Structural and biochemical studies of the transcription termination machinery

Peter Hsu

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor in Philosophy

University of Washington

2013

Reading committee:

Gabriele Varani, chair

Dustin Maly

Ronald Stenkamp

Program authorized to offer degree:

Chemistry

University of Washington

Abstract

Structural and biochemical studies of the transcription termination machinery

Peter Hsu

Chair of the Supervisory Committee:

Professor Gabriele Varani

Department of Chemistry and Biochemistry

RNA Polymerase II (PolII) dependent transcription of mRNAs is central to gene expression throughout eukaryotes. Transcription is a highly regulated process, with a defined initiation, elongation, and termination phase, all of which are controlled by multiple *trans*-acting protein factors and *cis*-acting elements on the template DNA and transcribed RNA. Extensive work has shown that the phosphorylation state of the C-terminal domain (CTD) of PolII plays a central role in the recruitment of *trans*-acting factors during all phases of transcription. Aberrant phosphorylation can result in a lethal phenotype in yeasts, therefore implying that correct control of phosphorylation by kinases and phosphatases specific for the CTD is critical for life. In addition to the polymerase, conserved *cis*-acting sequence elements near the 3'-end on pre-mRNAs help to define and recruit various RNA processing machines to the transcription elongation complex in

order to properly process and package the pre-mRNA for export to the cytoplasm for translation.

In this thesis, I first review current knowledge regarding PolII CTD phosphorylation and its effects on the transcription elongation complex. Additionally, in this first chapter, I will also provide an overview of the 3'-end mRNA processing/transcription termination machinery. In the second part of my thesis, I will describe my doctoral work on the structural and biochemical characterization of Rtr1, a unique PolII CTD phosphatase that represents a novel new member of this class of enzymes. In my studies, I show that Rtr1 is a *bona fide* phosphatase of unique sequence and structure that is allosterically regulated by its own C-terminus. Additionally, I show that Rtr1 is a dual specificity phosphatase, with activities against both serine and tyrosine residues on the CTD. In chapter 3 of this thesis, I describe my work on the *in vitro* reconstitution of the Cleavage Stimulation Factor (CstF) responsible for the recognition of sequences downstream of the polyadenylation site on pre-mRNAs that help to define the 3'-end processing reaction that occur at the end of genes. Using highly purified proteins, I show for the first time that CstF is a dimer of trimers, with two copies of each subunit in the entire assembly. In addition, I show that CstF, as a complex, can bind to G/U rich RNAs with nanomolar affinities, in stark contrast with previous studies showing that singly purified proteins from the complex binding with much weaker affinities.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Gabriele Varani, for his guidance and support over the course of the last five and a half years. He has provided me with endless opportunities and near complete freedom to pursue my scientific interests. I would also like to thank Dr. Ning Zheng, who has been practically a second advisor for my doctorate. Without his help and insights, many of my projects would not be where they are today. I would also like to thank my reading committee and supervisory committee for their comments and insights both during committee meetings and during the revision of this thesis. I also thank my collaborator, Dr. Amber Mosley and her lab, for providing useful discussion as well as experimental support on the Rtr1 project. Many thanks also go to multiple members of the Varani and Zheng labs for their day-to-day help and support over the last five years. Last, but not least, I extend my thanks to my family, friends, and wife. Without their support, encouragement, and love this work would never have been completed.

DEDICATION

To my family and friends.

Table of Contents

List of figures

List of tables

Chapter 1 . An introduction to transcription termination

RNA Polymerase II (PolII) is a large protein machine, consisting of 12 subunits, responsible for the transcription of multiple classes of RNAs, the largest class being mRNAs. Given its role in the genesis of protein-coding genes, PolII is the central player in gene expression in eukaryotes. Transcription of mRNAs have a well defined start, middle, and end that is regulated by multiple factors that exist either as sequence elements on the DNA template and/or pre-mRNA, or as *trans*-acting protein factors that interact with the transcription machinery. Due to PolII's highly processive activity, a defined mechanism must be in place for the polymerase to terminate transcription at the end of a gene, otherwise the polymerase would remain engaged on the DNA template beyond its stop point. Two major factors that help define transcription end are: the phosphorylation state of PolII's C-terminal domain (CTD), as well as conserved elements in the 3' untranslated region (UTR) of the pre-mRNA that recruit the 3'-end processing machinery to process and polyadenylate the RNA. In this chapter, I will describe the physiological relevance of each CTD modification, as well as factors that recognize and regulate it. I will also describe how the cleavage and polyadenylation reaction that occurs at the 3'-end of genes facilitates the proper termination of PolII mediated transcription.

1.1 Combinatorial modification of PolII's CTD provide a landing pad for transcriptional regulatory proteins

The C-terminal domain of PolII belongs to Rpb1, the largest subunit of the polymerase. The CTD is unique in structure and sequence in that it is comprised of a hepta-peptide repeat sequence of $Tyr_1Ser_2Pro_3Thr_4Ser_5Pro_6Ser_7$, which is repeated up to 26 times in the

yeast *Saccharomyces cerevisiae*, and 52 times in humans (Meinhart et al., 2005). While the CTD repeats are generally well conserved, more distal repeats relative to the body of PolII, especially in humans, can diverge from this hepta-peptide sequence; the function of these distal repeats have not been extensively studied.

In all crystal structures of PolII no electron density is observed for the CTD, further reinforcing the conclusion that the CTD is largely disordered. However, the position of the last observed residue in Rpb1 shows that the CTD can extend from the side of the polymerase, not far from the active site of the polymerase. Given the CTD's length, and presumed flexibility, the CTD can potentially assume a position near the exit channel of the polymerase, bringing in factors that can modify the elongating RNA (Fig. 1.1). While it is assumed that the CTD is largely unstructured, solution studies have suggested that the CTD can assume some level of local structure, potentially adopting a tight alternating β-turn structure, which upon modification can be expanded to a more linear structure to allow proteins to bind (Cagas and Corden, 1995; Meinhart and Cramer, 2004).

Of the seven residues on a CTD repeat, five (Tyr_1/Ser_2/Thr_4/Ser_5/Ser_7) can be potentially phosphorylated, and the two prolines (Pro_3/Pro_6) can be subject to *cis-trans* isomerization by prolyl isomerases. It has been shown that all five of the phosphorylable residues can be phosphorylated *in vivo*, and each mark has been associated with a functional event during transcription (Hsin and Manley, 2012). Furthermore, Pro_6 has been shown to be subject to isomerization by prolyl isomerases, changing the local structure around the phospho-serines to allow/prevent factors from recognizing the mark. In addition to the phosphorylated markers on a single repeat, it has been shown that the

functional unit of the CTD is a diheptad repeat, further expanding the combinatorial possibility of modifications (Stiller and Cook, 2004). Subunits of the Integrator complex have been shown to recognize an overlapping marker of phosphorylated serine 7 on heptad 1, and phosphorylated serine 2 on the proceeding heptad (Egloff et al., 2010).

These observations have led to the proposal that the CTD heptad and its multiple repeats encode a "CTD code", in which at various points during transcription the modifications on the CTD can be changed dynamically to actively recruit various transcription regulatory/RNA processing factors to the transcription complex (Table 1.1) (Buratowski, 2009). Below, I will describe each phospho-marker and its relevance in the transcription process as studied in yeasts (unless noted otherwise).

1.1.1 Phospho-tyrosine 1

The tyrosine 1 phospho-marker is the most recently characterized phospho-residue along the CTD. While it was initially described as being associated with the DNA damage response because it is heavily phosphorylated by the Abl kinase (Baskaran et al., 1993, 1997), tyrosine 1's relevance in transcription was only recently established with the generation of antibodies against phospho-tyr1 (Tyr1P). In a recent study characterizing this residue, Tyr1P was found to be enriched at 3' ends of actively transcribing genes, with little to none found at the start of transcription; build up of this marker begins only after the polymerase engages into elongation mode (Mayer et al., 2012).

In silico modeling of several transcription termination factors known to bind the CTD (such as Nrd1 and Pcf11) with their known phospho-marker and Tyr1P showed that the presence of the Tyr1P marker in addition to the other phospho-mark would prevent

binding of these known termination factors to the CTD (Mayer et al., 2012). This model has led to the currently accepted view that the physiological function of Tyr1P during transcription is to act as an anti-termination marker, preventing the premature association of termination factors with the active polymerase.

While older work has shown that, in higher eukaryotes, the Abl and Arg proteins are able to phosphorylate Tyr1 as part of a DNA damage response, the transcriptionally relevant Tyr1 kinase(s) and phosphatase(s) have yet to be identified.

1.1.2 Phospho-serine 2

Phophorylation of serine 2 on the CTD is one of the best studied markers along the CTD (the other being serine 5) and its role in transcription is well established. Both serine 2 and serine 5 play critical roles in transcription and are in a "see-saw" type equilibrium with each other. Canonically, at the start of transcription near the 5' end of genes, serine 5 is highly phosphorylated, while serine 2 is unphosphorylated. As PolII engages past the promoter, and begins to enter the elongation phase of transcription, the kinases Bur1/2 and/or CTDK-I (P-TEFb in higher eukaryotes) begin to increasingly phosphorylate serine 2, producing a bivalent Ser2P/Ser5P mark on the CTD, commonly associated with elongation (Cho et al., 2001; Jones et al., 2004; Qiu et al., 2009; Zhou et al., 2000).

During the elongation phase, the CTD is heavily and heterogeneously phosphorylated on numerous sites in addition to Ser2P and Ser5P (additional markers discussed above and below) and can influence the recruitment and activity of the spliceosome. While older studies have largely focused on the effects of hyper-phosphorylated PolII CTD on co-transcriptional splicing of mRNAs (David et al., 2011;

McCracken et al., 1997; Millhouse and Manley, 2005 and others), recent work using a Ser2 to Ala (S2A) mutant in yeast has suggested that Ser2P influences splicing directly. It was shown that a PolII CTD S2A mutant failed to recruit both the U2AF65 protein and the U2 snRNA to transcription sites, and as a result, splicing was impaired in these cell lines (Gu et al., 2013). While it is clear that Ser2P directly influences the splicing reaction, it is unclear if this is due to Ser2P exerting a direct effect on assembly and/or enzymatic reaction, or if other factors were disrupted in the background of this mutant. Additional work is needed to determine the biochemical mechanism of this phospho-marker on mRNA splicing.

As PolII moves towards the end of genes, serine 5 is largely dephosphorylated, and the predominant marker becomes Ser2P. Ser2P assists in the recruitment of several transcription termination factors, including the proteins Pcf11 and Rtt103, which bind to phospho-specific forms of the CTD via their CTD interacting domain (CID). Work from our lab has shown that tandem repeats of the CTD phosphorylated on serine 2 is necessary for the efficient recruitment of these CID containing termination factors, via a CID-CID homodimerization mediated by neighboring Ser2P CTD repeats; disruption of these homodimerization contacts *in vivo* resulted in improper termination of transcription (Lunde et al., 2010). Pcf11 is believed to participate in 3'-end processing of the pre-mRNA and then remains associated with the CTD to help recruit the mRNA export machinery to the processed mRNA (Johnson et al., 2009). Rtt103 is primarily responsible for recruiting the 5'-3' exonuclease complex Rat1/Rai1 to the termination complex, which degrades any RNA made after the polyA site, and then physically displaces PolII

from the DNA template (Kim et al., 2004). These termination events are effectively coupled with the 3'-end processing reaction (described below).

Ser2P, after participating in these steps in transcription, is then erased by the Ser2P specific phosphatase Fcp1. Crystal structures of Fcp1 with various phosphate analogs show a mechanism by which a conserved aspartate acts as the nucleophile (Ghosh et al., 2008). However, it is unclear how Fcp1 recognizes its substrate as no structure of Fcp1 with a CTD peptide has been solved. Erasure of Ser2P to reset/recycle RNA PolII to an unphosphorylated state is likely essential, as disruption of Fcp1 in *Drosophila melanogaster* causes cells to undergo apoptosis (Schauer et al., 2009; Tombácz et al., 2009).

1.1.3 Phospho-threonine 4

Similarly to Tyr1P, phosphorylation of threonine 4 (Thr4P) was only recently verified as a bona fide *in vivo* CTD marker. While Thr4P's existence has been shown in both yeasts and higher-order eukaryotes, its primary function in transcription has yet to be established. Initial experiments using a conditional Rpb1 knockout (KO)/plasmid complementation assay showed that DT40 cells (chicken derived cells) that were rescued with a Rpb1 mutant that had all Thr4 residues on the CTD mutated to valine (T4V) began to die after 24 hours of induction, suggesting that Thr4 played a role in cell viability. Follow up experiments in the same study showed that cells lacking Thr4 as a phospho-acceptor lacked properly processed histone mRNAs resulting in reduced histone production and likely, cell lethality (Hsin et al., 2011).

Subsequent work from an independent group also characterized Thr4P's existence in both lower and higher eukaryotes, but suggested that phosphorylation of threonine 4 plays a role in transcription elongation rather than a mRNA class specific function as previously studied. Their transcriptome wide studies in human cell lines using a T4A mutant showed accumulation of PolII in the body of genes across the transcriptome, with very low abundance at 3' ends, suggesting that PolII could not enter the elongation phase properly. Whether the discrepancies in these studies are due to different cell line backgrounds or mutant variants remains to be seen (Hintermair et al., 2012).

Published studies also are conflicted to the identity of the Thr4 kinase. Manley and colleagues showed that the human Ser2 kinase, P-TEFb, could potentially phosphorylate Thr4 since treatment with Cdk9 inhibitors reduced Thr4 phospho levels. However, later work contradicted these results showing that *in vitro*, Cdk9 displayed no activity towards Thr4, while Plk3 could phosphorylate both *in vitro* and *in vivo* (Hintermair et al., 2012; Hsin et al., 2011). The phosphatase responsible for the erasure of Thr4P remains unknown.

1.1.4 Phospho-serine 5

Serine 5 phosphorylation (Ser5P) is another well studied and well characterized phosphorylation mark on the CTD, and arguably the most important due to its role in the start of transcription. Phosphorylation of serine 5 occurs near the start of transcription and is performed by the Kin28 subunit of the general transcription factor TFIIH (Rodriguez et al., 2000; Trigon et al., 1998). Marking of serine 5 signals for promoter clearance of the polymerase by disrupting the interaction of PolII with the transcription initiation complex

(most notably the Mediator complex), thereby promoting polymerase escape from the promoter to enter transcription (Max et al., 2007).

The biogenesis of an mRNA undergoes several post-transcriptional modifications during maturation, the first being a 5' capping step that protects an mRNA from degradation in the nucleus. Via several enzymatic reactions, the capping machinery adds a guanine residue to the 5' end of the pre-mRNA, which is then further modified by a methylation at the N7 position on the guanine base (reviewed in Ghosh and Lima, 2010). Recruitment of several key enzymes in this essential step of mRNA synthesis are mediated by Ser5P, most notably the guanalyltransferase component of the capping complex for which structures have been determined of both yeast and animal homologs (Fabrega et al., 2003; Ghosh et al., 2011). Interestingly, despite the overall structure of the guanalyltransferase being conserved in both yeasts and animals, structures of both the yeast and animal versions show different modes of Ser5P binding (Fig 1.2). These results suggest that while the capping enzymes diverged evolutionarily, coupling of mRNA capping to Ser5P is an essential mechanism (Ghosh et al., 2011).

Following the capping step and shift of the polymerase into transcription elongation, Ser5P becomes increasingly dephosphorylated while serine 2 becomes increasingly phosphorylated. A number of phosphatases specific for Ser5P have been identified, including Ssu72 and Scp1. Ssu72 is a member of the APT complex, which is associated at the 3'-end of genes and assists in the 3'-end processing of mRNAs (Krishnamurthy et al., 2004; Nedea et al., 2003). Scp1 belongs to a family of small CTD phosphatases and is used to suppress the expression of neuronal genes in non-neuronal cells (Yeo et al., 2003), likely by removing Ser5P prior to promoter clearance. Neither of

these enzymes display the co-localization with PolII during transcription to act as a phosphatase to remove Ser5P in response to accumulation of Ser2P. More recently, a new phosphatase enzyme was discovered, named Rtr1, which showed the proper localization across genes with PolII, and could selectively dephosphorylate Ser5P in *in vitro* assays. It has been suggested that this is the elusive transcription transition phosphatase (Mosley et al., 2009). Additional work also shows that its human homolog, RPAP2, contains identical activity profiles against Ser5P (Egloff et al.). However, a crystal structure and accompanying biochemical work brought this conclusion into question as no active site was found, and no activity could be detected in highly purified recombinant protein (Xiang et al., 2012a). This subject will be discussed at length in chapter 2.

While a large population of Ser5P is removed during the transition to termination, a small population remains that must be removed prior to recycling of the polymerase. Ssu72, in association with Pta1 (Symplekin in animals), is likely the responsible phosphatase (Ghazy et al., 2009). Crystal structures of Ssu72 in complex with Symplekin and a Ser5P peptide (Fig. 1.3) reveal that Ssu72 recognizes a unique form of the CTD at the 3' end of genes: a *cis*-proline conformation on Pro_6 of the CTD is necessary for the association of Ssu72 (Xiang et al., 2010). Ess1, a prolyl isomerase, catalyzes this *cis* to *trans* conversion of proline; addition of this isomerase to *in vitro* phosphatase reactions increases the rate at which Ssu72 dephosphorylates Ser5P (Werner-Allen et al., 2011; Zhang et al., 2012b). While no influence of *cis-trans* prolines has been observed yet for Ser2P, it is clear from the work on serine 5 that the prolines can play a role in phosphorylation pattern of serines on the CTD, adding yet another layer of complexity to the CTD code.

1.1.5 Phospho-serine 7

Serine 7 phosphorylation (Ser7P) was identified as a transcriptionally relevant CTD phospho-marker in 2007. Initial studies identifying Ser7P's transcriptional role were done in HEK293 cells using complementation assays with either wild-type PolII or a mutant harboring all serine 7 residues mutated to Ala (S7A) (Chapman et al., 2007; Egloff et al., 2007). This work in S7A lines showed that while transcription of protein coding genes were unhampered by this mutation, clear defects in the processing of small nuclear RNAs (snRNA) were observed. Further supporting this finding, GST pulldown assays using *in vitro* phosphorylated CTD showed a loss of interaction with the S7A construct with the Integrator complex, a large protein complex involved in snRNA processing known to bind phosphorylated CTD.

Ser7P was suspected to be a metazoan-only transcription marker because the Integrator complex is only conserved in higher eukaryotes and not in yeasts. However, surprisingly, the transcriptionally relevant kinase responsible for phosphorylating serine 7 was identified as TFIIH in both yeasts and human cells, suggesting that the function of this marker is universally conserved (Akhtar et al., 2009; Boeing et al., 2010). Genome wide analysis of CTD phosphorylation in yeast also showed accumulation of Ser7P at introns and at sites in which the Nrd1 CTD interacting protein is enriched, suggesting a role for Ser7P in the processing of cryptic unspliced transcripts (CUTs) and snoRNAs (Kim et al., 2010a). Structural analysis of human TCERG1, a transcription elongation regulator known to bind both phospho-CTD and splicing factors, showed that TCERG1

requires the phosphorylation of all three serines on the CTD, potentially expanding the role of Ser7P into mRNA transcription (Liu et al., 2013).

