010). In my hands, overexpressed CstF2 eluted from Strep-tag affinity purification with a tightly associated contaminant (Paragraph 2.1.3). The co-eluting contaminant was visible on SDS PAGE in an almost one to one ratio with CstF2 and was even resistant towards high salt concentrations in washing steps. By in-gel mass spectrometry, this band was identified as a baculoviral protein, the so-called Spodoptera frugiperda Ascovirus (SfAV) ORF046. This open reading frame encodes a 63.4 kDa protein named 'multifunctional domains-SbcC/ATPase, SMC, Hec1, Reovirus-sigma1 and Intermediate filament protein domains' (source: Uniprot.org).

Based on scarce information from literature, this protein contains several very conserved multifunctional domains (SMC, Hec1) and shows high homology to an intermediate filament protein. Intermediate filament containing proteins are involved in virus maturation and release (Cudmore, Reckmann et al. 1997, Heath, Windsor et al. 2001). A potential role for the protein encoded by SfAV ORF046 is to serve as a scaffold for baculovirus assembly (Bideshi, Demattei et al. 2006). So far, it is not known, why this protein is so tightly associated to human CstF2 during purification even at high salt concentrations. Sequence alignment between human CstF2 and SfAV ORF046 (Figure 85 A) shows partially similar residues within the N-terminal region covering the first 200 residues of CstF2 (corresponding to the RRM and hinge domain) and the C-terminus (last 90 residues) of CstF2. The predicted unstructured middle part of CstF2 shows additional similarity to SfAV protein stopping at the MEARA/G repeats of CstF2. A Structure prediction of SfAV ORF046 using AlphaFold supposed a huge coiled-coil starting immediately at the N-terminus of the protein and an unstructured C-terminus (Figure 82 B). However, no information was available about the interaction of CstF2 and coiled-coil intermediate filament proteins, that could explain tight association between two proteins. A BLAST research with SfAV ORF046 did not reveal any significant hit known to interact with CstF2. Although the baculoviral protein was bound in a one to one ratio to CstF2 according to SDS PAGE (Figure 28, Paragraph 2.1.3), it did not interfere with CstF complex formation, structural studies by cryo-EM or biophysical measurements, because it was lost in later purification steps of the full CstF complex. Therefore, the tightly associated contamination band appearing in early steps of purification could be disregarded for downstream analysis.

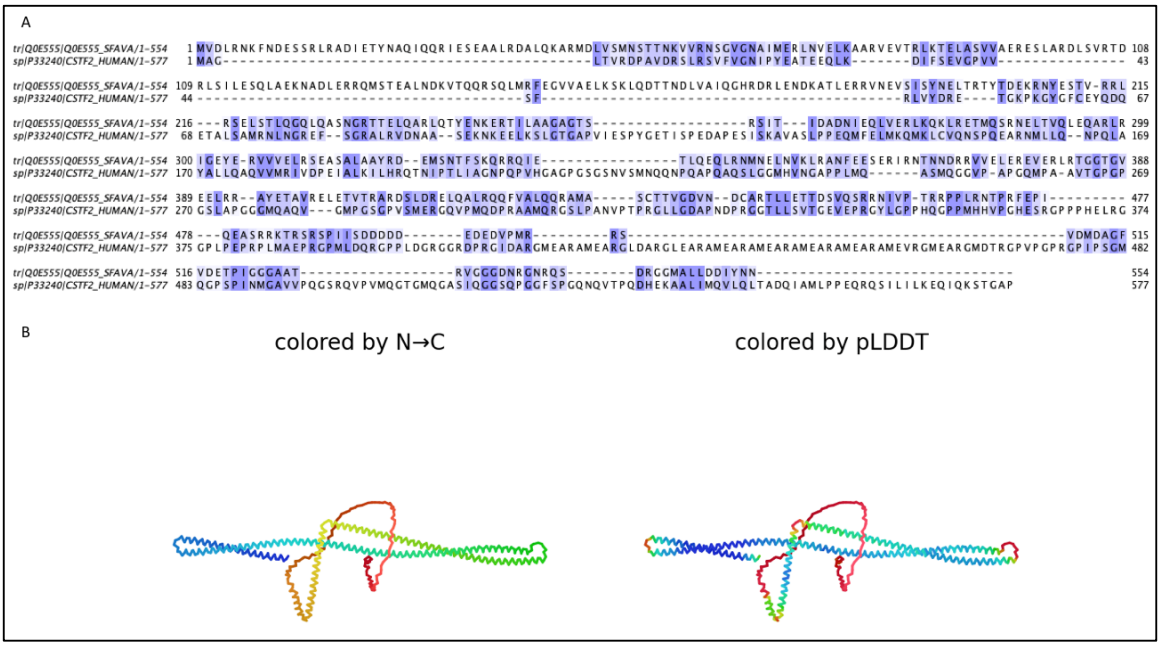


Figure 85. Sequence alignment and structure prediction of SfAV ORF046 protein. A) Sequence alignment of human CstF2 and baculoviral protein SfAV ORF046 created with Jalview. Residues are colored by similarity. B) AlphaFold prediction of SfAV ORF046 protein. Left panel: Colored from N-to C-terminus, whereas the N-terminus is depicted in blue. Right panel: Per-residue confidence metric. Lower confidence (red) is correlated with disordered regions.

3.2 Cryo-EM structure analyses of the full-length CstF complex and CstF1-CstF3 subcomplex were limited by complex instability during cryo-EM sample preparation and high conformational flexibility

The CstF complex consisting of its three subunits CstF1, CstF2 and CstF3 is an important protein factor in context of 3′-end processing of pre-mRNAs, as it is responsible for recognition of G/U-rich sequence elements on pre-mRNA and thereby assisting in definition of the cleavage site (Chen, MacDonald et al. 1995, Legrand, Pinaud et al. 2007). CstF was shown to adopt a trimeric structure of dimers by assembling two copies (2:2:2 stoichiometry) of each subunit (Legrand, Pinaud et al. 2007, Yang, Hsu et al. 2018). Previous structural studies of the CstF complex exist for isolated domains and subcomplexes only. Neither the overall structure of full CstF could be solved to date, nor a single, full-length subunit (Perez Canadillas and Varani 2003, Legrand, Pinaud et al. 2007, Yang, Hsu et al. 2018). With its molecular weight of 385 kDa, the CstF complex is a good target for structural studies using cryo-EM. In this thesis, I used recombinantly expressed and purified full-length proteins to reconstitute CstF with and without a G/U-rich RNA substrate for EM studies. Before the protein sample was prepared for cryo-EM, it was extensively screened and optimized in negative stain EM. First screening attempts clearly showed, that the CstF complex was either highly heterogenous regarding its particle size or disintegrated during grid preparation (Paragraph 2.4.1). Thus, complex

stabilization with chemical cross-linking was attempted, as it seemed impossible to overcome complex dissociation by just optimizing purification and buffer conditions. During the first cryo-EM screenings, it became clear that the CstF complex showed even higher degree of instability and heterogeneity in cryogenic conditions (Paragraph 2.4.2). Initially, it was not possible to detect intact particles of the full-length CstF complex plunged-frozen without cross-linker or RNA substrate (see paragraph 2.4.2, figure 62). Besides that, the complex showed great preference to bind to the carbon surface surrounding the holes of a grid, so that almost no particles were left in the holes.

Instability (e.g. due to contact with the air-water interface) or aggregation of the protein specimen when applied to an EM grid, is often encountered in the beginning of cryo-EM studies. Aggregation and disassembly, especially of multi-protein complexes, are most likely happening due to exposure to harsh physical conditions differing from optimal conditions (e.g. buffer) established for purification. Sample stability and distribution on the grid is influenced by the grid surface or the blotting procedure itself (e.g. contact with filter papers or the air-water interface). Additionally, protein concentrations in solution used for cryo-EM grid preparation plays an important role as well, because different samples tend to adsorb differentially to carbon grid surfaces. Another factor to be considered is the thickness of the vitreous ice. Some protein samples are preferably located in thick ice and disassemble or aggregate as soon as the continuous ice layer is too thin.

With all these factors in mind, grid preparation of the CstF complex was systematically optimized, considering sample-specific challenges appearing during sample preparation, data screening or data processing. Usually, the phenomenon of protein particle adsorption to the carbon layer can be addressed by either using grids with carbon support layers or by decreasing the glow-discharging time for a grid. Glow-discharging is usually applied to a grid to render the surface hydrophilic, thus allowing the protein solution to spread evenly over the grid surface. For the CstF complex, it turned out that short glow-discharging time (10 s) in combination with low protein concentration (260-520 nM) were the best combination to obtain suitable particle distribution in holes. However, this conditions only worked in combination with a short incubation step (20-30 s) after the protein sample was applied to the EM grid, directly before the blotting. Presence of additives (e.g. trehalose) or different detergents were tried out, but had no positive effects. Once a sample plunging protocol was established, it was reproducible and delivered grids with rather thin to medium thick ice in a shallow ice gradient. In contrast to reproducibility of sample plunging, the protein complex itself behaved completely different after variable ways of protein purification, cross-linking and complex composition. Cross-linking reagents used or procedure (e.g in-batch or GraFix) and purification strategy

(e.g. SEC or sucrose density gradient) strongly influenced conformational and compositional heterogeneity of the complex, visible in processing of cryo-EM data (Paragraph 2.4.3 and 2.4.4; Figures 64, 65 and 66). In general, usage of GA as cross-linking reagent delivered homogenous looking particles on the EM grid (Figure 63 A), but they turned out to result in blurry 2D classes in later processing of the data. Although characteristic features like the HAT dimer of CstF3 were visible in 2D classes for GA-cross-linked samples, secondary structure features were better resolved in datasets collected on BS3 cross-linked samples (Figure 62 C and D). Datasets of CstF cross-linked with BS3 showed secondary structure features and slightly different views of particles and conformations than data of GA-cross-linked CstF via GraFix (Paragraph 2.4.4). This indicates, that the heterogeneity apparent in 2D classes results from both, the high flexibility of the complex existing in several conformations, and from the complex falling apart on the grid. The full-length CstF complex likely partially disassembled on the grid, because density corresponding to CstF2 subunit was never visible in either of the datasets. Since density for a subcomplex consisting of CstF1 and CstF3 was easy to identify in processed data, a BS3-cross-linked CstF1-CstF3 subcomplex was prepared and cryo-EM datasets were collected (Paragraph 2.5). This sample turned out to deliver more homogenous data in 2D classification and a discrete set of complex conformations (Figure 69). However, a certain amount of conformational and compositional heterogeneity remained even in the final datasets. This was most likely the main reason, why no high-resolution reconstruction of this subcomplex was obtained. Conformational flexibility mainly resulted from the WD40 propellers of CstF1 adopting several conformations below the HAT dimer of CstF3 (Paragraph 2.4.4 and 2.5; Figures 66 A, 68 and 86).

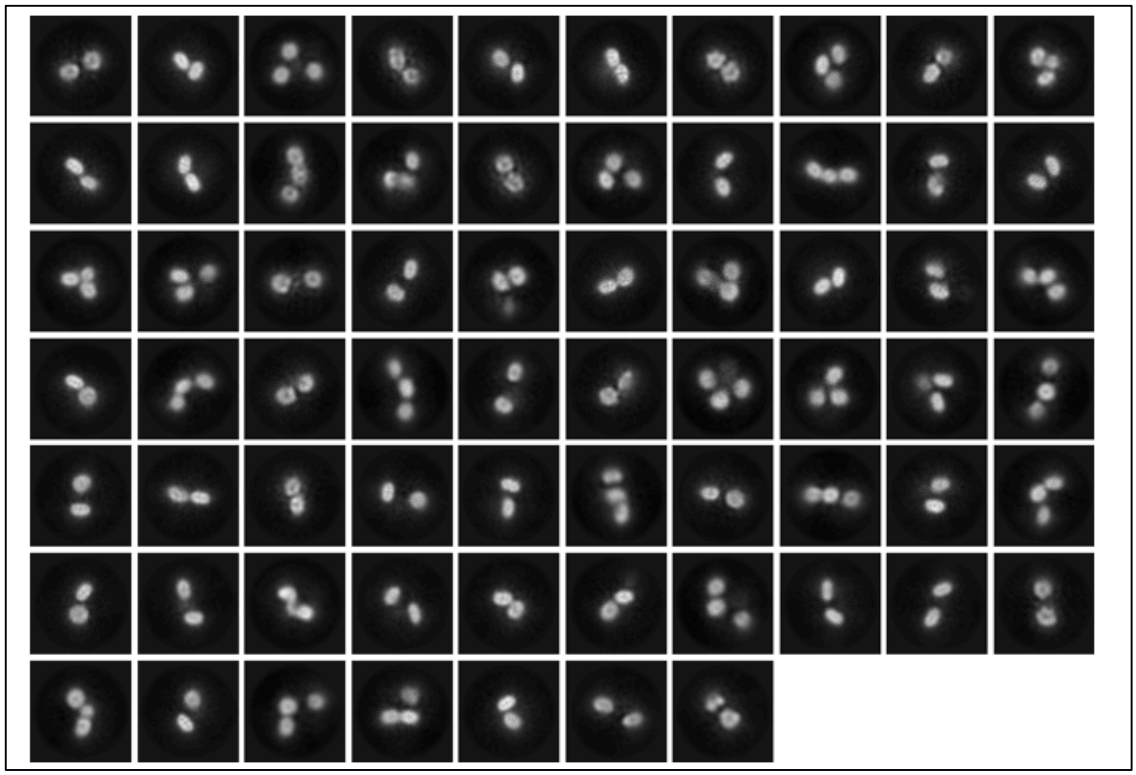


Figure 86. Different conformations of CstF1 WD40 domains. 2D classes show heterogenous conformations of CstF1 WD40 propellers in a subset of particles from a Krios dataset collected on the CstF complex cross-linked in batch with BS3. Particles for 2D classification were picked in CryoSparc using the Topaz implementation. Representative 2D classes show different views of the WD40 propellers. They adopt different conformations towards is each other by rotation and movement of the propellers.

In some classes, WD40s showed movement towards and away from each other, seen as varying distances between the two propellers was well as potentially tilted orientations with respect to each other (meaning that one of the WD40 propellers was rotated, Figure 86 and 87).

In case of the CstF complex, not only sample preparation or cryogenic conditions could be reason for sample heterogeneity as discussed above, but also missing factors of the 3′-end processing machinery. Since CstF is supposed to interact with several factors within the 3′-end processing machinery, presence of other protein complexes like the CPSF complex could help the CstF complex to adopt a defined conformation. Besides that, not only other protein factors of the 3′-end processing machinery, but also missing of a G/U-rich RNA substrate could be reason for conformational flexibility. When using the full-length CstF complex for structural analysis, presence of an RNA substrate could help to stabilize the complex in a certain defined confirmation.

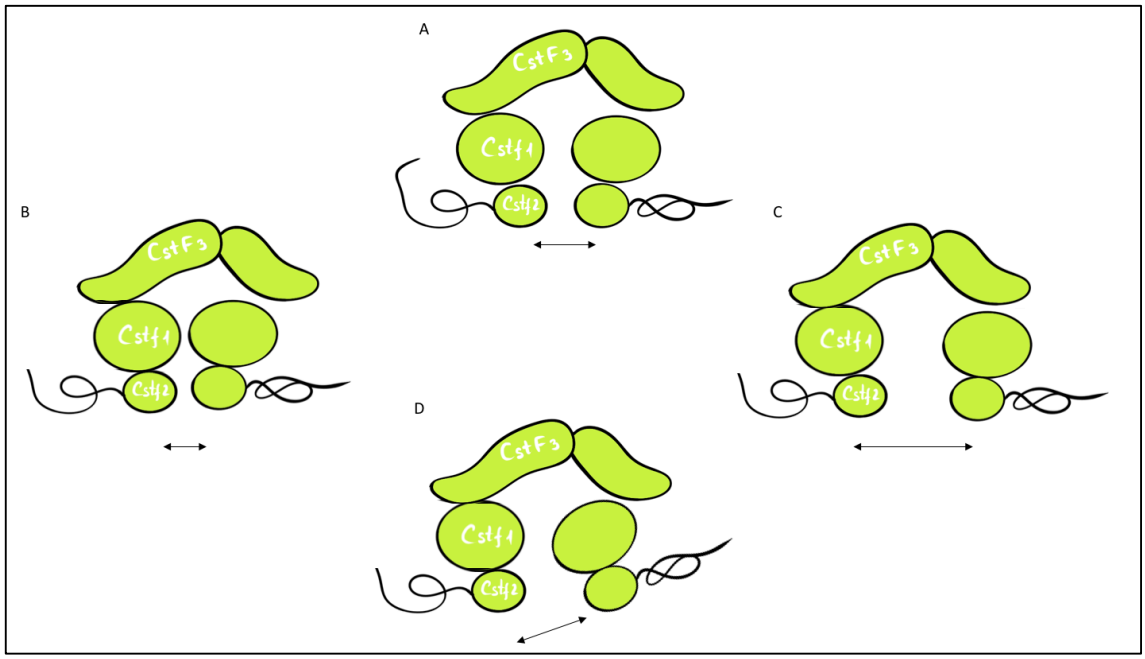


Figure 87. Cartoon representation of dynamic movement of CstF1 WD40 propellers within the CstF complex. A) WD40 propellers are located below CstF3 HAT dimer in a medium distance towards each other. B) WD40 propellers move closer towards each other thereby bringing the CstF2 RRMs in closer proximity. C) WD40s are separated from each other, which also creates more distance between both CstF2 RRMs. D) One WD40 propeller is flipped and rotated towards the other propeller and brings the CstF2 RRM in a flipped orientation.

In the CstF complex, residues 580-593 of CstF3 (peptide, Figure 88) interact with one WD40 propeller of CstF1 (based on a crystal structure pdb: 6P3X), followed by the monkeytail (residues 594-653) interacting with the CstF2 hinge domain (Yang, Hsu et al. 2018). Therefore, I asked whether movement of the WD40 propellers due to their flexibility would lead to movement of the RRM domain of CstF2. The CstF2 RRM domain is organized N-terminally of the hinge domain and is thereby connected indirectly to a WD40 propeller via CstF3 (Figure 88). Whether this assumption is true and which consequences the dynamic behavior of the CstF complex has in context of 3'-end processing and RNA binding, has to be further elucidated by structural and biochemical studies. To make the cartoon in figure 87 clearer, a model was created (see paragraph 2.6.2) showing a minimal CstF1-CstF2^{RH}-CstF3$^{HAT\text{-}MT}$ dimer and suggested positioning of subunits and domains (Figure 88). Rotation and movement of WD40s observed in cryo-EM studies in different directions could lead to different positions of CstF2 RRMs depending on how strong RRMs are positioned relative to the HAT-N domain by non-specific interactions.

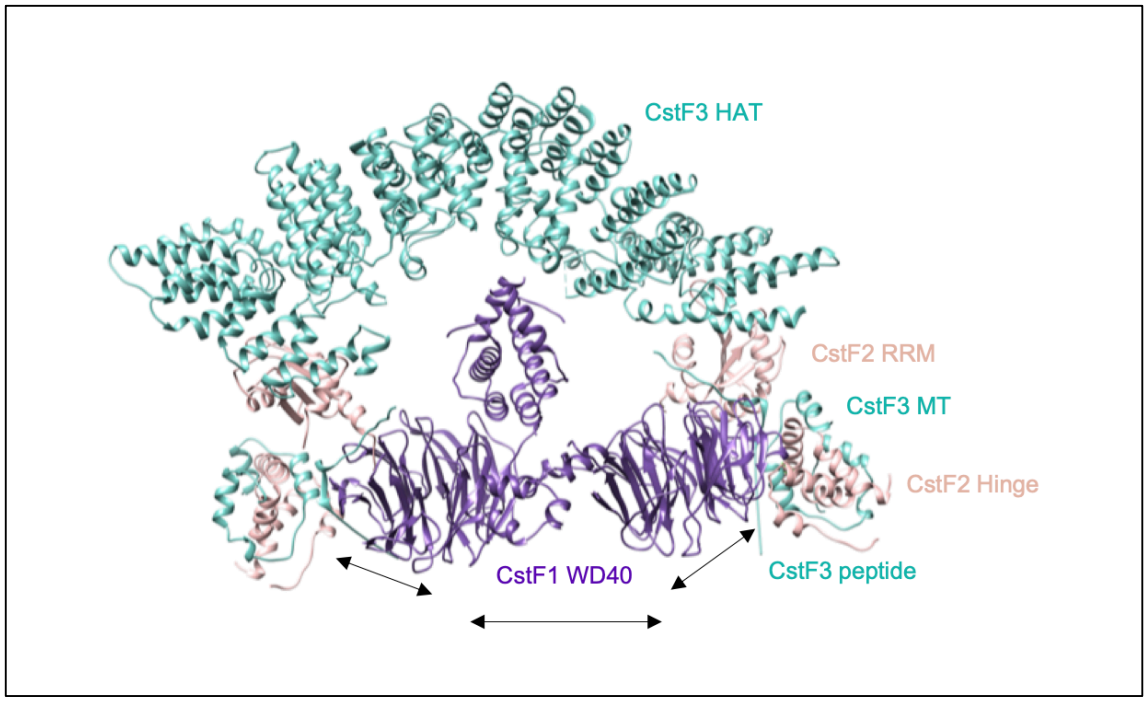


Figure 88. Model of a minimal CstF monomer depicting the indirect connection between CstF1 WD40 propeller and CstF2 RRM domain. CstF1 (residues 4-65 and 101-431) is shown in dark purple, CstF2 RRM and hinge domain (1-197) are shown in wheat and CstF3 (residues 25-550 and 576-660) is colored in deepteal. Models generated by AlphaFold were edited and colored in PyMol and the final model was assembled in Chimera based on cryo-EM reconstructions and XL-MS data. CstF1 WD40 propeller is closely bridged to CstF2 via CstF3. By movement of the WD40 towards and away from each other, CstF2 RRMs could also adopt different distances towards each other.

High flexibility of the CstF complex in cryo-EM studies in this thesis is in line with negative staining EM Zhang and co-workers (Zhang, Sun et al. 2020) performed on CstF within their study of the human 3′-end processing machinery.

3.3 Biochemical characterization of the RNA binding mechanism of the CstF complex hints to an unexpected role for the unstructured C-terminal part of CstF2

Early evidence that the CstF complex is involved in RNA binding was derived from UV-crosslinking of the CstF2 subunit to RNA containing a functional cleavage site (Wilusz and Shenk 1988). In later studies, the cross-linking site of CstF2 to two different pre-mRNAs was mapped on U-rich sequences downstream of the poly(A) site (MacDonald, Wilusz et al. 1994). It was postulated that the CstF complex is required for definition of the cleavage site by binding to G/U-rich elements on pre-mRNA, that are positioned within 30 nucleotides downstream of the poly(A) site (Zhao, Hyman et al. 1999). A N-terminal RNA binding domain was identified to specifically mediate the binding to these G/U-rich parts on the pre-mRNA (Takagaki, MacDonald et al. 1992, Beyer, Dandekar et al. 1997, Takagaki and Manley 1997). In more recent studies, RNA binding experiments were done using a recombinantly reconstituted

minimal CstF1-CstF2-CstF3 and CstF2-CstF3 complex, both containing truncated versions of $CstF2^{1-199}$ and $CstF3^{241-717}$, thereby identifying the contribution of single subunits to binding of $(GU)_n$ RNA stretches (Yang, Hsu et al. 2018). However, constructs used in the study were missing the whole C-terminal part (residues 200-577) of CstF2 and the N-terminal part (residues 1-240) of CstF3. In this thesis, I had a closer look into the amino acid sequence and domain organization of CstF2 before designing constructs for RNA binding experiments, in order not to exclude parts of the protein that could be important for RNA binding (Figure 89).

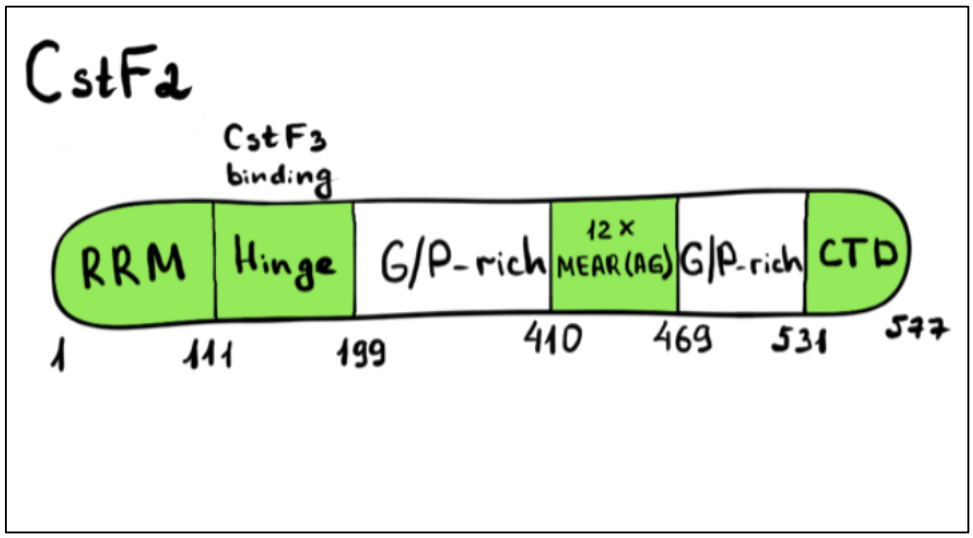


Figure 89. Domain organization of CstF2. CstF2 contains a N-terminal RRM domain, mediating binding to G/U-rich downstream elements on the pre-mRNA. 17 RG/RGG-like motifs are spanning over the G/P-rich region and the MEAR(AG) repeats, providing a second RNA binding motif potentially interacting with nucleotides around the G/U-rich sequences and thereby fine-tuning or enhancing CstF2-RNA interactions.

The RRM domain of CstF2 is covers the first 111 residues and adopts a canonical RRM fold (Perez Canadillas and Varani 2003). The region spanning over residues 112-199 was identified to be involved in interaction with CstF3 and Symplekin (Ruepp, Schweingruber et al. 2010, Moreno-Morcillo, Minvielle-Sebastia et al. 2011) and was therefore named hinge region (Figure 86). This region is very important for maintaining the 2:2:2 architecture of the CstF complex, since it is the only connection between CstF2 and CstF3. There is no direct contact between CstF1 and CstF2. Directly after the hinge region, a stretch rich in glycine and proline residues follows (residues 200-531), which was predicted to form around 10 β-turns (Takagaki, MacDonald et al. 1992). This long G/P-rich region is interrupted by almost 60 amino acids (residues 410-469) consisting of repetitive pentapeptides with the consensus sequence MEARA/G (Figure 86). This so far unique pentamer is repeated 12 times. CstF2 also contains 17 RG/RGG-like motifs, that are preceding and overlapping with the pentapeptide repeats. RG/RGG motifs are evolutionary conserved motifs, and a very common unstructured RNA-binding domain in many human proteins (Fornerod 2012, Rajyaguru and Parker 2012,

Gerstberger, Hafner et al. 2014, Beckmann, Castello et al. 2016, Jarvelin, Noerenberg et al. 2016).

This motif specifically occurs in proteins containing one or more RRMs and generally in proteins involved in RNA binding (Lischwe, Cook et al. 1985, Lischwe, Ochs et al. 1985, Kiledjian and Dreyfuss 1992, Corley and Gready 2008, Fornerod 2012, Rajyaguru and Parker 2012). RG/RGG motifs were identified to be recognition sites for protein arginine methyltransferases (PRMTs), and are therefore often modified post-transcriptionally (Boisvert, Chenard et al. 2005), altering RNA-binding behavior and preference for the ligand (Blackwell, Zhang et al. 2010). In contrast to RRM domains, RG/RGG motifs are disordered in absence of RNA substrates, which can facilitate target RNA selection due to conformational flexibility (Jarvelin, Noerenberg et al. 2016). Consequently, RNA binding by RG/RGG motifs is in most cases unspecific, but can facilitate RNA interactions of one or more RRMS, as it was shown in following studies: Disordered RG/RGG motifs contribute to RNA binding of RRM containing proteins either by orienting the RRM domains on the RNA or by binding to RNA themselves (Kiledjian and Dreyfuss 1992, Oberstrass, Auweter et al. 2005). So far, it is unknown if the RG/RGG motifs in the C-terminal part of CstF2 contribute in any way to its RNA binding. In line with this hypothesis, methylation of arginines 308, 468 and 475 was reported for CstF2 (Guo, Gu et al. 2014). These residues are located within the replicative RG/RGG motifs.

Based on literature, it could not be excluded that the unstructured C-terminal part of CstF2 contributes to RNA binding. I therefore decided to use the full-length CstF2 protein for RNA binding studies in contrast to the truncated version used by Yang and co-workers (Yang, Hsu et al. 2018). At the time of writing this thesis, there was no data available on RNA binding using a recombinantly purified CstF complex consisting of full-length proteins. The full-length CstF complex binds specifically to a G/U-rich RNA ligand (*CstF01* RNA) in FA experiments and discriminates against polyA or polyU RNA (Figure 43, Paragraph 2.3.1). However, the determined $K_D = 52.1 \pm 10.3$ nM slightly differed from the ones Yang and co-workers measured in their studies for a CstF1-CstF2$^{1\text{-}199}$-CstF3$^{241\text{-}717}$ complex (120nM for a GU_{14} RNA). Influence of buffer composition like high salt concentrations could be excluded, because the same buffer was used in both studies. Consequently, I assumed, that the difference in binding affinities might be due to presence of additional parts (e.g. the RG/RGG region) in the CstF complex. Although there was no particular function assigned to residues 200-577, it might be possible that the unstructured parts of CstF2 have a yet unknown role in RNA binding, additional to the N-terminal RRM, as stated above. To address this, single full-length subunits of the CstF complex were expressed and subjected to similar binding experiments using G/U RNAs (*CstF01* RNA). This experiment (Paragraph 2.3.4, Figure 47) confirmed, what was already

observed by Yang and co-workers, namely that presence of CstF3 has the strongest stimulating effect on the RNA binding affinity of the complex, shifting the K_D from a micromolar range for full-length CstF2 alone to a nanomolar range for the CstF2-CstF3 complex (Figure 47). This huge stimulatory effect of CstF3 is most likely achieved by bridging two CstF2 subunits, so that the two RRM domains are pre-arranged in close proximity. A small contribution of CstF1 on the overall affinity to G/U-rich RNA was found as well.

To determine if the difference in binding affinities to G/U-rich RNAs compared to studies of Yang, Hsu et al. 2018 arises from different complex composition, I used a minimal CstF (CstFdC) complex containing a truncated version of CstF2$^{1-204}$, consisting only of the N-terminal RRM followed by the hinge domain, and tested binding to the same G/U-rich RNA ligand. This measurement (Figure 49, Paragraph 2.3.4) delivered different RNA binding affinities for full CstF and CstFdC. The affinity of CstFdC was decreased about four-fold compared to full CstF. The $K_{D,\ CstFdC}$ = 196.6 $\pm$ 64 nM was in the range of binding affinities determined by Yang, Hsu et al. 2018 for the truncated CstF1-CstF2$^{1-199}$-CstF3$^{241-717}$ complex (120-220 nM). The difference in binding affinities depending on presence of the C-terminal part of CstF2 was also observed on the single protein level, when comparing full-length CstF2 to the RRM domain alone (Figure 49). Consequently, the question arose, if the unstructured RG/RGG motifs of CstF2 are involved in RNA recognition of the N-terminal RRM domain, like it was reported for other proteins (Schmidt, Knick et al. 2016).

Due to their unstructured nature, arginine residues in RG/RGG motifs can form direct interactions with RNA (Figure 90 A) via hydrogen-bonds and π-stacking (Chong, Vernon et al. 2018). The presence of multiple copies of RG/RGG motifs enhances the interaction. Since CstF2 harbors 17 repetitive RG/RGG motifs, the C-terminal region can theoretically provide a quite strong RNA binding platform. It is not clear, if RNA binding via RG/RGG motifs is sequence specific, but it is suggested, that this region might derive its specificity by its adaptable conformational behavior towards a set of RNA sequences. This means that upon RNA binding, the disordered RG/RGG motif region could get structured depending on the substrate bound (Figure 90 B). In case of the CstF complex, RRMs are highly specific for a bipartite G/U-rich DSE on the pre-RNA. However, it is questionable if these highly specific optimized target motifs are found in most of the 3′-UTRs. The role of DSE sequence recognition is important for the correct localization of the cleavage site and therefore for correct positioning of protein factors of the 3′-end processing machinery on pre-mRNA transcripts to perform cleavage and subsequent polyadenylation. By providing a second RNA binding motif, CstF complex could theoretically extend its RNA target spectrum to recognize a wide range of DSE regions on pre-mRNAs. The question is, if RG/RGG motifs of CstF2 ‘pre-select’ RNA

target sequences for the RRM or if they enhance RRM-DSE interactions by binding to nucleotides surrounding G/U-rich DSEs. In context of this thesis, these open questions could not be answered yet and further experiments are required to shed light on this hypothesis.

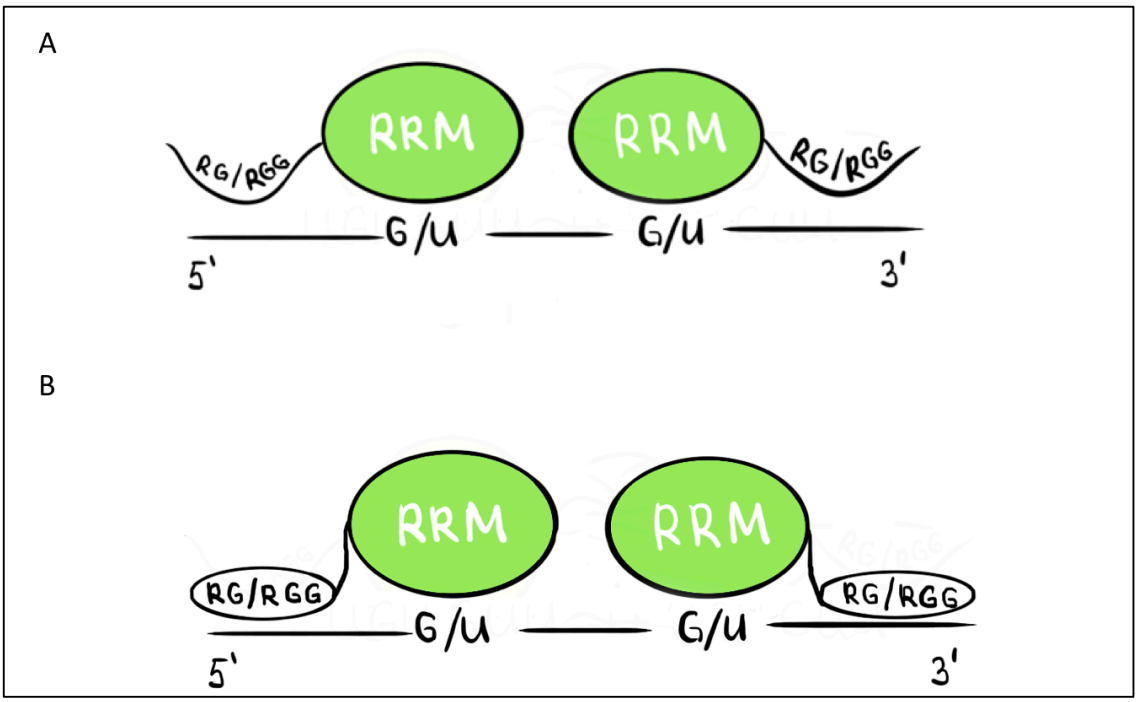


Figure 90. Model for RNA binding of CstF2 mediated via its RG/RGG motifs. A) Unstructured RG/RGG motifs can bind to RNA by themselves via repeated arginine residues. B) RG/RGG motifs could get structured depending on the RNA substrate, thereby assisting selection for CstF-specific sequences on pre-mRNA.

3.4 The full-length CstF complex preferably binds symmetric G/U-rich downstream element instead of asymmetric DSEs consisting of a proximal GU-rich part and a distal U-rich part

The exact location of the poly(A) site on a pre-mRNA is determined firstly by the distance between the very conserved hexameric poly(A) signal AAUAAA and G/U-rich DSE and, secondly, by the affinity of the CstF – RNA interaction (Deka, Rajan et al. 2005). In the past, many studies have been done on composition of the downstream element, but no consensus sequence was identified so far. Several studies described the DSE consisting of two parts, one proximal G/U-rich sequence element followed by a distal U-rich element (McDevitt, Hart et al. 1986, Gil and Proudfoot 1987, Zarudnaya, Kolomiets et al. 2003, Salisbury, Hutchison et al. 2006). Several SELEX experiments, however, challenged this assumption by identification of distinct G/U-rich binding patterns for CstF (Figure 91) only containing G/U-rich parts (Beyer, Dandekar et al. 1997, Takagaki and Manley 1997). Length and position of the sequence motifs on the mRNA are important for high affinity binding as well (Takagaki and Manley 1997).

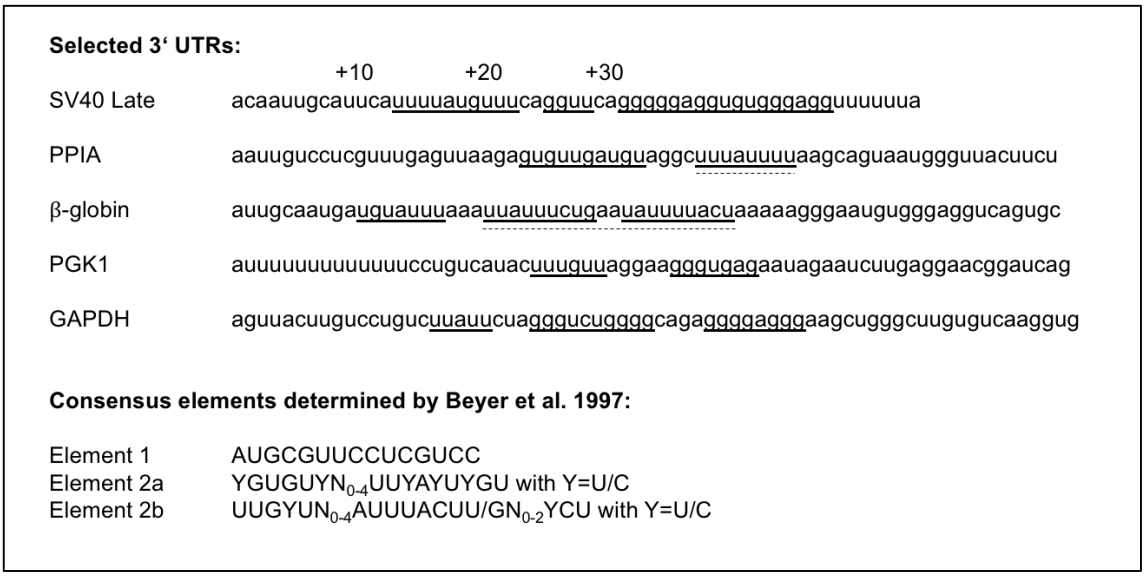


Figure 91. pre-mRNA sequences downstream of pre-mRNA poly(A) sites and G/U-rich consensus sequences identified by SELEX studies. RNA sequence starts after the cleavage site, meaning +10 indicates 10 nucleotides downstream the cleavage site. Upper row: Bipartite G/U-rich DSEs consisting of a proximal and a distal sequence element (underlined) are located within 30 nucleotides after the cleavage site. Distal U-rich sequence elements are indicated by dashed lines. Bottom row: Beyer et al. 1997 identified a set of G/U-rich consensus sequences selected by CstF in several rounds of SELEX.

One remaining open question is, how the CstF complex selects for its DSE targets without having a consensus sequence. Analysis of a NMR structure of the CstF2 RRM domain hinted at a molecular mechanism for specific recognition of a UU-dinucleotide, and that discrimination against other nucleotides is based on formation of a specific binding pocket which specifically fits UU (Perez Canadillas and Varani 2003). Requirement of UU-dinucleotides in DSE sequences would be basis of this hypothesis for optimal target sequence selection. Recent studies verified G/U-rich target specificity of an *in vitro* reconstituted CstF complex, containing truncated versions of CstF2 and CstF3, and showed that presence of a distal polyU motif results in lower binding affinity than for $(GU)_n$ stretches (Yang, Hsu et al. 2018).

RNA ligands used in this thesis were designed by combining information on DSE composition from several studies (Beyer, Dandekar et al. 1997, Graber, Cantor et al. 1999). Another hint to be considered when RNA ligands were designed, is the postulated dimeric structure of the CstF complex with two RRMs present for binding potentially two G/U-rich sequences simultaneously. Since there was no data available, whether the RNA binding mechanism of the CstF complex follows any sequential directionality, meaning that one RRM recognizes a certain sequence and by their proximity positions the second RRM on the RNA stretch, the RNA ligands in this thesis were designed in such a manner, that both G/U-rich sequence elements contain the same sequence in 5′ to 3′ and 3′ to 5′ direction (UGUGUU and UUGUGU). Therefore, I wanted to exclude preferable binding of one or the other RRM.

Additionally, G/U-rich sequence elements in this thesis were designed in such a way, that they contain a UU-dinucleotide and would therefore allow the postulated UU-dinucleotide based RNA recognition mechanism (Perez Canadillas and Varani 2003).

Recombinantly purified CstF complex containing the full-length subunits was shown to bind to the selected RNA ligand (referred to as *CstF01* RNA) with high affinity (Figure 41, Paragraph 2.3.1). The binding affinity was about four-fold higher than values determined by Yang and co-workers for binding of the truncated CstF1-CstF2$^{1-199}$-CstF3$^{241-717}$ complex to $(GU)_{14}$ RNA. Previous section of this chapter discussed the potential impact of a second yet unidentified RNA binding motif in the C-terminal part of CstF2 resulting in higher binding affinities, but there could be further reasons. Impact of buffer components and salt concentration were excluded, because both experiments were performed in 1x PBS buffer. Besides that, sequence of the G/U-rich binding motif itself may enhance the binding strength of CstF-RNA interactions. Previous studied postulated the specific recognition of a UU-dinucleotide by formation of a binding pocket especially fitting to UU. Amino acids outside the binding pocket are expected to participate in RNA binding by fine-tuning RNA-protein interactions, depending on the nucleotide sequence surrounding the G/U-rich binding motifs. In contrast to other RRM containing proteins, CstF2 has no additional RRMs assisting in RNA-protein complex formation (Wang and Tanaka Hall 2001). Due to the dimeric association of the CstF complex, the two RRMs in the complex might be sufficient to allow fine tuning of binding towards different RNAs via the respective DSEs and the surrounding nucleotide sequence.

Correct positioning of both RRMs on the RNA substrate and tight binding would be required to provide a stable interaction platform. Figure 91 shows five different 3′-UTRs of naturally occurring pre-mRNAs. The position of DSEs among 3′-UTR of genes is conserved, so that most G/U-rich DSE sequences are located within 30 nucleotides downstream the cleavage site (Figure 91). In the different examples listed in figure 91, the starting distance of DSEs differs from 10 (β-globin) to 24 (PGK1). Downstream sequences in selected 3′-UTRs depicted in figure 81 show different distances between the G/U-rich sequence elements. In case of PPIA and PGK1, where a bipartite G/U-sequence pattern can be observed, distance between both sequence motifs is in a range of 4-5 nucleotides. For the remaining 3′-UTRs, a set of three G/U-rich sequences can be detected, which are separated by 2-4 nucleotides. Having in mind, that the CstF complex has two RRMs, which can therefore bind to two G/U-rich sequence elements simultaneously, the distance between both motifs can vary depending on which one of them is bound by the CstF2 RRMs. Considering that in case of the listed 3′-UTRs with three potential sequence elements present (SV40 late, β-globin, GAPDH), CstF can select several combinations of those, also the first and the last one with maximum distancing between both.

In case of SV40 late 3′-UTR, if the first and third G/U-rich element is bound, the distance between both would be 8 nucleotides. If the first and the last G/U-rich element was selected in β-globin and GAPDH 3′-UTRs, the distance between both would be 14 and 17 nucleotides. Selected 3′-UTRs depicted in figure 91 show a clear variation in distance between potential G/U-rich sequence elements in mRNAs, so that I decided to perform experiments to test the binding affinities to *CstF01* RNA with varying length between both G/U-rich binding elements.

Experiments to determine the optimal distance between two G/U-rich binding motifs of *CstF01* RNA showed clear preference for shorter linker distances (Figure 44). Spacing from 2-8 nucleotides delivered moderate binding affinities between 100 and 256 nM for the full-length complex whereas upon spacing of 16 nucleotides, the RNA binding affinity drastically decreased to micromolar range. This result was in contrast to observation Yang, Hsu et al., 2018 made, when inserting polyA spacers (1-19 nucleotides) to separate a GU_{10} RNA into GUGUG (referred to as G/U-rich) and UGUGU (referred to as U-rich). In their work, all determined binding affinities for CstF1-CstF2$^{1-199}$-CstF3$^{241-717}$ were around 500 nM for all RNA ligands and spacer lengths tested, same for the CstF2$^{1-199}$-CstF3$^{241-717}$ subcomplex lacking CstF1. In this thesis, RNA binding of the full-length CstF complex already showed clear dependency on the distance between both G/U-rich binding motifs, and same behavior was observed for the full-length CstF2-CstF3 complex, too. To exclude that residues 200-577 of CstF2, which were truncated in the CstF complex Yang and co-workers reconstituted, mediated this spacer length dependency in an unknown way, I repeated experiments with a similar CstF1-CstF2$^{1-204}$-CstF3 complex (referred to as CstFdC; Figure 50). Again, decreasing binding affinities, this time already from a distance of 8 nucleotides between both G/U-rich binding motifs, were observed for the CstFdC complex.

Consequently, I assumed, that dependency of the distance between both binding motifs is connected to the nucleotide sequence itself and on strength of the interaction between RRMs and RNA. By providing a binding platform with similar sequences for both RRMs, *CstF01* RNA can be bound with very high affinity, no matter in which direction the RNA is recognized by the “first” RRM domain (Figure 92 A). Strong RNA-protein interactions might be lost upon a certain distance because the RRMs of CstF2 cannot be separated that far from each other. Due to the similar binding sequences, RNA-protein contact could be lost for either the “first” or the “second’ RRM, still providing same binding conditions for the RRM still associated to the RNA. Another scenario could be, that RNA binding is completely lost, when G/U-rich sequences cannot be reached by both RRMs anymore (Figure 92 C and D). Instead of sliding on the AC_n stretch separating both binding motifs, RRMs dissociate from the RNA ligand (Figure 92 B). If

one RRM remains bound to one of the G/U-rich motifs or both RRMs completely dissociate from the RNA is not clear.

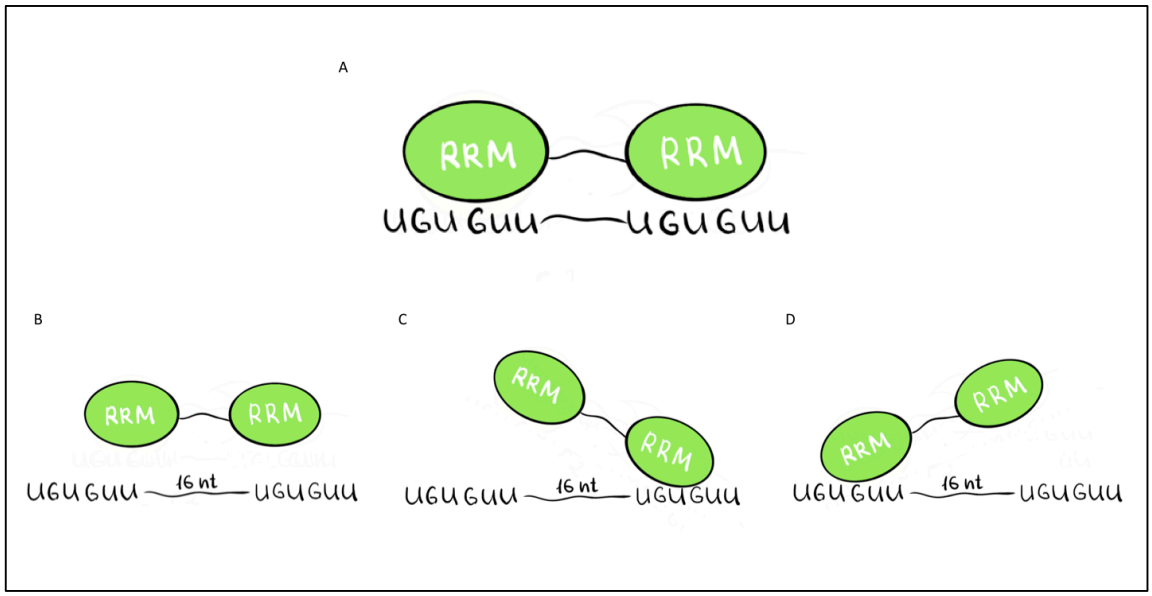


Figure 92. Schematic representation of a model explaining spacer dependency of CstF2 RRMs binding to G/U-rich sequence elements on the pre-mRNA. A) Two RRMs in close proximity tightly bind to G/U-rich RNA for a spacer length of 2 to 8 nucleotides between two G/U-rich elements on the RNA. Upon spacing of 16 nucleotides, contact between RRMs and RNA is either completely lost (B) or one of both RRMs remains associated to the RNA (C and D), whereas the other RRM is dissociating from the RNA ligand.

In case of a bipartite DSE consisting of a G/U-rich and a U-rich motif, RNA is only bound with medium to low affinity (Yang, Hsu et al. 2018). RNA binding affinities did not change upon increasing distance between both motifs, which could be explained by either one RRM stably associating with the G/U-rich motifs, whereas the other one is loosely attached to the U-rich motif not contributing much to the overall RNA binding (Figure 93 A). Secondly, by missing a UU-dinucleotide for base specific RNA recognition, it could be possible that RRMs are not stably anchored on the RNA and are moving along the RNA stretch independent of distance and sequence between both binding motifs (Figure 93 B).

To sum up, recombinantly purified CstF complexes preferably bound to DSEs containing symmetric sequence elements for both RRMs with high affinity. This interaction could be disrupted by increasing the distance between both binding elements, thereby loosing high-affinity interaction of one or both RRMs and the RNA target. Additional structural and biochemical characterization is necessary to further elucidate the mechanism of DSE binding and target sequence selection by the CstF2 subunit.

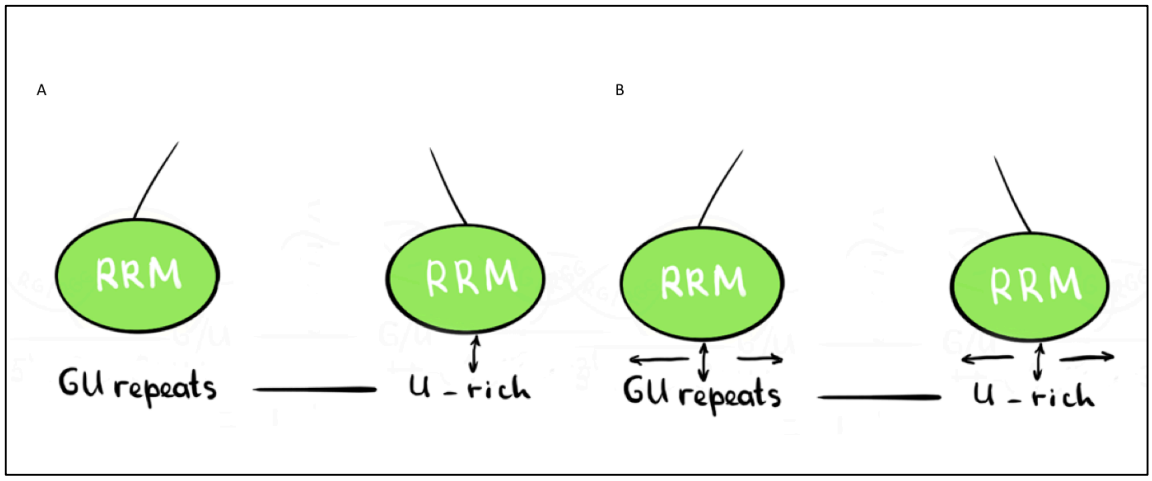


Figure 93. Schematic representation of weak and dynamic interactions between CstF2 RRM and RNA ligands containing a distal GU-repeated element and a proximal U-rich sequence element. A) "First" RRM is bound with higher affinity to the GU-repeats and the "second" RRM shows weak interaction to the polyU sequence and is not stably bound to the RNA. B) Both RRMs show medium to low interaction to a combination of distal GU-repeats and proximal polyU sequences and are not tightly bound to the RNA ligand allowing movement on the RNA and binding in several frames.

3.5 Biochemical characterization of CstF2 RRM mutants identified a dual role of Serine 17 in binding to G/U-rich RNA

A lot of studies have been done in the past to shed light on the mechanism how the RRM domain of CstF2 specifically recognizes G/U-containing DSEs on a pre-mRNA and thereby contributes to the strength of poly(A) signals, meaning frequency of their selection, and poly(A) site definition (MacDonald, Wilusz et al. 1994, Chen, MacDonald et al. 1995). Several RNA ligands were tested and binding affinities were determined for the single RRM domain and an *in vitro* reconstituted CstF complex (Takagaki and Manley 1997, Perez Canadillas and Varani 2003, Deka, Rajan et al. 2005, Yang, Hsu et al. 2018). By solving an NMR structure of the RRM and modelling it with UU-dinucleotide based on homology to the HuD-*cfos* structure, a molecular basis of UU-recognition was postulated (Wang and Tanaka Hall 2001, Perez Canadillas and Varani 2003).

RRMs are consensus RNA-binding domains harboring a central sequence of around eight amino acids, mostly aromatic and positively charged (K/R-G-F/Y-G/A-F/Y-V/I/L-X-F/Y), forming the so-called ribonucleoprotein (RNP) consensus sequence (Adam, Nakagawa et al. 1986, Swanson, Nakagawa et al. 1987, Maris, Dominguez et al. 2005). Another stretch of six conserved residues (V/I/L-F/Y-V/I/L-X-N/L) participating in RNA binding was identified later and therefore termed RNP-2 (Dreyfuss, Swanson et al. 1988). By adopting a canonical $\beta_1\alpha_1\beta_2\beta_3\alpha_2\beta_4\alpha_3$ fold, the RRM of CstF2 has a central β-sheet serving as RNA-binding platform, containing RNP-1 and RNP-2 in the middle strands β_1 and β_3. One special feature is

the presence of a C-terminal helix α_3 lying on top of the β-sheet, in this way covering the RNA-binding interface (Perez Canadillas and Varani 2003). Upon RNA folding, CstF2 RRM undergoes conformational changes and the C-terminal helix opens to make the β-sheet accessible for RNA. Several residues in RNP-1 and RNP-2 were identified to participate in maintaining the closed conformation, with helix C covering the β-sheet. After losing contact to helix C, the β-sheet can interact with RNA nucleotides. Base-specific recognition of U_1 and U_2 is achieved by Ser17 and Arg46 based on the RRM-UU dinucleotide model (Perez Canadillas and Varani 2003). However, importance of the selected residues was not addressed by mutational analysis so far. Consequently, I mutated conserved residues (S17, F19, F61, N91, N97) in the CstF2 RRM, located either in RNP-1, RNP-2 or in the C-terminal helix (Figure 35).

FA measurements were performed with the CstF complex containing full-length subunits with single or double mutations in the CstF2 RRM to calculate binding affinities to *CstF01* RNA (Figure 56, Paragraph 2.3.6). All mutants showed decreased binding affinities for the full complex to a G/U-rich RNA ligand, whereas the S17A mutation reduced the K_D about 55-fold. In contrast to wild type CstF2, presence of CstF1 and CstF3 could not stimulate the binding affinity of CstF2(S17A). Determined K_D values were in a low micromolar range independent of the presence of the other subunits (Figure 58, Paragraph 2.3.6). Measurements performed in paragraph 2.3.6 indicated that in context of the full-length CstF complex, S17A drastically decreased the RNA binding affinity of CstF2 RRM to a G/U-rich RNA substrate. To examine the impact of S17 on RNA binding in a simple setup mimicking the dimeric association of the CstF complex, a construct containing two RRM domains connected by a short linker was expressed and purified (Paragraph 2.2.2). Either two wild type RRMs were linked together, one or the other RRM carrying the S17A mutation or both RRMs were mutated (Figure 39). ITC experiments using G/U-rich *CstF01* RNA, however, delivered the opposite outcome for the RRM fusions compared to full CstF (Paragraph 2.3.6; Figure 60). The RNA binding affinity was increasing compared to wildtype RRMs, when mutating the “first” RRM domain in S17 position and was further enhanced around 1.3-fold when only the “second” RRM was mutated. With both RRMs mutated, RNA binding affinity increased about two-fold compared to the fusion of two wild type RRMs. This result is an initial indication of the existence of a directionality when both RRMs are bound on the same RNA ligand, since there was a difference if the “first”, “second” or “both” RRMs were mutated. The question about the underlying mechanism, how S17 can switch its role in presence or absence of the full complex was not answered by this experiment.

S17 is located in the beginning of the β_1-strand pointing towards the cleft formed between the β-sheet and helix C. Upon unfolding of helix C, S17 is exposed and can interact with the 5′-uracil of RNA ligands on the surface of the β-sheet. By replacing serine to alanine in this position, formation of H-bonds by the hydroxyl group of S17 is abolished. Increasing RNA binding affinities for the mutated RRM domain fusion could be explained in a way, that the RRM loses part of its specificity for recognition of the 5′-uracil (U_1) due to loss of the S17-U_1 interaction, leading to more “non-specific” binding to the G/U-rich sequence in several frames and not one locked position. Within the full-length CstF complex, C-terminal parts of CstF2 may contribute to RNA binding and positioning of the RRM in the right way on the RNA by interacting with nucleotides surrounding the two G/U binding motifs (Paragraph 3.3). The loss of recognition of the 5′-uracil by S17 within the complex might lead to disruption of the certain RNA-protein conformation and thereby result in a drastic drop in affinity. Two RRMs in close proximity, however, are lacking the C-terminal stimulatory influence of CstF2 on RNA binding. Besides that, they are missing a degree of flexibility due to the fixed linkage between both RRMs. By losing recognition of the 5′-uracil in either one or both RRMs, the optimal position on the RNA ligand might be lost, but remaining residues in the RRM could compensate this loss by recognizing nucleotides of the G/U-rich binding element in a less specific way, thereby creating several frames in which two RRMs could bind on RNA with a locked distance from each other.

4 Material and Methods

4.1 Materials

4.1.1 Chemicals and consumables

All common chemicals and reagents were purchased from Sigma-Aldrich or Carl Roth GmbH. Enzymes were used from Fermentas or New England Biolabs. Any exceptions are mentioned in the distinct paragraphs of the text.

4.1.1.1 Antibiotics

Table 6: Antibiotic solutions and concentrations

Antibiotic	*Stock concentration*	*Final concentration*
Ampicillin	100 mg/ml	100 μg/ml
Kanamycin	50 mg/ml	50 μg/ml
Gentamycin	7 mg/ml	7 μg/ml
Tetracyclin	10 mg/ml	10 μg/ml

4.1.1.2 Bacterial media

Table 7: Bacterial media

Media	*components*
LB-medium	1% Bacto Tryptone, 0.5% Bacto Yeast Extract, 1% NaCl
SOC-medium	0.5% Yeast Extract, 2% Tryptone, 10mM NaCl, 2.5mM KCl, 10mM $MgCl_2$, 10mM $MgSO_4$, 20mM Glucose
LB-agar	1% Bacto Tryptone, 0.5% Bacto Yeast Extract, 1% NaCl 1.5% Agar
LB-MultiBac plates	50 μg/ml kanamycin 7 μg/ml gentamycin 10 μg/ml tetracycline 100 μg/ml X-gal 40 μg/ml IPTG

4.1.1.3 Bacterial strains

Table 8: Bacterial strains

Bacterial strain	*Genotype*
XL1 blue	recA1 endA1 gyrA96 thi-1 hsdR17 supE44 relA1 lac [F' proAB lacI^{q}ZΔM15 Tn10 (Tet^{r})]
Omnimax	F′ [proAB^{+}lacI^{q}lacZΔM15 Tn10(Tet^{R}) Δ(ccdAB)] mcrA Δ(mrr-hsdRMS-mcrBC) φ80lacZΔM15 Δ(lacZYA-argF) U169 endA1 recA1 supE44 thi-1 gyrA96 (Nal^{R}) relA1 tonA panD
Bl21 (DE3) pLysS	F^{-} opmT hsdS(rB^{-} mB^{-}) gal dcm met^{-} λ(DE3) pLysS (Cam^{R})
DH10 EmbacY	F^{-}mcrA Δ(mrr-hsdRMS-mcrBC) φ80lacZΔM15 ΔlacX74 recA1 endA1 araD139 Δ(ara-leu)7697 galU galK λ^{-} rpsL nupG / bMON14272 / pMON7124

4.1.1.4 Plasmids and constructs for bacterial expression

Plasmid: pEC-A is a pBR322 derivate

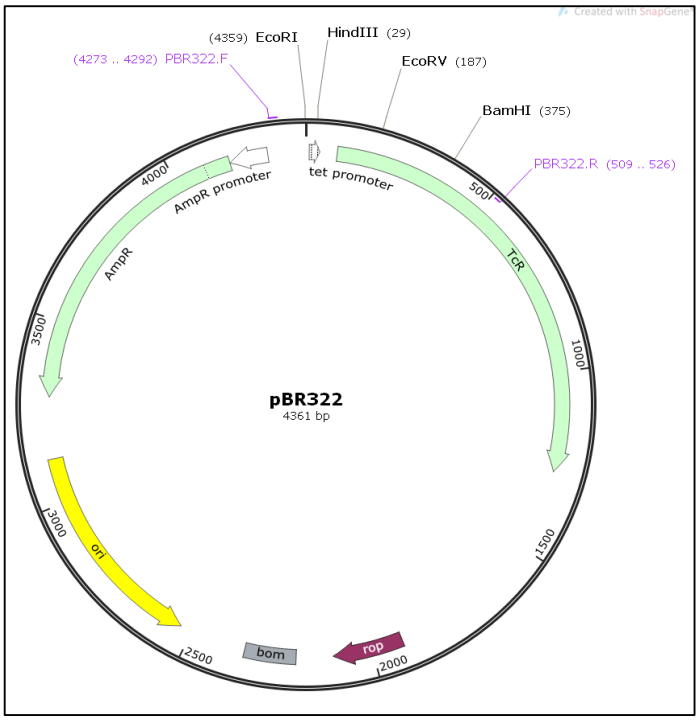


Figure 94: Vector map of pBR322. The map was obtained with SnapGene. Ori: origin of replication; AmpR: ampicillin resistance; TcR: Tetracycline resistance

Table 9: Constructs for bacterial expression system. A: Ampicillin resistance 3C: HRV-3C protease GS: Glycine-Serine linker

Construct name	*C-term. tag*	*Cleavage site*
pEC-A_RRM-GS-RRM-3C-Strep	Strep	3C
pEC-A_RRM-3C-Strep	Strep	3C
pEC-A_RRM(S17A)-GS-RRM-3C-Strep	Strep	3C
pEC-A_RRM-GS-RRM(S17A)-3C-Strep	Strep	3C
pEC-A_RRM(S17A)-GS-RRM(S17A)-3C-Strep	Strep	3C

4.1.1.5 Insect cell media

Insect cells were cultivated in serum free medium Sf-900™ II ordered from Gibco Life Technologies. Sf21 cells were only cultivated in Sf-900™ II prepared from powder with addition of Sf-900™ supplement and filtered through a 22µm Millipore sterile filter. High Five cells stocks were maintained in Sf-900™ II SFM ready-to-use medium. Large scale expressions were performed in Sf-900™ II prepared from powder.