Currently, the only phosphatase known to dephosphorylate Ser7P is the Ser5P phosphatase Ssu72. In yeast, depletion of Ssu72 resulted in an increase of the Ser7P marker in cells, and *in vitro* phosphatase assays using purified recombinant Ssu72 showed selective dephosphorylation of both Ser5P and Ser7P (Bataille et al., 2012; Zhang et al., 2012a). Crystal structures of a Symplekin-Ssu72-Ser7P peptide complex showed that Ssu72 binds to a Ser7P CTD peptide with an unusual and constrained geometry (Fig. 1.3), and that the phosphatase had a preference for selecting Ser5P over Ser7P as a substrate (Xiang et al., 2012b). These experiments suggest that serine 7 is not a preferred substrate and that potentially another phosphatase exists to target Ser7P.

1.2 mRNA 3'-end processing and transcription termination

Prior to the export of the mRNA to the cytoplasm, the pre-mRNA must undergo a final processing step that occurs co-transcriptionally, in which the nascent RNA is cleaved in the 3'-UTR and then polyadenylated; this process is termed 3'-end processing (Fig. 1.4). Polyadenylation is necessary for the efficient export, translation, and stability of the mRNA in the cytoplasm. While the reaction is relatively simple, consisting of an endonuclease which cleaves the RNA creating a 3'-OH end for Polyadenylate Polymerase (PAP) to synthesize the polyA tail, this process *in vivo* requires the recruitment of over a dozen characterized protein factors to the 3'-UTR. More recent mass spectrometry studies have implicated an even larger repertoire of proteins involved

in 3'-end processing, although many of these factors have yet to be fully characterized (Shi et al., 2009).

The 3'-end processing reaction also effectively helps to define and commit transcription into the termination phase (Fig. 1.4). The 3'-end cleavage reaction that frees the pre-mRNA from the elongating polymerase effectively creates a new 5'-end with a free 5' phosphate, which is then accessible to exonucleases to degrade. Below, I will describe in detail the complex protein machinery necessary to execute the 3'-end processing reaction, as well as the linked mechanism for the recruitment of transcription termination factors.

1.2.1 Protein-protein/protein-RNA interactions define 3'-end processing sites

Despite the relative simplicity of 3'-end maturation, a highly complex and coordinated protein machinery is necessary for the proper recognition of conserved RNA signals that help to define the position of the cleavage site on the pre-mRNA (Table 1.2). In animals, several conserved sequence elements along the pre-mRNA (*cis*-acting elements) help to define the position of the cleavage and polyadenylation reaction. Yeasts have similar sequence elements, however, they are less well conserved. For brevity, only animal sequences will be discussed.

The first and foremost is the canonical polyadenylation (polyA) site, which is a hexamer of the sequence AAUAAA. Genomic studies have shown that while the large majority of polyadenylated genes all carry and are processed at this signal (75%), a significant fraction of mRNAs also use AUUAAA as a polyadenylation site (~20%) (Tian et al., 2005). In addition to the second position in the hexamer being variable, most

mRNAs contain more than a single polyA site scattered across the 3'-UTR (Legendre and Gautheret, 2003; Mayr and Bartel, 2009; Ozsolak et al., 2010). Generally, the most distal (relative to the open reading frame) sites are used by default, but emerging evidence for alternative polyadenylation is showing that more proximal sites can be used as well resulting in longer or shorter 3'-UTRs depending on the site selected. 3'-UTRs have been shown to be binding sites for microRNA (miRNA) silencing of genes, and thus alternative polyadenylation can be used as a mechanism to escape regulation (Mayr and Bartel, 2009).

The polyA site also helps to define the position of the cleavage site on the pre-mRNA. Typically, the cleavage site is 10-30 nucleotides downstream of the polyA site (Fitzgerald and Shenk, 1981) and upstream of the G/U-rich positioning element (described below). The sequence surrounding the cleavage site is generally non-conserved (Gilmartin et al., 1995) and thus cleavage position is likely defined by the protein-protein/protein-RNA interactions of the 3'-end processing complex.

The G/U-rich downstream element (DSE) is located 20-30 nucleotides downstream of the cleavage site, and is necessary for efficient cleavage (McDevitt et al., 1986). Unlike the polyA site, the DSE lacks a consensus sequence of any sort, and is only defined as being rich in guanine and uracil residues (Cañadillas and Varani, 2003; Salisbury et al., 2006). Strength and selection of a polyA site has been linked to the G/U-rich DSE, where the percent content of uracil in the DSE is greater in "true" polyA sites, as opposed to weaker or random AAUAAA hexamers in the RNA sequence (Legendre and Gautheret, 2003). Due to its loose sequence requirements, the DSE is highly tolerant to mutations. However, the position of the element is more important as *in vitro* analyses

using extracts have shown that deletion or insertion of small sequence elements can drastically affect cleavage (McDevitt et al., 1986).

Historically, the polyA, DSE, and cleavage site defined the minimal necessary elements for cleavage and polyadenylation on model substrates, but an upstream sequence element (USE) relative to the polyA site has been identified in recent years that is functionally important for site selection as well. The USE is defined as being U-rich, although a strong consensus sequence of UGUA has also been reported (Hu et al., 2005; Yang et al., 2010).

While the overall 3'-end processing reaction is conserved between yeasts and animals, the architecture of the protein complexes involved are marginally divergent from one another (Fig. 1.5), likely due to the fact that the RNA elements they recognize are not well conserved between species (Zhao et al., 1999). In higher eukaryotes, the efficient cleavage of the pre-mRNA can be minimally reconstituted *in vitro* with the cleavage and polyadenylation specificity factor (CPSF), cleavage stimulatory factor (CstF), and cleavage factor I and II complexes (Takagaki et al., 1989), while the polyadenylation reaction needs only CPSF and PAP (Murthy and Manley, 1992).

Cleavage and Polyadenylation Specificity Factor (*CPSF)*

CPSF is a five subunit protein complex consisting of the proteins CPSF160, 100, 73, 30, and hFip1. CPSF was initially identified via fractionation experiments and was shown biochemically to be involved in both steps of 3'-end processing (Murthy and Manley, 1992; Takagaki et al., 1989). The largest factor, CPSF160, is a tri-β-propeller protein that shares much structural homology to DDB1, a member of the multisubunit cullin-ubiquitin

E3 ligases (Li et al., 2006). CPSF160 has been proposed to be the polyA site recognition protein via UV crosslinking experiments (Murthy and Manley, 1995). Further validating this finding, experiments with the *Saccharomyces cerevisiae* homolog of the protein, Yhh1, also showed direct binding to yeast processing signals (Dichtl et al., 2002). However, the molecular basis of polyA site recognition has remained elusive due to CPSF160 harboring no known RNA binding domains. In addition, this large protein also serves as the major scaffolding factor for numerous other proteins in CPSF, as well as bridging CPSF to CstF via a direct interaction with CstF77 (described below).

CPSF73 is the endonuclease responsible for the cleavage the pre-mRNA during 3'-end processing (Fig. 1.5). Structures of the N-terminus of CPSF73 reveal a β-CASP domain, known to bind nucleic acids, as well as a metallo-β-lactamase domain. The domains bind zinc ions in a manner reminiscent of RNAseZ and suggest a conserved mechanism of endonucleolytic cleavage of RNA (Fig. 1.6). Assays with purified protein against RNA show weak activity, in agreement with previous results with endogenously purified proteins that suggested multiple protein complexes were necessary for the efficient cleavage of substrate (Mandel et al., 2006).

Interestingly, CPSF100 bears striking homology to CPSF73. Structures of the yeast version of N-terminus of CPSF100 show a domain organization similar to that of CPSF73 (Fig. 1.6). However, CPSF100 lack the metal binding residues found in CPSF73 resulting in an enzymatically inactive protein (Mandel et al., 2006). Experiments using the yeast versions of the proteins show that CPSF100 can interact directly with CPSF160 and CPSF73, via its C-terminus (Kyburz et al., 2003). Evidence for interactions between the human versions also show an interaction that is mediated via the C-terminus of both

73 and 100, suggesting that these interactions are conserved across species (Dominski et al., 2005). Likely CPSF100 plays a bridging role between CPSF160 and 73, given that the distance between the polyA site and cleavage site can be considerable.

The smallest subunit, CPSF30, contains no known domains aside from several tandem CCCH-type zinc fingers in its C-terminus. These zinc fingers have been shown to bind polyU RNA sequences found near cleavage sites *in vitro* (Barabino et al., 1997). However, given the lack of RNA sequence conservation near the cleavage site, as well as the fact that CPSF30 deletion analysis show no loss of cleavage specificity, it is questionable how relevant this RNA binding activity is to 3'-end processing. CPSF30 is also known to bind hFip1, and a crystal structure has shown that CPSF30 can be subject to targeting by influenza virulence factors to suppress host anti-viral responses (Das et al., 2008).

hFip1 was initially identified as a factor in yeast that associated with PAP, and was only later identified as a bona fide member of the CPSF complex. A well conserved element mediates the interaction between hFip1 and PAP, while a human Fip1 only segment has been also shown to bind U-rich RNAs, similar to CPSF30. Additional protein-protein interactions have also been observed between hFip1 and CPSF160, CstF77, and CPSF30. hFip1 plays a role in partial role in the cleavage reaction, with hFip1 depleted extracts showing decreased cleavage. However, it plays an essential role in polyadenylation in which extracts lacking hFip1 showed no polyA activity (Kaufmann et al., 2004).

Cleavage Stimulatory Factor (CstF)

The CstF complex consists of three polypeptides of molecular weights 77 kDa, 64 kDa, and 50 kDa. Like CPSF, CstF was initially identified through biochemical fractionation experiments from tissues (Takagaki et al., 1989, 1990). While it remained associated with CPSF through most of the purifications, a single cation exchange column could separate the two complexes. Experiments using purified CPSF and CstF in an *in vitro* cleavage assay would show that increased cleavage activity would be associated with CstF. The major scaffolding factor from which the entire complex is organized is CstF77. The N-terminal 550 amino acids are composed of helical repeats known as half-a-TPR (HAT) that are known to mediate protein-protein interactions (Goebl and Yanagida, 1991; Preker and Keller, 1998). Biochemical studies showed strong self-association of CstF77, which was further validated by the crystal structure of the HAT domain (Fig. 1.7) (Bai et al., 2007; Legrand et al., 2007). Some evidence has suggested an interaction between CstF77 and CPSF160 via the HAT domain of 77 (Bai et al., 2007). Beyond the HAT domain, the remaining ~200 residues are heavily proline rich, and are responsible for scaffolding both CstF64 and CstF50 (Takagaki and Manley, 2000).

CstF64 was one of the earliest proteins identified to participate in 3'-end processing due to its UV crosslinking to RNA in a polyA-site dependent fashion (Wilusz and Shenk, 1988). Subsequent work would show that CstF64 bound the G/U-rich DSE, enhancing the specificity of the cleavage reaction (MacDonald et al., 1994). CstF64 is comprised of four distinct domains/regions, of which three have function associated with them. The N-terminus of 64 is a classic RNA recognition motif (RRM) domain used for binding G/U RNAs (Cañadillas and Varani, 2003; Pancevac et al., 2010; Takagaki and Manley, 1997). Following the RRM is a hinge domain unique to 64 that spans

approximately 100 residues that is used to bind to CstF77; some work has also shown that Symplekin, a large scaffolding protein well characterized in yeast (Pta1), also competes with CstF77 for binding to this region (Takagaki and Manley, 2000). A significant remainder of the protein is comprised of gly/pro-rich sequences, followed by MEARA repeats, none of which are predicted to have significant secondary structure. No biological function has yet to be assigned to this segment either. The final 50 residues form a 3-helical bundle, which has been shown to associate with other 3'-end processing factors (Qu et al., 2007).

CstF50 is a WD40-repeat protein with a unique dimerization domain (Fig. 1.7) located in the N-terminus of the protein, independent of the β-propeller structure (Moreno-Morcillo et al., 2011). Unlike CstF77 and 64, CstF50 does not have a known yeast homolog, suggesting that its function unique to higher eukaryotes. Studies have shown CstF50 interacts directly with BARD1 (Edwards et al., 2008; Kleiman and Manley, 1999), and a CstF-BARD-BRCA1 complex could be detected in cells treated with DNA damaging agents (Kleiman and Manley, 2001). Accumulation of unprocessed mRNA was also detected in these cells as well, suggesting a link between 3'-end processing and DNA damage repair through CstF50.

Cleavage Factor I and II (CFI_m/$CFII_m$)

Cleavage Factor I is comprised of a heterodimer of a 25kDa subunit, and either a 59 or 68kDa subunit; it is suspected that the 59 kDa component is a splice variant of the larger subunit. The crystal structure of the CFI_m25 nudix domain revealed that rather than possessing hydrolase activity, the domain had evolved in this complex to bind UGUA

RNA sequences found in the USE (Yang et al., 2010). Even more interestingly, the nudix domain was shown to stably dimerize, suggesting that CFI_m contained two copies of each subunit in the complex (Coseno et al., 2008). Crystal structures of CFI_m25 and CFI_m68's RRM show a heterotetramer, with the nudix domain of 25 serving as the major scaffold and the RRMs of 68 each binding to one copy of the nudix domain (Fig. 1.7) (Li et al., 2011; Yang et al., 2011). Uniquely, despite the presence of two RRMs from 68, the major contributor to RNA binding lay still in the nudix domains of 25. In accordance to its role in recognizing the USE, recent studies have implicated the CFIm complex in alternative polyA site selection (Katahira et al., 2013; Kim et al., 2010b; Kubo et al.).

Cleavage Factor II complex ($CFII_m$) is made of two conserved proteins, Pcf11 and Clp1. hPcf11 is a large 1555 amino acid protein, with little resemblance to the yeast homolog, with the exception of the N-terminal CID domain, and the C-terminus where it's interaction with Clp1 and members of the mRNA export machinery reside (Johnson et al., 2009; Noble et al., 2007). yPcf11 was initially identified as a binding partner with the yeast equivalents of CstF77/64, and mutants of yPcf11 displayed defects in both cleavage and polyadenylation (Amrani et al., 1997). Work has shown in yeast that Pcf11 is a bifunctional protein, with functions in both 3'-end processing and transcription termination, with termination activity being associated with the CID domain (Sadowski et al., 2003).

hClp1, along with hPcf11, was identified via purification of native $CFII_m$ from HeLa cell extracts. Clp1 was shown to be essential to the cleavage step of 3'-end processing, but dispensable for the polyadenylation step. The necessity of hClp1 (and $CFII_m$ in general) is likely to act as a bridging factor between CFI_m and CPSF, as

immunoprecipitation with Clp1 antibodies experiments show association with components of CPSF and CFI_m (de Vries et al., 2000).

1.2.2 Transcription termination is coupled to 3'-end processing

Following the end of transition of an open reading frame, PolII can remain engaged on the DNA template, even well past the polyA site, resulting in transcription run-on (TRO) which potentially interferes with the transcription of a neighboring gene (Shearwin et al., 2005). Thus, proper transcription termination is necessary to ensure the proper release of both polymerase and product RNA.

Multiple pathways for PolII transcription exist, depending on the class of RNA being transcribed. For mRNAs, the transcription termination pathway is tightly associated with the polyA site, and in general, the 3'-end processing machinery. Evidence of a polyA site-dependent pathway first appeared in the late '80s when it was found that impaired polyadenylation of reporter genes via usage of a poorly used polyA site or mutations in high efficiency polyA sites would result in improper PolII termination (Edwalds-Gilbert et al., 1993; Logan et al., 1987). A direct coupling of 3'-end processing to transcription was established much later when work in yeast showed temperature sensitive mutants of cleavage but not polyadenylation factors impaired transcription termination (Birse et al., 1998). Since then, a wealth of work has unveiled a tightly coupled association of 3'-end processing factors and PolII CTD phosphorylation to the transcription termination pathway. From this work, two major models for termination arose:

1. The allosteric model, in which transcription of the polyA signal would result in changes in the composition of the elongation complex, such as the dissociation of an anti-termination factor, and/or a conformation change in PolII itself.
2. The torpedo model, which posits that 3'-end cleavage of the pre-mRNA frees the product mRNA from the elongating polymerase. With the opening of a new 5' uncapped phosphate an exonuclease is recruited to digest the 3' cleavage product still tethered to PolII, where the rapid degradation by the nuclease allows it to catch up the polymerase and "torpedo" it from the template (Fig. 1.4) (reviewed in Rosonina et al., 2006).

While these models existed for many years as alternative possibilities, mounting evidence suggests they are not mutually exclusive and that termination likely represents a mix of these models (Luo et al., 2006).

Prior to the end of transcription a number of factors known to impede termination are known to associate with the PolII elongation complex, including the SR protein Npl3 (Bucheli and Buratowski, 2005; Bucheli et al., 2008; Deka et al., 2008) and Sub1 (Calvo and Manley, 2005). In addition to *trans*-acting protein factors, PolII CTD phosphorylated at tyrosine 1 has also been suggested to act as an anti-termination factor (Mayer et al., 2012). At the 3'-end of genes near the polyA site, Npl3 may be removed by either a competition mechanism with protein factors and/or a post-translational phosphorylation to allow for the binding of the 3'-end processing apparatus (Bucheli and Buratowski, 2005; Bucheli et al., 2008). Likely, dephosphorylation of Tyr1P occurs near or at these sites to allow for the binding of Ser2P interacting proteins such as Pcf11 and Rtt103. However, formal evidence for this has not yet been found.

Transcriptional pausing at or near the polyA site has also been observed (Orozco et al., 2002; Yonaha and Proudfoot, 1999), potentially allowing for the processing/termination complexes to assemble. Transcription of the polyA site, in addition to the dissociation/association of transcription termination factors, and the assembly of the 3'-end processing machinery then allows for the proper cleavage of the pre-mRNA by CPSF73 (described above). The known termination factor Pcf11 has been shown to directly associate with components of the processing assembly (Qu et al., 2007; Sadowski et al., 2003) as well as with Ser2P (Meinhart and Cramer, 2004; Noble et al., 2005). Work in yeast has shown that Pcf11 forms a central part of not only stabilizing the cleavage reaction by its interaction with Rna15 (CstF64 in animals), but is also responsible for the recruitment of the Rat1 nuclease, postulated to be the "torpedo" which disengages PolII from templates by digesting the 3' cleavage product (Fig. 1.4) still tethered to elongating PolII (Kim et al., 2004; Luo et al., 2006). The human homolog of the Rat1 nuclease, Xrn2, has also been shown to promote transcription termination, reinforcing the Rat1 torpedo hypothesis (Kaneko et al., 2007; West et al., 2004).

In addition to recruitment by Pcf11, Rat1 exists in a 1:1 complex with Rai1 (Xiang et al., 2009), and as a complex can also be recruited to sites of termination by the Ser2P interacting protein Rtt103 (Kim et al., 2004). It is unclear how this mechanism of recruitment is different from that of Pcf11, and if these work in tandem to facilitate the effective localization of Rat1 to termination sites. It is conceivable that Rat1 recruitment by Pcf11 is meant to couple 3'-end processing, CTD phosphorylation, and termination, but a remaining question then arises to why Rtt103 is needed given that both Pcf11 and Rtt103 can bind the same phospho-form of the CTD.