4.1.1.6 Insect cell lines

Sf21 cells came from the USDA Insect Pathology Laboratory and are common cell line for working with baculoviruses. They are derived from IPLBSF-21 cell line, which has its origin in pupal ovarian tissue of fall army worm *Spodoptera frugiperda* (Vaughn, Goodwin et al. 1977). In this thesis, Sf21 cells were usually used for transfection and production of recombinant virus, but sometimes also for protein expression.

High Five cell line originated from ovarian cells of the cabbage looper, *Trichoplusia ni* (Wickham, Davis et al. 1992) and were used as the common cell line for protein expression, since High five cells generally provided a higher level of protein expression than Sf21 cells (Wickham, Davis et al. 1992).

4.1.1.7 Plasmids and constructs for protein expression in insect cells

Vectors for expression in insect cells are from MultiBac system (Berger, Fitzgerald et al. 2004) and are called acceptor vectors (pAcceptor). Acceptor vector pACEBac1 was used in this thesis.

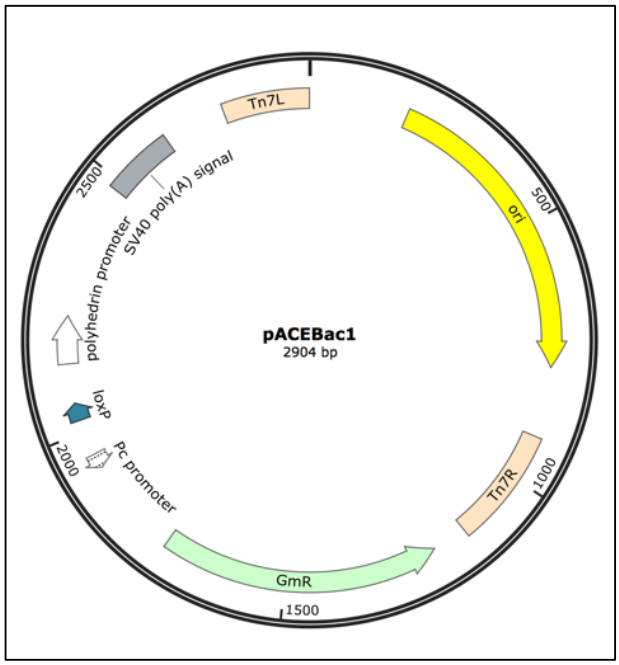


Figure 95: Vector map of pACEBac1. Vector map of the MultiBac acceptor plasmid (pAcceptor) pACEBac1. The map was generated with Snapgene. GmR: Gentamycin resistance; ori: origin of replication; Tn7L/Tn7R: transposon Tn7 site left (L) and right (R).

List of constructs, that were cloned for insect cell expression. Inserts were cloned into the LIC-site by Ligation Independent Cloning (Aslanidis and de Jong 1990).

Table 10: Construct for expression in insect cells

Construct name	*N-term. tag*	*C-term. tag*	*Cleavage site*
pFBDM_CstF (1,2,3)	Strep		
pACEBac1_TwinStrep-CstF2	TwinStrep		
pACEBac1_Strep-CstF3	Strep		3C
pACEBac1_Strep-CstF3-3C- His_8	Strep	His_8	TEV
pACEBac1_Strep-CstF3-TEV- His_8	Strep	His_8	
pACEBac1_Strep-CstF1	Strep		
pACEBac1_TwinStrep-CstF2(S17A)	TwinStrep		
pACEBac1_TwinStrep-CstF2(F19A)	TwinStrep		
pACEBac1_TwinStrep-CstF2(F61A)	TwinStrep		
pACEBac1_TwinStrep-CstF2(N91A-N97A)	TwinStrep		
pACEBac1_TwinStrep-CstF2ΔC	TwinStrep		
pACEBac1_CstF2(1-204)-Strep		Strep	3C

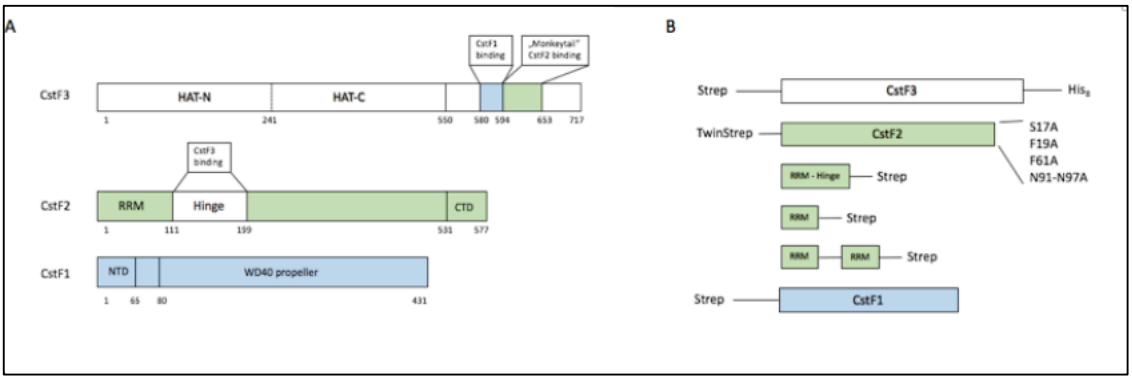


Figure 96. CstF subunit and construct scheme. Depiction of CstF subunits and constructs generated and used in this study to purify the hexameric CstF complex, distinct subcomplexes and the single RRM domain. A) Domain organization of the CstF subunits. CstF3 contains a Half a TPR (HAT) domain, divided in the N-terminal (HAT-N) and C-terminal (HAT-C) part, followed by the binding region for CstF1 and the monkeytail, which is binding to CstF2. CstF2 comprises of a N-terminal RNA Recognition Motif (RRM), followed by the hinge domain, which binds to CstF3 and a C-terminal domain (CTD). CstF1 has a N-terminal homodimerization domain (NTD) followed by a WD40 propeller. B) Constructs of the CstF complex. Full length CstF3 is N-terminally tagged with a Strep II tag and carries a C-terminal His_8 tag. Full length CstF2 carries a N-terminal TwinStrep tag, as well as all CstF2 mutants (S17A, F19A, F61A and N91A-N97A). The RRM-Hinge domain containing construct and the single RRM and RRM fusion construct carry a C-terminal Strep II tag. A N-terminal Strep II tag is fused to the full length CstF1 subunit.

4.1.1.8 Oligonucleotides for cloning

ATG: Start codon
TCA: Stop codon

Table 11: Primers for cloning

Primer name	*Sequence (5' - 3')*
LIC_TwinStrep-CstF2_fw	cgggcgcggaactcg**ATG**tggagccatccgcagtttgaaaaaggcggcggcagcggcggcggcagcggcggcagcgcgtggagccatccgcagtttgaaaaagcgggtttgactgtgagagac
LIC_CstF2_rev	cggaccggaaagTCAAGGTGCTCCAGTGGATTTCTGTATTTG
LIC_Strep-CstF3_fw	cgggcgcggaactcg**ATG**tggtctcacccacaatttgaaaaaTCAGGAGACGGAGCCACGGAG
LIC_CstF3_rev	cgggcgcggaactcg**ATG**TCAGGAGACGGAGCCACGGAG
LIC_CstF3-3C-His8_rev	cggaccggaaagTCAcggatcgccgtggtgatgatgatgatgatgagacgagtcgggcccctggaaCCGAATCCGCTTCTGCTGCCG
LIC_CstF3-TEV-His8_rev	cggaccggaaagTCAcggatcgccgtggtgatgatgatgatgatgagacgaggctgctccctggaaCCGAATCCGCTTCTGCTGCCG
LIC_Strep-CstF1_fw	cgggcgcggaactcg**ATG**tggtctcacccacaatttgaaaaaTACAGAACCAAAGTGGGCTTG
LIC_CstF1_rev	cggaccggaaagTCAGTCAGTGGTCGATCTCCGGTA
LIC_CstF2(1-204)_rev	cggaccggaaagTCAGTTGCCTGCAATCAGCGTTGG
LIC_CstF2(1-204)_3C_Str_rev	cggaccggaaagTCAtttttcaaattgtgggtgagaccacgagtcgggcccctggaa GTTGCCTGCAATCAGCGTTGG

4.1.1.9 Oligonucleotides for site directed mutagenesis to generate CstF2 RRM mutants

Table 12: Primers for site directed mutagenesis

Primer name	*Sequence (5' – 3')*
CstF2(S17A)_fw	TGGATCGTTCTCTACGTGCAGTGTTCGTGGGGAACATTCCT
CstF2(S17A)_rev	AGGAATGTTCCCCACGAACACTGCACGTAGAGAACGATCCA
CstF2(F19A)_fw	TCGTTCTCTACGTTCTGTGGCAGTGGGGAACATTCCTTATGAA
CstF2(F19A)_rev	TTCATAAGGAATGTTCCCCACTGCCACAGAACGTAGAGAACGA
CstF2(F61A)_fw	GCCAAAGGGTTATGGC GCA TGTGAATACCAAGACCAAGAG
CstF2(F61A)_rev	CTCTTGGTCTTGGTATTCACATGCGCCATAACCCTTTGGC
CstF2(N91A-N97A)_fw	CTTCGAGTGGACGCAGCTGCCAGTGAAAAGGCAAAAGAAGAG
CstF2(N91A-N97A)_rev	CTCTTCTTTTGCCTTTTCACTGGCAGCTGCGTCCACTCGAAG

4.1.1.10 RNA oligonucleotides for biochemical and structural studies

Table 13: RNA oligonucleotides for biochemical and biophysical assays

Oligo name	*characteristics*	*Sequence (5'-3')*
CstF01	SELEX RNA	UGU GUU UUU A UUG UGU
6-FAM-15U	polyU	5'Fl-UUU UUU UUU UUU UUU
6-FAM-15A	polyA	5'Fl-AAA AAA AAA AAA AAA
6-FAM-ARE	G/U-rich	5'Fl-UUU CUA UUU AUU UUG
6-FAM-CstF01		5'Fl-UGU GUU UUU A UUG UGU
CstF12	no linker	UGU GUU UUG UGU
CstF13	control	ACAACAACAACA
CstF14	2 nt linker	UGU GUU AC UUG UGU
CstF15	4 nt linker	UGU GUU ACA C UUG UGU
CstF16	6 nt linker	UGU GUU ACA CAC UUG UGU
CstF17	8 nt linker	UGU GUU ACA CAC AC UUG UGU
CstF18	16 nt linker	UGU GUU ACA CAC ACA ACA CAC C UUG UGU

4.1.1.11 Buffers for protein purification and biochemical assays

Table 14: Buffers for protein purification, biochemical and biophysical assays

Buffer	*Composition*	*Application*
CstF Lysis/Wash	50mM Sodium phosphate, pH 7.4 250mM NaCl 5mM $MgCl_2$	Lysis for purification of CstF and subcomplexes
CstF His-Tag Binding	50 mM Sodium phosphate, pH 7.4 250 mM NaCl 40 mM Imidazol	His-tag purification, equilibration and wash
CstF His-Tag Elution	50 mM Sodium phosphate, pH 7.4 250 mM NaCl 250 mM Imidazol	His-tag purification elution
SEC-4	20 mM Hepes, pH 7.4 250 mM NaCl 5 mM $MgCl_2$	Size exclusion bufffer
Strep-Tag Elution	50 mM Sodium phosphate, pH 7.4 250 mM NaCl 5 mM $MgCl_2$ 5 mM Desthiobiotin	Strep elution
CstF High	50 mM Sodium phosphate, pH 7.4 1M mM NaCl 5 mM $MgCl_2$	Strep high salt wash
1x PBS	5 mM $MgCl_2$	Fluorescence Anisotropy buffer
Sucrose Buffer (5% or 25%)	20 mM Hepes, pH 7.4 250 mM NaCl 5 mM $MgCl_2$ 5 % (w/v) sucrose or 25 % (w/v) sucrose	Sucrose density gradient
Heparin Binding	10 mM Sodium phosphate, pH 7.0	Heparin equilibration and wash
Heparin Elution	10 mM Sodium phosphate, pH 67.0 1M NaCl	Heparin gradient elution
CstF/RRM Lysis	50 mM Sodium phosphate, pH 7.5 250 mM NaCl 5 mM $MgCl_2$	RRM lysis
RRM Strep/SEC	20 mM Hepes, pH 7.5 120 mM NaCl	
RRM High	20 mM Hepes, pH 7.5 500 mM NaCl	
RRM Elution	20 mM Hepes, pH 7.5 120 mM NaCl 5 mM Desthiobiotin	
EMSA Binding Buffer	20 mM Hepes, pH 7.5 100 mM NaCl 5 mM $MgCl_2$	EMSA buffer

4.1.2 Lab equipment

Table 15: Lab equipment

Equipment	*Producer*
KingFisher Duo prime	Thermofisher Scientific, Waltham, Massachusetts, USA
100 ml DWK Life Sciences Kontes™ Dounce Homogenizer	DWK Life Sciences, Wertheim, Germany
Peristaltic pump	
AEKTA prime plus	Cytiva Life Sciences
AEKTA micro	Cytiva Life Sciences
AEKA avant	Cytiva Life Sciences
Typhoon™ Biomolecular Imager	Cytiva Life Sciences
Vitrobot Mark IV	FEI (ThermoFisher Scientific)
Talos Arctica TEM	FEI (ThermoFisher Scientific)
Titan Krios TEM	FEI (ThermoFisher Scientific)
Gatan K3 Specs camera (Titan Krios)	Gatan, Pleasanton, USA
Falcon 3EC Specs camera (Titan Halo and Talos Arctica)	FEI (ThermoFisher Scientific)
Titan Halo TEM	FEI (ThermoFisher Scientific)
TECAN Infinite Pro plate reader	Tecan Group, Maennerdorf, Switzerland

4.1.3 Computing software

Table 16: Computing software

Software for molecular biological analysis	*Developer/Supplier*
ApE	M. Wayne Davis
ExPASy	Artimo et al., 2012
Unicorn	GE Healthcare Life Science
SnapGene	Insightful Science; available at snapgene.com

Software for structural analysis	*Developer/Supplier*
AlphaFold	Hassabis et al., 2021
Pymol	Schroedinger and DeLano, 2020; available at www.pymol.org/pymol
UCSF Chimera	Pettersen et al., 2004

Software for Electron Microscopy	*Developer/Supplier*
EPU	FEI (ThermoFisher Scientific)
SerialEM	Mastronarde, 2005
CryoSparc	Punjani et al., 2017
Relion	Scheres, 2012
MotionCor2	Zheng et al., 2017
CTFFIND 4	Rohou and Grigorieff, 2015
Gautomatch	www.mrc-lmb.cam.ac.uk/kzhang
DeepEMhancer	Sanchez-Garcia et al., 2021
Focus	Biyani et al., 2017

TOPAZ	Bepler et al., 2020

Software for graphical illustration	*Developer/Supplier*
Colorlogical	Gramazio, et al., 2016
Viz Palette	https://projects.susielu.com/viz-palette
Sketchbook app for iPad	Sketchbook, Inc.

4.2 Methods

4.2.1 Ligation Independent Cloning (LIC) of constructs for insect cell expression

The MultiBac™ system (Berger, Fitzgerald et al. 2004) was used for co-expression of multi protein complexes in insect cells in this thesis. This cloning principle is based on so called transfer vectors, pFBDM and pUCDM, which contain a multiplication module M inserted between the two promotors. Assembly of multi gene cassettes works by using restriction enzymes to cut out the whole cassette containing gene A and B with their respective promotors and terminator (Figure 97). This cassette is then added into the multiplication module M of the vector containing gene C and D. The pFBDM plasmid containing all three subunits of CstF complex (Table 10) was generated following this principle by the collaborators in Halle and used as template for re-cloning of the subunits as described in the following section.

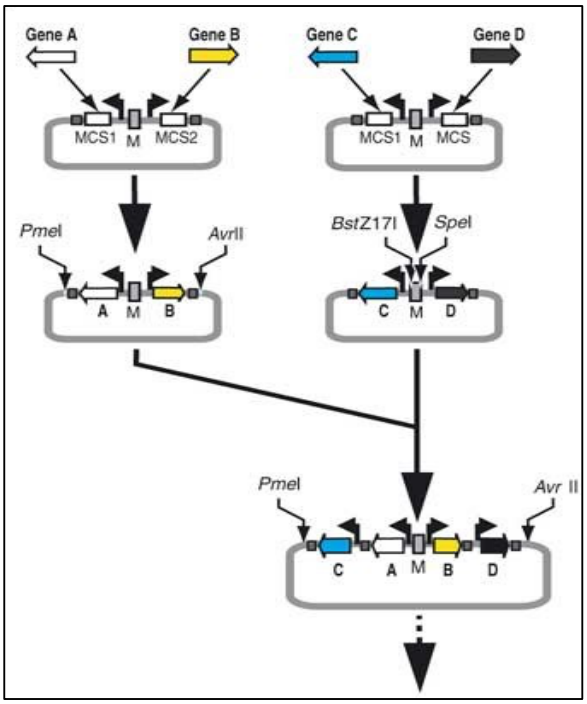


Figure 97: MultiBac system for expression of multiprotein complexes in insect cells (Berger, Fitzgerald et al. 2004).

DNA Templates and Polymerase Chain Reaction (PCR)

DNA templates for CstF complex were received from the group of Elmar Wahle (University Halle). Circular MultiBac vectors, which were used as templates for vector linearization by PCR, were available in the lab.

Genes of interest were amplified from templates by Polymerase Chain Reaction (PCR) using construct-specific primers (Table 11). PCR was carried out in an Eppendorf Mastercycler proS (Eppendorf, Hamburg, Germany) using standard conditions (see table 17) and 2x Phusion Flash HF DNA Polymerase Master Mix (ThermoFisher). Since the master mix already contains dNTPs, Phusion DNA Polymerase and Phusion buffer, it was directly added to the DNA template pre-mixed with insert-specific primers. For details about PCR reaction setup see table 17.

Table 17: PCR reaction setup

Component	*Stock concentration*	*Final concentration*	*Volume/amount*
Template DNA	50 - 500 ng/µl	5 – 10 ng	1 µl
Primer_fw	10 µM/10pmol	1 µM	1.5 µl
Primer_rev	10 µM/10pmol	1 µM	1.5 µl
Phusion MM	2x	1x	15 µl
dH_2O	-	-	up to 30 µl

PCR conditions were selected based on the insert to be amplify and on the optimal annealing temperatures for the respective primer pairs (using T_m Calculator online tool; ThermoFisher Scientific). Extension times were chosen based on the length of insert to be amplified, using 30sec/kb. For the detailed PCR cycle setup see Table 18.

Table 18: PCR cycle and condition

		aim	conditions
Initial denaturation		Complete denaturation of the dsDNA	30 sec., 98 °C
Repeat 30 cycles	Denaturation	Degradation of new synthesized strands from the matrix	10 sec., 98 °C
	Annealing	Annealing of the primers to the matrix	30 sec., $T_m \pm 3$ °C
	Elongation	Synthesis of complement strands	30 sec./ kb, 72 °C
Final elongation		The enzyme fills fragments	5 min., 72 °C
Final conservation		Sample conservation	infinite, 4 °C

PCR amplified inserts were analyzed by agarose gel electrophoresis using 1% agarose gels (1 % agarose w/v in 1x TBE (100 mM Tris base, 100 mM Boric acid, 2mM EDTA)), running in 1x TBE at a constant voltage of 110 V. Fragments of correct size were cut out of the gel and purified using the Wizard SV Gel and PCR Clean-Up System (Promega Corporation, Madison, Wisconsin, USA).

For vector amplification, PCR products of linearized pACEBac1 vector were purified using the Wizard SV Gel and PCR Clean-Up System (Promega Corporation, Madison, Wisconsin, USA).

Insert and Vector processing

PCR amplified inserts were cloned into MultiBac^{TM} pAcceptor (pACEBac1) vectors by LIC cloning, which can be used alternatively to restriction and ligation cloning. This technique is based on the 3´–5´ exonuclease activity of T4 DNA Polymerase, which thereby creates single stranded, complementary overhangs between insert and vector. By adding the sequence for appropriate extensions into primers, resulting vector and insert PCR products contain complementary overhangs. Treatment with T4 DNA Polymerase will thus create sticky ends and therefore make classic restriction & ligation unnecessary.

Before a PCR amplified vector could be processed with T4 Polymerase, it was digested with DpnI for 2.5 h at 37 °C. Dpn1 only cleaves methylated plasmid DNA and therefore only chews

up the template plasmid and not the newly synthesised linear vector DNA. Dpn1 digested reaction mixes were then separated on a 1% agarose gel, cut out and gel purified. Processing of insert and vector DNA using T4 DNA Polymerase and dNTP in order to create single stranded LIC overhangs was performed using the following pipetting scheme (Table 19). Reactions were incubated at room temperature for 30 minutes.

Table 19: Reaction mixture for T4 processing of insert and vector for LIC cloning

Vector processing		*Insert processing*	
component	amount	component	amount
Linearized vector	450 ng	Gel purified PCR product	600 ng
T4 DNA Pol buffer (10x)	3 µl	T4 DNA Pol. buffer (10x)	2 µl
dTTP (25 mM)	3 µl	dATP (25 mM)	2 µl
DTT (100 mM)	1.5 µl	DTT (100 mM)	1 µl
T4 DNA Pol. LIC qualified (Novagen)	0.6 µl	T4 DNA Pol. LIC qualified (Novagen)	0.4 µl
H_2O	to 30 µl	H_2O	to 20 µl

After incubation, the enzyme was inactivated for 20 minutes at 75°C, before vector and insert were mixed for the annealing reaction.

Annealing reaction and generation of MultiBac plasmids

Annealing of vector and insert was performed in a reaction volume of 10 µl following table 20. The mixture was incubated for 10 minutes at room temperature before, followed by addition of 1µl of 25 mM EDTA and another incubation at room temperature for 10 minutes.

Table 20: Mixture for LIC annealing reaction

Component	*amount*	*concentration*
insert	2 µl	10-40 ng/µl
vector	1 µl	15-50 ng/µl

2 µl of the annealing reaction was transformed in electro- or chemically competent *E.coli* strains, e.g. XL1 Blue and plated on agar containing the appropriate antibiotic. The next day, colonies were picked and grown in 2-3 ml LB medium overnight. Plasmids were isolated with the QIAprep Spin Miniprep Kit (Qiagen, Hilden, Germany) according to the manufacture's protocol and sent for sequencing (Eurofins).

4.2.2 Generation of RRM mutants by site directed mutagenesis

Single and double mutations were introduced into CstF2 coding sequence by PCR. Overlapping primers were designed such that both forward and reverse primer carried the desired mutation (Table 10). In each primer pair, mutations were flanked by 10-15 nucleotides and primers were designed such that their melting temperatures (T_m) were in very similar range. PCR reaction was set up according to table 21 for standard insert amplification and for the program used.

Table 21: PCR cycle setup for site directed mutagenesis

		aim	conditions
Initial denaturation		Complete denaturation of the dsDNA	1 min, 98 °C
Repeat 35 cycles	Denaturation	Denaturation of new synthesized strands from the matrix	30 sec., 98 °C
	Annealing	Hybridisation of the primers to the matrix	1 min, $T_m \pm 3$ °C
	Elongation	Synthesis of complement strands	30 sec./ kb, 72°C
Final elongation		The enzyme fills incomplete fragments	5 min., 72 °C
Final conservation		Sample conservation	infinite, 4 °C

After PCR amplification, 1 µl of DpnI was added to the reaction and incubated for 1 h at 37 °C. The digestion was directly transformed into *E.coli* XL 1 blue cells as described in 4.2.3. Positive clones were verified by sequencing (Eurofins).

4.2.3 Transformation of bacterial cells with recombinant DNA

Transformation of DNA into chemically competent *E.coli* cells was done by heat shock method. 0.1 µg of DNA was incubated with competent *E.coli* cells for 20 minutes on ice, followed by a heat shock for 45 seconds at 42 °C. Cells were placed on ice for 10 minutes. After addition of

270 µl SOC medium, transformations were incubated for 1h shaking at 37 °C before plating on selective agarose plates.
Electro-competent *E.coli* strains were thawed on ice for 1 min before approximately 0.2 µg of plasmid DNA was added. Cells were transferred to 0.1 cm electroporation cuvettes followed by electroporation using a Biorad MicroPulser (BioRad, Hercules, California, USA) and an electrical pulse of approximately 1.8 kV. 270 µl SOC medium was added to the bacteria suspension and cells were incubated at 37 °C shaking for 1h in a table top shaker, before plating them on agar containing the respective antibiotics and overnight incubation at 37 °C.

4.2.4 Bacmid isolation

Plasmids containing correctly tagged target sequences were transformed into chemically competent DH10EmBacY *E.coli* cells, which contained recipient baculoviral DNA (bacmid), a helper plasmid producing transposase to integrate plasmids into the resulting recombinant bacmid. The vector itself carried DNA elements to integrate into the baculovirus by T7 transposition. Transformed cells were plated on special MultiBac plates (Table 7) and positive colonies were selected based on blue/white screening principle. Very briefly, the principle of blue/white screening is that genes (PCR products) are inserted into a multiple cloning site (MCS) within a LacZ sequence. Consequently, upon successful integration of foreign DNA, a functional β-galactosidase can no longer be produced. Therefore, X-gal (present in the agar) cannot be hydrolyzed and yield the blue color.
For recombinant bacmid preparation 2 ml of LB medium containing kanamycin and gentamycin were inoculated with a single white single colony from the previous step 4.2.3 and grown at 37 °C over night. Cells were harvested by centrifugation for 15 minutes at 3220xg. Alkaline lysis was performed using the QIAprep Spin Miniprep Kit and recombinant bacmid DNA was isolated following the manufacture's protocol until a clear supernatant containing the DNA. Instead of using spin columns, DNA was precipitated by adding 800 µl of iso-propanol. Samples were placed on ice for 10 minutes to precipitate DNA. After another 15 minutes centrifugation step at room temperature, the DNA pellet could be located and the supernatant carefully removed. 800 µl of 70 % ethanol were now added to wash and sterilize the pellet. Work was now continued under sterile conditions. After a short centrifuge step of 5 minutes, ethanol was removed and the pellet was air-dried for 5 to 10 minutes. DNA was dissolved in 40 µl of sterile H_2O by tapping the tube. DNA concentrations were determined after 10 minutes with a NanoPhotometer.

4.2.5 Transfection and generation of baculoviruses

All baculovirus constructs were generated using transfer vectors from the MultiBac system (Berger et al., 2004), which can integrate genes via T7 transposition or Cre-lox site specific recombination in baculoviral DNA in *E.coli*. For some genes His- or Strep-tags were fused to the N- or C- termini. Generation of recombinant viruses was performed according to the MultiBac protocol (GenevaBiotech). Following baculoviruses coding for tagged or untagged human proteins were used in this thesis (Table 22).

Table 22: Recombinant baculovirus constructs.

#	*Virusname*	*encoded ORFs*
1	EmbacY-pACEBac1_TwinStrep-CstF2	TwinStrep-CstF2 (64k)
2	EmbacY-pACEBac1_Strep-CstF3-His_8	Strep-CstF3- His_8 (77k)
3	EmbacY-pFBDM_Strep-CstF1-CstF2	Strep-CstF1 (50k) CstF2 (64k)
4	EmbacY-pACEBac1_Strep-CstF1	Strep-CstF1 (50k)
5	EmbacY-pACEBac1_ CstF2-RH-Strep	CstF2 (1-204)-Strep (25k)
6	EmbacY-pACEBac1_ CstF2-RH-TwinStrep	CstF2 (1-204)-TwinStrep (25k)
7	EmbacY-pACEBac1_TwinStrep-CstF2(S17A)	TwinStrep-CstF2 (64k)
8	EmbacY-pACEBac1_TwinStrep-CstF2(S17A)	TwinStrep-CstF2 (64k)
9	EmbacY-pACEBac1_TwinStrep-CstF2(F19A)	TwinStrep-CstF2 (64k)
10	EmbacY-pACEBac1_TwinStrep-CstF2(F61A)	TwinStrep-CstF2 (64k)
11	EmbacY-pACEBac1_TwinStrep-CstF2(N91A-N97A)	TwinStrep-CstF2 (64k)
12	EmbacY_pFBDM_Strep-CstF1-CstF2-His-CstF3	His_8-CstF3 (77k) Strep-CstF1 (50k) CstF2 (64k)

All steps were performed in a cell culture room at 27°C.

For this work, insect cells were only transfected with single viruses.

To produce first-generation virus (P1 virus), 2 μg of bacmid-DNA were pre-incubated with 5μl FuGene HD Transfection Reagent (Promega, Madison, Wisconsin, USA) in 200μl serum free medium for 15 minutes, then added to $0.8x10^6$ adherent Sf21 cells in a 6-well tissue culture plate and incubated for 3-7 days. P1 virus containing supernatant was harvested and stored at 4 °C protected from light.

For first virus amplification (second generation or P2 virus), 5 µl P1 virus stock were added to 25 ml of Sf21 cells at $0.5x10^6$ cells/ml. After shaking at 85 rpm for 48 h, cells were counted and diluted to $0.8x10^6$ cells/ml with fresh medium. After another 48h, the supernatant containing P2 virus was harvested by centrifugation and the P2 virus stock was stored at 4 °C. The pellet obtained from centrifugation could be used to check for target protein expression
Final large-scale virus production (P3 virus) was performed in 100ml Sf21 cells at $0.5x10^6$ cells/ml, to which a final concentration of 0.1 % (v/v) of P2 virus was added. Cells were incubated shaking at 85 rpm for 48 h, counted and diluted again to $0.8x10^6$ cells/ml. After 48 h the supernatant containing P3 virus was collected and stored at 4 °C light protected.

4.2.6 Protein expression

4.2.6.1 Protein expression in bacterial cells

For a pre-culture, 100 ml of lysogeny broth (LB) medium containing the respective antibiotics were inoculated with an individual colony of *E.coli* BL21 (DE3) pLysS cells transformed with the respective plasmid and grown at 37 °C overnight. The next day, 6 × 500ml of terrific broth (TB) medium containing the respective antibiotics were inoculated with 1 ml of pre-culture per 100 ml of culture and cells were grown at 37 °C in a shaker at 120 rpm until OD_{600} reached 0.6-0.8. The shaker was then cooled down to 22 °C and protein expression was induced by adding 1 mM IPTG. Cells were harvested after 18 h on the next day by centrifugation for 10 minutes at 8000xg at 4 °C. Cell pellets could either directly be processed for purification or were frozen in liquid nitrogen and stored at -80 °C.

4.2.6.2 Small scale protein expression test in insect cells

For every virus of a new or already existing constructs, proteins expression was first tested expressed in small scale and virus to culture ratio and amount was optimized before those conditions were then transferred to large scale protein production in insect cells. For a standard expression test, Sf21 and High five were used since some proteins were better expressed in one or other cell type for expression. 25 ml culture of each cell line was set up in a 10 ml flask at 10^6 cells per ml culture and 1% (v/v) of target P2 or P3 virus was added. In case of co-infection with several viruses, I tested expression with an overall amount of 1% (v/v) virus added to the culture or with 1% (v/v) of each virus. Expression tests were incubated at 27 °C for 72 h on a shaker at 85 rpm. Cells were harvested by centrifugation at 800xg for 10 minutes and pellets were used for affinity pulldowns to check for the presence of the target proteins.

4.2.6.3 Affinity pulldown assay for analysis of recombinant protein expression from insect cells

Successful expression of different constructs was tested by affinity pulldown experiments using a magnetic bead processor (KingFisher Duo Prime, ThermoFisher Scientific, Waltham, Massachusetts, USA). Cell pellets from a 5 ml culture volume were resuspended in 1 ml of lysis Buffer (CstF lysis buffer, Tab. 14) and incubated on ice for 10 minutes with 1 mM AEBSF and 25U/1µl Benzonase. Lysates were cleared by centrifugation at highest speed (e.g. 16000xg) in a tabletop centrifuge for 15 min at 4 °C and supernatants were added afterwards into a 96-well plate according to the scheme in figure 98. Following magnetic affinity beads were used for His-tag and Strep-tag mediated pulldowns: Invitrogen Dynabeads™ His-Tag Isolation and Pulldown (ThermoFisher Scientific), MagStrep "type3" XT beads (IBA Lifesciences, Goettingen, Germany). Elutions were analyzed with SDS PAGE.

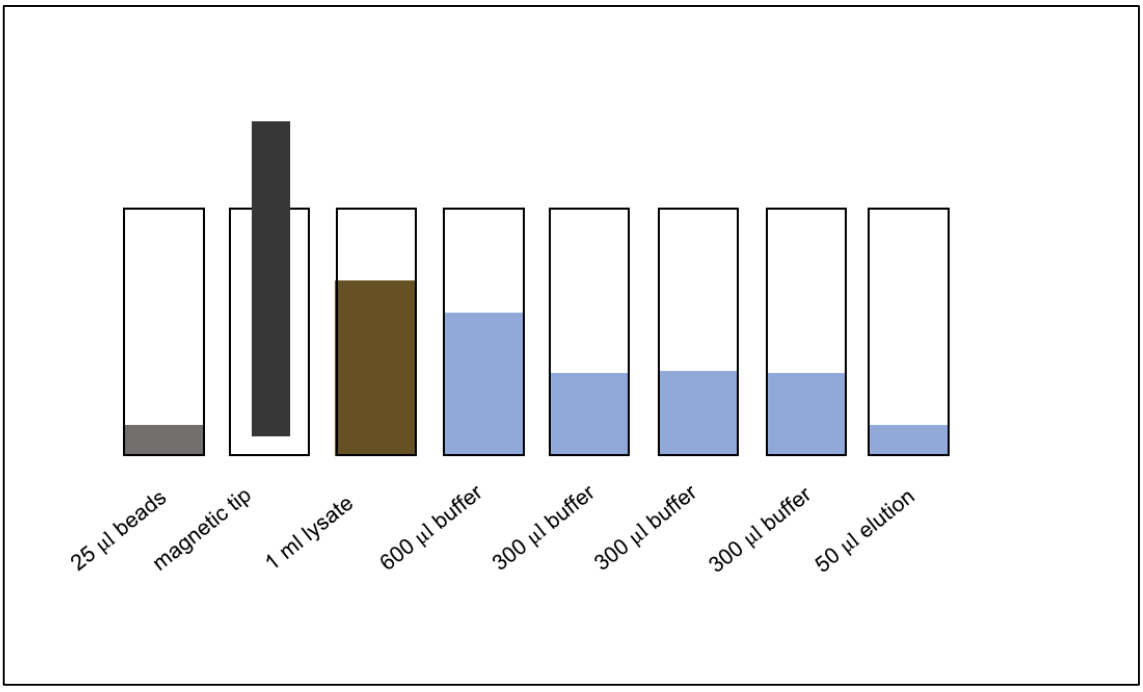


Figure 98: Pipetting scheme of an affinity pulldown performed with a King Fisher Duo prime magnetic bead processor. Pulldowns were performed in 96 deep well plates.

4.2.6.4 Large scale protein expression in insect cells

Large-scale protein expression was performed in High Five cells if not mentioned differently in the text. 1 L of 10^6 cells/ml of High Five cells were infected with 1% (v/v) of the specific P3 virus in 3 L wide bottom flasks. Cultures were kept at 27 °C, shaking for 72 h at 85 rpm. Cells

were harvested by centrifugation in a JLA 8.100 rotor at 800xg for 10 min, pellets were frozen in liquid nitrogen and stored at -80 °C for later protein purification.

4.2.7 Protein purification

4.2.7.1 Protein purification for cryo-EM studies

In general, CstF complexes were prepared by co-lysing combinations of insect cells that expressed individual subunits. Frozen cell pellets for each, CstF1, CstF2 and CstF3, were thawed on ice and resuspended in 20 ml of lysis Buffer (Table 14) per 10 g cell pellet, containing one Complete Protease Inhibitor tablet per 100 ml of lysis buffer, 750 U Benzonase (stock 750U/µl) and 1 mM AEBSF. To compensate for different expression levels, a 20 g pellet was used for CstF2 and the amount for the other subunits were adjusted in the following ratio 1:2:0.5 (CstF1:CstF2:CstF3). Cell lysis was performed using a 100 ml DWK Life Sciences Kontes™ Dounce Homogenizer (DWK Life Science, Wertheim, Germany) and ten up and down strokes. Crude homogenates were then clarified by centrifugation at 75600xg in a JA 25.50 rotor for 30 minutes at 10 °C. Protein containing supernatants were filtered through a 5 µm syringe filter and directly used for the first affinity purification step, depending on the type of affinity tag. 1 mM AEBSF was applied to all buffers and all steps were performed at 4 °C to prevent protein degradation.

Purification of the CstF apo complex

For recombinant CstF complex the co-lysate was loaded on a StrepTrap™ HP 5 ml column (Cytiva, Munich, Germany) with an IPC High Precision Multichannel Dispenser peristaltic pump (Ismatec, Wertheim, Germany). Flow-through was collected and columns were pre-washed with 10 column volumes (CV) of CstF wash/lysis Buffer. Further washing steps and elution of the protein was done by connecting the column to an AEKTAprime Plus system (Cytiva, Munich, Germany). The column was washed for 15 CV with CstF lysis/wash buffer, then 8 CV with CstF High buffer and again 4 CV with CstF lysis/wash buffer. Bound proteins were eluted with 5 ml of Strep elution buffer (Table 14). Protein containing fractions were analyzed by SDS PAGE. Usually, full CstF complex eluted from initial Strep affinity column already in quite stoichiometric amounts without major contaminants, so that it could already be used for further sample preparation for cryo-EM studies. In case of contaminants co-eluting with CstF from the StrepTrap, protein containing fractions were pooled and diluted 1:4 with Heparin binding buffer, before loading on a HiTrap 5 ml Heparin column (Cytiva, Munich, Germany). After washing with 10 CV of Heparin binding buffer (Table 14), the complex was eluted by a gradient of 0-100% of Heparin elution buffer over 10 CV. After analysis by a 12.5% SDS PAGE, all protein-containing fractions were pooled and concentrated in an Amicon Ultra 100 kDa MWCO concentrator (Merck Millipore, Burlington, Massachusetts, USA) to a final volume between 200

and 500 µl usually corresponding to a protein concentration between 5-10 mg/ml. Samples were directly used for preparation for cryo-EM grids as described in paragraph 4.2.10.2 or stored at -80 °C for later usage.

CstF1-CstF3 subcomplex

Recombinant CstF1-3 subcomplex was purified according to the protocol described for full CstF complex. For co-lysis, frozen cells from individual expression were mixed in a ratio of 2:1 (CstF1:CstF3) to have near stoichiometric amounts of each subunit (CstF3 was usually expressed at a twofold higher level than CstF1). Briefly, co-lysate was loaded on a StrepTrap™ HP 5 ml column, which was washed for 25 CV with CstF lysis/wash buffer, then 8 CV with CstF High and again 4 CV with CstF lysis/wash buffer. Bound proteins were eluted with 5 ml of Strep elution buffer (Table 14) and protein containing peak fractions were analyzed by a 12.5% SDS PAGE. All protein-containing fractions were pooled and concentrated in an Amicon Ultra 100 kDa MWCO concentrator to a final volume of 500 µl, corresponding to a protein concentration between 10-15 mg/ml. Samples were either stored at -80 °C or directly used for cryo-EM sample preparation (section 4.2.10.2).

4.2.7.2 Complex preparation for biophysical/biochemical studies

Cell lysis and clearance of cell lysate were done as described in paragraph 4.2.7.1. After initial affinity purification, Size Exclusion Chromatography (SEC) was used as last purification step to exchange purification buffer to desired conditions for further experiments. For all samples used for Fluorescence Anisotropy (FA) and Isothermal Titration Calorimetry (ITC) measurement, SEC was performed in 1xPBS buffer containing 5 mM $MgCl_2$.

CstF apo complex

Briefly described, for recombinant CstF complex, co-lysate was loaded on a StrepTrap™ column, which was initially washed for 25 CV with CstF lysis/wash buffer, then for 8 CV with CstF High and finally for 4 CV with CstF lysis/wash buffer. After elution of bound proteins with 5 ml of Strep elution buffer (Table 14), protein containing fractions were pooled and diluted 1:4 with Heparin binding buffer. After loading on a HiTrap Heparin column and washing with 10 CV of Heparin binding buffer (Table 14), the complex was eluted by a gradient of 0-100% of Heparin elution buffer over 10 CV. Peak fractions were analyzed by a 12.5% SDS PAGE and all protein-containing fractions were pooled and concentrated in an Amicon Ultra 100 kDa MWCO concentrator to a final volume of 500 µl, corresponding to a protein concentration of around 20-30 mg/ml. SEC was performed on an AEKTAavant (Cytiva, Munich, Germany). Concentrated complex was injected via a sample loop onto a Superose6 increase (S6i) 10/300

column (Cytiva, Munich, Germany) and eluted in buffer SEC-4 (Table 14) in 1 ml fractions. All peak fractions were loaded on a 12.5% SDS gel for final analysis to check presence of all proteins of the complex. Fractions containing intact complex with all three proteins present were then pooled, concentrated in an Amicon Ultra 100 kDa concentrator to stock concentrations between 10-15 mg/ml. Protein stocks were frozen in liquid nitrogen and stored at -80 °C or directly used for further analysis.

CstF1-CstF3 subcomplex and CstFmut

CstF1-CstF3 subcomplex and full CstF (CstFmut) carrying single or double mutations in the RRM domain of CstF2 were purified accordingly. In both cases, complexes were formed by co-lysis of individual pellets in different ratios as described in the paragraph above (2.2.7.1). Lysate was then loaded on a StrepTrap HP 5 ml column with a peristaltic pump. The flow-through was collected and columns were pre-washed with 10 CV of CstF wash/lysis buffer. Further washing steps and elution of the protein was done by connecting the column to an AEKTAprime plus system. The column was washed for another 10 CV with CstF lysis/wash buffer, then 10 CV with 500 mM NaCl (50% CstF High), 4 CV with 1 M NaCl (100 % CstF High) and finally 4 CV again with CstF lysis/wash buffer. The subcomplex was eluted from the StrepTrap column by injecting 5 ml of Strep elution buffer (Table 14). Peak fractions were pooled and loaded on a HisTrap™ FF 5 ml column (Cytiva, Munich, Germany) equilibrated with His binding buffer (Table 14). The column was pre-washed for 10 CV with His binding buffer, then connected to an AEKTAprime plus system. After further washing with His binding buffer for 25 CV, His-tagged proteins were eluted from the column in one step with 250 mM Imidazol. Peak fractions were analyzed by SDS PAGE, concentrated and loaded on a S6i 10/300. SEC was performed in 1xPBS containing 5 mM $MgCl_2$. Target complex was finally analyzed by 12.5% SDS PAGE, concentrated to stock concentrations between 10-15 mg/ml and stored at -80 °C.

CstF2 and CstF2-3 subcomplex

CstF2 and CstF2-CstF3 subcomplex were purified in a similar way. For CstF2 (N-terminal TwinStrep-tag), pellet from a 3 L large-scale expression from insect cells (corresponding to around 40 g of cell pellet) was needed to get reasonable amounts of protein. Cell lysis procedure was done as already described in chapter 4.2.7.1 and cleared cell lysate was loaded on a StrepTrap HP 5 ml column for the first affinity purification step. based on the N-terminal TwinStrep-tag on CstF2. The loaded StrepTrap column was then washed with 10 CV CstF lysis/wash buffer, followed by 10 CV of 500 mM NaCl, 4 CV of 1 M NaCl and finally another 4 CV of CstF lysis/wash buffer, before the single protein was eluted with 5 mM desthiobiotin.

CstF2 usually eluted with a contaminating protein running at a molecular weight of around 63.37 kDa, which was visible on the SDS PAGE in almost stoichiometric amounts compared to the target protein. To separate the contamination from CstF2 target protein, the Strep elution was diluted to below 100 mM NaCl with Heparin binding buffer and loaded on a HiTrap Heparin 5 ml column. Under these conditions, CstF2 was in the flow through and the contamination band remained bound to the column. CstF2 protein could then be concentrated in an Amicon Ultra 50 kDa concentrator and stored in the desired buffer used for RNA binding studies.
To form CstF2-CstF3 subcomplex, co-lysis of insect cell pellets which expressed TwinStrep tagged CstF2 and CstF3, N-terminally Strep-tagged and carrying a C-terminal His-tag, was done by mixing pellets in a ratio of 1:3 (CstF3:CstF2), because expression of CstF3 was roughly three times better. Cell lysis followed the protocol described for CstF apo-complex in paragraph 4.2.7.1, as well as clearing of the lysate and loading on the StrepTrap column. After pre-wash with 10 CV of CstF lysis/wash buffer (250 mM NaCl), the column was connected to the AEKTAprime plus system and further washed with 250 mM NaCl for 10 CV. Additional washing with 1 M NaCl for 10 CV and 250 mM NaCl for 5 CV was performed, before the subcomplex was eluted from the column with 5 mM desthiobiotin. Eluted fractions were diluted to 100 mM NaCl with Heparin binding buffer (Table 14), before loading on a Heparin 5 ml column. The column was washed with Heparin binding buffer (Table 14) for 10 CV and proteins were eluted with a 0-100% gradient for 20 CV using 100% Heparin elution buffer (1M NaCl). Peak fractions were analyzed by SDS PAGE, concentrated in an Amicon Ultra 100 kDa MWCO concentrator and stored in 1xPBS containing 5 mM $MgCl_2$ for Fluorescence Anisotropy measurements.

CstF2 RNA Recognition motifs (RRMs)

CstF2 single RRM was expressed as described in 4.2.6.1. Frozen cell pellets from 3l *E.coli* were thawed on ice and resuspended in 20 ml pellet of CstF lysis buffer (containing one Complete Protease Inhibitor tablet per 100 ml lysis buffer, 750 U Benzonase and 1 mM AEBSF, see Table 14) per 10 g pellet. Before performing cell lysis by sonification, 1 ml of 10 mg/ml lysozyme and 5 μg/ml DNaseI were added to the resuspended cells and stirred on ice for 10 minutes. Sonication was performed on ice using a Sonopuls HD 3200 and a VS-70T probe (both Bandelin). Crude extracts were then clarified by centrifugation at 75600xg in a JA 25.50 rotor for 30 minutes at 10 °C. Protein containing cell lysates were filtered through a 5 µm filter and directly loaded on a StrepTrap 5 ml column. The column was washed for 10 CV with RRM Strep/SEC buffer, then 6 CV with RRM High and finally 4 CV with RRM Strep/SEC buffer (Table 14). Single CstF2 RRM was eluted injecting 5 ml of RRM elution buffer via loop on the column. Peak fractions were analyzed by 15% SDS PAGE, concentrated in Amicon Ultra 3 kDa MWCO concentrator and loaded on a S6i 10/300 column. Single RRMs eluted in RRM

SEC buffer in a symmetric peak from the column and peak fractions were finally analyzed by 15 % SDS PAGE, concentrated again using an Amicon Ultra 3 kDa MWCO concentrator and stored in 1xPBS containing 5 mM $MgCl_2$. Expression and purification of RRM fusion constructs, both wild type and mutants, followed the same protocol, except that the StrepTrap 5 ml column was pre-equilibrated with the sodium phosphate based CstF lysis/wash buffer, instead of using a Hepes-based buffer as for single RRM. After loading protein containing cell lysates, the column was washed with 10 CV of CstF lysis/wash buffer, followed by a washing step with 500 mM NaCl and a high salt wash with 4 CV of CstF High buffer (Table 14). Before RRMs were eluted with 5 mM desthiobiotin, the column was again washed with 4 CV of CstF lysis/wash buffer. Peak fractions containing RRM fusion were concentrated in an Amicon Ultra 10 kDa MWCO concentrator and loaded on a S6i 10/300 SEC column. RRMS eluted in 1xPBS containing 5 mM $MgCl_2$ from SEC column and finally analyzed using a 15 % SDS PAGE. Peak fractions were pooled, concentrated in an Amicon Ultra 10 kDa MWCO concentrator and stored at -80°C for further experiments.

4.2.7.3 Protein cross-linking experiments

In-batch cross-linking studies

Cross-linking studies in this thesis were performed using glutaraldehyde (GA) or BS3 cross-linker reagents. BS3 stock solutions were prepared by resuspending BS3 sodium salt (ThermoFisher Scientific, Waltham, Massachusetts, USA) in water at room temperature to a final stock concentration of 50 mM. GA (Merck-SigmaAldrich, Darmstadt, Germany) was delivered as a 25% solution dissolved in water and was diluted to desired stock concentrations with water.

Initial cross-linker screening was performed in-batch, where protein samples were kept at a constant concentration of around 1-1.5 μM and concentration of cross-linker was increased in steps until an excess of 20x cross linker was reached. Cross-linking reactions with GA were incubated for 20 min at room temperature and those with BS3 for either 30 min at 26 °C or 5 min at 30 °C. All reactions were quenched with a final concentration of 25 mM Tris/HCl, pH 7.5 and loaded on an SDS PAGE for further analysis.

Gradient Fixation (GraFix)

Gradient Fixation (Stark 2010) is a combination of density gradient ultracentrifugation and cross-linking and was used in this thesis to stabilize protein complexes for cryo-EM single particle analysis.

Formation of a 5-25% gradient was performed by layering the two different sucrose buffers (Table 14) in a 14 x 95 mm SETON centrifugal tube (SETONScientific, Petaluma, USA). First, 6.5 ml of the less dense solution (5% sucrose, Tab. 14) was put into the tube, followed by

slowly adding the heavier solution (25% sucrose, Tab. 14) such that it settles underneath the less dense solution. Cross-linking reagents were always applied to the dense solution in a concentration range 0.05 – 0.2 % for glutaraldehyde and 2 mM for BS3. Tubes were placed into a Biocomp Piston IP gradient station (Biocomp Instruments, Fredericton, Canada), which formed a continuous gradient by slowly rotating the tubes. Tubes were equilibrated for at least one hour at 4 °C, before 200-400 µl of protein sample was carefully loaded on top of the gradient. Samples were then centrifuged in a SW-40 swing out rotor in a Beckman Optima XE (Beckman Coulter, Brea, California, USA) Ultracentrifuge for 18 h at 249375xg. Next, centrifugation gradients were fractionated from the top into 300 µl fractions. Protein concentrations were determined with a NanoPhotometer NP80 (Implen, Munich, Germany) and samples were analyzed by SDS PAGE.

Cross-linking - Mass Spectrometry studies (XL-MS)

XL-MS was performed with purified CstF1-CstF2-CstF3 complex (paragraph 4.2.7.1). After elution from the Strep column, the sample was concentrated to approximately 10mg/ml. Cross-linking was performed using 2 mM BS3 for 5 minutes at 30 °C and quenching with 25 mM Tris/HCl, pH 7.5. Mass Spectrometry analysis was performed by the in-house Mass Spectrometry Facility. Briefly, after cross-linking the CstF complex with BS3 in its native state, samples were enzymatically digested into peptides. Cross-linked peptides are then separated via liquid chromatography-tandem mass spectrometry (LC-MS/MS). MS-database analysis delivered cross-linked peptides and their cross-linking sites. In this thesis, cross-linking results were analyzed using UCSF Chimera containing the Xlink analyzer plugin (Kosinski et al 2015)

4.2.7.4 General analytic methods

Protein concentration determination

Concentration of protein samples were determined using a NanoPhotometer NP80 (Implen, Munich, Germany) and measuring the absorbance of the samples at 280 nm and 260 nm. For all measurements, relation between OD_{280} and protein concentration in mg/ml was 1:1.

SDS-PAGE

Quality and purity of proteins at different purification steps and final protein samples were analyzed by Sodium Dodecyl Sulfate Polyacrylamide Gel Electrophoresis (SDS PAGE; Laemmli at al., 1970). In general, homemade 12.5% or 15% polyacrylamide gels (see table 23) were run in a vertical electrophoresis chamber (BioRad, Munich, Germany) at constant voltage in standard SDS running buffer (25 mM Tris, 192 mM glycine, 0.1% w/v SDS). Samples were loaded after mixing with 2x SDS laoding buffer (100 mM Tris/HCl

pH 6.8, 200 mM dithiothreitol, 4% SDS, 0.2% bromophenol blue, 20% glycerol) and heating for 5 min at 95 °C. Gels were stained with InstantBlue commassie stain (Abcam, Cambridge, UK).

Table 23: Recipe for preparation of a 12.5 % SDS PAGE

Component	*Amount for resolving gels*	*Amount for stacking gels*
dH_2O	7 ml	5.2 ml
Tris II (0.375 M Tris, pH 8.8) Tris III (0.125M Tris, pH 6.8)	5.4 ml (Tris II)	2.4 ml (Tris III)
10% SDS	216 µl	92 µl
30% Acrylamide (37.5:1)	8.6 ml	1.54 ml
10% APS	214 µl	54 µl
TEMED (Tetramethylethylendiamin)	21.6 µl	27 µl

Native PAGE for RNA and protein binding assays

To analyze RNA binding properties of CstF complex and distinct subcomplexes, Electro Mobility Shift Assays (EMSA) were carried out as described in paragraph 4.2.8.3. Homemade 6% Tris-Borate-EDTA (TBE) gels (see table 24) were polymerized for 10 min at room temperature. Before samples were loaded, gels were pre-run in 0.5x TBE buffer for 30 minutes at 4 °C at a constant power of 2 W. After samples were loaded, gels were run for 30 minutes with same settings used for the pre-run.

Table 24: Recipe for preparation of a 6 % TBE gel

Component	*Amount for two 6 % TBE gels*
dH_2O	7.2 ml
5x TBE	2.4 ml
30% Acrylamide (37.5:1)	2.4 ml
10% APS	200 µl
TEMED (Tetramethylethylendiamin)	10 µl

Analytical Size Exclusion Chromatography (SEC)

Analytical gelfiltration was performed on an AEKTAmicro system (Cytiva, Munich, Germany) using a Superdex200i 3.2/300 or a Superose6i 3.2/300 column (Cytiva, Munich, Germany). Protein samples were diluted in buffer SEC-4 to the desired concentration and centrifuged for 10 minutes at 13400 xg and 4 °C in a tabletop centrifuge (Eppendorf, Hamburg, Germany), before 25 µl were injected via a 25 µl loop on the column. Runs were performed with a constant flow of 0.04 ml/min and 100 µl volume fractions were collected.

4.2.8 RNA binding studies

All RNA species used in this study (Table 13) were ordered from Biomers.net GmbH (Ulm, Germany) and delivered as lyophilized, dry pellets. RNA stocks were prepared by dissolving a specific RNA oligonucleotide in a small volume of water to get as high stock concentrations as possible. Stocks were stored at -20 °C. Working solutions were prepared by diluting the certain oligonucleotide to the desired working concentration in the buffer used for the experiment following. All RNAs carrying a 6'FAM label at their 5' ends were stored and handled such that they were protected from light.

4.2.8.1 Fluorescence Anisotropy

Fluorescence Anisotropy (FA) is a spectroscopy method, that can be used to study protein-ligand interactions (Mann and Krull 2003). In this thesis, protein - RNA interactions were studied by FA, meaning that fluorescently labeled RNAs served as ligands. The different RNAs (Table 13) used in this study had a fluorescein label at the 5'-end. Measurements were carried out in a TECAN Infinite Pro plate reader (Tecan Group, Maennedorf, Switzerland) at room temperature. In brief, for each combination of RNA and protein, a series of reactions was prepared where the RNA concentration was kept at a constant value of 11.75 nM and protein concentration was varied in discrete steps and the fluorescens signal was measured for each reaction. 1 µl of RNA at 117.5 nM was added to the 9 µl of the protein diluted in buffer 1xPBS + 5 mM $MgCl_2$ (Table 14) in a 96-well plate. After a 10 min incubation step at room temperature, measurements were started. They were all repeated independently from each other for three times. Change in free energy ΔG is calculated during each binding reaction according to the following formula:

$$\Delta G = \Delta H - T\,\Delta S$$

Together with the known temperature T, the K_D can be calculated using following equation:

$$K_D = e^{-\frac{\Delta G}{R \cdot T}}$$

with K_D = dissociation constant; T= temperature in [K]; R = 8.314 J mol^{-1} K^{-1} or 0.008314 kJ mol^{-1} K^{-1}.

Fluorescence Anisotropy measurements were carried out by Dr. Claire Basquin.

4.2.8.2 Isothermal Titration Calorimetry

Isothermal Titration Calorimetry (ITC) is a biophysical method to determine the dissociation constant of a complex by measuring the heat production or absorption upon combining a biological macromolecule and a ligand. The ITC200 (Malvern Panalytical, UK) consists of two cells, one reference cell and one sample cell, both containing one binding partner e.g. the protein sample. The second binding partner e.g. RNA ligand is filled into a syringe and titrated into the cell. Measurements consist of the time-dependent input of power required to maintain equal temperatures between the sample and reference cell upon titration of the ligand to the sample cell. Temperature difference between the cells caused by the reaction heat is compensated by either lowering (exothermic reaction) or increasing (endothermic reaction) the thermal power to maintain the temperature equilibrium.

Measurement of the heat change generated by the binding reaction is registered and permits to determine the kinetic parameters (ΔG, ΔH, ΔS and K_D, as well as the stoichiometry n) of the interaction by integrating the thermal power needed to keep constant temperatures in both cells over time. The released or absorbed heat decreases upon saturation of the substrate with increasing ligand concentration.

CstF01 RNA was used at a concentration of 300 μM in this experiment. It was titrated to a protein sample in the main cell at a concentration of 25 – 30 μM.

Different ITC experiments were carried out by Dr. Claire Basquin.

4.2.8.3 Electrophoretic Mobility Shift Assay

Electrophoretic Mobility Shift Assays (EMSA) were performed using fluorescently labeled RNA ligands and different CstF protein samples. RNA species from Table 13 containing a 5'-6-FAM-label were used in this study. According to following pipetting scheme (Table 25), RNA concentration was kept at a constant value of 10 pmol and concentrations of CstF complexes were increased stepwise from 0.6 – 96 pmol.

Table 25: Pipetting scheme for EMSA

RNA : protein	*1:0*	*1:0.5*	*1:1*	*1:2*	*1:4*	*1:10*	*1:20*	*1:40*	*1:80*
C_{RNA} /[pmol]	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2
$C_{protein}$ /[pmol]	0	0.6	1.2	2.4	4.8	12	24	48	96
$C_{yeast\ RNA}$ /[pmol]	12	12	12	12	12	12	12	12	12

Generally, proteins were diluted in EMSA Binding buffer (Table 14) to the working concentration and a 20-fold excess (24 pmol) of total yeast RNA extract was added. Specific target RNA was added and RNA-protein mixtures were incubated at 4 °C for 30 minutes. 2 µl of gel loading buffer (50 % glycerol, 0.3% (w/v) Orange G) was added to the samples, before they were loaded on a 6 % TBE gel. Electrophoresis was performed at 4 °C at a constant power of 2 W. Gels were imaged immediately after the run with an Amersham Tyhoon Biomolecular Imager (Cytiva Life Sciences, Freiburg, Germany) using the fluor stage and the Cy2 emission filter. The excitation was in a range from 515 to 535 nm.

4.2.9 Preparation of CstF complexes for Transmission Electron Microscopy

Recombinantly CstF complex purified according to the protocol described in paragraph 4.2.7.1, was directly prepared before every grid plunging session.

4.2.9.1 Preparation of apo-CstF complex by cross-linking via GraFix

Purified full CstF complex (see paragraph 4.2.7.1) was directly loaded on a 5-25% sucrose density gradient containing 0.01% of GA. GraFix tubes were prepared as described in section 4.2.7.3. Briefly, 200 µl of purified CstF complex at a concentration around 10 mg/ml were layered on top of the sucrose gradient and centrifuged for 18 h at 249375xg in a swing out rotor. Protein containing fractions were analyzed by SDS PAGE to check the cross-linking efficiency. The cross-linked protein was pooled and concentrated for a final analytical SEC to exchange the sucrose-based buffer to SEC-4. 25 µl of cross-linked CstF at a concentration of 1-2 mg/ml was injected on a S6i 3.2/300 (Cytiva, Munich, Germany). Peak fractions were analyzed by SDS PAGE and fraction B10 was directly used for cryo-EM grid preparation.

4.2.9.2 Preparation of RNA bound CstF complex by in-batch cross-linking with BS3

To reconstitute CstF complex with RNA for cryo-EM studies, the *CstF01* RNA was used (Table 13). SEC purified complex at a concentration of 0.5-0.6 µM, corresponding to a concentration of 0.2-0.25 mg/ml, was incubated with 2 µM *CstF01* RNA for 1 h at 4 °C. Directly before grid

preparation, RNA reconstituted CstF was cross-linked in batch with 2 mM. Therefore, 20 µl of RNA bound complex at a concentration of 0.5-0.6 µM was incubated with 2 mM BS3 for 5 min at 30 °C. The cross-linking reaction was quenched with 25 mM Tris/HCl, pH 7.5 and the sample immediately used for cryo-EM grid preparation.

4.2.9.3 Preparation of the CstF1-CstF3 subcomplex by density ultracentrifugation

For cryo-EM studies of CstF1-CstF3 subcomplex, purified protein complex was cross-linked with 2 mM BS3 for 5 minutes at 30 °C and quenched with 25 mM Tris/HCl, pH 7.5. The cross-linked sample was immediately used for density gradient ultracentrifugation with a 5-25% sucrose gradient as described in 4.2.7.3. 200 µl of protein sample was carefully loaded on top of the gradient and samples were then centrifuged in a SW-40Ti overnight. CstF1-CstF3 complex always eluted in a single peak from the gradient. Protein concentrations of the peak fractions were determined using a NanoPhotometer and samples were analyzed by SDS PAGE.

Fractions of the gradient containing cross-linked CstF1-CstF3 subcomplex were pooled, concentrated and loaded on an analytical S6i 3.2/300 to exchange the buffer from the sucrose gradient into buffer SEC-4, which was used for cryo-EM studies (Table 14). Eluting protein complex was collected in 100 µl fractions and fraction B10 was used for cryo-EM sample preparation.