1.3 Summary

Termination of mRNA transcription is an elaborate and well-regulated process. The proper phosphorylation patterns of the CTD and the assembly of multiple complexes must occur together both spatially and temporally in order to effectively complete synthesis of an mRNA. The work I will present in chapters 2 and 3, represent a step towards understanding the regulation and assembly of these factors. Rtr1, as described in 1.1.4, is a recently described CTD phosphatase with controversial function. In chapter 2, I will show that Rtr1 is a bona fide CTD phosphatase, despite reports to the contrary, and that it can also dephosphorylate Tyr1P *in* vitro in addition to removing Ser5P. My work presents not only the biochemical and kinetic characterization of a new phosphatase enzyme, but also provides an attractive candidate for the removal of a phospho "roadblock" to transcription termination.

Chapter 3 details the biochemical characterization of CstF, one of the essential factors required for the cleavage of pre-mRNAs as described in 1.2.1. Through my work, I resolve the controversy regarding CstF's stoichiometry and show that complex formation allows the assembly to recognize DSEs with high selectivity and affinity, further providing insight into polyA site selection by CstF. This work also provides a stepping stone into building larger recombinant complexes to allow for careful biochemical dissection of the 3'-end processing apparatus.

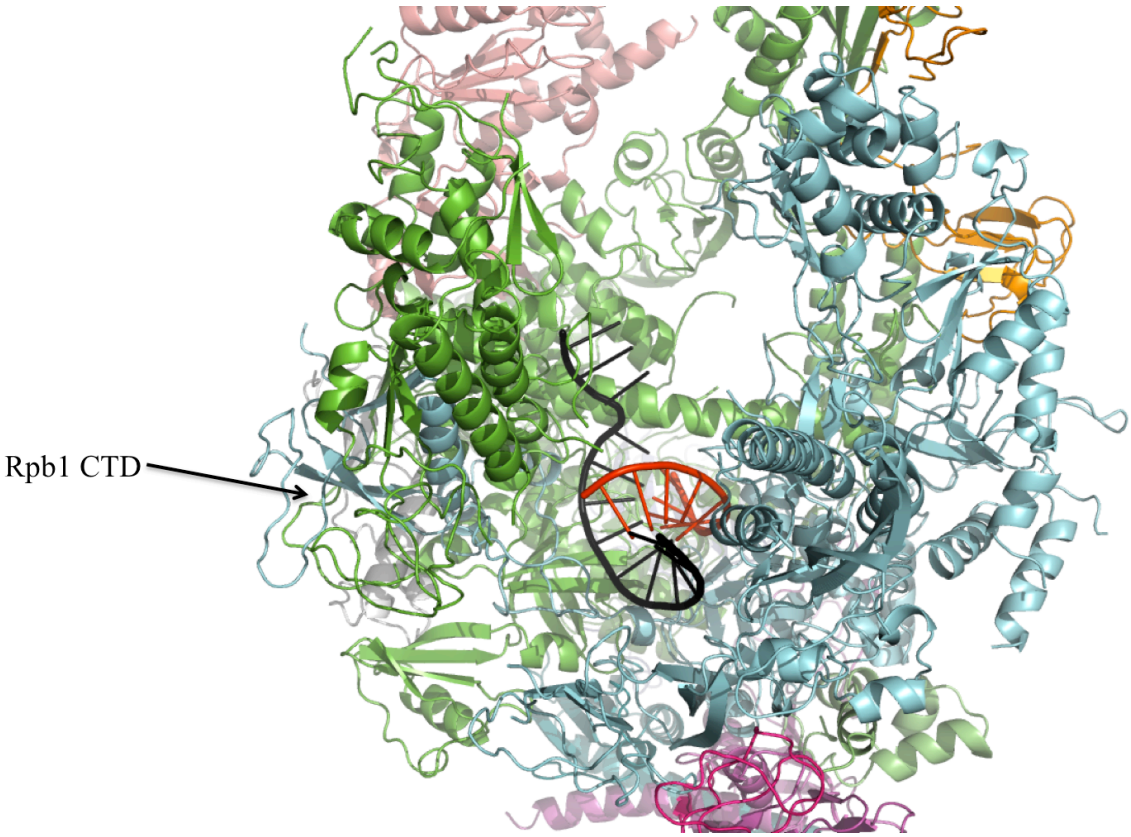

Figure 1.1 RNA Polymerase II elongation complex with template and transcript
Crystal structure of RNA PolII (PDB 1I6H) in complex with a DNA template (black) and RNA transcript (red). The position of the C-terminal end of Rpb1 in the crystal structure is indicated by the arrow. The CTD presumably is too flexible or was truncated during crystallization, thus no electron density is observed for it.

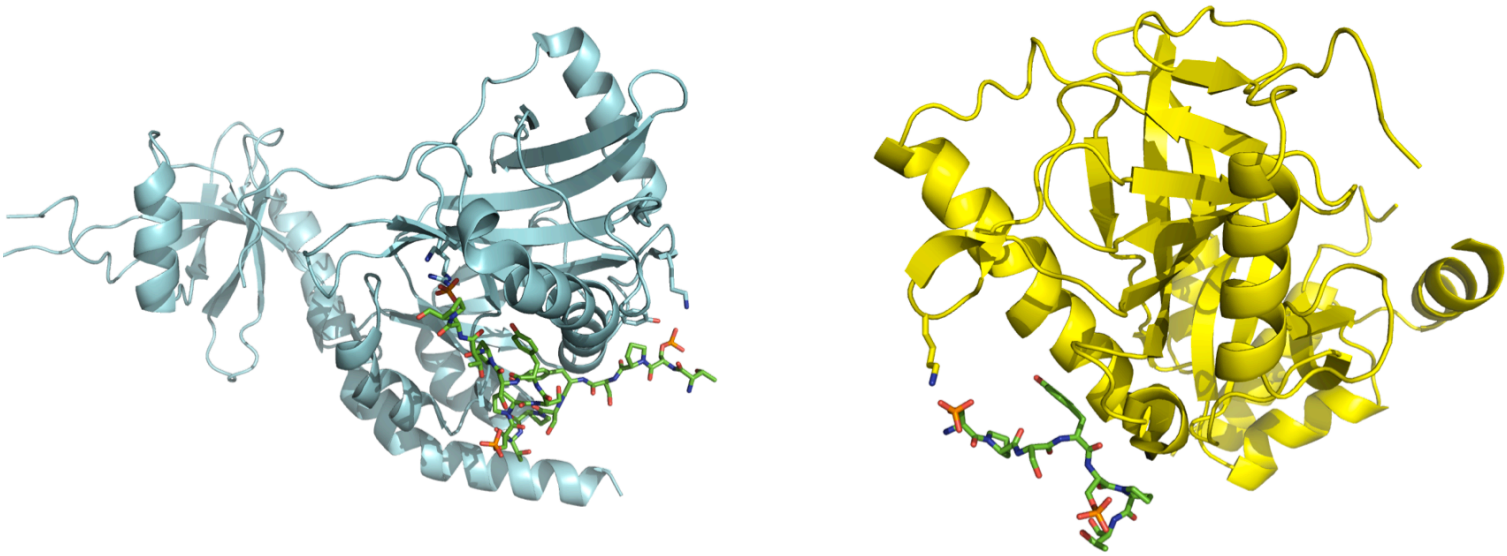

Figure 1.2 Divergent modes of Ser5P CTD recognition among conserved mRNA capping enzymes
Candida albicans Cgt1 (left, cyan, PDB 1P16) binds Ser5P CTD (shown in sticks) in an elongated fashion along a groove in its structure. Mouse Mce1 (right, yellow, PDB 3RTX), the vertebrate homolog, also binds Ser5P CTD (shown in sticks) along its surface, in a manner quite different from the yeast enzyme.

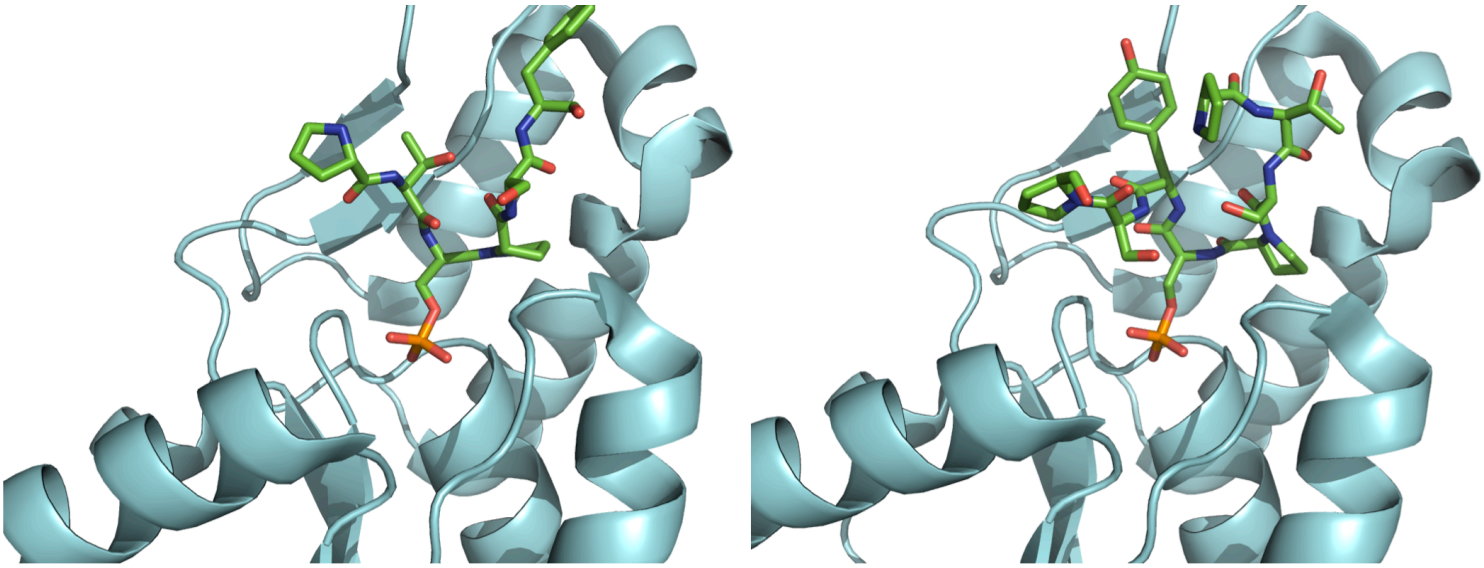

Figure 1.3 Ssu72 dephosphorylates both Ser5P and Ser7P CTD
Crystal structure of an Ssu72 active site mutant bound to a *cis*-proline Ser5P CTD (left, 3O2Q). Structure of Ssu72 bound to a Ser7P CTD peptide (right, 4H3H). Comparison of the two structures shows a constrained geometry of the Ser7P peptide relative to the Ser5P substrate.

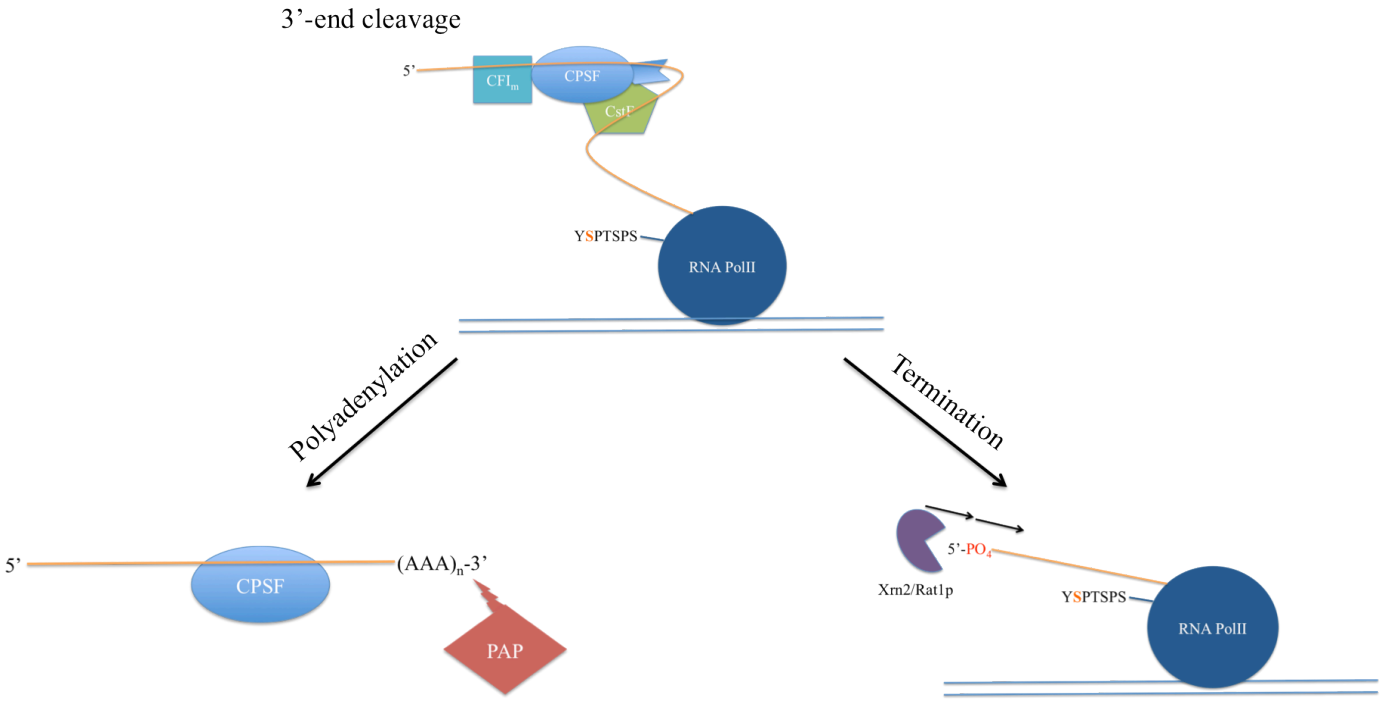

Figure 1.4 Co-transcriptional mRNA 3’-end cleavage is coupled to polyadenylation and transcription termination
The pre-mRNA (orange) is cleaved co-transcriptionally at designated positions near polyA sites by the 3’-end processing machinery; only several components are shown for clarity (top). Following cleavage, two events may occur simultaneously: the polyadenylation at the newly formed 3’-end of the pre-mRNA by PAP (lower left), and/or the termination of transcription intitiated by the Xrn2/Rat1p exonuclease. Not shown are the termination factors Pcf11 and Rtt103 that bind the Ser2P form of the CTD.

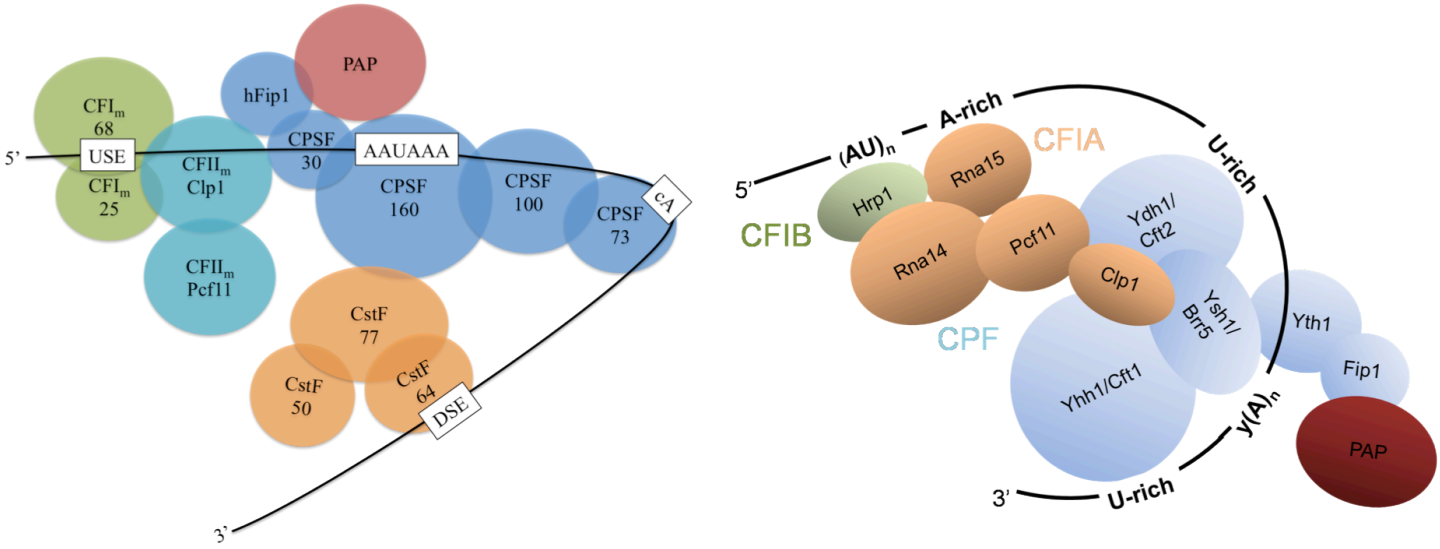

Figure 1.5 Overall 3'-end processing machineries of metazoans and yeasts
Both metazoan (left) and yeast (right) contain homologous protein complexes, however, their overall arrangement and organization differ. The yeast CPF complex is homologous to animal CPSF, and components of CFIA are homologous to CstF and $CFII_m$. CFI_m (animal) and CFIB (yeast) are unique to their respective organisms. Organization of these complexes around the RNA differ due to divergence of 3'-end processing signals in metazoans and yeasts.

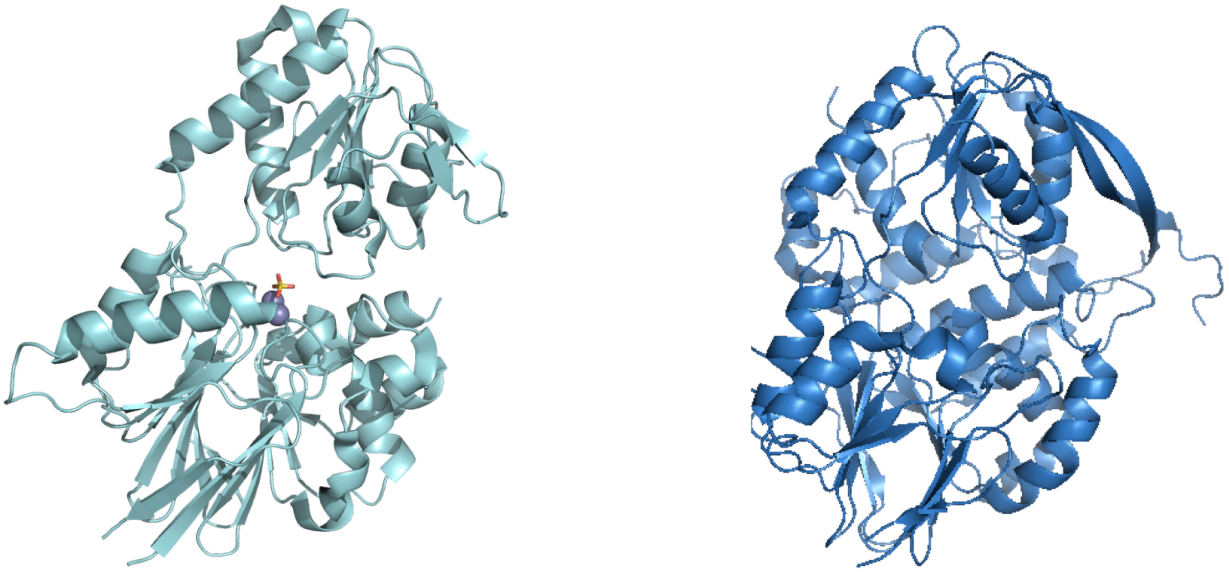

Figure 1.6 Structures of CPSF73 and CPSF100 are highly homologous structures
CPSF73 (left, cyan, 2I7T) shows an open structure with two lobes sandwiching in the center two zinc ions coordinating a sulfate molecule (derived from crystallization conditions), likely mimicking the phosphate backbone of substrate RNA. Yeast CPSF100 (right, blue, 2I7X) shows a similar architecture with two distinct halves. However, despite the homology, the overall structure is "closed" relative to CPSF73 and cannot coordinate the zinc ions necessary for endonuclease activity.

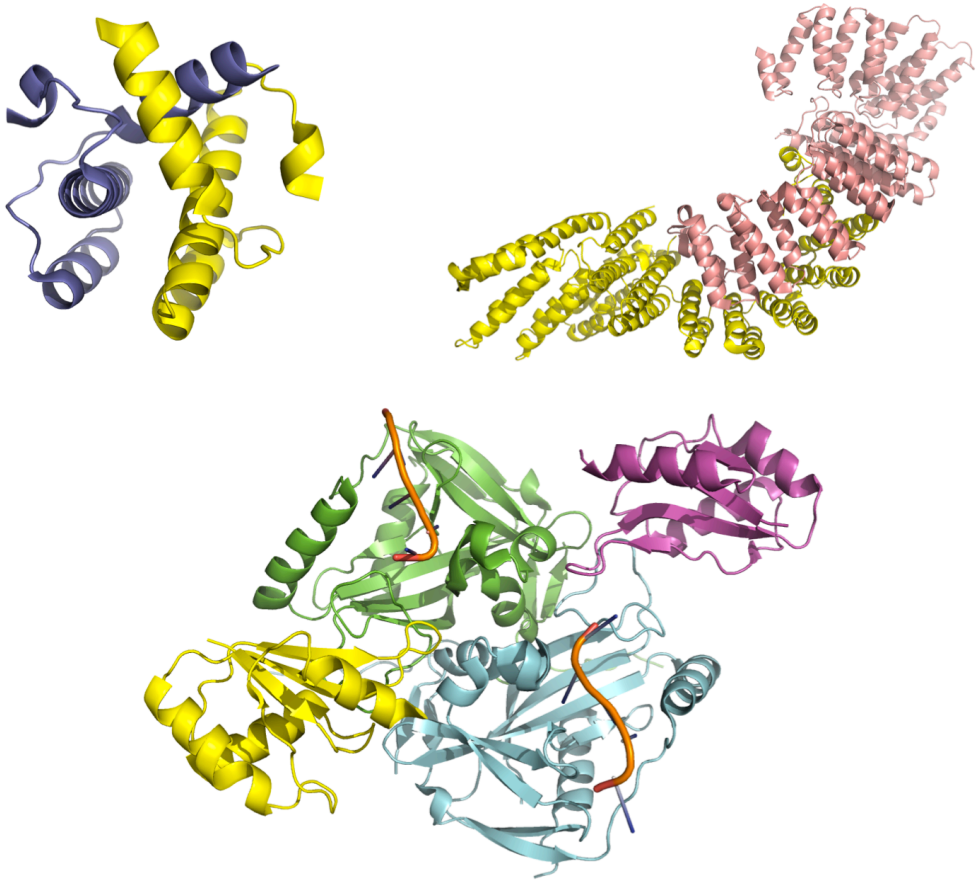

Figure 1.7 Dimeric structures in 3'-end processing
Both CstF50 (upper left, 2XZ2) and CstF77 (2OOE) homodimerize via well conserved domains, although the implications of this self-association in the whole CstF complex and their impact on its function are currently unknown. CFI_m forms a heterotetramer via (central bottom, 3QT2) via homodimerization of its CFI_m25 subunit (green and cyan).

Table 1.1 RNA Polymerase II CTD phosphorylation – regulation and function

Residue	Kinase(s)	Phosphatase(s)	Function
Tyrosine 1	Abl/Arg (metazoan), transcriptional relevance unknown	Unknown	Anti-termination signal, prevents termination factors from binding CTD
Serine 2	Bur1/2 and CTDK-I (yeast) P-TEFb (metazoan)	Fcp1	Transcription elongation/termination signal, assists in pre-mRNA processing
Threonine 4	P-TEFb and/or Plk3 (metazoan)	Unknown	Histone mRNA processing and/or transcription elongation signal
Serine 5	TFIIH	Rtr1, Scp1, Ssu72	Transcription initiation, mRNA capping, transcription elongation
Serine 7	TFIIH	Ssu72 (putative)	snRNA/snoRNA processing, possible role in pre-mRNA splicing

Table 1.2 Mammalian 3'-end processing factors and their known yeast homologs

Protein	Yeast homolog	Complex	Function
CPSF160	Yhh1/Cft1	CPSF	Recognizes universal polyA (AAUAAA) signal
CPSF100	Ydh1/Cft2	CPSF	Unknown, likely bridges CPSF160 and 73
CPSF73	Ysh1/Brr5	CPSF	Endonuclease responsible for cleavage
CPSF30	Yth1	CPSF	Participates in cleavage via unknown mechanism
hFip1	Fip1p	CPSF	Necessary for polyadenylation
CstF77	Rna14	CstF	Scaffolding protein, bridges CstF with CPSF
CstF64	Rna15	CstF	Recognizes G/U-rich DSE sequences
CstF50	Unknown	CstF	Recruits BARD1/BRCA1 to 3'-end assembly
CFI_m68	Unknown	Cleavage Factor I	Participates in recognition of USE with CFI_m25
CFI_m25	Unknown	Cleavage Factor I	Primary RNA recognition protein for USE
Pcf11	Pcf11p	Cleavage Factor II	Couples termination to 3'-end processing
Clp1	Clp1p	Cleavage Factor II	Required for cleavage via unknown mechanism
PAP	PAP1p	Associates w/CPSF	Polyadenylates free 3'-OH after cleavage

1.4 References

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

Amrani, N., Minet, M., Wyers, F., Dufour, M.E., Aggerbeck, L.P., and Lacroute, F. (1997). PCF11 encodes a third protein component of yeast cleavage and polyadenylation factor I. Mol. Cell. Biol. *17*, 1102–1109.

Bai, Y., Auperin, T.C., Chou, C.-Y., Chang, G.-G., Manley, J.L., and Tong, L. (2007). Crystal structure of murine CstF-77: dimeric association and implications for polyadenylation of mRNA precursors. Mol. Cell *25*, 863–875.

Barabino, S.M., Hübner, W., Jenny, A., Minvielle-Sebastia, L., and Keller, W. (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*, 1703–1716.

Baskaran, R., Dahmus, M.E., and Wang, J.Y. (1993). Tyrosine phosphorylation of mammalian RNA polymerase II carboxyl-terminal domain. Proc. Natl. Acad. Sci. *90*, 11167–11171.

Baskaran, R., Chiang, G.G., Mysliwiec, T., Kruh, G.D., and Wang, J.Y. (1997). Tyrosine phosphorylation of RNA polymerase II carboxyl-terminal domain by the Abl-related gene product. J. Biol. Chem. *272*, 18905–18909.

Bataille, A.R., Jeronimo, C., Jacques, P.-É., Laramée, L., Fortin, M.-È., Forest, A., Bergeron, M., Hanes, S.D., and Robert, F. (2012). A universal RNA polymerase II CTD cycle is orchestrated by complex interplays between kinase, phosphatase, and isomerase enzymes along genes. Mol. Cell *45*, 158–170.

Birse, C.E., Minvielle-Sebastia, L., Lee, B.A., Keller, W., and Proudfoot, N.J. (1998). Coupling Termination of Transcription to Messenger RNA Maturation in Yeast. Science *280*, 298–301.

Boeing, S., Rigault, C., Heidemann, M., Eick, D., and Meisterernst, M. (2010). RNA Polymerase II C-terminal Heptarepeat Domain Ser-7 Phosphorylation Is Established in a Mediator-dependent Fashion. J. Biol. Chem. *285*, 188–196.

Bucheli, M.E., and Buratowski, S. (2005). Npl3 is an antagonist of mRNA 3' end formation by RNA polymerase II. EMBO J. *24*, 2150–2160.

Bucheli, M.E., Dermody, J.L., Dreyfuss, J.M., Villén, J., Ogundipe, B., Gygi, S.P., Park, P.J., Ponticelli, A.S., Moore, C.L., and Buratowski, S. (2008). Unphosphorylated SR-like protein Npl3 stimulates RNA polymerase II elongation. PLoS ONE *3*, e3273.

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

Cagas, P.M., and Corden, J.L. (1995). Structural studies of a synthetic peptide derived from the carboxyl-terminal domain of RNA polymerase II. Proteins *21*, 149–160.

Calvo, O., and Manley, J.L. (2005). The transcriptional coactivator PC4/Sub1 has multiple functions in RNA polymerase II transcription. EMBO J. *24*, 1009–1020.

Cañadillas, J.M.P., and Varani, G. (2003). Recognition of GU-rich polyadenylation regulatory elements by human CstF-64 protein. EMBO J. *22*, 2821–2830.

Chapman, R.D., Heidemann, M., Albert, T.K., Mailhammer, R., Flatley, A., Meisterernst, M., Kremmer, E., and Eick, D. (2007). Transcribing RNA polymerase II is phosphorylated at CTD residue serine-7. Science *318*, 1780–1782.

Cho, E.J., Kobor, M.S., Kim, M., Greenblatt, J., and Buratowski, S. (2001). Opposing effects of Ctk1 kinase and Fcp1 phosphatase at Ser 2 of the RNA polymerase II C-terminal domain. Genes Dev. *15*, 3319–3329.

Coseno, M., Martin, G., Berger, C., Gilmartin, G., Keller, W., and Doublié, S. (2008). Crystal structure of the 25 kDa subunit of human cleavage factor Im. Nucleic Acids Res. *36*, 3474–3483.

Das, K., Ma, L.-C., Xiao, R., Radvansky, B., Aramini, J., Zhao, L., Marklund, J., Kuo, R.-L., Twu, K.Y., Arnold, E., et al. (2008). Structural basis for suppression of a host antiviral response by influenza A virus. Proc. Natl. Acad. Sci. U. S. A. *105*, 13093–13098.

David, C.J., Boyne, A.R., Millhouse, S.R., and Manley, J.L. (2011). The RNA polymerase II C-terminal domain promotes splicing activation through recruitment of a U2AF65-Prp19 complex. Genes Dev. *25*, 972–983.

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

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

Dominski, Z., Yang, X., Purdy, M., Wagner, E.J., and Marzluff, W.F. (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*, 1489–1500.

Edwalds-Gilbert, G., Prescott, J., and Falck-Pedersen, E. (1993). 3' RNA processing efficiency plays a primary role in generating termination-competent RNA polymerase II elongation complexes. Mol. Cell. Biol. *13*, 3472–3480.

Edwards, R.A., Lee, M.S., Tsutakawa, S.E., Williams, R.S., Nazeer, I., Kleiman, F.E., Tainer, J.A., and Glover, J.N.M. (2008). The BARD1 C-terminal domain structure and

interactions with polyadenylation factor CstF-50. Biochemistry (Mosc.) *47*, 11446–11456.

Egloff, S., O'Reilly, D., Chapman, R.D., Taylor, A., Tanzhaus, K., Pitts, L., Eick, D., and Murphy, S. (2007). Serine 7 of the RNA polymerase II CTD is specifically required for snRNA gene expression. Science *318*, 1777–1779.

Egloff, S., Szczepaniak, S.A., Dienstbier, M., Taylor, A., Knight, S., and Murphy, S. (2010). The Integrator Complex Recognizes a New Double Mark on the RNA Polymerase II Carboxyl-terminal Domain. J. Biol. Chem. *285*, 20564 –20569.

Egloff, S., Zaborowska, J., Laitem, C., Kiss, T., and Murphy, S. Ser7 Phosphorylation of the CTD Recruits the RPAP2 Ser5 Phosphatase to snRNA Genes. Mol. Cell.

Fabrega, C., Shen, V., Shuman, S., and Lima, C.D. (2003). Structure of an mRNA capping enzyme bound to the phosphorylated carboxy-terminal domain of RNA polymerase II. Mol. Cell *11*, 1549–1561.

Fitzgerald, M., and Shenk, T. (1981). The sequence 5'-AAUAAA-3'forms parts of the recognition site for polyadenylation of late SV40 mRNAs. Cell *24*, 251–260.

Ghazy, M.A., He, X., Singh, B.N., Hampsey, M., and Moore, C. (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*, 2296–2307.

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

Ghosh, A., Shuman, S., and Lima, C.D. (2008). The structure of Fcp1, an essential RNA polymerase II CTD phosphatase. Mol. Cell *32*, 478–490.

Ghosh, A., Shuman, S., and Lima, C.D. (2011). Structural insights to how mammalian capping enzyme reads the CTD code. Mol. Cell *43*, 299–310.

Gilmartin, G.M., Fleming, E.S., Oetjen, J., and Graveley, B.R. (1995). CPSF recognition of an HIV-1 mRNA 3'-processing enhancer: multiple sequence contacts involved in poly(A) site definition. Genes Dev. *9*, 72–83.

Goebl, M., and Yanagida, M. (1991). The TPR snap helix: a novel protein repeat motif from mitosis to transcription. Trends Biochem. Sci. *16*, 173–177.

Gu, B., Eick, D., and Bensaude, O. (2013). CTD serine-2 plays a critical role in splicing and termination factor recruitment to RNA polymerase II in vivo. Nucleic Acids Res. *41*, 1591–1603.

Hintermair, C., Heidemann, M., Koch, F., Descostes, N., Gut, M., Gut, I., Fenouil, R., Ferrier, P., Flatley, A., Kremmer, E., et al. (2012). Threonine-4 of mammalian RNA

polymerase II CTD is targeted by Polo-like kinase 3 and required for transcriptional elongation. EMBO J.

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

Hsin, J.-P., Sheth, A., and Manley, J.L. (2011). RNAP II CTD Phosphorylated on Threonine-4 Is Required for Histone mRNA 3′ End Processing. Science *334*, 683 –686.

Hu, J., Lutz, C.S., Wilusz, J., and Tian, B. (2005). Bioinformatic identification of candidate cis-regulatory elements involved in human mRNA polyadenylation. RNA New York N *11*, 1485–1493.

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

Jones, J.C., Phatnani, H.P., Haystead, T.A., MacDonald, J.A., Alam, S.M., and Greenleaf, A.L. (2004). C-terminal repeat domain kinase I phosphorylates Ser2 and Ser5 of RNA polymerase II C-terminal domain repeats. J. Biol. Chem. *279*, 24957–24964.

Kaneko, S., Rozenblatt-Rosen, O., Meyerson, M., and Manley, J.L. (2007). The multifunctional protein p54nrb/PSF recruits the exonuclease XRN2 to facilitate pre-mRNA 3' processing and transcription termination. Genes Dev. *21*, 1779–1789.

Katahira, J., Okuzaki, D., Inoue, H., Yoneda, Y., Maehara, K., and Ohkawa, Y. (2013). Human TREX component Thoc5 affects alternative polyadenylation site choice by recruiting mammalian cleavage factor I. Nucleic Acids Res. *41*, 7060–7072.

Kaufmann, I., Martin, G., Friedlein, A., Langen, H., and Keller, W. (2004). Human Fip1 is a subunit of CPSF that binds to U-rich RNA elements and stimulates poly(A) polymerase. EMBO J. *23*, 616–626.

Kim, H., Erickson, B., Luo, W., Seward, D., Graber, J.H., Pollock, D.D., Megee, P.C., and Bentley, D.L. (2010a). Gene-specific RNA polymerase II phosphorylation and the CTD code. Nat. Struct. Mol. Biol.

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

Kim, S., Yamamoto, J., Chen, Y., Aida, M., Wada, T., Handa, H., and Yamaguchi, Y. (2010b). Evidence that cleavage factor Im is a heterotetrameric protein complex controlling alternative polyadenylation. Genes Cells Devoted Mol. Cell. Mech. *15*, 1003–1013.

Kleiman, F.E., and Manley, J.L. (1999). Functional interaction of BRCA1-associated BARD1 with polyadenylation factor CstF-50. Science *285*, 1576–1579.

Kleiman, F.E., and Manley, J.L. (2001). The BARD1-CstF-50 interaction links mRNA 3' end formation to DNA damage and tumor suppression. Cell *104*, 743–753.

Krishnamurthy, S., He, X., Reyes-Reyes, M., Moore, C., and Hampsey, M. (2004). Ssu72 Is an RNA polymerase II CTD phosphatase. Mol. Cell *14*, 387–394.

Kubo, T., Wada, T., Yamaguchi, Y., Shimizu, A., and Handa, H. Knock-down of 25 kDa subunit of cleavage factor Im in Hela cells alters alternative polyadenylation within 3′-UTRs. Nucleic Acids Res. *34*, 6264 –6271.

Kyburz, A., Sadowski, M., Dichtl, B., and Keller, W. (2003). The role of the yeast cleavage and polyadenylation factor subunit Ydh1p/Cft2p in pre-mRNA 3?-end formation. Nucleic Acids Res. *31*, 3936–3945.

Legendre, M., and Gautheret, D. (2003). Sequence determinants in human polyadenylation site selection. BMC Genomics *4*, 7.

Legrand, P., Pinaud, N., Minvielle-Sébastia, L., and Fribourg, S. (2007). The structure of the CstF-77 homodimer provides insights into CstF assembly. Nucleic Acids Res. *35*, 4515–4522.

Li, H., Tong, S., Li, X., Shi, H., Ying, Z., Gao, Y., Ge, H., Niu, L., and Teng, M. (2011). Structural basis of pre-mRNA recognition by the human cleavage factor Im complex. Cell Res. *21*, 1039–1051.

Li, T., Chen, X., Garbutt, K.C., Zhou, P., and Zheng, N. (2006). Structure of DDB1 in complex with a paramyxovirus V protein: viral hijack of a propeller cluster in ubiquitin ligase. Cell *124*, 105–117.

Liu, J., Fan, S., Lee, C.-J., Greenleaf, A.L., and Zhou, P. (2013). Specific interaction of the transcription elongation regulator TCERG1 with RNA polymerase II requires simultaneous phosphorylation at Ser2, Ser5, and Ser7 within the carboxyl-terminal domain repeat. J. Biol. Chem. *288*, 10890–10901.

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

Lunde, B.M., Reichow, S.L., Kim, M., Suh, H., Leeper, T.C., Yang, F., Mutschler, H., Buratowski, S., Meinhart, A., and Varani, G. (2010). Cooperative interaction of transcription termination factors with the RNA polymerase II C-terminal domain. Nat. Struct. Mol. Biol. *17*, 1195–1201.

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

MacDonald, C.C., Wilusz, J., and Shenk, T. (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*, 6647–6654.

Mandel, C.R., Kaneko, S., Zhang, H., Gebauer, D., Vethantham, V., Manley, J.L., and Tong, L. (2006). Polyadenylation factor CPSF-73 is the pre-mRNA 3'-end-processing endonuclease. Nature *444*, 953–956.

Max, T., Søgaard, M., and Svejstrup, J.Q. (2007). Hyperphosphorylation of the C-terminal Repeat Domain of RNA Polymerase II Facilitates Dissociation of Its Complex with Mediator. J. Biol. Chem. *282*, 14113–14120.