4.2.10 Transmission Electron Microscopy (TEM) and single particle analysis of the CstF complex

4.2.10.1 Negative stain sample preparation of the full length CstF complex

Negative staining is a transmission electron microscopy technique. Protein samples are adsorbed to a copper support grid covered with a thin continuous carbon layer. Adsorbed biological samples are then embedded in an amorphous layer of heavy metal salt, in this case uranyl acetate. The heavy metal salt strongly scatters electrons, producing a dark background, while macromolecules scatter electrons only weakly and appear as bright areas. Negative staining can be used to evaluate the sample homogeneity and concentration, quantity and quality of particles and protein complex stability and distribution on the grid.

To prepare negative stain samples, homemade copper grids coated with a thin carbon support film were used. In order to create a hydrophilic surface, grids were glow discharged for 30 seconds in a GloQube Plus Glow discharger (Quantum Design GmbH, Darmstadt, Germany). 5 µl of a protein sample at 100nM was then applied to the grid and incubated for 1 minute to allow adsorption. Excess sample was removed by blotting with a Whatman filter and the grid

was washed three times by touching a drop of 5 µl water and immediate blotting, followed by a 5 µl drop of 1% uranyl acetate stain. The stain was usually left on the grid for 30 seconds before blotting and the stained grid was dried in open air. Prepared negative stain grids could either be stored for later screening or immediately imaged on a Titan Halo (FEI) operating at 300 kV. Images were usually recorded with a pixel size of 1.85 Å/pix with a Falcon 3EC (ThermoFisher Scientific) direct electron detector.

4.2.10.2 Cryogenic grid preparation

Samples used for cryo-EM studies were purified as described in paragraph 4.2.7.1 and plunge frozen immediately after analytical SEC using a FEI Vitrobot Mark IV (ThermoFisher Scientific) equilibrated to 95% humidity and 4 °C. Holey carbon grids (Quantifoil R2/1 on Cu 200 mesh) were glow discharged for 10 s in a GloQube Plus Glow discharger (Quantum Design GmbH, Darmstadt, Germany). Directly before plunging, 0.04% detergent, in this case β-Octyl glucoside (β-OG), was added to the protein sample. 4 µl of each sample was applied on the grid, incubated for 15-20 seconds, blotted for 4 seconds with a blot force of 3.5 and plunged into liquid ethane/propane (30:70). Grids were then transferred to a cryo grid box and stored in liquid nitrogen. Screening was done on a Talos Arctica (FEI) TEM operating at 200 kV and datasets were collected with a Falcon 3EC (ThermoFisher Scientific) direct electron detector at a pixel size of 1.99 Å/pix using the software EPU (FEI).

4.2.10.3 Cryo-EM data collection of the CstF complex prepared by GraFix

For data collection of full-length CstF prepared as described in paragraph 4.2.9.1, roughly 2100 movies were collected in counting mode on a Titan Krios TEM operating at 300 kV, equipped with an energy filter and a K2 direct electron detector at a magnification of 105 kx, which corresponds to a pixel size of 1.34 $\text{Å}/\text{pix}$. A total dose of 38.88 $\text{e}^{-}/\text{Å}^2$ was applied to each movie containing 43 single frames and defocus values ranged between -1.5 to -3.5 micron. After correcting movies for beam-induced sample motion using MotionCor2, single frame micrographs were imported into Relion and the CTF was estimated with CTFFIND4. CTF-corrected micrographs were then pre-selected manually based on the estimated CTF resolution, so that micrographs with an estimated resolution worse than 5 Å were discarded, leaving 2085 micrographs. Particle coordinates from template-free particle picking by Gautomatch were imported into Relion and particles were extracted with a box sizes of 192 pix. After several rounds of 2D classification, 125 000 particles were left.

4.2.10.4 Cryo-EM data collection of a BS3 cross-linked CstF1-CstF2-CstF3

The dataset described in paragraph 2.4.4 was collected on a CstF1-CstF2-CstF3 complex, prepared as described in Material and Methods 4.2.9.2. Cryo-EM data were collected at a pixel size of 1.05 $\text{Å}/\text{pix}$ on a Titan Krios TEM operating at 300 kV, equipped with an energy filter and a K2 direct electron detector. In total, 4474 movies were collected, each consisting of 50 frames with a total dose of 72.29 $\text{e}^{-}/\text{Å}^2$. Defocus values ranged from -1.5 to 3 micron. MotionCor2 was applied to correct movie frames for beam-induced sample motion and align them to single frame micrographs, which were then used for template-free particle picking with Gautomatch. In parallel, micrographs were imported into Relion, where CTF estimation was performed using CTFFIND4. A total number of around 850000 particle coordinates were picked, extracted within Relion and used for several rounds of 2D classification, which led to a final stack of around 596000 particles.

In parallel, motion corrected micrographs were imported into CryoSparc and CTFFIND4 was used for CTF estimation as well. Particles were picked using the Topaz extract (Bepler et al., 2020) implementation with a pretrained model (ResNet16). Around 1096630 particles were picked from all micrographs, extracted and after several rounds of 2D classification, a subset of 70800 particles was used to train a Topaz model on the CstF complex for improved particle picking. With this model, around 1747000 particles were picked, extracted and used for 2D classification. After several rounds of 2D classification, a particle stack of 1500000 particles was left. Using the 1500000 particles, 3D ab initio reconstructions were generated and the 3D ab inito model with best resolution and highest particle content was used for 3D classification.

4.2.10.5 Cryo-EM data collection and analysis of CstF1-CstF3 complex

Final datasets resulting in the 5.3 and 6.5 Å resolution reconstructions described in 2.5.2 were collected on a sample that was purified as described in paragraph 4.2.7.1. Data collection was done on a Titan Krios operating at 300 kV equipped with a Gatan K3 (Gatan, Pleasanton, USA) direct detection camera and an energy filter (DQE 5-40 e^{-}/px/s; sensor pixel size 5 μm). An overview of the two final Krios-datasets collected on CstF1-CstF3 subcomplex are summarized in table 26.

Table 26: Krios datasets of CstF1-CstF3 subcomplex

Dataset	*Number of movies*	*Total dose / $[e^- / Å^2]$*	*Pixel size I $[Å]$*	*Generated maps/$[Å]$*	*Number of particles*
CstF1-CstF3 Krios I	9253	77.9	0.8512	6.5 5.3	31 327 105203
CstF1-CstF3 Krios II	14636	64.2	0.8512	3.43 HAT	90 822

4.2.10.5.1 Processing of the CstF1-CstF3 dataset I

Data were collected over several days resulting in 9253 movies (each 40 frames) with a magnified pixel size of 0.8512 Å/pix and a total dose of 77.9 $e^-/Å^2$. Before importing into a data processing software, the 40 frames in each movie were motion corrected and aligned to give a single summed micrograph, using Focus (Biyani, Righetto et al. 2017) software pipeline and MotionCor2 (Zheng, Palovcak et al. 2017). These micrographs were imported into Relion (Scheres 2012) and CTF estimation was performed using CTFFIND4 (Rohou and Grigorieff 2015). A summarized processing scheme is depicted in figure 99.

Initially, particles were picked using the template free 'blob picker' function in CryoSparc (Punjani, Rubinstein et al. 2017) and a subset of 350 000 particles was extracted and used for several rounds of 2D classification until a subset of particles representing several views of the protein complex was generated to be used as templates for particle picking with Gautomatch and training of the Topaz model (Bepler, Kelley et al. 2020) in CryoSparc. The trained model was then used to pick on all micrographs, which resulted in 2 250 000 particle coordinates. Particles were extracted with a box size of 384 pix and subjected to several rounds of 2D classification in CryoSparc until a final particle stack of 1 975 000 particles was left (Figure 86). In parallel, template-based particle picking was performed with Gautomatch using the 2D templates generated in CryoSparc. Picked coordinates were imported into Relion and corresponding particles were extracted with a 368 pix box size and re-scaled to 96 pix. Extracted particles were subjected for several rounds of 2D classification resulting in a final stack of 598 000 particles. Based on this particle stack, further 2D classification was performed, where only particles were selected, that showed clear density for the CstF3 HAT domain and the WD40 propeller of CstF1. Six 3D initial models were reconstructed in Relion from a subset of 145 000 particles and the model containing most of the particles and highest resolution was used as input for 3D classification into four classes. Only classes, that fit size and shape wise to the overall structure of the CstF1-CstF3 complex, were kept and subjected to another round of 3D classification. After iterative clean-up in 3D classification, the best resolved class was used as input for the last round of 3D classification using unbinned data and a tight mask generated in Relion. The last run delivered a class with densities for WD40s

and HAT dimer containing 31 327 particles, which was then refined and postprocessed to 6.5 Å. Final resolution was calculated using Relion gold standard FSC weighting and the global resolution corresponding to the 0.143 FSC cutoff is 6.6 Å

To avoid user bias by extensive reduction of particles in 2D classification, the particle stack containing 598 000 particles was used for another round of 3D classification using the refined model as input. The best resolved output class contained to 17.3% of the particles, was unbinned, refined and postprocessed to a final resolution of 5.3 Å. Final resolution was calculated using Relion gold standard FSC weighting and the global resolution corresponding to the 0.143 FSC cutoff is 5.26 Å.

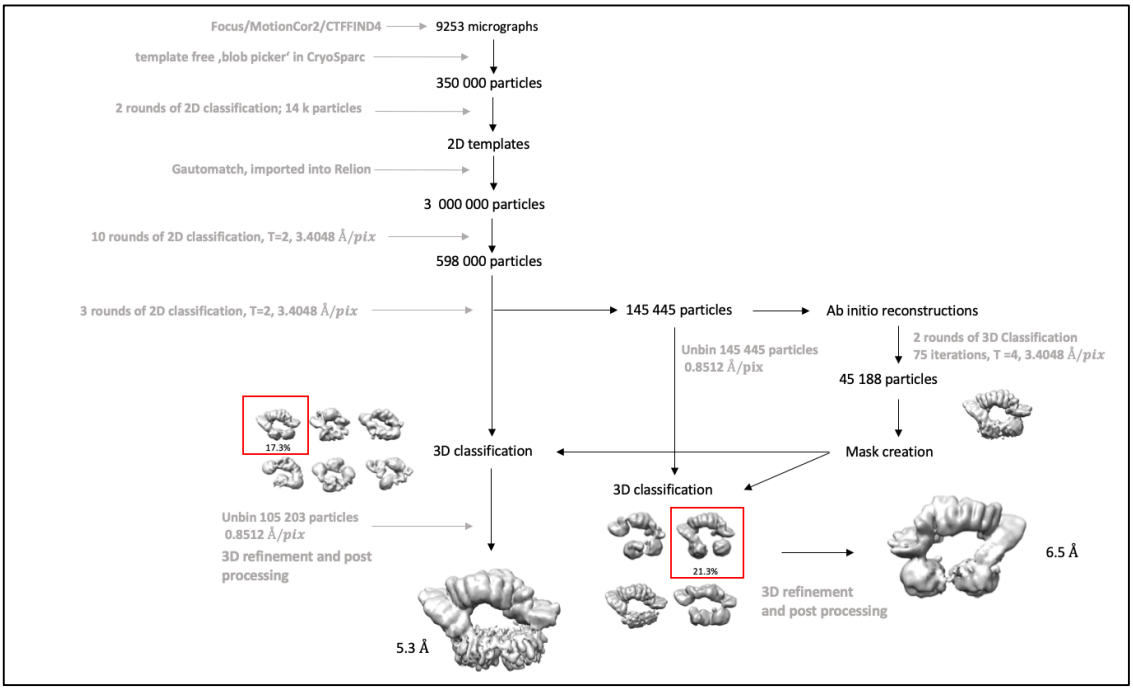


Figure 99: Processing scheme of the first CstF1-CstF3 Krios dataset which led to two 3D reconstructions at medium resolution.

4.2.10.5.2 Processing of the CstF1-CstF3 dataset II

Besides the first dataset discussed in the text above, another Krios dataset was collected on the same batch of sample of CstF1-CstF3, identically plunged and prepared. Data collection was set up over several days resulting in 14636 movie frames with a pixel size of 0.8512 Å/pix and a total dose of 64.2 $e^-/Å^2$. A graphical summary with the corresponding processing scheme is depicted in figure 100.

Directly after being collected, all 40 frames were motion corrected and aligned to a single frame by Focus software. Resulting single frame micrographs were imported into CryoSparc, where CTF estimation was performed using CTFFIND4. Particles were picked on a subset of 4000

micrographs using the TOPAZ extract implementation within CryoSparc, and extracted with a box size of 384 pix. After several rounds of 2D classification, selecting for clear density for HAT dimer and CstF1 WD40s connected to it, a few classes were chosen, representing different views of CstF1-CstF3 complex and containing a pool of around 10 000 particles. This selection was used as input to train the TOPAZ model within CryoSparc on a subset of 4000 micrographs. The trained model was then used to pick particles from the whole micrograph stack, resulting in 1 750 903 particles, which were again extracted with a box size of 384 pix, rescaled to 96 pix and used for 2D classification. Classes were kept only, where a clear density for both of the complex forming proteins was visible. After several rounds of clean-up in 2D, resulting 559 000 particles were used to generate three initial models in CryoSparc. Two models with highest particle content were used as input for 3D classification into four classes, which yielded in separation of HAT dimer alone from HAT domain with some additional density for CstF1 WD40 propellers (Figure 100).

Particles containing HAT domain alone, were processed separately and subjected to another round of 3D classification, where only classes that deliver secondary structure features and high-resolution estimation better than 5 Å were kept. Those particles were fed into 3D refinement with application of a tight mask created in CryoSparc with a treshold of 0.36, a dilation Radius of 3 pix and a soft padding width of 3 pi45_Cl. Refinement output was further iteratively improved by global and local CTF refinement and resulting 3D reconstruction was finally sharpened and postprocessed using DeepEMhancer (Sanchez-Garcia, Gomez-Blanco et al. 2021) to a final resolution of 3.43 Å (Figure 100). Final resolution was calculated using the CryoSparc gold standard FSC weighting and the global resolution corresponding to the 0.143 FSC cutoff was 3.4 Å.

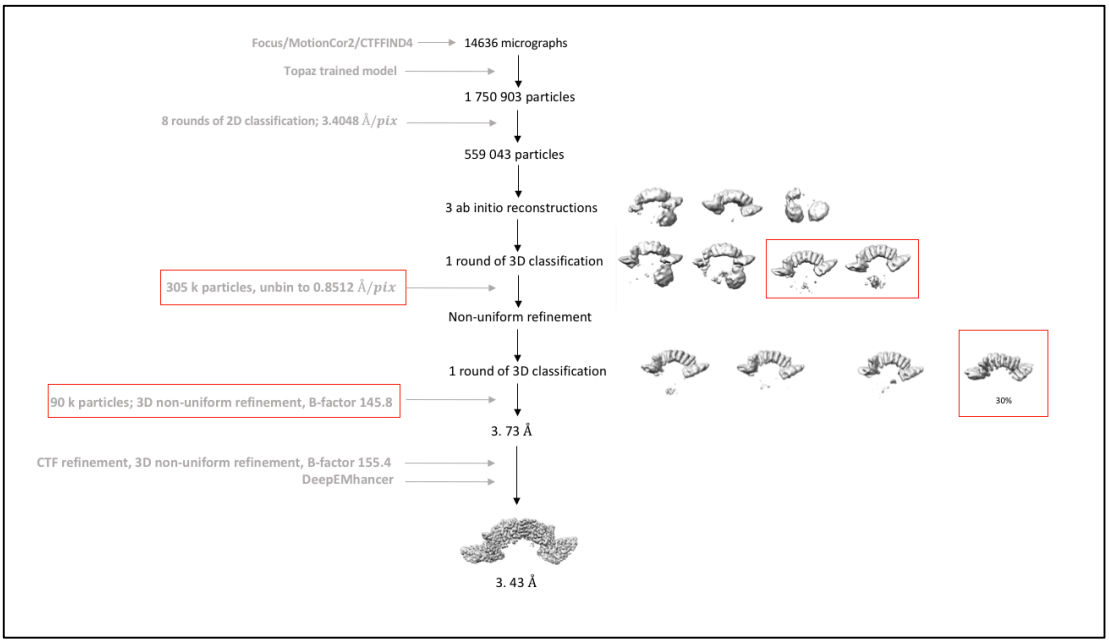


Figure 100. Processing scheme of the second Krios dataset of the CstF1-CstF3 subcomplex resulting in a high-resolution 3D reconstruction of the CstF3 HAT dimer.

Based on the 559000-particle stack from 2D classification, further rounds of 2D classification were performed only selecting for classes that showed clear density for the HAT domain and the WD40 propellers until a subset of 16 000 particles was selected for four ab-inito reconstructions (Figure 101). All four initial models were subjected to the first round of 3D classification in CryoSparc using the 559000-particle stack obtained by 2D classification. After selecting the classes showing initial secondary structure features for the HAT domain and clear density for one or both WD40 propeller, a second round of 3D classification was started using remaining 327 000 particles. Final 3D classes still showed heterogeneous sub-conformations of CstF1-CstF3 subcomplex. Due to low particle content in all classes, 3D refinement did not improve the resolution. Each class contained between 26 000 and 77 000 particles (Figure 101) and medium resolution calculated by CryoSparc was 8-10 Å.

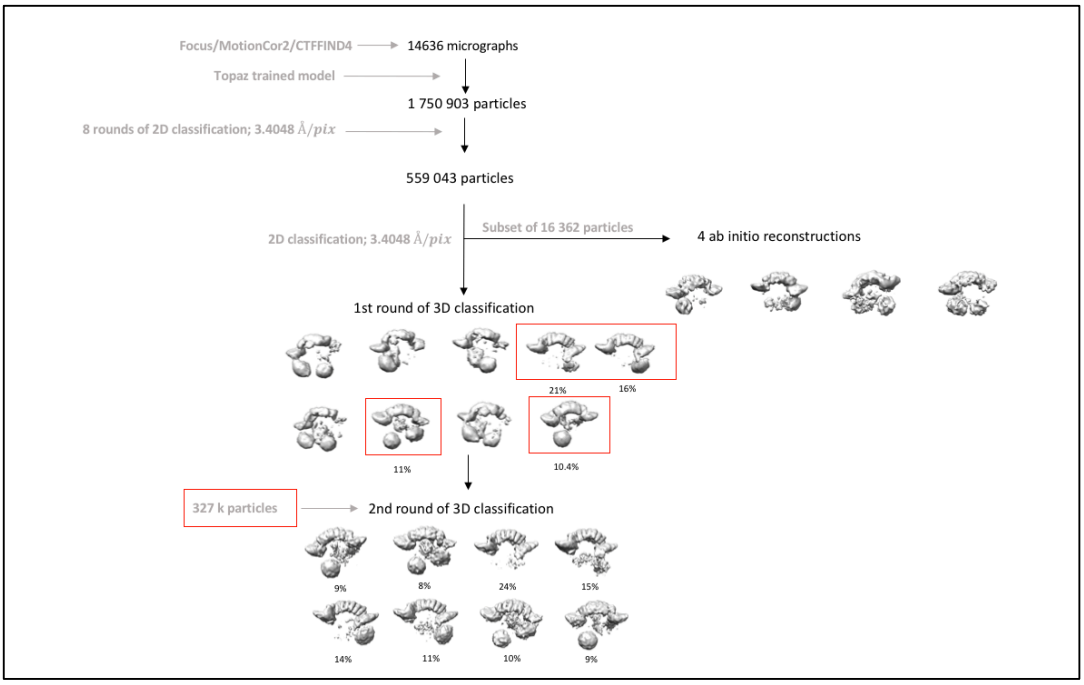


Figure 101. Processing scheme of the second Krios dataset of CstF1-CstF3 subcomplex resulting 3D reconstructions of the complex in different conformations.

Appendix

RNA	linker /[nt]	K_D/[μM]
CstF01	4	0.151
CstF12	/	0.592
CstF13	control	1.9
CstF14	2	0.643
CstF15	4	0.571
CstF16	6	0.703
CstF17	8	0.793
CstF18	16	1.4

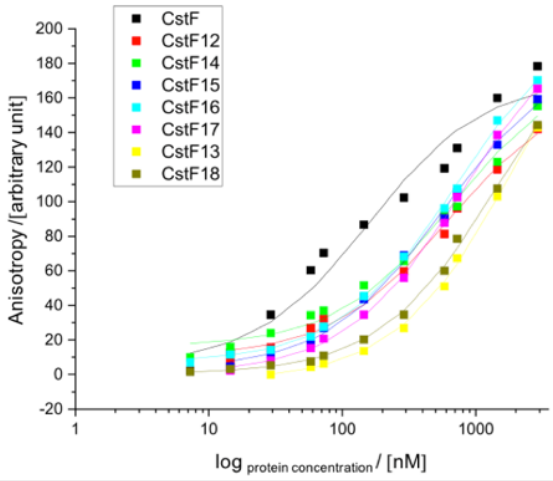


Figure 102. Determination of linker preference between G/U-rich downstream elements of human CstF2-CstF3 complex by FA measurements. RNAs were designed based on the G/U-rich binding motifs of *CstF01* RNA with a repetitive AC-linker in between spanning from two to 16 nucleotides. Measurements were repeated three times each resulting in K_D values listed in the table and depicted in the graph on the right. The graph shows anisotropy plotted in dependency of logarithmic protein concentration.

Abbreviations

3C	HRV-3C-Protease
6-FAM	6-Carboxyfluorescin
A	Adenine
Amp	Ampicillin resistance
APA	Alternative polyadenylation
BS	Branch site
BS3	bis(sulfosuccinimidyl)suberate
C	Cytosine
CF II_m	Cleavage Factor II
CF I_m	Cleavage Factor I
CID	Pol II interacting domain
CPA	Cleavage and polyadenylation
CPSF	Cleavage and Polyadenylation Specificity Factor
CstF	Cleavage Stimulation Factor
CTD	C-terminal domain
CV	Column volumes
CV	column volumes
CyPSF	cytoplasmic CPSF
DNA	deoxyribonucleic acid
DSE	downstream element
EC	elongation complex
EJC	Exon Junction Complex
EM	Electron Microscopy
FA	Fluorescence Anisotropy
fw	forward
G	Guanine
GA	Glutaraldehyde
GraFix	Gradient Fixation
GS	Glycine-Serine linker
GTF	general transcription factors
HAT	Half a TPR
Hi5	High five cells
IPTG	isopropyl-β-d-thiogalactopyranoside
ITC	Isothermal Titration Calorimetry

K_D	dissociation constant
LIC	Ligation independent cloning
M	Molar
$MgCl_2$	Magnesium Chloride
MLLE	mademoiselle domain
mRNP	matured ribonucleoprotein
MS	Mass Spectrometry
NaCl	Sodium Chloride
NaCl	Sodium chloride
ncPAP	non-canonical poly(A) polymerases
NGD	No-go decay
NMD	Nonsense-mediated decay
NMR	nuclear magnetic resonance
NSD	non-stop decay
NSL	Nuclear localization signal
NTD	N-terminal domain
NTD	Nucleotidyl transferase domain
PAM2	poly(A) interacting motif 2
PAP	Poly(A) polymerase
PAS	Poly(A) signals
PIC	pre-initiation process
PIPKIα	phosphatidyl inositol phosphate kinase Iα
PRMT	protein arginine methyltransferase
PRR	proline rich region
RBD	RNA Binding Domain
rev	reverse
RFM	Rossmann fold methyltransferase
RNA	ribonucleic acid
RNA Pol II	RNA Polymerase II
RNP	ribonucleoprotein particle
RRM	RNA Recognition Motif
S6i	Superose 6 increase
SDS PAGE	Sodium dodecylsulfate polyacrylamide gel electrophoresis
SEC	Size Exclusion Chromatography
SELEX	Systematic Evolution of Ligands by Exponential enrichment
SPA	Single particle analysis
SS	Splice site

STAR-PAP	Speckle Targeted PIPKα Regulated Poly(A) Polymerase
T	Thymine
TBE	Tris-Borate-EDTA
TEM	Transmission Electron Microscope
TEV	Tobacco Etch Virus Protease
TREX	Transcription Export
TSS	Transcription start site
U	Uracile
USE	upstream element
Wt	wild type
XL	cross-linking
ZF	zinc finger
β-OG	Octyl-beta-Glucoside

List of tables

List of figures

List of figures

List of figures

Acknowledgements

This PhD project would not have been possible without the support and contribution of many people. I firstly would like to gratefully thank my doctoral supervisor Prof. Dr. Elena Conti for her expertise guidance, endless support and encouragement during the past years. Deepest gratitude for offering me the opportunity to work on this project, for skillful supervision and for always having an open office door and open ear for everything I was dealing with during my PhD time.

Besides that, I would also like to thank the other members of my TAC committee, Prof. Dr. Foerstemann and Dr. Andreas Bracher for many useful scientific discussions and input driving the project forward and made me produce results in the present form.

A very very special thanks goes to my supervisor Dr. Christian Benda, who initially started this project and who was the most supporting and encouraging person within the last years. He generously took time out of his schedules to participate in my project and without his countless support, suggestions and unwavering believe in me and this project, I would not have grown professional. Thanks for especially motivating me again at the very end of the thesis to squeeze out the best results of this project and my data.

I gratefully recognize the contribution of MPIB cryo-EM Facility, Dr. Daniel Bollschweiler and Dr. Tillman Schaefer, who offered endless support and technical and knowhow regarding cryo-EM data screening and data collection. Thank you so much for all those big Krios datasets and for keeping all the microscopes running and in parallel providing us user tutorials and continuous guidance, so that everyone of us is not afraid to touch a microscope anymore. Regarding cryo-EM, I also have to thank Dr. Rajan Prabu for being our IT-specialist and setting up and maintaining all our data processing computing infrastructure. I was so thankful endless support you provided, whenever I encountered a problem at 10 am in the evening, when my limited computing skills needed your professional assistance. Another big thanks go to MPIB Core Facility and to Dr. Nagarjuna Nagaraj, for providing a perfect scientific infrastructure and technical support for various biochemical requests and MS-data analysis.

Very big thanks go to several members of the Conti lab, who directly or indirectly contributed to this work. Thanks to Dr. Claire Basquin for all those hundreds of FA measurements we did within the past years and dealing with my special wishes about graphs and figures. Thanks to Ingmar for his expertise and knowledge and supportive presence in the background. A big thanks goes to Peter for keeping all the Aektas running and spending some time in the cold

room with me, whenever there was again a problem with one of the Aektas. Thank you, Judith, Petra and Ulli, who contribute so much in the smooth way the department runs. Finally, thanks to Marc for keeping the insect cell facility running and providing enough cells for all my protein expressions.

And one of the biggest thanks is for all Conti people – Postdocs, PhDs and technicians – for creating this unique atmosphere in the lab, for all your input and questions during 'whatever kind of' meetings, for your help and assistance once I needed it. Thanks to Jana, for being my pink-purple-girly bay mate, who I could always talk to (and sometimes did too much) and share sweets with. Thanks Mahesh, for being such a close friend, when I needed one- for all our dinners, endless chats and running sessions and whatever we have been through the last years. I also want to thank Achim for our 'deep talks' while harvesting cells or sometimes even with a beer on the balcony. Thanks to Lukas and Felix for many exhausting and funny bouldering sessions and Felix especially for being part of the poly(A) team.
The biggest thank you goes to Iuliia, firstly for your sharing your scientific experience with me, for endless input and ideas. Thanks for all our coffee breaks or walks catching fresh air and discussing scientific problems. Secondly for proofreading this thesis and illustrating these amazing figures. But most importantly, for being one of my closest friends I can always talk to and rely on!

There are also several people outside Conti lab, I have to thank for their continuous support and patience within the last years. My huge amazing and unique girls crew, who creates this stable network I can rely on every second of my life! Although I was often very busy the last years, I know that I can totally count on you and you are the family I chose to surround myself with. For sure, thanks to the most supportive and understanding partner, a girl can wish for! Thanks for everything within the last years to my whole family – it was not always easy, but we managed, all together! And finally, the biggest contributions to everything in my life comes from my closest person – my sister! Thanks for being my twin by heart, my first person to talk to, sharing my big passion with me, being always there in my darkest moments and especially thanks for moving back to Munich for me! Having a sister like you is priceless…

References

Abaza, I. and F. Gebauer (2008). "Trading translation with RNA-binding proteins." RNA **14**(3): 404-409.
Adam, S. A., T. Nakagawa, M. S. Swanson, T. K. Woodruff and G. Dreyfuss (1986). "mRNA polyadenylate-binding protein: gene isolation and sequencing and identification of a ribonucleoprotein consensus sequence." Mol Cell Biol **6**(8): 2932-2943.
Aguilera, A. (2005). "Cotranscriptional mRNP assembly: from the DNA to the nuclear pore." Curr Opin Cell Biol **17**(3): 242-250.
Amrani, N., M. Minet, F. Wyers, M. E. Dufour, L. P. Aggerbeck and F. Lacroute (1997). "PCF11 encodes a third protein component of yeast cleavage and polyadenylation factor I." Mol Cell Biol **17**(3): 1102-1109.
Andersen, P. K., T. H. Jensen and S. Lykke-Andersen (2013). "Making ends meet: coordination between RNA 3'-end processing and transcription initiation." Wiley Interdiscip Rev RNA **4**(3): 233-246.
Andrade, M. A., C. Petosa, S. I. O'Donoghue, C. W. Muller and P. Bork (2001). "Comparison of ARM and HEAT protein repeats." J Mol Biol **309**(1): 1-18.
Apponi, L. H., S. W. Leung, K. R. Williams, S. R. Valentini, A. H. Corbett and G. K. Pavlath (2010). "Loss of nuclear poly(A)-binding protein 1 causes defects in myogenesis and mRNA biogenesis." Hum Mol Genet **19**(6): 1058-1065.
Araki, Y., S. Takahashi, T. Kobayashi, H. Kajiho, S. Hoshino and T. Katada (2001). "Ski7p G protein interacts with the exosome and the Ski complex for 3'-to-5' mRNA decay in yeast." EMBO J **20**(17): 4684-4693.
Aravind, L. (1999). "An evolutionary classification of the metallo-beta-lactamase fold proteins." In Silico Biol **1**(2): 69-91.
Aslanidis, C. and P. J. de Jong (1990). "Ligation-independent cloning of PCR products (LIC-PCR)." Nucleic Acids Res **18**(20): 6069-6074.
Auweter, S. D., F. C. Oberstrass and F. H. Allain (2006). "Sequence-specific binding of single-stranded RNA: is there a code for recognition?" Nucleic Acids Res **34**(17): 4943-4959.
Avis, J. M., F. H. Allain, P. W. Howe, G. Varani, K. Nagai and D. Neuhaus (1996). "Solution structure of the N-terminal RNP domain of U1A protein: the role of C-terminal residues in structure stability and RNA binding." J Mol Biol **257**(2): 398-411.
Badis, G., C. Saveanu, M. Fromont-Racine and A. Jacquier (2004). "Targeted mRNA degradation by deadenylation-independent decapping." Mol Cell **15**(1): 5-15.
Baer, B. W. and R. D. Kornberg (1983). "The protein responsible for the repeating structure of cytoplasmic poly(A)-ribonucleoprotein." J Cell Biol **96**(3): 717-721.
Bai, Y., T. C. Auperin, C. Y. Chou, G. G. Chang, J. L. Manley and L. Tong (2007). "Crystal structure of murine CstF-77: dimeric association and implications for polyadenylation of mRNA precursors." Mol Cell **25**(6): 863-875.
Bai, Y., T. C. Auperin and L. Tong (2007). "The use of in situ proteolysis in the crystallization of murine CstF-77." Acta Crystallogr Sect F Struct Biol Cryst Commun **63**(Pt 2): 135-138.
Balbo, P. B. and A. Bohm (2007). "Mechanism of poly(A) polymerase: structure of the enzyme-MgATP-RNA ternary complex and kinetic analysis." Structure **15**(9): 1117-1131.
Balbo, P. B., J. Toth and A. Bohm (2007). "X-ray crystallographic and steady state fluorescence characterization of the protein dynamics of yeast polyadenylate polymerase." J Mol Biol **366**(5): 1401-1415.
Barabino, S. M., W. Hubner, A. Jenny, L. Minvielle-Sebastia and W. Keller (1997). "The 30-kD subunit of mammalian cleavage and polyadenylation specificity factor and its yeast homolog are RNA-binding zinc finger proteins." Genes Dev **11**(13): 1703-1716.
Barabino, S. M., M. Ohnacker and W. Keller (2000). "Distinct roles of two Yth1p domains in 3'-end cleavage and polyadenylation of yeast pre-mRNAs." EMBO J **19**(14): 3778-3787.
Bard, J., A. M. Zhelkovsky, S. Helmling, T. N. Earnest, C. L. Moore and A. Bohm (2000). "Structure of yeast poly(A) polymerase alone and in complex with 3'-dATP." Science **289**(5483): 1346-1349.