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

Mayr, C., and Bartel, D.P. (2009). Widespread shortening of 3′UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. Cell *138*, 673.

McCracken, S., Fong, N., Yankulov, K., Ballantyne, S., Pan, G., Greenblatt, J., Patterson, S.D., Wickens, M., and Bentley, D.L. (1997). The C-terminal domain of RNA polymerase II couples mRNA processing to transcription. Nature *385*, 357–361.

McDevitt, M.A., Hart, R.P., Wong, W.W., and Nevins, J.R. (1986). Sequences capable of restoring poly(A) site function define two distinct downstream elements. EMBO J. *5*, 2907–2913.

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

Meinhart, A., Kamenski, T., Hoeppner, S., Baumli, S., and Cramer, P. (2005). A structural perspective of CTD function. Genes Dev. *19*, 1401–1415.

Millhouse, S., and Manley, J.L. (2005). The C-terminal domain of RNA polymerase II functions as a phosphorylation-dependent splicing activator in a heterologous protein. Mol. Cell. Biol. *25*, 533–544.

Moreno-Morcillo, M., Minvielle-Sébastia, L., Mackereth, C., and Fribourg, S. (2011). Hexameric architecture of CstF supported by CstF-50 homodimerization domain structure. RNA New York N *17*, 412–418.

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

Murthy, K.G., and Manley, J.L. (1992). Characterization of the multisubunit cleavage-polyadenylation specificity factor from calf thymus. J. Biol. Chem. *267*, 14804–14811.

Murthy, K.G., and Manley, J.L. (1995). The 160-kD subunit of human cleavage-polyadenylation specificity factor coordinates pre-mRNA 3'-end formation. Genes Dev. *9*, 2672–2683.

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

Noble, C.G., Hollingworth, D., Martin, S.R., Ennis-Adeniran, V., Smerdon, S.J., Kelly, G., Taylor, I.A., and Ramos, A. (2005). Key features of the interaction between Pcf11 CID and RNA polymerase II CTD. Nat. Struct. Mol. Biol. *12*, 144–151.

Noble, C.G., Beuth, B., and Taylor, I.A. (2007). Structure of a nucleotide-bound Clp1-Pcf11 polyadenylation factor. Nucleic Acids Res. *35*, 87–99.

Orozco, I.J., Kim, S.J., and Martinson, H.G. (2002). The poly(A) signal, without the assistance of any downstream element, directs RNA polymerase II to pause in vivo and then to release stochastically from the template. J. Biol. Chem. *277*, 42899–42911.

Ozsolak, F., Kapranov, P., Foissac, S., Kim, S.W., Fishilevich, E., Monaghan, A.P., John, B., and Milos, P.M. (2010). Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. Cell *143*, 1018–1029.

Pancevac, C., Goldstone, D.C., Ramos, A., and Taylor, I.A. (2010). Structure of the Rna15 RRM-RNA complex reveals the molecular basis of GU specificity in transcriptional 3'-end processing factors. Nucleic Acids Res.

Preker, P.J., and Keller, W. (1998). The HAT helix, a repetitive motif implicated in RNA processing. Trends Biochem. Sci. *23*, 15–16.

Qiu, H., Hu, C., and Hinnebusch, A.G. (2009). Phosphorylation of the Pol II CTD by KIN28 enhances BUR1/BUR2 recruitment and Ser2 CTD phosphorylation near promoters. Mol. Cell *33*, 752–762.

Qu, X., Perez-Canadillas, J.-M., Agrawal, S., De Baecke, J., Cheng, H., Varani, G., and Moore, C. (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*, 2101–2115.

Rodriguez, C.R., Cho, E.-J., Keogh, M.-C., Moore, C.L., Greenleaf, A.L., and Buratowski, S. (2000). Kin28, the TFIIH-Associated Carboxy-Terminal Domain Kinase, Facilitates the Recruitment of mRNA Processing Machinery to RNA Polymerase II. Mol. Cell. Biol. *20*, 104–112.

Rosonina, E., Kaneko, S., and Manley, J.L. (2006). Terminating the transcript: breaking up is hard to do. Genes Dev. *20*, 1050–1056.

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

Salisbury, J., Hutchison, K.W., and Graber, J.H. (2006). A multispecies comparison of the metazoan 3'-processing downstream elements and the CstF-64 RNA recognition motif. BMC Genomics *7*, 55.

Schauer, T., Tombácz, I., Ciurciu, A., Komonyi, O., and Boros, I.M. (2009). Misregulated RNA Pol II C-terminal domain phosphorylation results in apoptosis. Cell. Mol. Life Sci. CMLS *66*, 909–918.

Shearwin, K.E., Callen, B.P., and Egan, J.B. (2005). Transcriptional interference--a crash course. Trends Genet. TIG *21*, 339–345.

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

Stiller, J.W., and Cook, M.S. (2004). Functional unit of the RNA polymerase II C-terminal domain lies within heptapeptide pairs. Eukaryot. Cell *3*, 735–740.

Takagaki, Y., and Manley, J.L. (1997). RNA recognition by the human polyadenylation factor CstF. Mol. Cell. Biol. *17*, 3907–3914.

Takagaki, Y., and Manley, J.L. (2000). Complex protein interactions within the human polyadenylation machinery identify a novel component. Mol. Cell. Biol. *20*, 1515–1525.

Takagaki, Y., Ryner, L.C., and Manley, J.L. (1989). Four factors are required for 3'-end cleavage of pre-mRNAs. Genes Dev. *3*, 1711–1724.

Takagaki, Y., Manley, J.L., MacDonald, C.C., Wilusz, J., and Shenk, T. (1990). A multisubunit factor, CstF, is required for polyadenylation of mammalian pre-mRNAs. Genes Dev. *4*, 2112–2120.

Tian, B., Hu, J., Zhang, H., and Lutz, C.S. (2005). A large-scale analysis of mRNA polyadenylation of human and mouse genes. Nucleic Acids Res. *33*, 201–212.

Tombácz, I., Schauer, T., Juhász, I., Komonyi, O., and Boros, I. (2009). The RNA Pol II CTD phosphatase Fcp1 is essential for normal development in Drosophila melanogaster. Gene *446*, 58–67.

Trigon, S., Serizawa, H., Conaway, J.W., Conaway, R.C., Jackson, S.P., and Morange, M. (1998). Characterization of the Residues Phosphorylated in Vitro by Different C-terminal Domain Kinases. J. Biol. Chem. *273*, 6769–6775.

De Vries, H., Rüegsegger, U., Hübner, W., Friedlein, A., Langen, H., and Keller, W. (2000). Human pre-mRNA cleavage factor II(m) contains homologs of yeast proteins and bridges two other cleavage factors. EMBO J. *19*, 5895–5904.

Werner-Allen, J.W., Lee, C.-J., Liu, P., Nicely, N.I., Wang, S., Greenleaf, A.L., and Zhou, P. (2011). cis-Proline-mediated Ser(P)5 Dephosphorylation by the RNA Polymerase II C-terminal Domain Phosphatase Ssu72. J. Biol. Chem. *286*, 5717–5726.

West, S., Gromak, N., and Proudfoot, N.J. (2004). Human 5' --> 3' exonuclease Xrn2 promotes transcription termination at co-transcriptional cleavage sites. Nature *432*, 522–525.

Wilusz, J., and Shenk, T. (1988). A 64 kd nuclear protein binds to RNA segments that include the AAUAAA polyadenylation motif. Cell *52*, 221–228.

Xiang, K., Nagaike, T., Xiang, S., Kilic, T., Beh, M.M., Manley, J.L., and Tong, L. (2010). Crystal structure of the human symplekin-Ssu72-CTD phosphopeptide complex. Nature *467*, 729–733.

Xiang, K., Manley, J.L., and Tong, L. (2012a). The yeast regulator of transcription protein Rtr1 lacks an active site and phosphatase activity. Nat. Commun. *3*, 946.

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

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

Yang, Q., Gilmartin, G.M., and Doublié, S. (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*, 10062–10067.

Yang, Q., Coseno, M., Gilmartin, G.M., and Doublié, S. (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. Struct. Lond. Engl. 1993 *19*, 368–377.

Yeo, M., Lin, P.S., Dahmus, M.E., and Gill, G.N. (2003). A Novel RNA Polymerase II C-terminal Domain Phosphatase That Preferentially Dephosphorylates Serine 5. J. Biol. Chem. *278*, 26078–26085.

Yonaha, M., and Proudfoot, N.J. (1999). Specific Transcriptional Pausing Activates Polyadenylation in a Coupled In Vitro System. Mol. Cell *3*, 593–600.

Zhang, D.W., Mosley, A.L., Ramisetty, S.R., Rodriguez-Molina, J.B., Washburn, M.P., and Ansari, A.Z. (2012a). Ssu72 phosphatase dependent erasure of phospho-Ser7 marks

on the RNA Polymerase II C-terminal domain is essential for viability and transcription termination. J. Biol. Chem.

Zhang, M., Wang, X.J., Chen, X., Bowman, M.E., Luo, Y., Noel, J.P., Ellington, A.D., Etzkorn, F.A., and Zhang, Y. (2012b). Structural and kinetic analysis of prolyl-isomerization/phosphorylation cross-talk in the CTD code. ACS Chem. Biol. *7*, 1462–1470.

Zhao, J., Hyman, L., and Moore, C. (1999). Formation of mRNA 3' ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis. Microbiol. Mol. Biol. Rev. MMBR *63*, 405–445.

Zhou, M., Halanski, M.A., Radonovich, M.F., Kashanchi, F., Peng, J., Price, D.H., and Brady, J.N. (2000). Tat modifies the activity of CDK9 to phosphorylate serine 5 of the RNA polymerase II carboxyl-terminal domain during human immunodeficiency virus type 1 transcription. Mol. Cell. Biol. *20*, 5077–5086.

Chapter 2 . Rtr1 is a structurally novel phosphatase that dephosphorylates tyrosine 1 and serine 5 on the RNA Polymerase II CTD[1]

2.1 Introduction

The phosphorylation state of the C-terminal domain (CTD) of RNA Polymerase II (PolII) Rpb1 subunit controls transcription (Hsin and Manley, 2012). The CTD consists of a highly conserved heptapeptide ($Y_1S_2P_3T_4S_5P_6S_7$) repeated between 26 times in *Saccharomyces cerevisiae* and 52 times in humans. Ser2 and Ser5 are reversibly phosphorylated, while the prolines are subject to cis-trans isomerization facilitated by isomerases such as Ess1 (Kubicek et al., 2012; Meinhart et al., 2005; Morris et al., 1999; Xiang et al., 2010; Zhang et al., 2012). In addition, the Tyr1, Thr4 and Ser7 residues can also be phosphorylated, although the impact and scope of these modifications is less well understood (Chapman et al., 2007; Hsin et al., 2011; Mayer et al., 2012). The dynamic combination of post-translational modifications constitutes a 'CTD code' which helps recruit or activate various factors to the polymerase during the transcription cycle (Buratowski, 2003, 2009; Schwer and Shuman, 2011).

High levels of phosphorylation of Ser5 (Ser5P) on the CTD occur at or near the promoter and help recruit mRNA capping and transcription elongation factors (Ghosh et al., 2011; Komarnitsky et al., 2000; Mayer et al., 2010). This modification can also act as a signal for the snoRNA/snRNA termination pathway via the Nrd1-Nab3-Sen1 complex

[1] The contents of this chapter are currently under review for publication in Cell Reports

Hsu P, Yang F, Smith-Kinnaman W, Yang W, Zheng N, Mosley A, Varani G

Rtr1 is a dual specificity phosphatase that dephosphorylates Tyr1 and Ser5 on the RNA Polymerase II CTD

in yeasts (Vasiljeva et al., 2008). Ser5P is progressively dephosphorylated as the polymerase progresses into the elongation and termination phases of transcription. In contrast, Ser2 phosphorylation (Ser2P) levels are low at the start of transcription and increase as the polymerase moves along a gene, peaking near the 3' end of genes, where this modification signals the recruitment and/or activation of transcription termination factors (Ahn et al., 2004; Gu et al., 2013; Lunde et al., 2010).

Multiple Ser2/5 kinases and phosphatases have been identified (Hsin and Manley, 2012), but the identity of the phosphatase responsible for the critical transition from Ser5P to Ser2P during transcriptional elongation remains unclear. Yeast Rtr1 was recently proposed to be the Ser5P phosphatase responsible for this transition (Mosley et al., 2009), a hypothesis further supported by the independent observation that its human orthologue (RPAP2) has phosphatase activity with identical selectivity profile: active on Ser5P, but not upon Ser2P nor Ser7P (Egloff et al., 2012; Mosley et al., 2009). However, this attribution was negated by the lack of *in vitro* phosphatase activity in *Kluyveromyces lactis* Rtr1, whose crystal structure also failed to reveal a canonical active site observed in other phosphatases (Xiang et al., 2012). It was proposed that the phosphatase activity detected for Rtr1 might arise from the co-purification of an *E.coli* phosphatase enzyme, although it would appear unlikely that the accidental presence of a recombinant protein from bacterial sources would yield an enzyme that selectively dephosphorylates a substrate without equivalents in bacteria.

Here I resolve this controversy by reporting that Rtr1 is active as a phosphatase and that its enzymatic activity is functional: mutation in a single absolutely conserved residue that significantly reduces catalytic activity *in vitro* also abolishes its function *in*

vivo. My collaborators and I further show that Rtr1 can target and dephosphorylate PolII CTD repeats carrying both Ser5P and the newly described anti-termination Tyr1 phosphorylation marker, providing additional evidence that Rtr1 is the phosphatase that promotes the transition from initiation to the elongation and termination phases of transcription.

2.2 Results

2.2.1 Rtr1 is a phosphatase

I independently determined the crystal structure of the *K.lactis* Rtr1 (KlRtr1) NTD (amino acids 1-156, Table 2.1), which is nearly identical to the previously determined structure (Xiang et al., 2012) (Cα RMSD = 0.35Å) (Fig. 2.1). Purification of the full-length KlRtr1 protein using standard protocols (Fig. 2.2A, upper flow) resulted in preparations that lacked activity when assayed against both phosphorylated GST-CTD (data not shown) and the acid phosphatase substrate 6,8-difluoro-4-methylumbelliferyl phosphate (DiFMUP) (Fig. 2.2B), a classical phosphatase substrate. However, a closer examination of purification protocols in light of reports that some phosphatases are inhibited by very low concentrations of divalent metal ions (Wilson et al., 2012), prompted us to consider the possibility that activity was abolished by an inhibitory metal. Thus, I re-purified the KlRtr1 protein with one additional step: washing the protein with EDTA prior to the final gel filtration step (Fig. 2.2A, lower flow). This EDTA-treated protein exhibited robust activity against the phosphatase substrate (Fig. 2.2B) and the GST-CTD (see below). Structures solved with preparations from both purification

methods still contain zinc bound by the CCCH zinc finger, implying that Rtr1 retains zinc in a manner that is inaccessible to EDTA.

Mass spectrometry did not reveal the presence of any other co-purified protein in our preparations, but very low levels of contamination cannot be ruled out, leaving open the possibility that phosphatase activity could arise from an *E.coli* protein (Xiang et al., 2012). To demonstrate that catalytic activity resides in Rtr1, I introduced point mutants within highly conserved residues, including the zinc finger motif. Mutations of the conserved Cys residues resulted in obviously inactive, insoluble unfolded proteins, consistent with the critical role of zinc coordination in protein folding. More revealing, a conservative mutation of the strictly conserved Glu66 (Fig. 2.2C) to Gln along helix 4 of Rtr1 produced soluble folded protein with significantly reduced activity against DiFMUP (Fig. 2.2B). The loss of activity following a structurally conservative Glu-Gln substitution conclusively demonstrates that the enzymatic activity originates with Rtr1 and not from co-purified contaminants.

To investigate whether the phosphatase activity of Rtr1 is functionally important in cells, my collaborators introduced these mutants into yeast *in vivo*. While Rtr1 is not an essential gene, cells lacking folded Rtr1 (Fig. 2.3, Cys mutants) grow poorly at 37^{o}C under stress in the presence of formamide (Gibney et al., 2008). Strikingly, cells bearing the E66A mutant behave similarly to the Cys mutants. Thus, abrogating the enzymatic activity of the protein by a single amino acid change leads to a functional defect comparable to that observed with an obviously non-functional misfolded protein.

The enzymatic activity is sensitive to specific phosphatase inhibitors. The classical competitive phosphatase inhibitor orthovanadate inhibited Rtr1 with an

inhibition constant K_i of approximately 0.8μM (Fig. 2.4A). However, when I performed the same experiment using another phosphatase inhibitor, β-glycerophosphate (BGP), I observed no inhibitory activity at comparable concentrations. Exhaustive efforts to soak or co-crystallize a wide range of known phosphatase inhibitors and/or peptides to obtain an enzyme:substrate complex were unsuccessful. Since a simple treatment with EDTA yielded an active protein, it suggested to me that Rtr1 was likely pulling down an inhibitory divalent metal during purification. I performed the assay in the presence of Mg, Ni and Ca (Fig. 2.4B), common divalent cations often found in biological systems, to identify a possible metal inhibitor. However, none had an effect on the activity of the protein and the molecular basis of Rtr1's inhibition remains unknown.

I conclude that Rtr1 is an active phosphatase, which is unrelated structurally to other known such enzymes, and that the function of the protein *in vivo* is directly related to its enzymatic activity.

2.2.2 The phosphatase activity resides within the conserved N-terminal domain and is regulated by the C-terminal region

Rtr1 is a highly conserved protein in all eukaryotes, even if it is a nonessential gene (Gibney et al., 2008). Based on the crystal structure and the sequence information, yeast Rtr1 proteins can be divided into a highly conserved N-terminal domain (NTD) of approximately 150 residues and a less conserved region in its C-terminus, referred to as the C-terminal region (CTR) (Fig. 2.5A). Metazoan Rtr1 proteins (RPAP2) tend to be much larger proteins of approximately 600 residues, with the only significant conservation found within the NTD.

Given that the crystal structure of the most conserved domain revealed no putative active site, and that activity was greatly weakened by a single point mutant, I considered the possibility that activity could originate at the interface between the two domains. I thus divided KlRtr1 into two halves, purified the constructs, and assayed their activities. The CTR exhibited no activity above background, while both NTD and full protein had robust activity and similar affinities for the substrate (Table 2.2). Interestingly, the NTD exhibited a nearly 30% increase in V_{max} relative to the full protein (Fig. 2.5B), suggesting that the CTR regulates enzymatic activity via an allosteric mechanism. Thus, phosphatase activity resides completely within the conserved NTD, which constitutes a new structure for such enzymes.

Based on this allosteric inhibition of the NTD by the CTR, I asked if the previously observed inhibition observed during purification is alleviated with the deletion of the CTR. I thus purified the NTD using both purification methods as illustrated in Fig. 2.2A and measured its kinetics against DiFMUP. KlRtr1 NTD purified using either method yielded active protein with similar kinetic parameters against the substrate (Fig. 2.5C). I hypothesize, based on these observations that CTR inactivation of the phosphatase activity is due to the presence of a divalent metal "gluing" the CTR into a conformation that inactivates the NTD.

2.2.3 Allosteric auto-inhibition confines the active site

I was unable to crystallize the complete *K.lactis* Rtr1 protein, and NMR analysis by my colleague on *S.cerevisiae* Rtr1 (ScRtr1) yields spectra of mixed quality. Peaks from the NTD are well dispersed, as expected for a domain with a well-defined fold, while peaks

originating from the CTR cluster within a narrow spectral region (8-8.5 ppm) and are more intense, suggesting that this region of the protein is only partially structured (Fig. 2.6A). Backbone dynamics as measured by NMR also indicate that the CTR is partially structured, with the extreme C-terminus displaying great flexibility (Fig. 2.6B). Unambiguous NOEs were nonetheless observed between residues belonging to the CTR and the NTD (Fig. 2.7A, left), indicating an interaction between these two regions. Significantly, the residues contacted by the CTR occur near the zinc finger on the "back" face of the protein (Fig. 2.7A, right), on the opposite side of the invariant Glu66 residue that drastically reduces activity (Fig. 2.2B, Fig. 2.3 and Fig. 2.7A, right). Furthermore, in a partially refined NMR structure of the complete ScRtr1 protein, the CTR assumes a position on the back face of the NTD, occupying a large portion on the back face of the NTD (Fig. 2.7B). However, since kinetic data from Fig. 2.5B suggest that the CTR is an allosteric noncompetitive regulator of NTD activity the active site of Rtr1 cannot reside on this back face of the protein. Due to the NMR data having been collected on an inactive sample, and my kinetic experiments from Fig. 2.5C, these data together suggest that the conformation of the CTR shown in the NMR structure reflects an inactive form of the protein. Likely the contaminating metal helps stabilize the CTR in this conformation, locking Rtr1 into an inactive state.

Sequence alignment of Rtr1 (Fig. 2.8A) reveals an invariant (in yeasts) glutamate (KlRtr1 Glu197) near the extreme C-terminus of the enzyme, as well as a number of highly conserved residues surrounding this glutamate. Several of these conserved residues display NOE connectivies to the NTD (Fig. 2.7A, left). Given its invariance and structural connectivity to a key structural element in the NTD, I hypothesized that this

conserved residue of the protein regulates enzyme activity. When Glu197 was conservatively mutated to Gln, we observed an increase in K_{cat} relative to the wild type full-length enzyme. Comparison of the kinetic parameters of this mutant with the NTD construct shows a similar $V_{max}/K_{cat,}$ suggesting abrogation of allosteric auto regulation by mutation of a single conserved glutamate (Fig. 2.8B, Table 2.2).

I also crystallized and solved the structure of full-length ScRtr1, although to a much lower resolution compared to KlRtr1's NTD (~4Å) (Table 2.1). The low resolution structure revealed no additional electron density for the CTR aside from a single helix stabilized by packing interactions. While the Sc NTD structure is nearly identical to the *K.lactis* protein, I observed very clear features of a loop (residues ~70-100) in the low-resolution electron density map of ScRtr1 (Fig. 2.9A), not seen in the crystal structure of the KlRtr1 NTD. Backbone dynamics conducted by NMR relaxation methods confirmed that the loop is flexible (Figure 2.6B). This loop is stabilized by a packing interaction with a neighboring molecule in the crystal (Fig. 2.9A) and forms a V-shaped crevasse with helices 4 (including Glu66) and 5, potentially forming a structurally dynamic active site.

To investigate the role of this loop in enzymatic activity, I introduced a series of internal deletions in this loop in our KlRtr1 NTD construct to assay for activity. Assays of NTDΔ90-99 show a near 40% drop in V_{max}; an even larger deletion (NTDΔ85-99) shows a more significant decrease in V_{max} (near 70%) (Fig. 2.9B). Gel filtration of these constructs show elution volumes similar to that of the wildtype NTD, suggesting that deletion of this loop did not affect the folding of this domain (data not shown), and that the abrogation of activity is solely attributed to the loss of the loop. The decrease in V_{max},

also suggest that the shape of the active site has been altered by our loop deletions, resulting in a loss of overall catalytic activity.

To summarize (Figure 2.9C), the noncompetitive nature of the regulation of the NTD suggests that the CTR does not mask the active site. Given the CTR's connection with the NTD via the back side of the domain, and an activity-disrupting mutant on the opposite side next to this conserved loop, I conclude that the active site is located within the front side of this protein near this loop.

2.2.4 Rtr1 targets both Ser5P and Tyr1P for dephosphorylation

In order to establish the specificity of purified KlRtr1 protein on the CTD, my collaborators carried out phosphatase assays using GST-CTD phosphorylated with purified TFIIH (Fig. 2.10A). KlRtr1 dephosphorylates Ser5P but not Ser2P and Ser7P. Quantitation of the signals from the blots clearly shows robust dephosphorylation of Ser5P with increasing enzyme (Fig. 2.10B).

Recent work identified Tyr1 phosphorylation within the *S.cerevisiae* CTD as an anti-termination marker (Mayer et al., 2012). Namely, chromatin immunoprecipitation profiles of Tyr1P show an enrichment of the phospho-marker during the elongation phase of transcription, in agreement with its putative role in preventing the premature recruitment of transcription termination factors. Given the role of Rtr1 in dephosphorylating Ser5P during transcription to generate the predominant Ser2P form observed in late phases of transcription, we wondered whether Rtr1 would dephosphorylate Tyr1P as well.

The kinase activity of TFIIH is only weakly active on tyrosines *in vitro* (Fig. 2.10A, lower left panels). Thus, we used Abl kinase and TFIIH together to effectively phosphorylate serines and tyrosines on the CTD *in vitro* (Baskaran et al., 1997). We then assayed the resulting modified polypeptide with an anti-Tyr1P antibody (Chromotek) in parallel with the other phospho-specific antibodies. As shown in Fig. 2.10A, Rtr1 dephosphorylates Tyr1P as well: the signal for the Tyr1P-specific antibody decreases with increasing amounts of Rtr1. The presence of the Tyr1P marker does not disrupt the ability of the enzyme to dephosphorylate Ser5P (Fig. 2.10A,B), since quantitation shows that dephosphorylation is comparable for Tyr1 and Ser5, but the signals for Ser2P and Ser7P remain constant after Rtr1 treatment. In addition to highlighting a new and unexpected phosphatase specificity, this result suggests that Rtr1 does not recognize and bind to the CTD like other well-characterized CTD-interacting proteins, which are repelled by the presence of the Tyr1P marker (Mayer et al., 2012).

Finally, I tested the ability of Rtr1 to dephosphorylate CTD peptide mimics to quantify its activity against near-native substrates using commercially available calf intestinal phosphatase (CIP) as a positive control. Surprisingly, Rtr1 could not dephosphorylate a 10-mer Ser5P peptide, nor a four repeat CTD (Ser5P) peptide, or a Tyr1P-containing peptide even at high concentrations of enzyme (20μM). As a positive control, peptides treated with CIP showed robust phosphate release after reaction termination by the addition of malachite green (Fig. 2.10C). The activity of Ssu72 is stimulated by the Pro cis-trans isomerase Ess1 (Werner-Allen et al., 2011; Xiang et al., 2010), but addition of Ess1 did not stimulate Rtr1 to dephosphorylate the Ser5P peptides either (data not shown). These results are in apparent contrast with the data on the

complete CTD and raise two non-mutually exclusive explanations: either the enzyme is highly processive and only act on multiple repeats, or a specific combination of phospho-markers are needed for Rtr1 to recognize its substrate, which can only be achieved combinatorially on a long enough polypeptide to present the relevant marker.

2.3 Discussion

The interplay of kinases and phosphatases that act upon the C-terminal domain of RNA PolII regulates and times the synthesis and biogenesis of cellular RNAs. However, the critical transition phosphatase that removes the Ser5P marker and shift the polymerase to the elongation and termination mode remains to be firmly established. Rtr1 (RPAP2 in vertebrates), a highly conserved protein in all eukaryotes, was proposed to be such a phosphatase in two independent studies showing that Rtr1 in both yeasts and vertebrates can specifically dephosphorylate the CTD Ser5P (Egloff et al., 2012; Mosley et al., 2009), but this conclusion was negated by the report that a highly purified, crystallized *K. lactis* Rtr1 was inactive (Xiang et al., 2012). I demonstrate here conclusively that Rtr1 is a phosphatase of new structure and attribute previous results on the lack of enzymatic activity to the absence of EDTA in the purification protocol, which resulted in an inactive protein (Fig. 2.2A,B). I further show that the phosphatase activity of Rtr1 is functionally important. Mutation of the absolutely conserved Glu66 to Gln reduces catalytic activity significantly and leads to the same phenotype *in vivo* observed for mutations of the zinc coordinating Cys residues (Fig. 2.3), which generate an unfolded, obviously non-functional protein.

The use of EDTA during purification was understandably overlooked because Rtr1 requires a single structural zinc ion to maintain its fold. However, Rtr1 maintains its hold on the structural zinc ion, once expressed, even in the presence of high concentrations of chelating agents, and in my hands required no additional zinc to be added to the growth media, or purification solutions (Xiang et al., 2012). While I was able to crystallize and obtain crystals of full length protein that diffracted to 4Å, I did not identify an obvious contaminating metal, likely due to the low resolution.

Rtr1 displays linear rates of product formation over a one hour time course as assayed by both fluorescence (observation of product), or malachite green (observation of phosphate release) and is inhibited specifically by a classical phosphatase inhibitor. However, it is an inefficient enzyme when compared with other Ser5P phosphatases, such as Ssu72 and Scp1 (Zhang et al., 2006, 2011). Even when compared to the Ser2P phosphatase Fcp1 (Hausmann and Shuman, 2002), the slowest known CTD phosphatase, Rtr1 is nearly 400 times slower, at about $1x10^{-3}$ s^{-1} *in vitro* against DiFMUP. The poor turnover rate of Rtr1 is likely a reflection of its structure, which lacks a well-defined pocket or groove to serve as an active site. I provide this conjecture also as an explanation for my inability to crystallize an enzyme-inhibitor complex despite exhaustive attempts.

Why would nature choose such a slow and inefficient enzyme as the Ser5 transition phosphatase? ChIP data show that Rtr1 associates with PolII at the start of transcription and enrichment gradually declines as Ser5P levels decline as well (Mosley et al., 2009). Data on the recruitment of human Rtr1 also suggest that the enzyme is recruited during the assembly of PolII in the nucleus (Forget et al., 2013). I hypothesize

that the premature dephosphorylation of Ser5P by an efficient enzyme would prevent recruitment of the mRNA capping machinery, a critical modification for mRNAs (Rodriguez et al., 2000; Schwer and Shuman, 2011), and/or favor premature termination. Rtr1 could be made more active as transcription progresses by an effector signal via post-translational modifications, interactions with RNA PolII, or an as yet identified protein partner(s).

Kinetic measurements revealed an allosteric auto-inhibitory function for the partially conserved CTR of Rtr1, mediated by a conserved glutamate located near the C-terminus of the protein and by nearby residues. Mutation of this single residue alleviated the partial inhibition of Rtr1 (Fig. 2.8B). While suppression of enzymatic activity is only ~30%, this observation is consistent with the hypothesis that Rtr1 has naturally evolved to be a kinetically slow enzyme to control the timing of Ser5 dephosphorylation. In my *in vitro* characterization, I cannot rule out that Rtr1's inhibition by a divalent metal is not reflective of a native state where a regulatory mechanism is in place to activate the protein by removing the inhibitory metal.

My collaborators and I also observe that Rtr1 is a dual specificity phosphatase, which acts not only on Ser5P but on Tyr1P, a new anti-termination marker (Mayer et al., 2012) which has until now not been associated with a known phosphatase that would erase it as transcription progresses. The activity towards Tyr1P is specific: similar levels of dephosphorylation were observed for both Tyr1P and Ser5P, while levels of Ser2P and Ser7P were not affected at all (Fig. 2.10A,B). Interestingly, ChIP data show that the levels of Rtr1 do not decrease until the end of transcription (Mosley et al., 2009), in coincidence with the decline of the Tyr1P marker (Mayer et al., 2012).

Rtr1 is active on long CTD repeats, but it displays no activity towards synthetic CTD phosphopeptide mimics up to four repeats, phosphorylated on either Tyr1, Ser2 and/or Ser5 (Fig. 2.10C). This activity is not stimulated by Ess1 either, suggesting that Rtr1 does not require a *cis* proline conformation for dephosphorylation (data not shown). The poor rate of catalysis, as well as the difference in lengths and phosphorylation patterns of long GST-CTD substrates compared to well defined synthetic peptides, suggest that the activity of Rtr1 is stimulated by an as of yet undetermined phosphorylation pattern. Perhaps Rtr1 requires highly phosphorylated repeats such as those generated by our *in vitro* kinase assays in which Ser2, Ser5, Ser7, and Tyr1 were phosphorylated in a likely mix of combinations *in vitro*. This hypothesis is supported by studies on RPAP2 that suggest that Ser7 phosphorylation is required for RPAP2 activity (Egloff et al., 2012) and studies on Rtr1 that show interaction with both the Ser5 and Ser2 forms of PolII (Mosley et al., 2013). During the elongation phase of transcription, the CTD is heavily modified, with Ser2/5 and Tyr1 as known phospho-markers, and potential marks at Thr4 and Ser7 as well. Overlapping phospho-marks between neighboring repeats can also specify a recruitment signal for CTD interacting proteins (Egloff et al., 2010). In addition, Rtr1 could be a processive enzyme with significant activity over longer phosphorylated substrates.

I propose a model in which Rtr1 is recruited to PolII during assembly of the polymerase (Forget et al., 2013). Due to its slow kinetics and inability to dephosphorylate isolated Tyr1 and Ser5 phospho-marks, Rtr1 remains largely inactive during transcriptional initiation, capping and promoter clearance (Fig. 2.11 top). As the polymerase shifts fully into processive elongation and additional phospho-markers are

deposited along the CTD, Rtr1 is activated via a specific but unknown 'CTD code' and more efficiently removes the Tyr1 and Ser5 markers, progressively setting the polymerase into the transcription termination mode (Fig. 2.11 bottom). This model does not rule out additional factors that could stimulate Rtr1 either by direct binding or post translational modifications. Future work will be needed to systematically identify the exact CTD substrate recognized and dephosphorylated by Rtr1.

In conclusion, the data presented here demonstrate that Rtr1 is a phosphatase of novel structure that removes the Tyr1P and Ser5P markers from the PolII CTD, albeit inefficiently. Thus, it is the phosphatase responsible for the transition to the elongation and termination phase of transcription. Future work will be needed to elucidate its exact CTD substrate, as well as its catalytic mechanism, which may be distinct from that of known phosphatases to which Rtr1 bears no structural homology.

2.4 Materials and methods

2.4.1 Protein expression and purification

Saccharomyces cerevisiae (Sc) and *Kluoverymyces lactis* (Kl) Rtr1 proteins were cloned into a modified pET-28a (Novagen) vector with a Protein G B1 domain (GB1) fused to the N-terminus to facilitate expression. Plasmids encoding the gene were transformed into Rosetta DE3 *E.coli*, shaken at 37^{o}C until induction with IPTG and expressed overnight at 18^{o}C. Cells were harvested the next morning and resuspended in lysis buffer (50mM HEPES pH7.5, 200mM NaCl, 30mM imidazole, 5mM βME), lysed by sonication and cleared by high-speed centrifugation. Lysate was applied to a HisTrap column (GE Healthcare) equilibrated in lysis buffer. Bound protein was eluted from the column by a

linear gradient against elution buffer (lysis buffer + 500mM imidazole). Protein-containing fractions were pooled and placed into dialysis buffer (20mM HEPES pH7.5, 100mM NaCl, 5mM βME); TEV protease was incubated overnight to remove the His-GB1 tag.

Dialyzed material was collected and re-applied to a HisTrap column equilibrated in dialysis buffer to remove the tag, TEV protease, and any uncleaved protein. Prior to applying the protein to the gel filtration column, 10mM EDTA was added to the protein to remove any residual divalent metals that may have co-purified with the protein. After incubation with EDTA, the protein was applied to a Superdex 75 (GE Healthcare) equilibrated in storage buffer (dialysis buffer but 5mM DTT substituted the βME). The protein eluted at a volume consistent with a monomer. Protein-containing fractions were concentrated to 10-20mg/mL and flash frozen using liquid nitrogen. For the KlRtr1 1-156 (NTD) construct, the protein was only concentrated to ~2mg/mL, due to more limited solubility before storage.

2.4.2 Site directed mutagenesis

Point mutants were generated using the QuikChange kit (Stratagene). All mutants were verified by sequencing. Expression and purification of the mutants were done exactly as for the wild type protein.

2.4.3 NMR sample preparation and experiments

NMR samples were prepared by growing Rosetta DE3 transformed with ScRtr1 in M9 minimal media supplemented with 0.5 g/L $^{15}NH_4Cl$, 2 g/L ^{13}C-glucose and 0-100% D_2O

(Sigma-Aldrich) as needed. Selective methyl labeling of ILV residues was done as described(Rosen et al., 1996). Samples were purified as described above, but the final storage buffer was different (20mM Bistris pH6.5, 50mM NaCl, 2mM DTT).

NMR spectra were recorded at 298K on Bruker Avance 600 and Avance 800 spectrometers equipped with triple-resonance cryoprobes and pulse field gradients. Data were processed with NMRpipe (Delaglio et al., 1995) and analyzed with CCPNMR(Vranken et al., 2005). Rtr1 backbone assignments were obtained using TROSY, trHNCA, trHN(CO)CA, trHN(CO)CACB, trHNCACB, trHNCO and trHN(CA)CO spectra on a ^{15}N, ^{13}C, ^{2}H labeled protein in 90% H2O, 10%D2O. Methyl assignments of Ile, Leu and Val were obtained using (H)CC(CO)NH and H(CC)(CO)NH spectra recorded on perdeuterated Rtr1 retaining ^{1}H, ^{13}C labels at the Ile, Leu, and Val methyl positions.

2.4.4 Crystallization, data collection, structure determination and refinement

Initial crystals of KlRtr1 were obtained by adding trypsin (Sigma) to full-length protein at a ratio of 1:10000 w/w and incubating for 30 minutes at room temperature prior to setting up drops. Crystals were obtained by hanging drop by mixing one volume of sample with an equal volume of precipitant (0.1M Bicine pH8.5, 10-20% MPD) at 4^{o}C. Crystals appeared in ~2 days and matured to final size after a week. Crystals were cryoprotected by mother liquor supplemented with 30% MPD, flash frozen in liquid nitrogen and harvested for data collection.

All datasets were collected at the Advanced Light Source at the Lawrence Berkeley National Laboratory at beam lines BL5.0.1, 5.0.2 and 5.0.3. Datasets were

indexed, integrated, and scaled with the HKL2000 package (Otwinowski and Minor, 1997). Initial phases were determined from a SAD dataset using a single SeMet derivatized crystal. The SAD dataset was collected at BL5.0.1 at a wavelength of 0.98Å and phases were determined by the Solve/Resolve program (Terwilliger and Berendzen, 1999). Initial model building and refinement were done by Coot and CNS (Brünger et al., 1998; Emsley and Cowtan, 2004).

Using the initial model as a template, we re-crystallized KlRtr1 NTD in similar conditions and obtained phases using the SeMet model as a molecular replacement solution. Final model building and refinement were done on this construct using Coot and Refmac5 as part of the CCP4 package (Winn et al., 2011). The final model had 97.8% of all residues in the favored region of the Ramachandran plot, and 2.2% in the allowed region.