Barilla, D., B. A. Lee and N. J. Proudfoot (2001). "Cleavage/polyadenylation factor IA associates with the carboxyl-terminal domain of RNA polymerase II in Saccharomyces cerevisiae." Proc Natl Acad Sci U S A **98**(2): 445-450.
Barlow, C. A., R. S. Laishram and R. A. Anderson (2010). "Nuclear phosphoinositides: a signaling enigma wrapped in a compartmental conundrum." Trends Cell Biol **20**(1): 25-35.
Barnard, D. C., K. Ryan, J. L. Manley and J. D. Richter (2004). "Symplekin and xGLD-2 are required for CPEB-mediated cytoplasmic polyadenylation." Cell **119**(5): 641-651.
Bartkowiak, B., A. L. Mackellar and A. L. Greenleaf (2011). "Updating the CTD Story: From Tail to Epic." Genet Res Int **2011**: 623718.
Baumann, M., J. Pontiller and W. Ernst (2010). "Structure and basal transcription complex of RNA polymerase II core promoters in the mammalian genome: an overview." Mol Biotechnol **45**(3): 241-247.
Beaudoing, E., S. Freier, J. R. Wyatt, J. M. Claverie and D. Gautheret (2000). "Patterns of variant polyadenylation signal usage in human genes." Genome Res **10**(7): 1001-1010.
Bebrone, C. (2007). "Metallo-beta-lactamases (classification, activity, genetic organization, structure, zinc coordination) and their superfamily." Biochem Pharmacol **74**(12): 1686-1701.
Beckmann, B. M., A. Castello and J. Medenbach (2016). "The expanding universe of ribonucleoproteins: of novel RNA-binding proteins and unconventional interactions." Pflugers Arch **468**(6): 1029-1040.
Belloc, E. and R. Mendez (2008). "A deadenylation negative feedback mechanism governs meiotic metaphase arrest." Nature **452**(7190): 1017-1021.
Bentley, D. L. (2005). "Rules of engagement: co-transcriptional recruitment of pre-mRNA processing factors." Curr Opin Cell Biol **17**(3): 251-256.
Bepler, T., K. Kelley, A. J. Noble and B. Berger (2020). "Topaz-Denoise: general deep denoising models for cryoEM and cryoET." Nat Commun **11**(1): 5208.
Berger, I., D. J. Fitzgerald and T. J. Richmond (2004). "Baculovirus expression system for heterologous multiprotein complexes." Nat Biotechnol **22**(12): 1583-1587.
Berget, S. M., C. Moore and P. A. Sharp (1977). "Spliced segments at the 5' terminus of adenovirus 2 late mRNA." Proc Natl Acad Sci U S A **74**(8): 3171-3175.
Beyer, K., T. Dandekar and W. Keller (1997). "RNA ligands selected by cleavage stimulation factor contain distinct sequence motifs that function as downstream elements in 3'-end processing of pre-mRNA." J Biol Chem **272**(42): 26769-26779.
Bideshi, D. K., M. V. Demattei, F. Rouleux-Bonnin, K. Stasiak, Y. Tan, S. Bigot, Y. Bigot and B. A. Federici (2006). "Genomic sequence of Spodoptera frugiperda Ascovirus 1a, an enveloped, double-stranded DNA insect virus that manipulates apoptosis for viral reproduction." J Virol **80**(23): 11791-11805.
Bienroth, S., W. Keller and E. Wahle (1993). "Assembly of a processive messenger RNA polyadenylation complex." EMBO J **12**(2): 585-594.
Bienroth, S., E. Wahle, C. Suter-Crazzolara and W. Keller (1991). "Purification of the cleavage and polyadenylation factor involved in the 3'-processing of messenger RNA precursors." J Biol Chem **266**(29): 19768-19776.
Bilger, A., C. A. Fox, E. Wahle and M. Wickens (1994). "Nuclear polyadenylation factors recognize cytoplasmic polyadenylation elements." Genes Dev **8**(9): 1106-1116.
Biyani, N., R. D. Righetto, R. McLeod, D. Caujolle-Bert, D. Castano-Diez, K. N. Goldie and H. Stahlberg (2017). "Focus: The interface between data collection and data processing in cryo-EM." J Struct Biol **198**(2): 124-133.
Blackwell, E., X. Zhang and S. Ceman (2010). "Arginines of the RGG box regulate FMRP association with polyribosomes and mRNA." Hum Mol Genet **19**(7): 1314-1323.
Blobel, G. (1973). "A protein of molecular weight 78,000 bound to the polyadenylate region of eukaryotic messenger RNAs." Proc Natl Acad Sci U S A **70**(3): 924-928.
Boisvert, F. M., C. A. Chenard and S. Richard (2005). "Protein interfaces in signaling regulated by arginine methylation." Sci STKE **2005**(271): re2.
Borman, A. M., Y. M. Michel and K. M. Kean (2000). "Biochemical characterisation of cap-poly(A) synergy in rabbit reticulocyte lysates: the eIF4G-PABP interaction increases the functional affinity of eIF4E for the capped mRNA 5'-end." Nucleic Acids Res **28**(21): 4068-4075.

Braun, J. E., V. Truffault, A. Boland, E. Huntzinger, C. T. Chang, G. Haas, O. Weichenrieder, M. Coles and E. Izaurralde (2012). "A direct interaction between DCP1 and XRN1 couples mRNA decapping to 5' exonucleolytic degradation." Nat Struct Mol Biol **19**(12): 1324-1331.
Braun, K. A. and E. T. Young (2014). "Coupling mRNA synthesis and decay." Mol Cell Biol **34**(22): 4078-4087.
Brody, E. and J. Abelson (1985). "The "spliceosome": yeast pre-messenger RNA associates with a 40S complex in a splicing-dependent reaction." Science **228**(4702): 963-967.
Brown, K. M. and G. M. Gilmartin (2003). "A mechanism for the regulation of pre-mRNA 3' processing by human cleavage factor Im." Mol Cell **12**(6): 1467-1476.
Bunce, M. W., K. Bergendahl and R. A. Anderson (2006). "Nuclear PI(4,5)P(2): a new place for an old signal." Biochim Biophys Acta **1761**(5-6): 560-569.
Buratowski, S. (2003). "The CTD code." Nat Struct Biol **10**(9): 679-680.
Burd, C. G., E. L. Matunis and G. Dreyfuss (1991). "The multiple RNA-binding domains of the mRNA poly(A)-binding protein have different RNA-binding activities." Mol Cell Biol **11**(7): 3419-3424.
Callebaut, I., D. Moshous, J. P. Mornon and J. P. de Villartay (2002). "Metallo-beta-lactamase fold within nucleic acids processing enzymes: the beta-CASP family." Nucleic Acids Res **30**(16): 3592-3601.
Calvo, O. and J. L. Manley (2001). "Evolutionarily conserved interaction between CstF-64 and PC4 links transcription, polyadenylation, and termination." Molecular Cell **7**(5): 1013-1023.
Carninci, P., A. Sandelin, B. Lenhard, S. Katayama, K. Shimokawa, J. Ponjavic, C. A. Semple, M. S. Taylor, P. G. Engstrom, M. C. Frith, A. R. Forrest, W. B. Alkema, S. L. Tan, C. Plessy, R. Kodzius, T. Ravasi, T. Kasukawa, S. Fukuda, M. Kanamori-Katayama, Y. Kitazume, H. Kawaji, C. Kai, M. Nakamura, H. Konno, K. Nakano, S. Mottagui-Tabar, P. Arner, A. Chesi, S. Gustincich, F. Persichetti, H. Suzuki, S. M. Grimmond, C. A. Wells, V. Orlando, C. Wahlestedt, E. T. Liu, M. Harbers, J. Kawai, V. B. Bajic, D. A. Hume and Y. Hayashizaki (2006). "Genome-wide analysis of mammalian promoter architecture and evolution." Nat Genet **38**(6): 626-635.
Casanal, A., A. Kumar, C. H. Hill, A. D. Easter, P. Emsley, G. Degliesposti, Y. Gordiyenko, B. Santhanam, J. Wolf, K. Wiederhold, G. L. Dornan, M. Skehel, C. V. Robinson and L. A. Passmore (2017). "Architecture of eukaryotic mRNA 3'-end processing machinery." Science **358**(6366): 1056-1059.
Chan, S., E. A. Choi and Y. Shi (2011). "Pre-mRNA 3'-end processing complex assembly and function." Wiley Interdiscip Rev RNA **2**(3): 321-335.
Chan, S. L., I. Huppertz, C. Yao, L. Weng, J. J. Moresco, J. R. Yates, 3rd, J. Ule, J. L. Manley and Y. Shi (2014). "CPSF30 and Wdr33 directly bind to AAUAAA in mammalian mRNA 3' processing." Genes Dev **28**(21): 2370-2380.
Chanarat, S., C. Burkert-Kautzsch, D. M. Meinel and K. Strasser (2012). "Prp19C and TREX: interacting to promote transcription elongationand mRNA export." Transcription **3**(1): 8-12.
Chang, C. T., N. Bercovich, B. Loh, S. Jonas and E. Izaurralde (2014). "The activation of the decapping enzyme DCP2 by DCP1 occurs on the EDC4 scaffold and involves a conserved loop in DCP1." Nucleic Acids Res **42**(8): 5217-5233.
Chen, C. Y. and A. B. Shyu (2011). "Mechanisms of deadenylation-dependent decay." Wiley Interdiscip Rev RNA **2**(2): 167-183.
Chen, C. Y., N. Xu and A. B. Shyu (1995). "mRNA decay mediated by two distinct AU-rich elements from c-fos and granulocyte-macrophage colony-stimulating factor transcripts: different deadenylation kinetics and uncoupling from translation." Mol Cell Biol **15**(10): 5777-5788.
Chen, F., C. C. MacDonald and J. Wilusz (1995). "Cleavage site determinants in the mammalian polyadenylation signal." Nucleic Acids Res **23**(14): 2614-2620.
Chen, Z., Y. Li and R. M. Krug (1999). "Influenza A virus NS1 protein targets poly(A)-binding protein II of the cellular 3'-end processing machinery." EMBO J **18**(8): 2273-2283.
Chi, B., Q. Wang, G. Wu, M. Tan, L. Wang, M. Shi, X. Chang and H. Cheng (2013). "Aly and THO are required for assembly of the human TREX complex and association of TREX components with the spliced mRNA." Nucleic Acids Res **41**(2): 1294-1306.

Chong, P. A., R. M. Vernon and J. D. Forman-Kay (2018). "RGG/RG Motif Regions in RNA Binding and Phase Separation." J Mol Biol **430**(23): 4650-4665.
Chow, L. T., R. E. Gelinas, T. R. Broker and R. J. Roberts (1977). "An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA." Cell **12**(1): 1-8.
Christofori, G. and W. Keller (1988). "3' cleavage and polyadenylation of mRNA precursors in vitro requires a poly(A) polymerase, a cleavage factor, and a snRNP." Cell **54**(6): 875-889.
Clement, S. L. and J. Lykke-Andersen (2006). "No mercy for messages that mess with the ribosome." Nature Structural & Molecular Biology **13**(4): 299-301.
Clerici, M., M. Faini, R. Aebersold and M. Jinek (2017). "Structural insights into the assembly and polyA signal recognition mechanism of the human CPSF complex." Elife **6**.
Clerici, M., M. Faini, L. M. Muckenfuss, R. Aebersold and M. Jinek (2018). "Structural basis of AAUAAA polyadenylation signal recognition by the human CPSF complex." Nat Struct Mol Biol **25**(2): 135-138.
Clery, A., M. Blatter and F. H. Allain (2008). "RNA recognition motifs: boring? Not quite." Curr Opin Struct Biol **18**(3): 290-298.
Colgan, D. F. and J. L. Manley (1997). "Mechanism and regulation of mRNA polyadenylation." Genes Dev **11**(21): 2755-2766.
Colgan, D. F., K. G. Murthy, C. Prives and J. L. Manley (1996). "Cell-cycle related regulation of poly(A) polymerase by phosphorylation." Nature **384**(6606): 282-285.
Colgan, D. F., K. G. Murthy, W. Zhao, C. Prives and J. L. Manley (1998). "Inhibition of poly(A) polymerase requires p34cdc2/cyclin B phosphorylation of multiple consensus and non-consensus sites." EMBO J **17**(4): 1053-1062.
Coll, O., A. Villalba, G. Bussotti, C. Notredame and F. Gebauer (2010). "A novel, noncanonical mechanism of cytoplasmic polyadenylation operates in Drosophila embryogenesis." Genes Dev **24**(2): 129-134.
Connelly, S. and J. L. Manley (1988). "A functional mRNA polyadenylation signal is required for transcription termination by RNA polymerase II." Genes Dev **2**(4): 440-452.
Corley, S. M. and J. E. Gready (2008). "Identification of the RGG box motif in Shadoo: RNA-binding and signaling roles?" Bioinform Biol Insights **2**: 383-400.
Coseno, M., G. Martin, C. Berger, G. Gilmartin, W. Keller and S. Doublie (2008). "Crystal structure of the 25 kDa subunit of human cleavage factor Im." Nucleic Acids Res **36**(10): 3474-3483.
Cramer, P. (2004). "RNA polymerase II structure: from core to functional complexes." Curr Opin Genet Dev **14**(2): 218-226.
Cramer, P., D. A. Bushnell and R. D. Kornberg (2001). "Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution." Science **292**(5523): 1863-1876.
Cudmore, S., I. Reckmann and M. Way (1997). "Viral manipulations of the actin cytoskeleton." Trends Microbiol **5**(4): 142-148.
D'Souza, V. and M. F. Summers (2004). "Structural basis for packaging the dimeric genome of Moloney murine leukaemia virus." Nature **431**(7008): 586-590.
Danckwardt, S., M. W. Hentze and A. E. Kulozik (2008). "3 ' end mRNA processing: molecular mechanisms and implications for health and disease." Embo Journal **27**(3): 482-498.
Danckwardt, S., I. Kaufmann, M. Gentzel, K. U. Foerstner, A. S. Gantzert, N. H. Gehring, G. Neu-Yilik, P. Bork, W. Keller, M. Wilm, M. W. Hentze and A. E. Kulozik (2007). "Splicing factors stimulate polyadenylation via USEs at non-canonical 3' end formation signals." EMBO J **26**(11): 2658-2669.
Davies, J. F., 2nd, R. J. Almassy, Z. Hostomska, R. A. Ferre and Z. Hostomsky (1994). "2.3 A crystal structure of the catalytic domain of DNA polymerase beta." Cell **76**(6): 1123-1133.
de Vries, H., U. Ruegsegger, W. Hubner, A. Friedlein, H. Langen and W. Keller (2000). "Human pre-mRNA cleavage factor II(m) contains homologs of yeast proteins and bridges two other cleavage factors." EMBO J **19**(21): 5895-5904.
Deka, P., P. K. Rajan, J. M. Perez-Canadillas and G. Varani (2005). "Protein and RNA dynamics play key roles in determining the specific recognition of GU-rich polyadenylation regulatory elements by human Cstf-64 protein." J Mol Biol **347**(4): 719-733.

Derti, A., P. Garrett-Engele, K. D. Macisaac, R. C. Stevens, S. Sriram, R. Chen, C. A. Rohl, J. M. Johnson and T. Babak (2012). "A quantitative atlas of polyadenylation in five mammals." Genome Res **22**(6): 1173-1183.
Dettwiler, S., C. Aringhieri, S. Cardinale, W. Keller and S. M. Barabino (2004). "Distinct sequence motifs within the 68-kDa subunit of cleavage factor Im mediate RNA binding, protein-protein interactions, and subcellular localization." J Biol Chem **279**(34): 35788-35797.
Di Giammartino, D. C., K. Nishida and J. L. Manley (2011). "Mechanisms and consequences of alternative polyadenylation." Mol Cell **43**(6): 853-866.
Dickson, A. M. and J. Wilusz (2010). "Polyadenylation: alternative lifestyles of the A-rich (and famous?)." EMBO J **29**(9): 1473-1474.
Dickson, K. S., S. R. Thompson, N. K. Gray and M. Wickens (2001). "Poly(A) polymerase and the regulation of cytoplasmic polyadenylation." J Biol Chem **276**(45): 41810-41816.
Doidge, R., S. Mittal, A. Aslam and G. S. Winkler (2012). "Deadenylation of cytoplasmic mRNA by the mammalian Ccr4-Not complex." Biochem Soc Trans **40**(4): 896-901.
Doma, M. K. and R. Parker (2006). "Endonucleolytic cleavage of eukaryotic mRNAs with stalls in translation elongation." Nature **440**(7083): 561-564.
Dominguez, C., J. F. Fisette, B. Chabot and F. H. Allain (2010). "Structural basis of G-tract recognition and encaging by hnRNP F quasi-RRMs." Nat Struct Mol Biol **17**(7): 853-861.
Dominski, Z. (2007). "Nucleases of the metallo-beta-lactamase family and their role in DNA and RNA metabolism." Crit Rev Biochem Mol Biol **42**(2): 67-93.
Dominski, Z. and W. F. Marzluff (1999). "Formation of the 3' end of histone mRNA." Gene **239**(1): 1-14.
Dominski, Z. and W. F. Marzluff (2007). "Formation of the 3' end of histone mRNA: getting closer to the end." Gene **396**(2): 373-390.
Dominski, Z., X. C. Yang and W. F. Marzluff (2005). "The polyadenylation factor CPSF-73 is involved in histone-pre-mRNA processing." Cell **123**(1): 37-48.
Dominski, Z., X. C. Yang, M. Purdy, E. J. Wagner and W. F. Marzluff (2005). "A CPSF-73 homologue is required for cell cycle progression but not cell growth and interacts with a protein having features of CPSF-100." Mol Cell Biol **25**(4): 1489-1500.
Doughman, R. L., A. J. Firestone and R. A. Anderson (2003). "Phosphatidylinositol phosphate kinases put PI4,5P(2) in its place." J Membr Biol **194**(2): 77-89.
Dreyfuss, G., M. S. Swanson and S. Pinol-Roma (1988). "Heterogeneous nuclear ribonucleoprotein particles and the pathway of mRNA formation." Trends Biochem Sci **13**(3): 86-91.
Dunckley, T. and R. Parker (1999). "The DCP2 protein is required for mRNA decapping in Saccharomyces cerevisiae and contains a functional MutT motif." EMBO J **18**(19): 5411-5422.
Edmonds, M. and R. Abrams (1960). "Polynucleotide biosynthesis: formation of a sequence of adenylate units from adenosine triphosphate by an enzyme from thymus nuclei." J Biol Chem **235**: 1142-1149.
Egloff, S., M. Dienstbier and S. Murphy (2012). "Updating the RNA polymerase CTD code: adding gene-specific layers." Trends Genet **28**(7): 333-341.
Engel, C., S. Neyer and P. Cramer (2018). "Distinct Mechanisms of Transcription Initiation by RNA Polymerases I and II." Annual Review of Biophysics, Vol 47 **47**: 425-446.
Epshtein, V., C. J. Cardinale, A. E. Ruckenstein, S. Borukhov and E. Nudler (2007). "An allosteric path to transcription termination." Mol Cell **28**(6): 991-1001.
Erson-Bensan, A. E. and T. Can (2016). "Alternative Polyadenylation: Another Foe in Cancer." Mol Cancer Res **14**(6): 507-517.
Fitzgerald, M. and T. Shenk (1981). "The sequence 5'-AAUAAA-3'forms parts of the recognition site for polyadenylation of late SV40 mRNAs." Cell **24**(1): 251-260.
Fornerod, M. (2012). "RS and RGG repeats as primitive proteins at the transition between the RNA and RNP worlds." Nucleus **3**(1): 4-5.
Frendewey, D. and W. Keller (1985). "Stepwise assembly of a pre-mRNA splicing complex requires U-snRNPs and specific intron sequences." Cell **42**(1): 355-367.

Frischmeyer, P. A., A. van Hoof, K. O'Donnell, A. L. Guerrerio, R. Parker and H. C. Dietz (2002). "An mRNA surveillance mechanism that eliminates transcripts lacking termination codons." Science **295**(5563): 2258-2261.
Fuda, N. J., M. B. Ardehali and J. T. Lis (2009). "Defining mechanisms that regulate RNA polymerase II transcription in vivo." Nature **461**(7261): 186-192.
Ganem, C., F. Devaux, C. Torchet, C. Jacq, S. Quevillon-Cheruel, G. Labesse, C. Facca and G. Faye (2003). "Ssu72 is a phosphatase essential for transcription termination of snoRNAs and specific mRNAs in yeast." EMBO J **22**(7): 1588-1598.
Gavin, A. C., M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A. M. Michon, C. M. Cruciat, M. Remor, C. Hofert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M. A. Heurtier, R. R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer and G. Superti-Furga (2002). "Functional organization of the yeast proteome by systematic analysis of protein complexes." Nature **415**(6868): 141-147.
Ge, H., D. Zhou, S. Tong, Y. Gao, M. Teng and L. Niu (2008). "Crystal structure and possible dimerization of the single RRM of human PABPN1." Proteins **71**(3): 1539-1545.
Gerstberger, S., M. Hafner and T. Tuschl (2014). "A census of human RNA-binding proteins." Nat Rev Genet **15**(12): 829-845.
Ghazy, M. A., J. M. Gordon, S. D. Lee, B. N. Singh, A. Bohm, M. Hampsey and C. Moore (2012). "The interaction of Pcf11 and Clp1 is needed for mRNA 3'-end formation and is modulated by amino acids in the ATP-binding site." Nucleic Acids Res **40**(3): 1214-1225.
Ghazy, M. A., X. He, B. N. Singh, M. Hampsey and C. Moore (2009). "The essential N terminus of the Pta1 scaffold protein is required for snoRNA transcription termination and Ssu72 function but is dispensable for pre-mRNA 3'-end processing." Mol Cell Biol **29**(8): 2296-2307.
Ghosh, A. and C. D. Lima (2010). "Enzymology of RNA cap synthesis." Wiley Interdiscip Rev RNA **1**(1): 152-172.
Gieselmann, V., A. Polten, J. Kreysing and K. von Figura (1989). "Arylsulfatase A pseudodeficiency: loss of a polyadenylylation signal and N-glycosylation site." Proc Natl Acad Sci U S A **86**(23): 9436-9440.
Gil, A. and N. J. Proudfoot (1987). "Position-dependent sequence elements downstream of AAUAAA are required for efficient rabbit beta-globin mRNA 3' end formation." Cell **49**(3): 399-406.
Gilmartin, G. M., E. S. Fleming, J. Oetjen and B. R. Graveley (1995). "CPSF recognition of an HIV-1 mRNA 3'-processing enhancer: multiple sequence contacts involved in poly(A) site definition." Genes Dev **9**(1): 72-83.
Gilmartin, G. M. and J. R. Nevins (1989). "An ordered pathway of assembly of components required for polyadenylation site recognition and processing." Genes Dev **3**(12B): 2180-2190.
Gilmartin, G. M. and J. R. Nevins (1991). "Molecular analyses of two poly(A) site-processing factors that determine the recognition and efficiency of cleavage of the pre-mRNA." Mol Cell Biol **11**(5): 2432-2438.
Goldstrohm, A. C. and M. Wickens (2008). "Multifunctional deadenylase complexes diversify mRNA control." Nat Rev Mol Cell Biol **9**(4): 337-344.
Gordon, J. M., S. Shikov, J. N. Kuehner, M. Liriano, E. Lee, W. Stafford, M. B. Poulsen, C. Harrison, C. Moore and A. Bohm (2011). "Reconstitution of CF IA from overexpressed subunits reveals stoichiometry and provides insights into molecular topology." Biochemistry **50**(47): 10203-10214.
Graber, J. H., C. R. Cantor, S. C. Mohr and T. F. Smith (1999). "In silico detection of control signals: mRNA 3'-end-processing sequences in diverse species." Proc Natl Acad Sci U S A **96**(24): 14055-14060.
Gross, S. and C. Moore (2001). "Five subunits are required for reconstitution of the cleavage and polyadenylation activities of Saccharomyces cerevisiae cleavage factor I." Proc Natl Acad Sci U S A **98**(11): 6080-6085.

Gruber, A. J., R. Schmidt, A. R. Gruber, G. Martin, S. Ghosh, M. Belmadani, W. Keller and M. Zavolan (2016). "A comprehensive analysis of 3' end sequencing data sets reveals novel polyadenylation signals and the repressive role of heterogeneous ribonucleoprotein C on cleavage and polyadenylation." Genome Res **26**(8): 1145-1159.
Gu, M. and C. D. Lima (2005). "Processing the message: structural insights into capping and decapping mRNA." Curr Opin Struct Biol **15**(1): 99-106.
Guo, A., H. Gu, J. Zhou, D. Mulhern, Y. Wang, K. A. Lee, V. Yang, M. Aguiar, J. Kornhauser, X. Jia, J. Ren, S. A. Beausoleil, J. C. Silva, V. Vemulapalli, M. T. Bedford and M. J. Comb (2014). "Immunoaffinity enrichment and mass spectrometry analysis of protein methylation." Mol Cell Proteomics **13**(1): 372-387.
Haddad, R., F. Maurice, N. Viphakone, F. Voisinet-Hakil, S. Fribourg and L. Minvielle-Sebastia (2012). "An essential role for Clp1 in assembly of polyadenylation complex CF IA and Pol II transcription termination." Nucleic Acids Res **40**(3): 1226-1239.
Halbach, F., P. Reichelt, M. Rode and E. Conti (2013). "The yeast ski complex: crystal structure and RNA channeling to the exosome complex." Cell **154**(4): 814-826.
Hall-Pogar, T., S. Liang, L. K. Hague and C. S. Lutz (2007). "Specific trans-acting proteins interact with auxiliary RNA polyadenylation elements in the COX-2 3'-UTR." RNA **13**(7): 1103-1115.
Hanada, T., S. Weitzer, B. Mair, C. Bernreuther, B. J. Wainger, J. Ichida, R. Hanada, M. Orthofer, S. J. Cronin, V. Komnenovic, A. Minis, F. Sato, H. Mimata, A. Yoshimura, I. Tamir, J. Rainer, R. Kofler, A. Yaron, K. C. Eggan, C. J. Woolf, M. Glatzel, R. Herbst, J. Martinez and J. M. Penninger (2013). "CLP1 links tRNA metabolism to progressive motor-neuron loss." Nature **495**(7442): 474-480.
Hatton, L. S., J. J. Eloranta, L. M. Figueiredo, Y. Takagaki, J. L. Manley and K. O'Hare (2000). "The Drosophila homologue of the 64 kDa subunit of cleavage stimulation factor interacts with the 77 kDa subunit encoded by the suppressor of forked gene." Nucleic Acids Res **28**(2): 520-526.
Hautbergue, G. M., M. L. Hung, A. P. Golovanov, L. Y. Lian and S. A. Wilson (2008). "Mutually exclusive interactions drive handover of mRNA from export adaptors to TAP." Proc Natl Acad Sci U S A **105**(13): 5154-5159.
He, F., A. Celik, C. Wu and A. Jacobson (2018). "General decapping activators target different subsets of inefficiently translated mRNAs." Elife **7**.
He, F., C. Li, B. Roy and A. Jacobson (2014). "Yeast Edc3 targets RPS28B mRNA for decapping by binding to a 3' untranslated region decay-inducing regulatory element." Mol Cell Biol **34**(8): 1438-1451.
He, X., A. U. Khan, H. Cheng, D. L. Pappas, Jr., M. Hampsey and C. L. Moore (2003). "Functional interactions between the transcription and mRNA 3' end processing machineries mediated by Ssu72 and Sub1." Genes Dev **17**(8): 1030-1042.
Heath, C. G., N. Viphakone and S. A. Wilson (2016). "The role of TREX in gene expression and disease." Biochem J **473**(19): 2911-2935.
Heath, C. M., M. Windsor and T. Wileman (2001). "Aggresomes resemble sites specialized for virus assembly." J Cell Biol **153**(3): 449-455.
Heidemann, M., C. Hintermair, K. Voss and D. Eick (2013). "Dynamic phosphorylation patterns of RNA polymerase II CTD during transcription." Biochim Biophys Acta **1829**(1): 55-62.
Hetzer, M. W. and S. R. Wente (2009). "Border control at the nucleus: biogenesis and organization of the nuclear membrane and pore complexes." Dev Cell **17**(5): 606-616.
Higgs, D. R., S. E. Goodbourn, J. Lamb, J. B. Clegg, D. J. Weatherall and N. J. Proudfoot (1983). "Alpha-thalassaemia caused by a polyadenylation signal mutation." Nature **306**(5941): 398-400.
Hill, C. H., V. Boreikaite, A. Kumar, A. Casanal, P. Kubik, G. Degliesposti, S. Maslen, A. Mariani, O. von Loeffelholz, M. Girbig, M. Skehel and L. A. Passmore (2019). "Activation of the Endonuclease that Defines mRNA 3' Ends Requires Incorporation into an 8-Subunit Core Cleavage and Polyadenylation Factor Complex." Mol Cell **73**(6): 1217-1231 e1211.
Hirose, Y. and J. L. Manley (1998). "RNA polymerase II is an essential mRNA polyadenylation factor." Nature **395**(6697): 93-96.