Crystals of full length ScRtr1 were obtained using one volume of reductively methylated protein (Walter et al., 2006) with one volume of precipitant (0.04M Bicine pH8.5, 150mM LiCl, 20mM hexammine cobalt(III) chloride, 22-30% MPD). Crystals were cryoprotected with mother liquor +35% MPD and flash frozen in liquid nitrogen prior to data collection. The best crystals diffracted to 4Å at the Advanced Light Source. Phases were obtained by molecular replacement using our KlRtr1 NTD structure as a search model.

2.4.5 Phosphatase assays

Steady state kinetic assays were performed with varying concentrations of DiFMUP (Life Technologies) and 10μM KlRtr1 (various constructs as described in the text) in 50mM

MES, pH 5.5 at 30°C. Monitoring of product formation was observed by either extracting aliquots of the phosphatase reaction at fixed time points and quenching with Biomol Green reagent (Enzo Sciences), or by continuous observation of fluorescence emission spectra at a wavelength of 450nm. Product formation/phosphate release was determined by comparison against a standard curve of either DiFMU or phosphate. Data were analyzed and fit to the Michaelis-Menten equation using GraphPad Prism.

Assays using CTD peptides (Genscript) were done with 2mM peptide in 50mM MES at pH5.5, with varying concentrations of KlRtr1 at 30°C for 1 hour before quenching with Biomol Green. Calf intestinal phosphatase (New England Biolabs) was used as a positive control. Background was determined by performing reactions using inactive KlRtr1 purified via the upper scheme diagrammed in Fig. 2.2A

IC_{50} experiments were performed using sodium orthovanadate (New England Biolabs) and β-glycerophosphate (Sigma-Aldrich) as inhibitors. Data were plotted and analyzed using GraphPad Prism. Inhibition constants (K_i) were determined by converting IC_{50} values using the Cheng-Prusoff equation.

2.4.6 *In vivo* experiments

All yeast strains used in this study were derived from BY4741. To create Rtr1 mutant strains, a RTR1 fragment (-266 to +678) was amplified from BY4741 yeast genomic DNA and cloned into the Nco I site of pBS1539 to create RTR1-TAP (Puig et al., 2001). Mutant plasmids were created by site directed mutagenesis using a Quickchange lightning kit according to the manufacturer's instructions (Agilent). All resulting plasmids were verified by DNA sequencing. Yeast strains were created by transforming

BY4741 with a PCR product amplified from wild type or mutant RTR1-TAP plasmids and selecting transformants on complete synthetic media lacking uracyl. Expression of the mutant constructs was confirmed by western blotting. For the growth assays, 5-fold serial dilutions of yeast were spotted on YPD or YPD + 2% formamide as previously described (Gibney et al., 2008).

2.4.7 GST-CTD phosphatase reactions

GST-CTD phosphatase reactions were performed as previously described (Mosley et al., 2009). Briefly, purified recombinant GST-CTD was phosphorylated *in vitro* by TFIIH and/or Abl kinase in the presence of ATP. Unreacted ATP was removed by gel filtration. For each phosphatase reaction, approximately 5pmol of modified GST-CTD was used as a substrate in the presence of increasing concentrations of KlRtr1 as indicated. Reactions were quenched with 2X SDS-PAGE loading buffer followed by western blot analysis using the CTD phosphorylation antibodies described above.

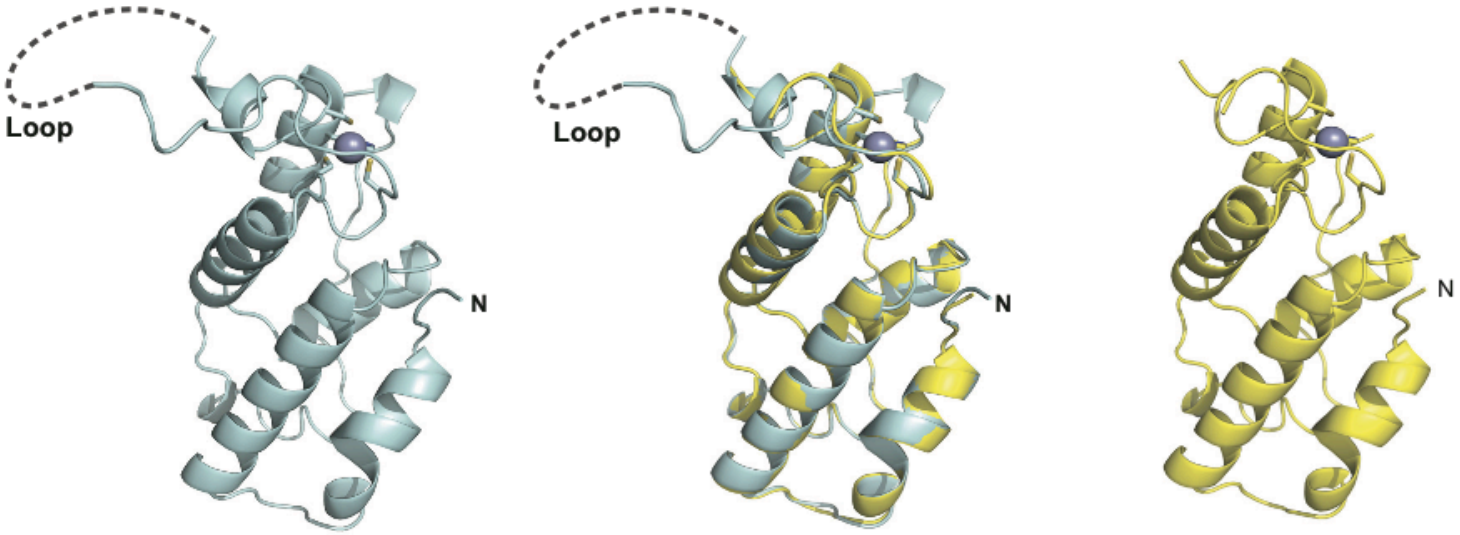

Figure 2.1 Crystal structure of K.lactis Rtr1 NTD
The crystal structure of our independently determined KlRtr1 NTD is shown on the left (cyan). The dotted line outlines a missing loop unseen in the electron density maps. A single zinc ion (grey sphere, both structures) is coordinated by a unique CCCH zinc finger. The previously reported crystal structure KlRtr1's NTD (PDB 4FC8) is shown on the right (yellow) and an overlay of both crystal structures is shown in the center.

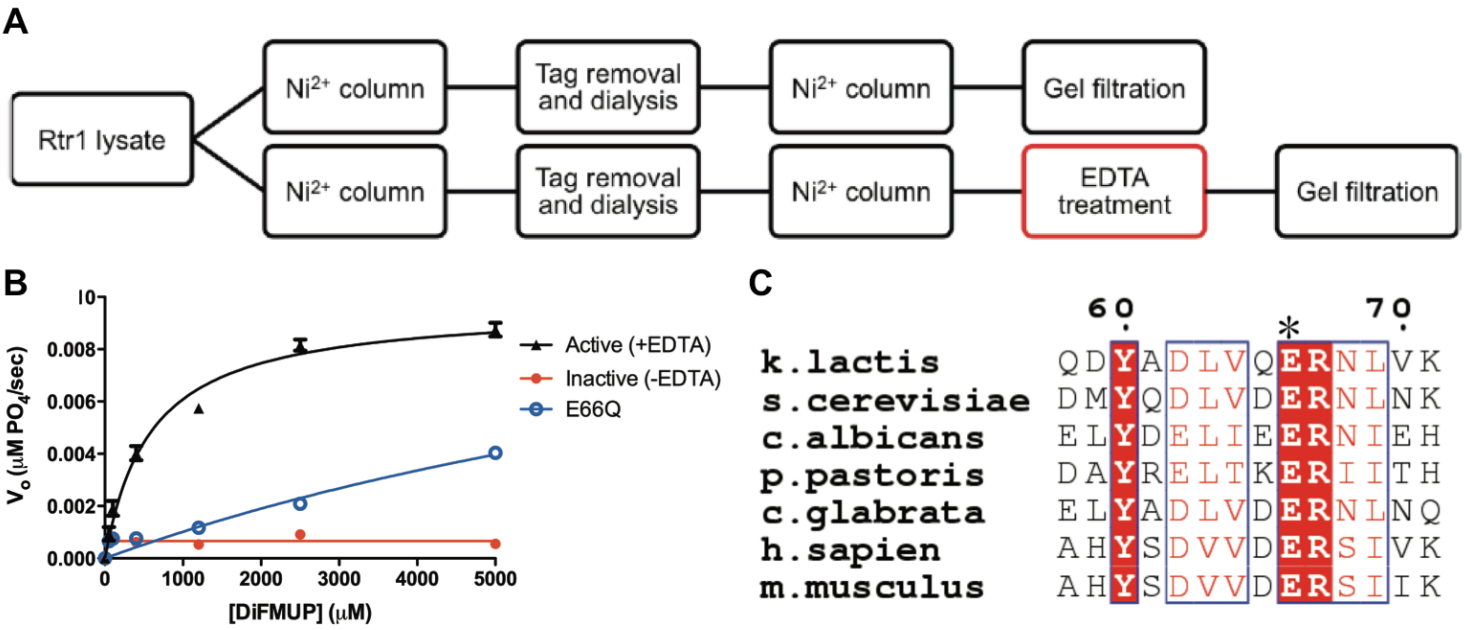

Figure 2.2 Rtr1 is an active phosphatase

A) Purification scheme adopted in this study to obtain active recombinant Rtr1 expressed in *E.coli*. The determining step in the purification necessary to obtain active enzyme is highlighted in red.

B) Steady state phosphatase assays (n = 3) (with DiFMUP substrate) performed using KlRtr1 proteins (10μM) obtained from both purification schemes, as diagrammed in part A (+EDTA in black, -EDTA in red), and the E66Q mutant (blue)

C) Sequence alignment of Rtr1 (numbering based on *K.lactis*) across yeasts and vertebrates highlighting the strictly conserved Glu66 residue (asterisked).

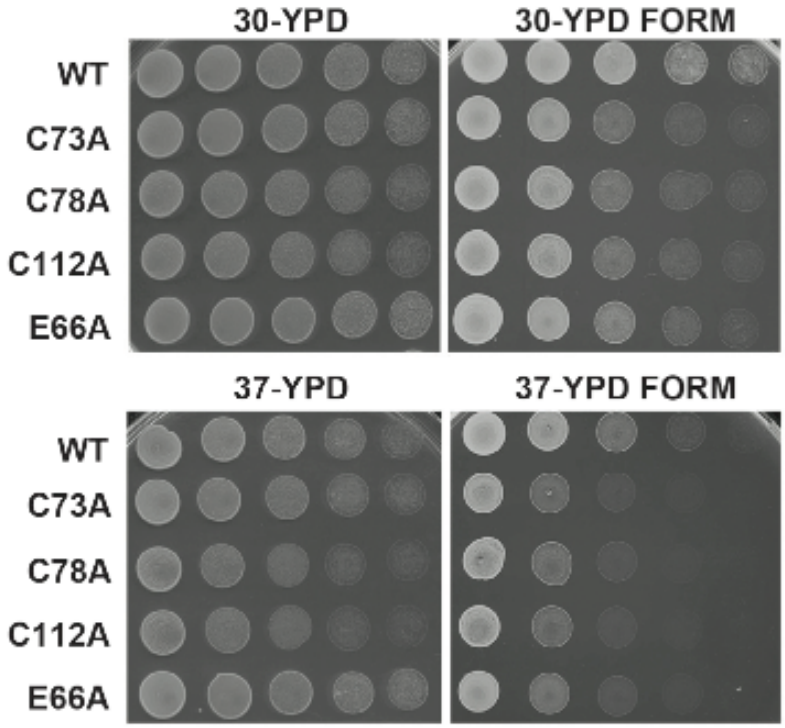

Figure 2.3 Disruption of phosphatase activity results in an observable phenotype Fitness of four mutants (E66A, C73A, C78A, C112A) and of WT ScRtr1 grown under permissive and elevated temperatures, both in the presence and absence of formamide to induce stress. The phenotypes observed with a single mutant that reduces catalytic activity are comparable to those observed when the protein is unfolded by disrupting the zinc finger.

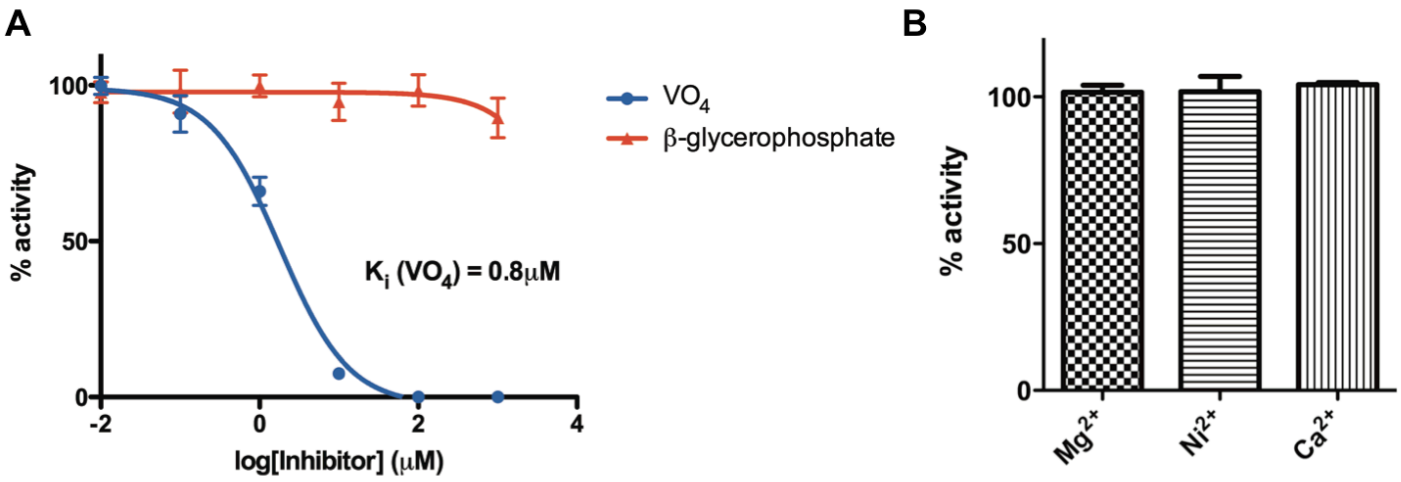

Figure 2.4 Rtr1 is inhibited by traditional phosphatase inhibitors

A) Inhibition experiments (n = 3) performed using 10μM KlRtr1 and 1mM DiFMUP against two traditional competitive phosphatase inhibitors (vanadate in blue, BGP in red). The K_i of vanadate against KlRtr1 is approximately 0.8μM.

B) Activity of KlRtr1 in the presence of divalent metal ions (n = 3) (10mM). Percent activities were normalized against reactions performed with buffer.

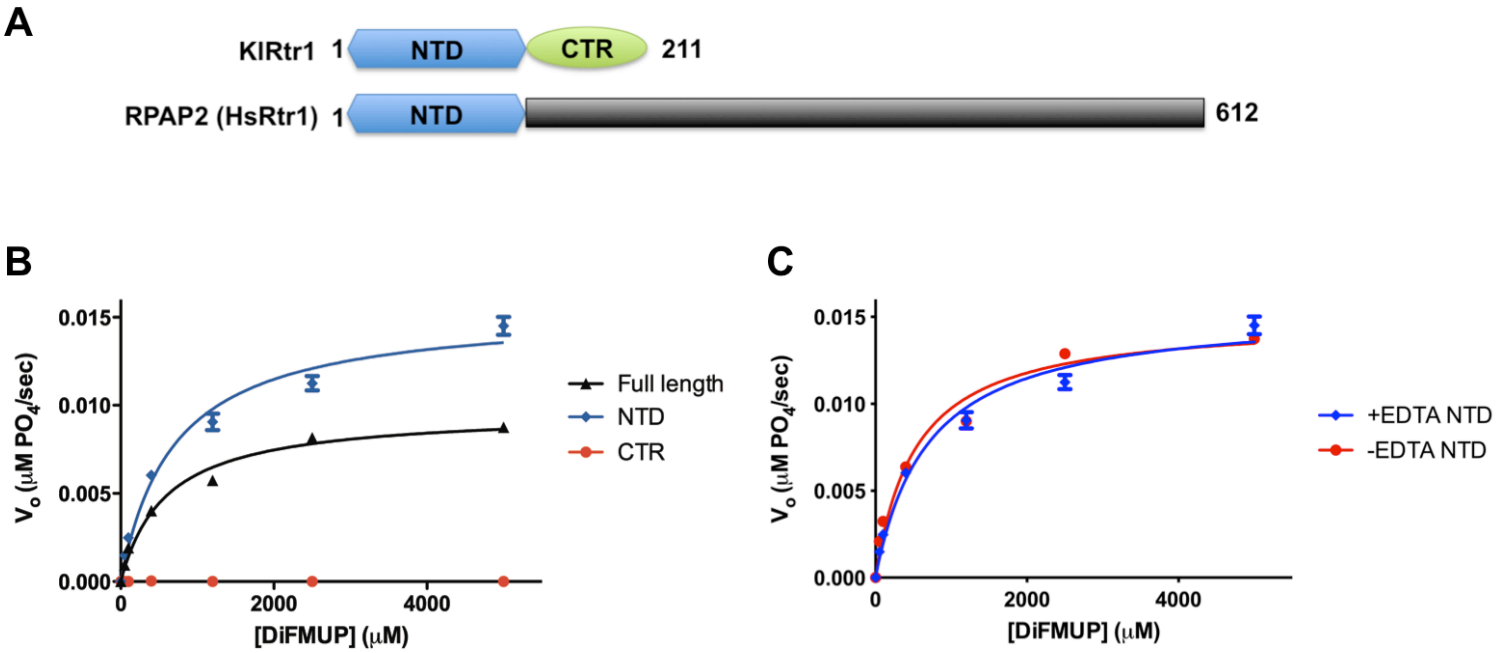

Figure 2.5 The N-terminal domain of Rtr1 is the functional phosphatase domain

A) Domain breakdown of Rtr1 in yeasts and vertebrates. Yeast Rtr1 proteins are smaller proteins with conserved N-terminal (NTD) and C-terminal domains (CTR). Vertebrate Rtr1 (RPAP2) proteins are typically larger with the only conservation to yeast Rtr1 proteins found within the NTD.

B) Steady state phosphatase assays (n = 3) of KlRtr1 against DiFMUP for full length (black), NTD (blue), and CTR (red) (10μM protein, all constructs).

C) Steady state phosphatase assays (n =3) of KlRtr1 NTD purified using methods outlined in Fig. 2.2A against DiFMUP (+EDTA in blue, -EDTA in red).