Hirose, Y. and J. L. Manley (2000). "RNA polymerase II and the integration of nuclear events." Genes Dev **14**(12): 1415-1429.
Hockert, J. A., H. J. Yeh and C. C. MacDonald (2010). "The hinge domain of the cleavage stimulation factor protein CstF-64 is essential for CstF-77 interaction, nuclear localization, and polyadenylation." J Biol Chem **285**(1): 695-704.
Hofmann, I., M. Schnolzer, I. Kaufmann and W. W. Franke (2002). "Symplekin, a constitutive protein of karyo- and cytoplasmic particles involved in mRNA biogenesis in Xenopus laevis oocytes." Mol Biol Cell **13**(5): 1665-1676.
Houseley, J. and D. Tollervey (2009). "The many pathways of RNA degradation." Cell **136**(4): 763-776.
Hsin, J. P. and J. L. Manley (2012). "The RNA polymerase II CTD coordinates transcription and RNA processing." Genes Dev **26**(19): 2119-2137.
Hu, J., C. S. Lutz, J. Wilusz and B. Tian (2005). "Bioinformatic identification of candidate cis-regulatory elements involved in human mRNA polyadenylation." RNA **11**(10): 1485-1493.
Hudson, B. P., M. A. Martinez-Yamout, H. J. Dyson and P. E. Wright (2004). "Recognition of the mRNA AU-rich element by the zinc finger domain of TIS11d." Nat Struct Mol Biol **11**(3): 257-264.
Hurt, E., K. Strasser, A. Segref, S. Bailer, N. Schlaich, C. Presutti, D. Tollervey and R. Jansen (2000). "Mex67p mediates nuclear export of a variety of RNA polymerase II transcripts." J Biol Chem **275**(12): 8361-8368.
Hwang, H. W., C. Y. Park, H. Goodarzi, J. J. Fak, A. Mele, M. J. Moore, Y. Saito and R. B. Darnell (2016). "PAPERCLIP Identifies MicroRNA Targets and a Role of CstF64/64tau in Promoting Non-canonical poly(A) Site Usage." Cell Rep **15**(2): 423-435.
Ito, S., A. Sakai, T. Nomura, Y. Miki, M. Ouchida, J. Sasaki and K. Shimizu (2001). "A novel WD40 repeat protein, WDC146, highly expressed during spermatogenesis in a stage-specific manner." Biochem Biophys Res Commun **280**(3): 656-663.
Jackson, R. J., C. U. Hellen and T. V. Pestova (2010). "The mechanism of eukaryotic translation initiation and principles of its regulation." Nat Rev Mol Cell Biol **11**(2): 113-127.
Jacobson, A. and S. W. Peltz (1996). "Interrelationships of the pathways of mRNA decay and translation in eukaryotic cells." Annu Rev Biochem **65**: 693-739.
Jarvelin, A. I., M. Noerenberg, I. Davis and A. Castello (2016). "The new (dis)order in RNA regulation." Cell Commun Signal **14**: 9.
Jasnovidova, O. and R. Stefl (2013). "The CTD code of RNA polymerase II: a structural view." Wiley Interdiscip Rev RNA **4**(1): 1-16.
Jensen, T. H., J. Boulay, J. R. Olesen, J. Colin, M. Weyler and D. Libri (2004). "Modulation of transcription affects mRNP quality." Mol Cell **16**(2): 235-244.
Jensen, T. H., K. Dower, D. Libri and M. Rosbash (2003). "Early formation of mRNP: license for export or quality control?" Mol Cell **11**(5): 1129-1138.
Ji, Z., J. Y. Lee, Z. Pan, B. Jiang and B. Tian (2009). "Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development." Proc Natl Acad Sci U S A **106**(17): 7028-7033.
Johnson, S. A., G. Cubberley and D. L. Bentley (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.
Johnson, S. A., H. Kim, B. Erickson and D. L. Bentley (2011). "The export factor Yra1 modulates mRNA 3' end processing." Nat Struct Mol Biol **18**(10): 1164-1171.
Jonas, S., M. Christie, D. Peter, D. Bhandari, B. Loh, E. Huntzinger, O. Weichenrieder and E. Izaurralde (2014). "An asymmetric PAN3 dimer recruits a single PAN2 exonuclease to mediate mRNA deadenylation and decay." Nat Struct Mol Biol **21**(7): 599-608.
Jonas, S. and E. Izaurralde (2013). "The role of disordered protein regions in the assembly of decapping complexes and RNP granules." Genes Dev **27**(24): 2628-2641.
Jumper, J., R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Zidek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli and D. Hassabis

(2021). "Highly accurate protein structure prediction with AlphaFold." Nature **596**(7873): 583-589.
Karamyshev, A. L. and Z. N. Karamysheva (2018). "Lost in Translation: Ribosome-Associated mRNA and Protein Quality Controls." Frontiers in Genetics **9**.
Kashiwabara, S., T. Zhuang, K. Yamagata, J. Noguchi, A. Fukamizu and T. Baba (2000). "Identification of a novel isoform of poly(A) polymerase, TPAP, specifically present in the cytoplasm of spermatogenic cells." Dev Biol **228**(1): 106-115.
Katahira, J. (2012). "mRNA export and the TREX complex." Biochim Biophys Acta **1819**(6): 507-513.
Katahira, J., K. Strasser, A. Podtelejnikov, M. Mann, J. U. Jung and E. Hurt (1999). "The Mex67p-mediated nuclear mRNA export pathway is conserved from yeast to human." EMBO J **18**(9): 2593-2609.
Katahira, J. and Y. Yoneda (2009). "Roles of the TREX complex in nuclear export of mRNA." RNA Biol **6**(2): 149-152.
Kaufmann, I., G. Martin, A. Friedlein, H. Langen and W. Keller (2004). "Human Fip1 is a subunit of CPSF that binds to U-rich RNA elements and stimulates poly(A) polymerase." EMBO J **23**(3): 616-626.
Keller, R. W., U. Kuhn, M. Aragon, L. Bornikova, E. Wahle and D. G. Bear (2000). "The nuclear poly(A) binding protein, PABP2, forms an oligomeric particle covering the length of the poly(A) tail." J Mol Biol **297**(3): 569-583.
Keller, W., S. Bienroth, K. M. Lang and G. Christofori (1991). "Cleavage and polyadenylation factor CPF specifically interacts with the pre-mRNA 3' processing signal AAUAAA." EMBO J **10**(13): 4241-4249.
Kennedy, S. A., M. L. Frazier, M. Steiniger, A. M. Mast, W. F. Marzluff and M. R. Redinbo (2009). "Crystal structure of the HEAT domain from the Pre-mRNA processing factor Symplekin." J Mol Biol **392**(1): 115-128.
Keon, B. H., S. Schafer, C. Kuhn, C. Grund and W. W. Franke (1996). "Symplekin, a novel type of tight junction plaque protein." J Cell Biol **134**(4): 1003-1018.
Kerwitz, Y., U. Kuhn, H. Lilie, A. Knoth, T. Scheuermann, H. Friedrich, E. Schwarz and E. Wahle (2003). "Stimulation of poly(A) polymerase through a direct interaction with the nuclear poly(A) binding protein allosterically regulated by RNA." EMBO J **22**(14): 3705-3714.
Khatter, H., M. K. Vorlander and C. W. Muller (2017). "RNA polymerase I and III: similar yet unique." Current Opinion in Structural Biology **47**: 88-94.
Kiledjian, M. and G. Dreyfuss (1992). "Primary structure and binding activity of the hnRNP U protein: binding RNA through RGG box." EMBO J **11**(7): 2655-2664.
Kim, H., J. H. Lee and Y. Lee (2003). "Regulation of poly(A) polymerase by 14-3-3epsilon." EMBO J **22**(19): 5208-5219.
Kim, H. and Y. Lee (2001). "Interaction of poly(A) polymerase with the 25-kDa subunit of cleavage factor I." Biochem Biophys Res Commun **289**(2): 513-518.
Kim, M., S. H. Ahn, N. J. Krogan, J. F. Greenblatt and S. Buratowski (2004). "Transitions in RNA polymerase II elongation complexes at the 3' ends of genes." EMBO J **23**(2): 354-364.
Kim, M., N. J. Krogan, L. Vasiljeva, O. J. Rando, E. Nedea, J. F. Greenblatt and S. Buratowski (2004). "The yeast Rat1 exonuclease promotes transcription termination by RNA polymerase II." Nature **432**(7016): 517-522.
Kohler, A. and E. Hurt (2007). "Exporting RNA from the nucleus to the cytoplasm." Nat Rev Mol Cell Biol **8**(10): 761-773.
Kolev, N. G. and J. A. Steitz (2005). "Symplekin and multiple other polyadenylation factors participate in 3'-end maturation of histone mRNAs." Genes Dev **19**(21): 2583-2592.
Kolev, N. G., T. A. Yario, E. Benson and J. A. Steitz (2008). "Conserved motifs in both CPSF73 and CPSF100 are required to assemble the active endonuclease for histone mRNA 3'-end maturation." EMBO Rep **9**(10): 1013-1018.
Kosinski, J., A. von Appen, A. Ori, K. Karius, C. W. Muller and M. Beck (2015). "Xlink Analyzer: software for analysis and visualization of cross-linking data in the context of three-dimensional structures." J Struct Biol **189**(3): 177-183.

Kubo, T., T. Wada, Y. Yamaguchi, A. Shimizu and H. Handa (2006). "Knock-down of 25 kDa subunit of cleavage factor Im in Hela cells alters alternative polyadenylation within 3'-UTRs." Nucleic Acids Res **34**(21): 6264-6271.
Kuhn, U., M. Gundel, A. Knoth, Y. Kerwitz, S. Rudel and E. Wahle (2009). "Poly(A) tail length is controlled by the nuclear poly(A)-binding protein regulating the interaction between poly(A) polymerase and the cleavage and polyadenylation specificity factor." J Biol Chem **284**(34): 22803-22814.
Kuhn, U., A. Nemeth, S. Meyer and E. Wahle (2003). "The RNA binding domains of the nuclear poly(A)-binding protein." J Biol Chem **278**(19): 16916-16925.
Kuhn, U. and T. Pieler (1996). "Xenopus poly(A) binding protein: functional domains in RNA binding and protein-protein interaction." J Mol Biol **256**(1): 20-30.
Kuhn, U. and E. Wahle (2004). "Structure and function of poly(A) binding proteins." Biochim Biophys Acta **1678**(2-3): 67-84.
Kumar, A., M. Clerici, L. M. Muckenfuss, L. A. Passmore and M. Jinek (2019). "Mechanistic insights into mRNA 3'-end processing." Curr Opin Struct Biol **59**: 143-150.
Kyburz, A., A. Friedlein, H. Langen and W. Keller (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.
Kyburz, A., M. Sadowski, B. Dichtl and W. Keller (2003). "The role of the yeast cleavage and polyadenylation factor subunit Ydh1p/Cft2p in pre-mRNA 3'-end formation." Nucleic Acids Res **31**(14): 3936-3945.
Kyriakopoulou, C. B., H. Nordvarg and A. Virtanen (2001). "A novel nuclear human poly(A) polymerase (PAP), PAP gamma." J Biol Chem **276**(36): 33504-33511.
Labno, A., R. Tomecki and A. Dziembowski (2016). "Cytoplasmic RNA decay pathways - Enzymes and mechanisms." Biochim Biophys Acta **1863**(12): 3125-3147.
Lackford, B., C. Yao, G. M. Charles, L. Weng, X. Zheng, E. A. Choi, X. Xie, J. Wan, Y. Xing, J. M. Freudenberg, P. Yang, R. Jothi, G. Hu and Y. Shi (2014). "Fip1 regulates mRNA alternative polyadenylation to promote stem cell self-renewal." EMBO J **33**(8): 878-889.
Laishram, R. S. (2014). "Poly(A) polymerase (PAP) diversity in gene expression--star-PAP vs canonical PAP." FEBS Lett **588**(14): 2185-2197.
Lamb, J. R., S. Tugendreich and P. Hieter (1995). "Tetratrico peptide repeat interactions: to TPR or not to TPR?" Trends Biochem Sci **20**(7): 257-259.
Lau, N. C., A. Kolkman, F. M. van Schaik, K. W. Mulder, W. W. Pijnappel, A. J. Heck and H. T. Timmers (2009). "Human Ccr4-Not complexes contain variable deadenylase subunits." Biochem J **422**(3): 443-453.
Le, Y. J., H. Kim, J. H. Chung and Y. Lee (2001). "Testis-specific expression of an intronless gene encoding a human poly(A) polymerase." Mol Cells **11**(3): 379-385.
Lee, Y. J., Y. Lee and J. H. Chung (2000). "An intronless gene encoding a poly(A) polymerase is specifically expressed in testis." FEBS Lett **487**(2): 287-292.
Leeper, T. C., X. Qu, C. Lu, C. Moore and G. Varani (2010). "Novel protein-protein contacts facilitate mRNA 3'-processing signal recognition by Rna15 and Hrp1." J Mol Biol **401**(3): 334-349.
Legrand, P., N. Pinaud, L. Minvielle-Sebastia and S. Fribourg (2007). "The structure of the CstF-77 homodimer provides insights into CstF assembly." Nucleic Acids Res **35**(13): 4515-4522.
Li, D. and R. Roberts (2001). "WD-repeat proteins: structure characteristics, biological function, and their involvement in human diseases." Cell Mol Life Sci **58**(14): 2085-2097.
Li de la Sierra-Gallay, I., L. Zig, A. Jamalli and H. Putzer (2008). "Structural insights into the dual activity of RNase J." Nat Struct Mol Biol **15**(2): 206-212.
Li, H., S. Tong, X. Li, H. Shi, Z. Ying, Y. Gao, H. Ge, L. Niu and M. Teng (2011). "Structural basis of pre-mRNA recognition by the human cleavage factor Im complex." Cell Res **21**(7): 1039-1051.
Li, W., R. S. Laishram, Z. Ji, C. A. Barlow, B. Tian and R. A. Anderson (2012). "Star-PAP control of BIK expression and apoptosis is regulated by nuclear PIPKIalpha and PKCdelta signaling." Mol Cell **45**(1): 25-37.

Li, Y. and M. Kiledjian (2010). "Regulation of mRNA decapping." Wiley Interdiscip Rev RNA **1**(2): 253-265.
Licatalosi, D. D., G. Geiger, M. Minet, S. Schroeder, K. Cilli, J. B. McNeil and D. L. Bentley (2002). "Functional interaction of yeast pre-mRNA 3' end processing factors with RNA polymerase II." Mol Cell **9**(5): 1101-1111.
Lischwe, M. A., R. G. Cook, Y. S. Ahn, L. C. Yeoman and H. Busch (1985). "Clustering of glycine and NG,NG-dimethylarginine in nucleolar protein C23." Biochemistry **24**(22): 6025-6028.
Lischwe, M. A., R. L. Ochs, R. Reddy, R. G. Cook, L. C. Yeoman, E. M. Tan, M. Reichlin and H. Busch (1985). "Purification and partial characterization of a nucleolar scleroderma antigen (Mr = 34,000; pI, 8.5) rich in NG,NG-dimethylarginine." J Biol Chem **260**(26): 14304-14310.
Liu, H. and M. Kiledjian (2006). "Decapping the message: a beginning or an end." Biochem Soc Trans **34**(Pt 1): 35-38.
Logan, J., E. Falck-Pedersen, J. E. Darnell, Jr. and T. Shenk (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.
Luo, W., A. W. Johnson and D. L. Bentley (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.
Lutz, C. S. (2008). "Alternative polyadenylation: a twist on mRNA 3' end formation." ACS Chem Biol **3**(10): 609-617.
Lykke-Andersen, S. and T. H. Jensen (2015). "Nonsense-mediated mRNA decay: an intricate machinery that shapes transcriptomes." Nat Rev Mol Cell Biol **16**(11): 665-677.
MacDonald, C. C., J. Wilusz and T. Shenk (1994). "The 64-kilodalton subunit of the CstF polyadenylation factor binds to pre-mRNAs downstream of the cleavage site and influences cleavage site location." Mol Cell Biol **14**(10): 6647-6654.
Madeira, F., Y. M. Park, J. Lee, N. Buso, T. Gur, N. Madhusoodanan, P. Basutkar, A. R. N. Tivey, S. C. Potter, R. D. Finn and R. Lopez (2019). "The EMBL-EBI search and sequence analysis tools APIs in 2019." Nucleic Acids Res **47**(W1): W636-W641.
Malik, S. and R. G. Roeder (2010). "The metazoan Mediator co-activator complex as an integrative hub for transcriptional regulation." Nat Rev Genet **11**(11): 761-772.
Mandel, C. R., Y. Bai and L. Tong (2008). "Protein factors in pre-mRNA 3'-end processing." Cell Mol Life Sci **65**(7-8): 1099-1122.
Mandel, C. R., S. Kaneko, H. Zhang, D. Gebauer, V. Vethantham, J. L. Manley and L. Tong (2006). "Polyadenylation factor CPSF-73 is the pre-mRNA 3'-end-processing endonuclease." Nature **444**(7121): 953-956.
Mangus, D. A., M. C. Evans and A. Jacobson (2003). "Poly(A)-binding proteins: multifunctional scaffolds for the post-transcriptional control of gene expression." Genome Biol **4**(7): 223.
Mann, T. L. and U. J. Krull (2003). "Fluorescence polarization spectroscopy in protein analysis." Analyst **128**(4): 313-317.
Maris, C., C. Dominguez and F. H. Allain (2005). "The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression." FEBS J **272**(9): 2118-2131.
Martin, G., A. R. Gruber, W. Keller and M. Zavolan "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.
Martin, G., P. Jeno and W. Keller (1999). "Mapping of ATP binding regions in poly(A) polymerases by photoaffinity labeling and by mutational analysis identifies a domain conserved in many nucleotidyltransferases." Protein Sci **8**(11): 2380-2391.
Martin, G. and W. Keller (1996). "Mutational analysis of mammalian poly(A) polymerase identifies a region for primer binding and catalytic domain, homologous to the family X polymerases, and to other nucleotidyltransferases." EMBO J **15**(10): 2593-2603.
Martin, G., W. Keller and S. Doublie (2000). "Crystal structure of mammalian poly(A) polymerase in complex with an analog of ATP." EMBO J **19**(16): 4193-4203.

Martin, G., A. Moglich, W. Keller and S. Doublie (2004). "Biochemical and structural insights into substrate binding and catalytic mechanism of mammalian poly(A) polymerase." J Mol Biol **341**(4): 911-925.
Marzluff, W. F., E. J. Wagner and R. J. Duronio (2008). "Metabolism and regulation of canonical histone mRNAs: life without a poly(A) tail." Nat Rev Genet **9**(11): 843-854.
Masamha, C. P. and E. Wagner (2018). "Multiple Mechanisms Driving Alternative Polyadenylation of Cyclin D1 (CCND1) pre-mRNA Processing." Faseb Journal **32**(1).
Mason, P. J., J. A. Elkington, M. M. Lloyd, M. B. Jones and J. G. Williams (1986). "Mutations downstream of the polyadenylation site of a Xenopus beta-globin mRNA affect the position but not the efficiency of 3' processing." Cell **46**(2): 263-270.
Mayr, C. and D. P. Bartel (2009). "Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells." Cell **138**(4): 673-684.
McCracken, S., N. Fong, K. Yankulov, S. Ballantyne, G. Pan, J. Greenblatt, S. D. Patterson, M. Wickens and D. L. Bentley (1997). "The C-terminal domain of RNA polymerase II couples mRNA processing to transcription." Nature **385**(6614): 357-361.
McDevitt, M. A., R. P. Hart, W. W. Wong and J. R. Nevins (1986). "Sequences capable of restoring poly(A) site function define two distinct downstream elements." EMBO J **5**(11): 2907-2913.
McGrew, L. L. and J. D. Richter (1990). "Translational control by cytoplasmic polyadenylation during Xenopus oocyte maturation: characterization of cis and trans elements and regulation by cyclin/MPF." EMBO J **9**(11): 3743-3751.
Meinhart, A. and P. Cramer (2004). "Recognition of RNA polymerase II carboxy-terminal domain by 3'-RNA-processing factors." Nature **430**(6996): 223-226.
Meinke, G., C. Ezeokonkwo, P. Balbo, W. Stafford, C. Moore and A. Bohm (2008). "Structure of yeast poly(A) polymerase in complex with a peptide from Fip1, an intrinsically disordered protein." Biochemistry **47**(26): 6859-6869.
Mellman, D. L., M. L. Gonzales, C. Song, C. A. Barlow, P. Wang, C. Kendziorski and R. A. Anderson (2008). "A PtdIns4,5P2-regulated nuclear poly(A) polymerase controls expression of select mRNAs." Nature **451**(7181): 1013-1017.
Mendez, R., K. G. Murthy, K. Ryan, J. L. Manley and J. D. Richter (2000). "Phosphorylation of CPEB by Eg2 mediates the recruitment of CPSF into an active cytoplasmic polyadenylation complex." Mol Cell **6**(5): 1253-1259.
Merkley, E. D., S. Rysavy, A. Kahraman, R. P. Hafen, V. Daggett and J. N. Adkins (2014). "Distance restraints from crosslinking mass spectrometry: mining a molecular dynamics simulation database to evaluate lysine-lysine distances." Protein Sci **23**(6): 747-759.
Merrick, W. C. (2004). "Cap-dependent and cap-independent translation in eukaryotic systems." Gene **332**: 1-11.
Meyer, S., C. Temme and E. Wahle (2004). "Messenger RNA turnover in eukaryotes: pathways and enzymes." Crit Rev Biochem Mol Biol **39**(4): 197-216.
Milac, A. L., E. Bojarska and A. Wypijewska del Nogal (2014). "Decapping Scavenger (DcpS) enzyme: advances in its structure, activity and roles in the cap-dependent mRNA metabolism." Biochim Biophys Acta **1839**(6): 452-462.
Mildvan, A. S., Z. Xia, H. F. Azurmendi, V. Saraswat, P. M. Legler, M. A. Massiah, S. B. Gabelli, M. A. Bianchet, L. W. Kang and L. M. Amzel (2005). "Structures and mechanisms of Nudix hydrolases." Arch Biochem Biophys **433**(1): 129-143.
Millevoi, S. and S. Vagner (2010). "Molecular mechanisms of eukaryotic pre-mRNA 3' end processing regulation." Nucleic Acids Res **38**(9): 2757-2774.
Minvielle-Sebastia, L., P. J. Preker and W. Keller (1994). "RNA14 and RNA15 proteins as components of a yeast pre-mRNA 3'-end processing factor." Science **266**(5191): 1702-1705.
Minvielle-Sebastia, L., P. J. Preker, T. Wiederkehr, Y. Strahm and W. Keller (1997). "The major yeast poly(A)-binding protein is associated with cleavage factor IA and functions in premessenger RNA 3'-end formation." Proc Natl Acad Sci U S A **94**(15): 7897-7902.
Moore, C. L., J. Chen and J. Whoriskey (1988). "Two proteins crosslinked to RNA containing the adenovirus L3 poly(A) site require the AAUAAA sequence for binding." EMBO J **7**(10): 3159-3169.

Moore, M. J. (2005). "From birth to death: the complex lives of eukaryotic mRNAs." Science **309**(5740): 1514-1518.
Moore, M. J. and N. J. Proudfoot (2009). "Pre-mRNA processing reaches back to transcription and ahead to translation." Cell **136**(4): 688-700.
Moreno-Morcillo, M., L. Minvielle-Sebastia, S. Fribourg and C. D. Mackereth (2011). "Locked tether formation by cooperative folding of Rna14p monkeytail and Rna15p hinge domains in the yeast CF IA complex." Structure **19**(4): 534-545.
Moreno-Morcillo, M., L. Minvielle-Sebastia, C. Mackereth and S. Fribourg "Hexameric architecture of CstF supported by CstF-50 homodimerization domain structure." RNA **17**(3): 412-418.
Moteki, S. and D. Price (2002). "Functional coupling of capping and transcription of mRNA." Mol Cell **10**(3): 599-609.
Mugridge, J. S., J. Coller and J. D. Gross (2018). "Structural and molecular mechanisms for the control of eukaryotic 5'-3' mRNA decay." Nat Struct Mol Biol **25**(12): 1077-1085.
Murthy, K. G. and J. L. Manley (1992). "Characterization of the multisubunit cleavage-polyadenylation specificity factor from calf thymus." J Biol Chem **267**(21): 14804-14811.
Murthy, K. G. and J. L. Manley (1995). "The 160-kD subunit of human cleavage-polyadenylation specificity factor coordinates pre-mRNA 3'-end formation." Genes Dev **9**(21): 2672-2683.
Nag, A., K. Narsinh and H. G. Martinson (2007). "The poly(A)-dependent transcriptional pause is mediated by CPSF acting on the body of the polymerase." Nat Struct Mol Biol **14**(7): 662-669.
Nagai, K., C. Oubridge, T. H. Jessen, J. Li and P. R. Evans (1990). "Crystal structure of the RNA-binding domain of the U1 small nuclear ribonucleoprotein A." Nature **348**(6301): 515-520.
Nagaike, T. and J. L. Manley (2011). "Transcriptional activators enhance polyadenylation of mRNA precursors." RNA Biol **8**(6): 964-967.
Nagarajan, V. K., C. I. Jones, S. F. Newbury and P. J. Green (2013). "XRN 5'-->3' exoribonucleases: structure, mechanisms and functions." Biochim Biophys Acta **1829**(6-7): 590-603.
Nagy, E. and L. E. Maquat (1998). "A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance." Trends in Biochemical Sciences **23**(6): 198-199.
Nemeth, A., S. Krause, D. Blank, A. Jenny, P. Jeno, A. Lustig and E. Wahle (1995). "Isolation of genomic and cDNA clones encoding bovine poly(A) binding protein II." Nucleic Acids Res **23**(20): 4034-4041.
Neuwald, A. F. and A. Poleksic (2000). "PSI-BLAST searches using hidden markov models of structural repeats: prediction of an unusual sliding DNA clamp and of beta-propellers in UV-damaged DNA-binding protein." Nucleic Acids Res **28**(18): 3570-3580.
Nissan, T., P. Rajyaguru, M. She, H. Song and R. Parker (2010). "Decapping activators in Saccharomyces cerevisiae act by multiple mechanisms." Mol Cell **39**(5): 773-783.
Noble, C. G., B. Beuth and I. A. Taylor (2007). "Structure of a nucleotide-bound Clp1-Pcf11 polyadenylation factor." Nucleic Acids Res **35**(1): 87-99.
Noble, C. G., P. A. Walker, L. J. Calder and I. A. Taylor (2004). "Rna14-Rna15 assembly mediates the RNA-binding capability of Saccharomyces cerevisiae cleavage factor IA." Nucleic Acids Res **32**(11): 3364-3375.
Nunes, N. M., W. Li, B. Tian and A. Furger (2010). "A functional human Poly(A) site requires only a potent DSE and an A-rich upstream sequence." EMBO J **29**(9): 1523-1536.
Oberstrass, F. C., S. D. Auweter, M. Erat, Y. Hargous, A. Henning, P. Wenter, L. Reymond, B. Amir-Ahmady, S. Pitsch, D. L. Black and F. H. Allain (2005). "Structure of PTB bound to RNA: specific binding and implications for splicing regulation." Science **309**(5743): 2054-2057.
Orkin, S. H., T. C. Cheng, S. E. Antonarakis and H. H. Kazazian, Jr. (1985). "Thalassemia due to a mutation in the cleavage-polyadenylation signal of the human beta-globin gene." EMBO J **4**(2): 453-456.

Palacios, I. M., D. Gatfield, D. St Johnston and E. Izaurralde (2004). "An eIF4AIII-containing complex required for mRNA localization and nonsense-mediated mRNA decay." Nature **427**(6976): 753-757.
Pan, Q., O. Shai, L. J. Lee, B. J. Frey and B. J. Blencowe (2008). "Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing." Nat Genet **40**(12): 1413-1415.
Pancevac, C., D. C. Goldstone, A. Ramos and I. A. Taylor (2010). "Structure of the Rna15 RRM-RNA complex reveals the molecular basis of GU specificity in transcriptional 3'-end processing factors." Nucleic Acids Res **38**(9): 3119-3132.
Pandey, N. B., N. Chodchoy, T. J. Liu and W. F. Marzluff (1990). "Introns in histone genes alter the distribution of 3' ends." Nucleic Acids Res **18**(11): 3161-3170.
Paris, J., H. B. Osborne, A. Couturier, R. Le Guellec and M. Philippe (1988). "Changes in the polyadenylation of specific stable RNA during the early development of Xenopus laevis." Gene **72**(1-2): 169-176.
Parker, R. (2012). "RNA degradation in Saccharomyces cerevisae." Genetics **191**(3): 671-702.
Parker, R. and H. Song (2004). "The enzymes and control of eukaryotic mRNA turnover." Nat Struct Mol Biol **11**(2): 121-127.
Passmore, L. A. and J. Coller (2022). "Roles of mRNA poly(A) tails in regulation of eukaryotic gene expression." Nat Rev Mol Cell Biol **23**(2): 93-106.
Passos, D. O., M. K. Doma, C. J. Shoemaker, D. Muhlrad, R. Green, J. Weissman, J. Hollien and R. Parker (2009). "Analysis of Dom34 and Its Function in No-Go Decay." Molecular Biology of the Cell **20**(13): 3025-3032.
Paulson, A. R. and L. Tong (2012). "Crystal structure of the Rna14-Rna15 complex." RNA **18**(6): 1154-1162.
Perez Canadillas, J. M. and G. Varani (2003). "Recognition of GU-rich polyadenylation regulatory elements by human CstF-64 protein." EMBO J **22**(11): 2821-2830.
Perumal, K., K. Sinha, D. Henning and R. Reddy (2001). "Purification, characterization, and cloning of the cDNA of human signal recognition particle RNA 3'-adenylating enzyme." J Biol Chem **276**(24): 21791-21796.
Pettersen, E. F., T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng and T. E. Ferrin (2004). "UCSF Chimera--a visualization system for exploratory research and analysis." J Comput Chem **25**(13): 1605-1612.
Piccirillo, C., R. Khanna and M. Kiledjian (2003). "Functional characterization of the mammalian mRNA decapping enzyme hDcp2." RNA **9**(9): 1138-1147.
Pique, M., J. M. Lopez, S. Foissac, R. Guigo and R. Mendez (2008). "A combinatorial code for CPE-mediated translational control." Cell **132**(3): 434-448.
Pirngruber, J. and S. A. Johnsen (2010). "Induced G1 cell-cycle arrest controls replication-dependent histone mRNA 3' end processing through p21, NPAT and CDK9." Oncogene **29**(19): 2853-2863.
Porrua, O. and D. Libri (2015). "Transcription termination and the control of the transcriptome: why, where and how to stop." Nat Rev Mol Cell Biol **16**(3): 190-202.
Preker, P. J. and W. Keller (1998). "The HAT helix, a repetitive motif implicated in RNA processing." Trends Biochem Sci **23**(1): 15-16.
Preker, P. J., J. Lingner, L. Minvielle-Sebastia and W. Keller (1995). "The FIP1 gene encodes a component of a yeast pre-mRNA polyadenylation factor that directly interacts with poly(A) polymerase." Cell **81**(3): 379-389.
Preker, P. J., M. Ohnacker, L. Minvielle-Sebastia and W. Keller (1997). "A multisubunit 3' end processing factor from yeast containing poly(A) polymerase and homologues of the subunits of mammalian cleavage and polyadenylation specificity factor." EMBO J **16**(15): 4727-4737.
Proudfoot, N. (1991). "Poly(A) signals." Cell **64**(4): 671-674.
Proudfoot, N. (2004). "New perspectives on connecting messenger RNA 3' end formation to transcription." Curr Opin Cell Biol **16**(3): 272-278.
Proudfoot, N. and J. O'Sullivan (2002). "Polyadenylation: a tail of two complexes." Curr Biol **12**(24): R855-857.

Proudfoot, N. J. (2011). "Ending the message: poly(A) signals then and now." Genes Dev **25**(17): 1770-1782.
Puhringer, T., U. Hohmann, L. Fin, B. Pacheco-Fiallos, U. Schellhaas, J. Brennecke and C. Plaschka (2020). "Structure of the human core transcription-export complex reveals a hub for multivalent interactions." Elife **9**.
Punjani, A., J. L. Rubinstein, D. J. Fleet and M. A. Brubaker (2017). "cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination." Nat Methods **14**(3): 290-296.
Qu, X., J. M. Perez-Canadillas, S. Agrawal, J. De Baecke, H. Cheng, G. Varani and C. Moore (2007). "The C-terminal domains of vertebrate CstF-64 and its yeast orthologue Rna15 form a new structure critical for mRNA 3'-end processing." J Biol Chem **282**(3): 2101-2115.
Raabe, T., F. J. Bollum and J. L. Manley (1991). "Primary structure and expression of bovine poly(A) polymerase." Nature **353**(6341): 229-234.
Raabe, T., K. G. Murthy and J. L. Manley (1994). "Poly(A) polymerase contains multiple functional domains." Mol Cell Biol **14**(5): 2946-2957.
Radford, H. E., H. A. Meijer and C. H. de Moor (2008). "Translational control by cytoplasmic polyadenylation in Xenopus oocytes." Biochim Biophys Acta **1779**(4): 217-229.
Rajyaguru, P. and R. Parker (2012). "RGG motif proteins: modulators of mRNA functional states." Cell Cycle **11**(14): 2594-2599.
Ramanathan, A., G. B. Robb and S. H. Chan (2016). "mRNA capping: biological functions and applications." Nucleic Acids Res **44**(16): 7511-7526.
Reed, R. and H. Cheng (2005). "TREX, SR proteins and export of mRNA." Curr Opin Cell Biol **17**(3): 269-273.
Ren, F., N. Zhang, L. Zhang, E. Miller and J. J. Pu (2020). "Alternative Polyadenylation: a new frontier in post transcriptional regulation." Biomark Res **8**(1): 67.
Richardson, J. M., K. W. McMahon, C. C. MacDonald and G. I. Makhatadze (1999). "MEARA sequence repeat of human CstF-64 polyadenylation factor is helical in solution. A spectroscopic and calorimetric study." Biochemistry **38**(39): 12869-12875.
Richter, J. D. (2007). "CPEB: a life in translation." Trends Biochem Sci **32**(6): 279-285.
Rodrigues, J. P., M. Rode, D. Gatfield, B. J. Blencowe, M. Carmo-Fonseca and E. Izaurralde (2001). "REF proteins mediate the export of spliced and unspliced mRNAs from the nucleus." Proc Natl Acad Sci U S A **98**(3): 1030-1035.
Rodriguez-Navarro, S. and E. Hurt (2011). "Linking gene regulation to mRNA production and export." Curr Opin Cell Biol **23**(3): 302-309.
Rohou, A. and N. Grigorieff (2015). "CTFFIND4: Fast and accurate defocus estimation from electron micrographs." Journal of Structural Biology **192**(2): 216-221.
Rondon, A. G., S. Jimeno and A. Aguilera (2010). "The interface between transcription and mRNP export: from THO to THSC/TREX-2." Biochim Biophys Acta **1799**(8): 533-538.
Rosado-Lugo, J. D. and M. Hampsey (2014). "The Ssu72 phosphatase mediates the RNA polymerase II initiation-elongation transition." J Biol Chem **289**(49): 33916-33926.
Rose, R., M. Weyand, M. Lammers, T. Ishizaki, M. R. Ahmadian and A. Wittinghofer (2005). "Structural and mechanistic insights into the interaction between Rho and mammalian Dia." Nature **435**(7041): 513-518.
Ruegsegger, U., K. Beyer and W. Keller (1996). "Purification and characterization of human cleavage factor Im involved in the 3' end processing of messenger RNA precursors." J Biol Chem **271**(11): 6107-6113.
Ruegsegger, U., D. Blank and W. Keller (1998). "Human pre-mRNA cleavage factor Im is related to spliceosomal SR proteins and can be reconstituted in vitro from recombinant subunits." Mol Cell **1**(2): 243-253.
Ruepp, M. D., D. Schumperli and S. M. Barabino (2011). "mRNA 3' end processing and more--multiple functions of mammalian cleavage factor I-68." Wiley Interdiscip Rev RNA **2**(1): 79-91.
Ruepp, M. D., C. Schweingruber, N. Kleinschmidt and D. Schumperli (2010). "Interactions of CstF-64, CstF-77, and symplekin: implications on localisation and function." Mol Biol Cell **22**(1): 91-104.

Ruepp, M. D., C. Schweingruber, N. Kleinschmidt and D. Schumperli (2011). "Interactions of CstF-64, CstF-77, and symplekin: implications on localisation and function." Mol Biol Cell **22**(1): 91-104.
Ryan, K., O. Calvo and J. L. Manley (2004). "Evidence that polyadenylation factor CPSF-73 is the mRNA 3' processing endonuclease." RNA **10**(4): 565-573.
Sadowski, M., B. Dichtl, W. Hubner and W. Keller (2003). "Independent functions of yeast Pcf11p in pre-mRNA 3' end processing and in transcription termination." EMBO J **22**(9): 2167-2177.
Salisbury, J., K. W. Hutchison and J. H. Graber (2006). "A multispecies comparison of the metazoan 3'-processing downstream elements and the CstF-64 RNA recognition motif." BMC Genomics **7**: 55.
Salles, F. J., M. E. Lieberfarb, C. Wreden, J. P. Gergen and S. Strickland (1994). "Coordinate initiation of Drosophila development by regulated polyadenylation of maternal messenger RNAs." Science **266**(5193): 1996-1999.
Sanchez-Garcia, R., J. Gomez-Blanco, A. Cuervo, J. M. Carazo, C. O. S. Sorzano and J. Vargas (2021). "DeepEMhancer: a deep learning solution for cryo-EM volume post-processing." Commun Biol **4**(1): 874.
Santos-Rosa, H., H. Moreno, G. Simos, A. Segref, B. Fahrenkrog, N. Pante and E. Hurt (1998). "Nuclear mRNA export requires complex formation between Mex67p and Mtr2p at the nuclear pores." Mol Cell Biol **18**(11): 6826-6838.
Sartini, B. L., H. Wang, W. Wang, C. F. Millette and D. L. Kilpatrick (2008). "Pre-messenger RNA cleavage factor I (CFIm): potential role in alternative polyadenylation during spermatogenesis." Biol Reprod **78**(3): 472-482.
Schafer, I. B., M. Rode, F. Bonneau, S. Schussler and E. Conti (2014). "The structure of the Pan2-Pan3 core complex reveals cross-talk between deadenylase and pseudokinase." Nat Struct Mol Biol **21**(7): 591-598.
Schafer, I. B., M. Yamashita, J. M. Schuller, S. Schussler, P. Reichelt, M. Strauss and E. Conti (2019). "Molecular Basis for poly(A) RNP Architecture and Recognition by the Pan2-Pan3 Deadenylase." Cell **177**(6): 1619-1631 e1621.
Schafer, P., C. Tuting, L. Schonemann, U. Kuhn, T. Treiber, N. Treiber, C. Ihling, A. Graber, W. Keller, G. Meister, A. Sinz and E. Wahle (2018). "Reconstitution of mammalian cleavage factor II involved in 3' processing of mRNA precursors." RNA **24**(12): 1721-1737.
Scheres, S. H. (2012). "RELION: implementation of a Bayesian approach to cryo-EM structure determination." J Struct Biol **180**(3): 519-530.
Schmidt, T., P. Knick, H. Lilie, S. Friedrich, R. P. Golbik and S. E. Behrens (2016). "Coordinated Action of Two Double-Stranded RNA Binding Motifs and an RGG Motif Enables Nuclear Factor 90 To Flexibly Target Different RNA Substrates." Biochemistry **55**(6): 948-959.
Schmidt, T. G., L. Batz, L. Bonet, U. Carl, G. Holzapfel, K. Kiem, K. Matulewicz, D. Niermeier, I. Schuchardt and K. Stanar (2013). "Development of the Twin-Strep-tag(R) and its application for purification of recombinant proteins from cell culture supernatants." Protein Expr Purif **92**(1): 54-61.
Schonemann, L., U. Kuhn, G. Martin, P. Schafer, A. R. Gruber, W. Keller, M. Zavolan and E. Wahle (2014). "Reconstitution of CPSF active in polyadenylation: recognition of the polyadenylation signal by WDR33." Genes Dev **28**(21): 2381-2393.
Schroeder, S. C., B. Schwer, S. Shuman and D. Bentley (2000). "Dynamic association of capping enzymes with transcribing RNA polymerase II." Genes Dev **14**(19): 2435-2440.
Scrima, A., R. Konickova, B. K. Czyzewski, Y. Kawasaki, P. D. Jeffrey, R. Groisman, Y. Nakatani, S. Iwai, N. P. Pavletich and N. H. Thoma (2008). "Structural basis of UV DNA-damage recognition by the DDB1-DDB2 complex." Cell **135**(7): 1213-1223.
Segref, A., K. Sharma, V. Doye, A. Hellwig, J. Huber, R. Luhrmann and E. Hurt (1997). "Mex67p, a novel factor for nuclear mRNA export, binds to both poly(A)+ RNA and nuclear pores." EMBO J **16**(11): 3256-3271.
Shatkin, A. J. (1976). "Capping of eucaryotic mRNAs." Cell **9**(4 PT 2): 645-653.
Shatkin, A. J. and J. L. Manley (2000). "The ends of the affair: capping and polyadenylation." Nat Struct Biol **7**(10): 838-842.

Sheets, M. D., S. C. Ogg and M. P. Wickens (1990). "Point mutations in AAUAAA and the poly (A) addition site: effects on the accuracy and efficiency of cleavage and polyadenylation in vitro." Nucleic Acids Res **18**(19): 5799-5805.
Shell, S. A., C. Hesse, S. M. Morris, Jr. and C. Milcarek (2005). "Elevated levels of the 64-kDa cleavage stimulatory factor (CstF-64) in lipopolysaccharide-stimulated macrophages influence gene expression and induce alternative poly(A) site selection." J Biol Chem **280**(48): 39950-39961.
Shen, V., H. Liu, S. W. Liu, X. Jiao and M. Kiledjian (2008). "DcpS scavenger decapping enzyme can modulate pre-mRNA splicing." RNA **14**(6): 1132-1142.
Shi, Y., D. C. Di Giammartino, D. Taylor, A. Sarkeshik, W. J. Rice, J. R. Yates, 3rd, J. Frank and J. L. Manley (2009). "Molecular architecture of the human pre-mRNA 3' processing complex." Mol Cell **33**(3): 365-376.
Shi, Y., P. Kirwan and F. J. Livesey (2012). "Directed differentiation of human pluripotent stem cells to cerebral cortex neurons and neural networks." Nat Protoc **7**(10): 1836-1846.
Shi, Y. and J. L. Manley (2015). "The end of the message: multiple protein-RNA interactions define the mRNA polyadenylation site." Genes Dev **29**(9): 889-897.
Shimazu, T., S. Horinouchi and M. Yoshida (2007). "Multiple histone deacetylases and the CREB-binding protein regulate pre-mRNA 3'-end processing." J Biol Chem **282**(7): 4470-4478.
Shoemaker, C. J. and R. Green (2012). "Translation drives mRNA quality control." Nat Struct Mol Biol **19**(6): 594-601.
Shuman, S. (2001). "Structure, mechanism, and evolution of the mRNA capping apparatus." Prog Nucleic Acid Res Mol Biol **66**: 1-40.
Sikorski, T. W. and S. Buratowski (2009). "The basal initiation machinery: beyond the general transcription factors." Curr Opin Cell Biol **21**(3): 344-351.
Simms, C. L., E. N. Thomas and H. S. Zaher (2017). "Ribosome-based quality control of mRNA and nascent peptides." Wiley Interdiscip Rev RNA **8**(1).
Singh, G., G. Pratt, G. W. Yeo and M. J. Moore (2015). "The Clothes Make the mRNA: Past and Present Trends in mRNP Fashion." Annu Rev Biochem **84**: 325-354.
Siwaszek, A., M. Ukleja and A. Dziembowski (2014). "Proteins involved in the degradation of cytoplasmic mRNA in the major eukaryotic model systems." RNA Biol **11**(9): 1122-1136.
Smith, T. F., C. Gaitatzes, K. Saxena and E. J. Neer (1999). "The WD repeat: a common architecture for diverse functions." Trends Biochem Sci **24**(5): 181-185.
Stark, H. (2010). "GraFix: stabilization of fragile macromolecular complexes for single particle cryo-EM." Methods Enzymol **481**: 109-126.
Stewart, M. (2010). "Nuclear export of mRNA." Trends Biochem Sci **35**(11): 609-617.
Stewart, M. (2019). "Polyadenylation and nuclear export of mRNAs." J Biol Chem **294**(9): 2977-2987.
Stirnimann, C. U., E. Petsalaki, R. B. Russell and C. W. Muller "WD40 proteins propel cellular networks." Trends Biochem Sci **35**(10): 565-574.
Stojko, J., A. Dupin, S. Chaignepain, L. Beaurepaire, A. Vallet-Courbin, A. Van Dorsselaer, J. M. Schmitter, L. Minvielle-Sebastia, S. Fribourg and S. Cianferani (2017). "Structural characterization of the yeast CF IA complex through a combination of mass spectrometry approaches." International Journal of Mass Spectrometry **420**: 57-66.
Strambio-De-Castillia, C., M. Niepel and M. P. Rout (2010). "The nuclear pore complex: bridging nuclear transport and gene regulation." Nat Rev Mol Cell Biol **11**(7): 490-501.
Strasser, K. and E. Hurt (2000). "Yra1p, a conserved nuclear RNA-binding protein, interacts directly with Mex67p and is required for mRNA export." EMBO J **19**(3): 410-420.
Strasser, K. and E. Hurt (2001). "Splicing factor Sub2p is required for nuclear mRNA export through its interaction with Yra1p." Nature **413**(6856): 648-652.
Strasser, K., S. Masuda, P. Mason, J. Pfannstiel, M. Oppizzi, S. Rodriguez-Navarro, A. G. Rondon, A. Aguilera, K. Struhl, R. Reed and E. Hurt (2002). "TREX is a conserved complex coupling transcription with messenger RNA export." Nature **417**(6886): 304-308.
Sullivan, K. D., M. Steiniger and W. F. Marzluff (2009). "A core complex of CPSF73, CPSF100, and Symplekin may form two different cleavage factors for processing of poly(A) and histone mRNAs." Mol Cell **34**(3): 322-332.

Sun, Y., K. Hamilton and L. Tong (2020). "Recent molecular insights into canonical pre-mRNA 3'-end processing." Transcription **11**(2): 83-96.
Sun, Y., Y. Zhang, W. S. Aik, X. C. Yang, W. F. Marzluff, T. Walz, Z. Dominski and L. Tong (2020). "Structure of an active human histone pre-mRNA 3'-end processing machinery." Science **367**(6478): 700-703.
Sun, Y., Y. Zhang, K. Hamilton, J. L. Manley, Y. Shi, T. Walz and L. Tong (2018). "Molecular basis for the recognition of the human AAUAAA polyadenylation signal." Proc Natl Acad Sci U S A **115**(7): E1419-E1428.
Swanson, M. S., T. Y. Nakagawa, K. LeVan and G. Dreyfuss (1987). "Primary structure of human nuclear ribonucleoprotein particle C proteins: conservation of sequence and domain structures in heterogeneous nuclear RNA, mRNA, and pre-rRNA-binding proteins." Mol Cell Biol **7**(5): 1731-1739.
Tacahashi, Y., S. Helmling and C. L. Moore (2003). "Functional dissection of the zinc finger and flanking domains of the Yth1 cleavage/polyadenylation factor." Nucleic Acids Res **31**(6): 1744-1752.
Takagaki, Y., C. C. MacDonald, T. Shenk and J. L. Manley (1992). "The human 64-kDa polyadenylylation factor contains a ribonucleoprotein-type RNA binding domain and unusual auxiliary motifs." Proc Natl Acad Sci U S A **89**(4): 1403-1407.
Takagaki, Y. and J. L. Manley (1994). "A polyadenylation factor subunit is the human homologue of the Drosophila suppressor of forked protein." Nature **372**(6505): 471-474.
Takagaki, Y. and J. L. Manley (1997). "RNA recognition by the human polyadenylation factor CstF." Mol Cell Biol **17**(7): 3907-3914.
Takagaki, Y. and J. L. Manley (2000). "Complex protein interactions within the human polyadenylation machinery identify a novel component." Mol Cell Biol **20**(5): 1515-1525.
Takagaki, Y., J. L. Manley, C. C. MacDonald, J. Wilusz and T. Shenk (1990). "A multisubunit factor, CstF, is required for polyadenylation of mammalian pre-mRNAs." Genes Dev **4**(12A): 2112-2120.
Takagaki, Y., L. C. Ryner and J. L. Manley (1988). "Separation and characterization of a poly(A) polymerase and a cleavage/specificity factor required for pre-mRNA polyadenylation." Cell **52**(5): 731-742.
Takagaki, Y., L. C. Ryner and J. L. Manley (1989). "Four factors are required for 3'-end cleavage of pre-mRNAs." Genes Dev **3**(11): 1711-1724.
Tarun, S. Z., Jr. and A. B. Sachs (1996). "Association of the yeast poly(A) tail binding protein with translation initiation factor eIF-4G." EMBO J **15**(24): 7168-7177.
Terry, L. J., E. B. Shows and S. R. Wente (2007). "Crossing the nuclear envelope: hierarchical regulation of nucleocytoplasmic transport." Science **318**(5855): 1412-1416.
Tian, B., J. Hu, H. Zhang and C. S. Lutz (2005). "A large-scale analysis of mRNA polyadenylation of human and mouse genes." Nucleic Acids Res **33**(1): 201-212.
Tian, B. and J. L. Manley (2013). "Alternative cleavage and polyadenylation: the long and short of it." Trends Biochem Sci **38**(6): 312-320.
Tian, B. and J. L. Manley (2017). "Alternative polyadenylation of mRNA precursors." Nat Rev Mol Cell Biol **18**(1): 18-30.
Topalian, S. L., S. Kaneko, M. I. Gonzales, G. L. Bond, Y. Ward and J. L. Manley (2001). "Identification and functional characterization of neo-poly(A) polymerase, an RNA processing enzyme overexpressed in human tumors." Mol Cell Biol **21**(16): 5614-5623.
Tresaugues, L., P. Stenmark, H. Schuler, S. Flodin, M. Welin, T. Nyman, M. Hammarstrom, M. Moche, S. Graslund and P. Nordlund (2008). "The crystal structure of human cleavage and polyadenylation specific factor-5 reveals a dimeric Nudix protein with a conserved catalytic site." Proteins **73**(4): 1047-1052.
Tsuboi, T., K. Kuroha, K. Kudo, S. Makino, E. Inoue, I. Kashima and T. Inada (2012). "Dom34:Hbs1 Plays a General Role in Quality-Control Systems by Dissociation of a Stalled Ribosome at the 3 ' End of Aberrant mRNA." Molecular Cell **46**(4): 518-529.
Tutucci, E. and F. Stutz (2011). "Keeping mRNPs in check during assembly and nuclear export." Nat Rev Mol Cell Biol **12**(6): 377-384.
Vannini, A. and P. Cramer (2012). "Conservation between the RNA polymerase I, II, and III transcription initiation machineries." Mol Cell **45**(4): 439-446.

Varani, G. and K. Nagai (1998). "RNA recognition by RNP proteins during RNA processing." Annu Rev Biophys Biomol Struct **27**: 407-445.
Vaughn, J. L., R. H. Goodwin, G. J. Tompkins and P. Mccawley (1977). "Establishment of 2 Cell Lines from Insect Spodoptera-Frugiperda (Lepidoptera-Noctuidae)." In Vitro-Journal of the Tissue Culture Association **13**(4): 213-217.
Venkataraman, K., K. M. Brown and G. M. Gilmartin (2005). "Analysis of a noncanonical poly(A) site reveals a tripartite mechanism for vertebrate poly(A) site recognition." Genes Dev **19**(11): 1315-1327.
Villalba, A., O. Coll and F. Gebauer (2011). "Cytoplasmic polyadenylation and translational control." Curr Opin Genet Dev **21**(4): 452-457.
Voss, S. and A. Skerra (1997). "Mutagenesis of a flexible loop in streptavidin leads to higher affinity for the Strep-tag II peptide and improved performance in recombinant protein purification." Protein Eng **10**(8): 975-982.
Wahl, M. C., C. L. Will and R. Luhrmann (2009). "The spliceosome: design principles of a dynamic RNP machine." Cell **136**(4): 701-718.
Wahle, E. (1991). "A novel poly(A)-binding protein acts as a specificity factor in the second phase of messenger RNA polyadenylation." Cell **66**(4): 759-768.
Wahle, E. (1991). "Purification and characterization of a mammalian polyadenylate polymerase involved in the 3' end processing of messenger RNA precursors." J Biol Chem **266**(5): 3131-3139.
Wahle, E. (1995). "Poly(A) tail length control is caused by termination of processive synthesis." J Biol Chem **270**(6): 2800-2808.
Wahle, E. and W. Keller (1992). "The biochemistry of 3'-end cleavage and polyadenylation of messenger RNA precursors." Annu Rev Biochem **61**: 419-440.
Wahle, E., A. Lustig, P. Jeno and P. Maurer (1993). "Mammalian poly(A)-binding protein II. Physical properties and binding to polynucleotides." J Biol Chem **268**(4): 2937-2945.
Wahle, E. and U. Ruegsegger (1999). "3'-End processing of pre-mRNA in eukaryotes." FEMS Microbiol Rev **23**(3): 277-295.
Wahle, E. and G. S. Winkler (2013). "RNA decay machines: deadenylation by the Ccr4-not and Pan2-Pan3 complexes." Biochim Biophys Acta **1829**(6-7): 561-570.
Walther, T. N., T. H. Wittop Koning, D. Schumperli and B. Muller (1998). "A 5'-3' exonuclease activity involved in forming the 3' products of histone pre-mRNA processing in vitro." RNA **4**(9): 1034-1046.
Wang, L., C. R. Eckmann, L. C. Kadyk, M. Wickens and J. Kimble (2002). "A regulatory cytoplasmic poly(A) polymerase in Caenorhabditis elegans." Nature **419**(6904): 312-316.
Wang, X. and T. M. Tanaka Hall (2001). "Structural basis for recognition of AU-rich element RNA by the HuD protein." Nat Struct Biol **8**(2): 141-145.
Waterhouse, A. M., J. B. Procter, D. M. Martin, M. Clamp and G. J. Barton (2009). "Jalview Version 2--a multiple sequence alignment editor and analysis workbench." Bioinformatics **25**(9): 1189-1191.
Weill, L., E. Belloc, F. A. Bava and R. Mendez (2012). "Translational control by changes in poly(A) tail length: recycling mRNAs." Nat Struct Mol Biol **19**(6): 577-585.
Weitzer, S. and J. Martinez (2007). "The human RNA kinase hClp1 is active on 3' transfer RNA exons and short interfering RNAs." Nature **447**(7141): 222-226.
Wells, D. and L. Kedes (1985). "Structure of a human histone cDNA: evidence that basally expressed histone genes have intervening sequences and encode polyadenylylated mRNAs." Proc Natl Acad Sci U S A **82**(9): 2834-2838.
Weng, T., J. Ko, C. P. Masamha, Z. Xia, Y. Xiang, N. Y. Chen, J. G. Molina, S. Collum, T. C. Mertens, F. Luo, K. Philip, J. Davies, J. Huang, C. Wilson, R. A. Thandavarayan, B. A. Bruckner, S. S. Jyothula, K. A. Volcik, L. Li, L. Han, W. Li, S. Assassi, H. Karmouty-Quintana, E. J. Wagner and M. R. Blackburn (2019). "Cleavage factor 25 deregulation contributes to pulmonary fibrosis through alternative polyadenylation." J Clin Invest **129**(5): 1984-1999.
West, S., N. Gromak and N. J. Proudfoot (2004). "Human 5' --> 3' exonuclease Xrn2 promotes transcription termination at co-transcriptional cleavage sites." Nature **432**(7016): 522-525.

West, S. and N. J. Proudfoot (2008). "Human Pcf11 enhances degradation of RNA polymerase II-associated nascent RNA and transcriptional termination." Nucleic Acids Res **36**(3): 905-914.
Whitelaw, E. and N. Proudfoot (1986). "Alpha-thalassaemia caused by a poly(A) site mutation reveals that transcriptional termination is linked to 3' end processing in the human alpha 2 globin gene." EMBO J **5**(11): 2915-2922.
Wickham, T. J., T. Davis, R. R. Granados, M. L. Shuler and H. A. Wood (1992). "Screening of insect cell lines for the production of recombinant proteins and infectious virus in the baculovirus expression system." Biotechnol Prog **8**(5): 391-396.
Will, C. L. and R. Luhrmann (2011). "Spliceosome structure and function." Cold Spring Harb Perspect Biol **3**(7).
Wilusz, J. and T. Shenk (1988). "A 64 kd nuclear protein binds to RNA segments that include the AAUAAA polyadenylation motif." Cell **52**(2): 221-228.
Wilusz, J., T. Shenk, Y. Takagaki and J. L. Manley (1990). "A multicomponent complex is required for the AAUAAA-dependent cross-linking of a 64-kilodalton protein to polyadenylation substrates." Mol Cell Biol **10**(3): 1244-1248.
Wittkopp, N., E. Huntzinger, C. Weiler, J. Sauliere, S. Schmidt, M. Sonawane and E. Izaurralde (2009). "Nonsense-mediated mRNA decay effectors are essential for zebrafish embryonic development and survival." Mol Cell Biol **29**(13): 3517-3528.
Wolf, J., E. Valkov, M. D. Allen, B. Meineke, Y. Gordiyenko, S. H. McLaughlin, T. M. Olsen, C. V. Robinson, M. Bycroft, M. Stewart and L. A. Passmore (2014). "Structural basis for Pan3 binding to Pan2 and its function in mRNA recruitment and deadenylation." EMBO J **33**(14): 1514-1526.
Xia, Z., L. A. Donehower, T. A. Cooper, J. R. Neilson, D. A. Wheeler, E. J. Wagner and W. Li (2014). "Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3'-UTR landscape across seven tumour types." Nat Commun **5**: 5274.
Xiang, K., T. Nagaike, S. Xiang, T. Kilic, M. M. Beh, J. L. Manley and L. Tong (2010). "Crystal structure of the human symplekin-Ssu72-CTD phosphopeptide complex." Nature **467**(7316): 729-733.
Xie, J. W., G. Kozlov and K. Gehring (2014). "The "tale" of poly(A) binding protein: The MLLE domain and PAM2-containing proteins." Biochimica Et Biophysica Acta-Gene Regulatory Mechanisms **1839**(11): 1062-1068.
Xu, R., H. Zhao, R. D. Dinkins, X. Cheng, G. Carberry and Q. Q. Li (2006). "The 73 kD subunit of the cleavage and polyadenylation specificity factor (CPSF) complex affects reproductive development in Arabidopsis." Plant Mol Biol **61**(4-5): 799-815.
Xu, X. Q., N. Perebaskine, L. Minvielle-Sebastia, S. Fribourg and C. D. Mackereth (2015). "Chemical shift assignments of a new folded domain from yeast Pcf11." Biomolecular Nmr Assignments **9**(2): 421-425.
Yamashita, A., T. C. Chang, Y. Yamashita, W. Zhu, Z. Zhong, C. Y. Chen and A. B. Shyu (2005). "Concerted action of poly(A) nucleases and decapping enzyme in mammalian mRNA turnover." Nat Struct Mol Biol **12**(12): 1054-1063.
Yan, C., R. Wan and Y. Shi (2019). "Molecular Mechanisms of pre-mRNA Splicing through Structural Biology of the Spliceosome." Cold Spring Harb Perspect Biol **11**(1).
Yang, Q., M. Coseno, G. M. Gilmartin and S. Doublie (2011). "Crystal structure of a human cleavage factor CFI(m)25/CFI(m)68/RNA complex provides an insight into poly(A) site recognition and RNA looping." Structure **19**(3): 368-377.
Yang, Q. and S. Doublie (2011). "Structural biology of poly(A) site definition." Wiley Interdiscip Rev RNA **2**(5): 732-747.
Yang, Q., G. M. Gilmartin and S. Doublie (2010). "Structural basis of UGUA recognition by the Nudix protein CFI(m)25 and implications for a regulatory role in mRNA 3' processing." Proc Natl Acad Sci U S A **107**(22): 10062-10067.
Yang, W., P. L. Hsu, F. Yang, J. E. Song and G. Varani (2018). "Reconstitution of the CstF complex unveils a regulatory role for CstF-50 in recognition of 3 '-end processing signals." Nucleic Acids Research **46**(2): 493-503.

Yang, X. C., K. D. Sullivan, W. F. Marzluff and Z. Dominski (2009). "Studies of the 5' exonuclease and endonuclease activities of CPSF-73 in histone pre-mRNA processing." Mol Cell Biol **29**(1): 31-42.
Yang, X. C., B. Xu, I. Sabath, L. Kunduru, B. D. Burch, W. F. Marzluff and Z. Dominski (2011). "FLASH is required for the endonucleolytic cleavage of histone pre-mRNAs but is dispensable for the 5' exonucleolytic degradation of the downstream cleavage product." Mol Cell Biol **31**(7): 1492-1502.
Ye, K. and D. J. Patel (2005). "RNA silencing suppressor p21 of Beet yellows virus forms an RNA binding octameric ring structure." Structure **13**(9): 1375-1384.
Yeh, H. S. and J. Yong (2016). "Alternative Polyadenylation of mRNAs: 3'-Untranslated Region Matters in Gene Expression." Mol Cells **39**(4): 281-285.
Yudkovsky, N., J. A. Ranish and S. Hahn (2000). "A transcription reinitiation intermediate that is stabilized by activator." Nature **408**(6809): 225-229.
Zaret, K. S. and F. Sherman (1982). "DNA sequence required for efficient transcription termination in yeast." Cell **28**(3): 563-573.
Zarudnaya, M. I., I. M. Kolomiets, A. L. Potyahaylo and D. M. Hovorun (2003). "Downstream elements of mammalian pre-mRNA polyadenylation signals: primary, secondary and higher-order structures." Nucleic Acids Res **31**(5): 1375-1386.
Zenklusen, D., P. Vinciguerra, J. C. Wyss and F. Stutz (2002). "Stable mRNP formation and export require cotranscriptional recruitment of the mRNA export factors Yra1p and Sub2p by Hpr1p." Mol Cell Biol **22**(23): 8241-8253.
Zhang, D. W., J. B. Rodriguez-Molina, J. R. Tietjen, C. M. Nemec and A. Z. Ansari (2012). "Emerging Views on the CTD Code." Genet Res Int **2012**: 347214.
Zhang, Y. X., Y. D. Sun, Y. S. Shi, T. Walz and L. Tong (2020). "Structural Insights into the Human Pre-mRNA 3 '-End Processing Machinery." Molecular Cell **77**(4): 800-+.
Zhang, Z., J. Fu and D. S. Gilmour (2005). "CTD-dependent dismantling of the RNA polymerase II elongation complex by the pre-mRNA 3'-end processing factor, Pcf11." Genes Dev **19**(13): 1572-1580.
Zhang, Z. and D. S. Gilmour (2006). "Pcf11 is a termination factor in Drosophila that dismantles the elongation complex by bridging the CTD of RNA polymerase II to the nascent transcript." Mol Cell **21**(1): 65-74.
Zhao, J., L. Hyman and C. Moore (1999). "Formation of mRNA 3' ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis." Microbiol Mol Biol Rev **63**(2): 405-445.
Zhao, W. and J. L. Manley (1998). "Deregulation of poly(A) polymerase interferes with cell growth." Mol Cell Biol **18**(9): 5010-5020.
Zhelkovsky, A., Y. Tacahashi, T. Nasser, X. He, U. Sterzer, T. H. Jensen, H. Domdey and C. Moore (2006). "The role of the Brr5/Ysh1 C-terminal domain and its homolog Syc1 in mRNA 3'-end processing in Saccharomyces cerevisiae." RNA **12**(3): 435-445.
Zheng, S. Q., E. Palovcak, J. P. Armache, K. A. Verba, Y. Cheng and D. A. Agard (2017). "MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy." Nat Methods **14**(4): 331-332.
Zhu, Y., X. Y. Wang, E. Forouzmand, J. S. Jeong, F. Qiao, G. A. Sowd, A. N. Engelman, X. H. Xie, K. J. Hertel and Y. S. Shi (2018). "Molecular Mechanisms for CFIm-Mediated Regulation of mRNA Alternative Polyadenylation." Molecular Cell **69**(1): 62-+.