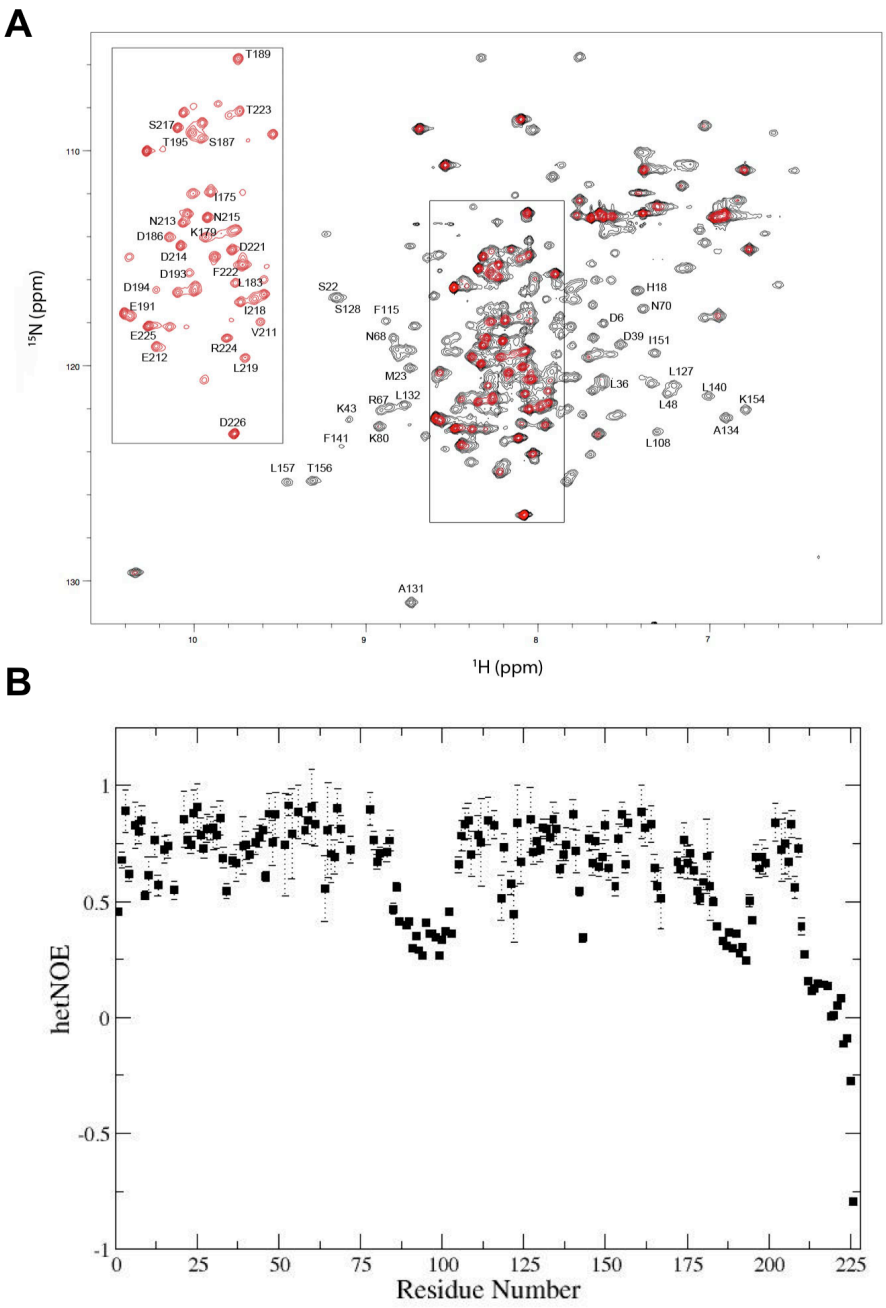

Figure 2.6 NMR analysis of *S.cerevisiae* Rtr1

A) ^{1}H-^{15}N HSQC spectra of ScRtr1. The black peaks correspond to a lower contour level, while the red peaks correspond to a much higher contour level. Partial assignments are shown on the spectra. The window highlights the more intense C-terminal peaks of ScRtr1.

B) Backbone dynamics of ScRtr1 as observed by measuring (^{1}H-^{15}N) heteroNOE (Nuclear Overhauser Effect). The dip for residues 75-100 indicates increased mobility for a potentially critical flexible loop disordered in the crystal structure of the *K.lactis* protein (see Fig. 2.1).

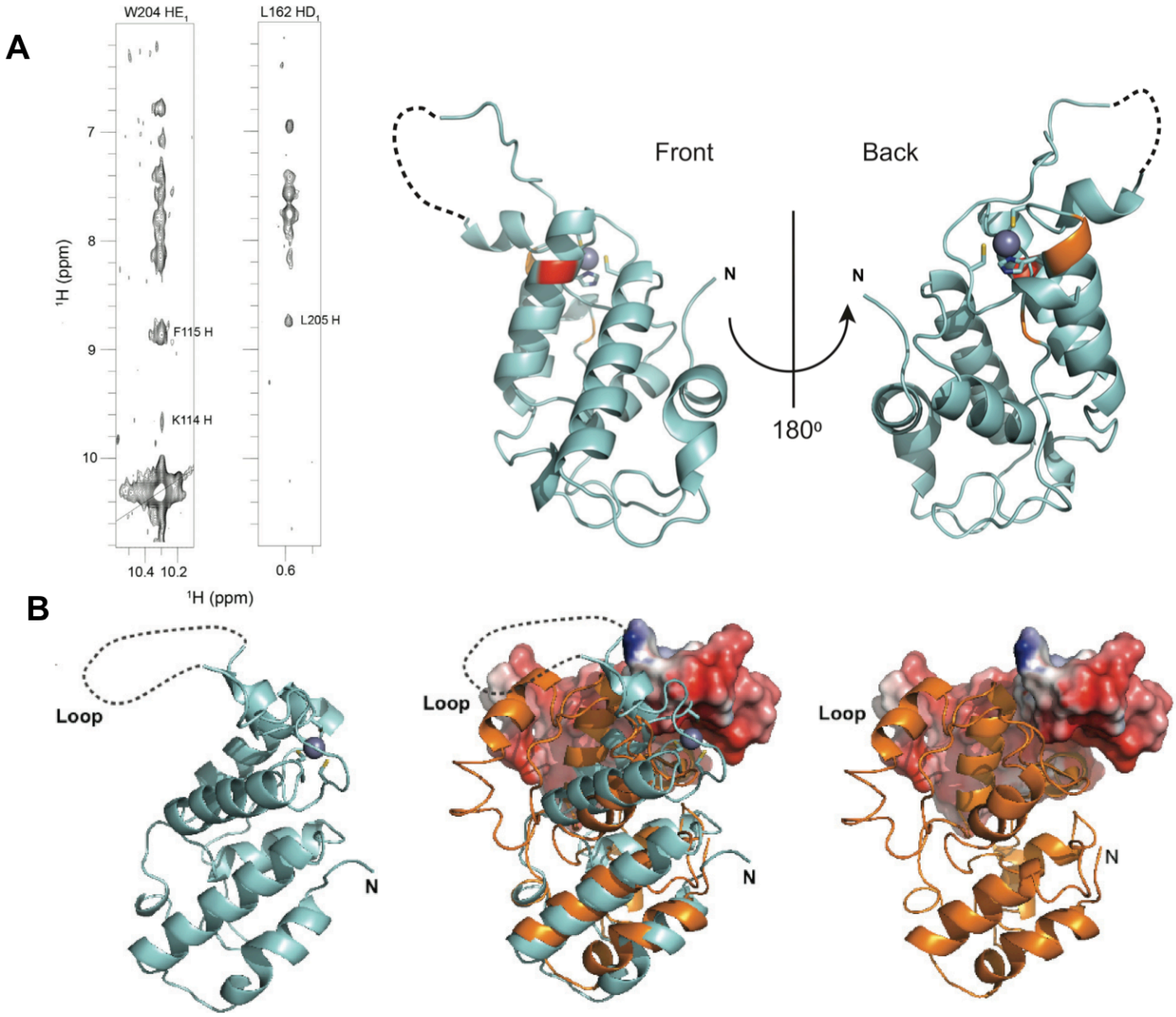

Figure 2.7 The CTR is positioned on the back side of the Rtr1 phosphatase domain

A) NOE connectivities between residues in the CTR of ScRtr1 to residues in the NTD (left panel). The equivalent residues within the NTD that the CTR makes contacts with are highlighted in orange on the crystal structure of the *K.lactis* protein. The invariant Glu66 residue is highlighted in red.

B) Crystal (left, cyan) and partially refined NMR (right, orange) structures. The NTD folds of both KlRtr1 and ScRtr1 are shown in cartoon form, while the CTR of ScRtr1 is shown as a surface diagram. The zinc ion is shown as a grey sphere and the side chains of the coordinating residues are shown in the KlRtr1 structure. Overlay of the crystal and NMR structures is shown in the center panel. The position of the N-terminus and flexible loop are noted on the structures.

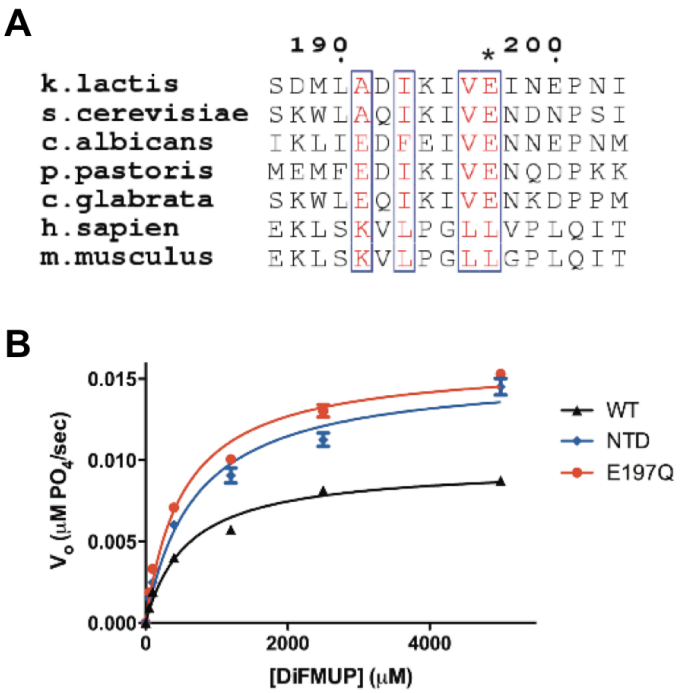

Figure 2.8 A conserved glutamate modulates auto-regulation of activity

A) Sequence alignment of Rtr1 (numbering based on *K.lactis*) across yeasts and vertebrates, focusing on the extreme C-terminus of the protein. Glu197 is asterisked.

B) Steady state phosphatase assay (n = 3) against DiFMUP of KlRtr1 E197Q (red) compared with full length protein (black) and with the NTD phosphatase domain (blue).

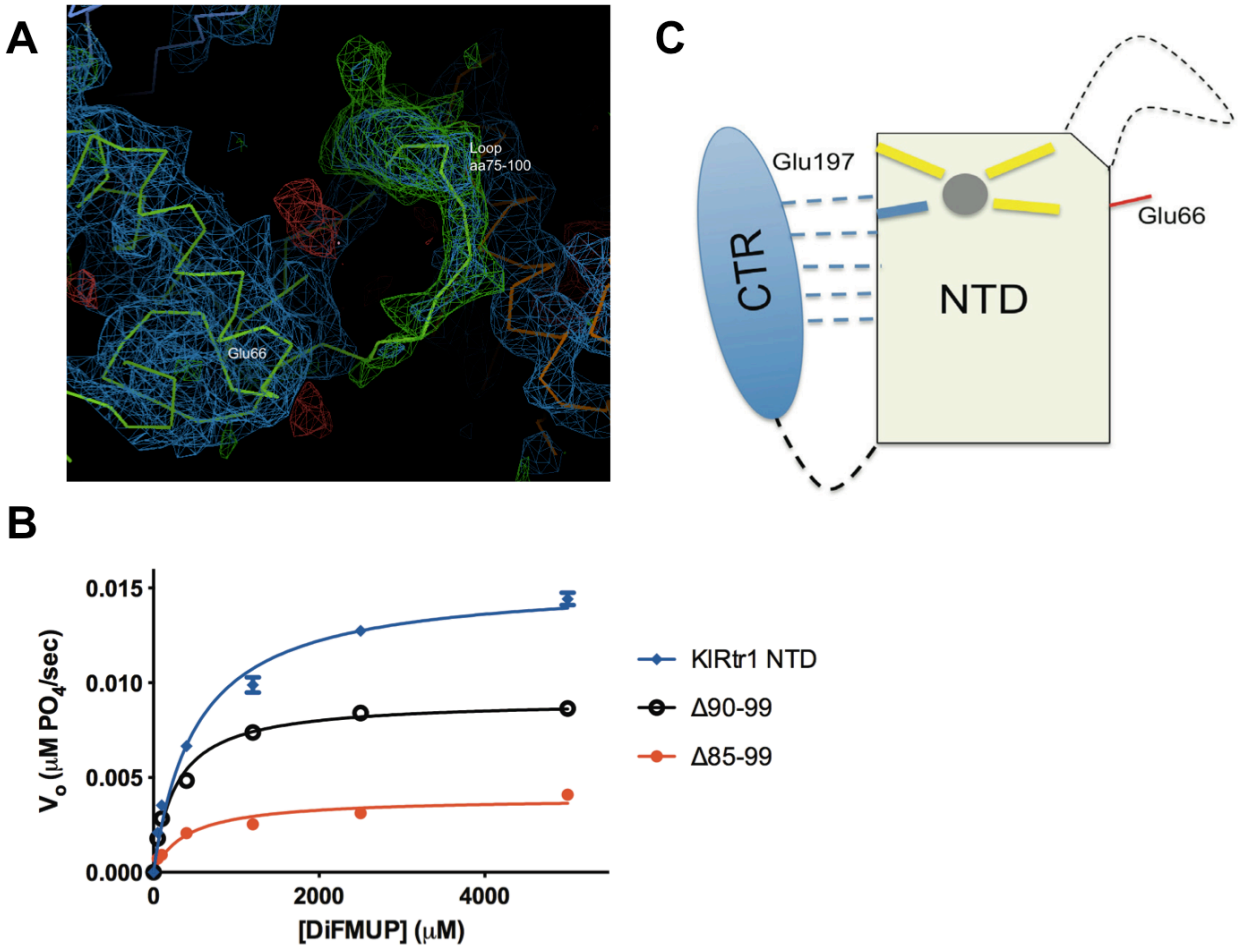

Figure 2.9 A dynamic loop helps define a potentially cryptic active site on Rtr1

A) 4Å F_o-F_c map of ScRtr1 contoured at 2.5σ, displaying a flexible loop (approximately aa 75-100) between helices 4 and 5 in the NTD not seen in the KlRtr1 crystal structure, but stabilized in the ScRtr1 structure by a packing interaction with a neighboring molecule.

B) Steady state phosphatase assays (n = 3) against DiFMUP of KlRtr1 NTD with numerous internal deletions (black and red) compared against WT KlRtr1 NTD (blue).

C) Cartoon illustration of the regulation of Rtr1 activity by its C-terminus. Zinc is shown as a grey sphere and the structurally dynamic loop is shown as a dashed line. Glu66 is denoted by a red line on the same face as the active site forming loop. The CTR weakly interacts with the NTD in the absence of a metal, as shown by the dashed lines; among interacting residues is the critical Glu197 necessary for regulation.

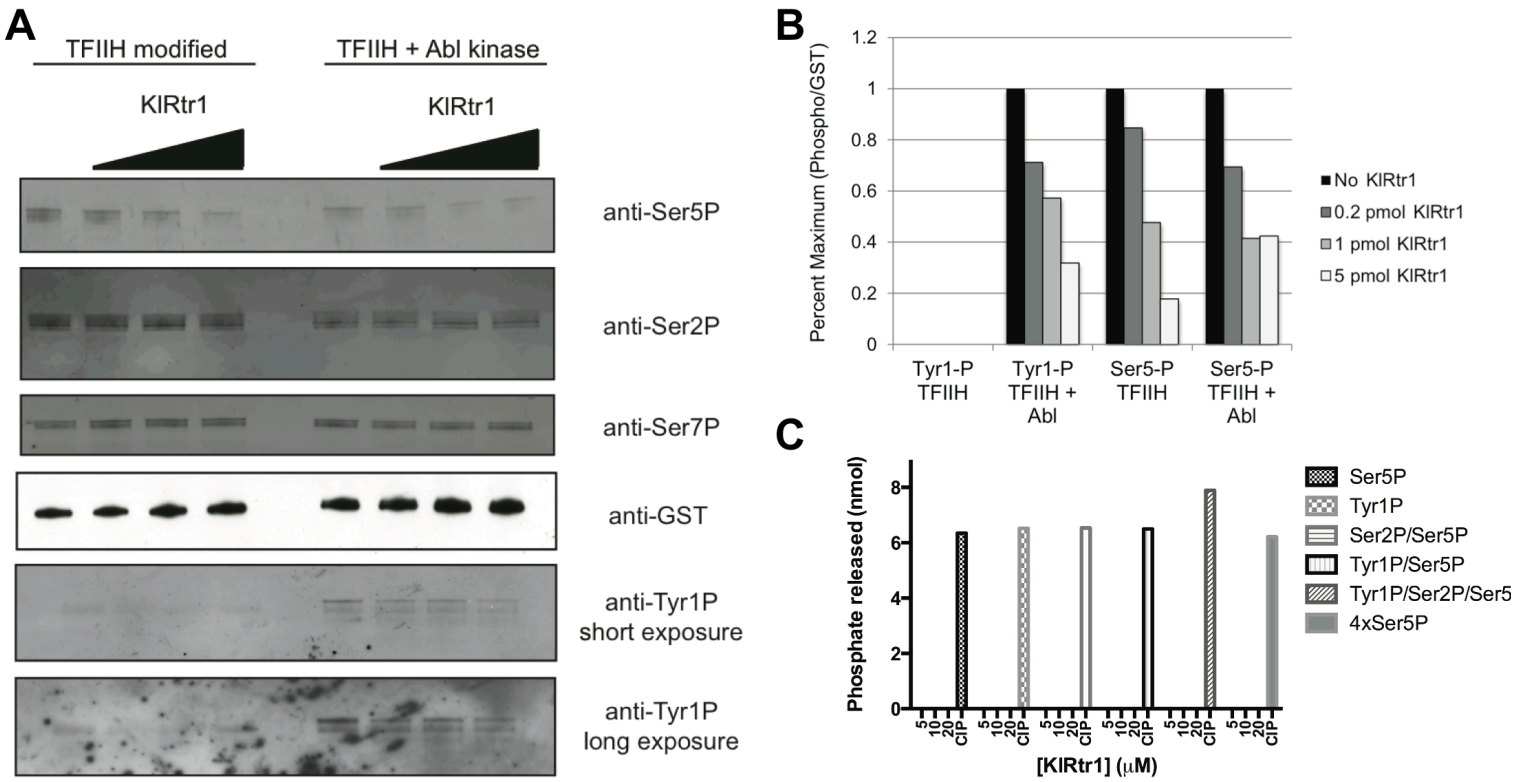

Figure 2.10 Rtr1 is a dual specificity phosphatase that acts on both Tyr1 and Ser5

A) GST-CTD phosphorylated with either purified TFIIH (left) or TFIIH/Abl kinase (right) were used as substrates for increasing amounts of KlRtr1, then probed with antibodies against Ser2, Ser5, Ser7, and Tyr1. Two exposures for Tyr1 are shown for clarity.

B) Quantitation of phospho-signals from the Ser5P and Tyr1P blots on GST-CTD modified by TFIIH or TFIIH plus Abl kinase.

C) Activity of full length KlRtr1 against CTD peptides phosphorylated on Tyr1, Ser2, Ser5 and combinations of these markers, as measured by malachite green (n = 3). CIP was used as a positive control in all experiments. 4xSer5P denotes a 4 repeat heptapeptide.

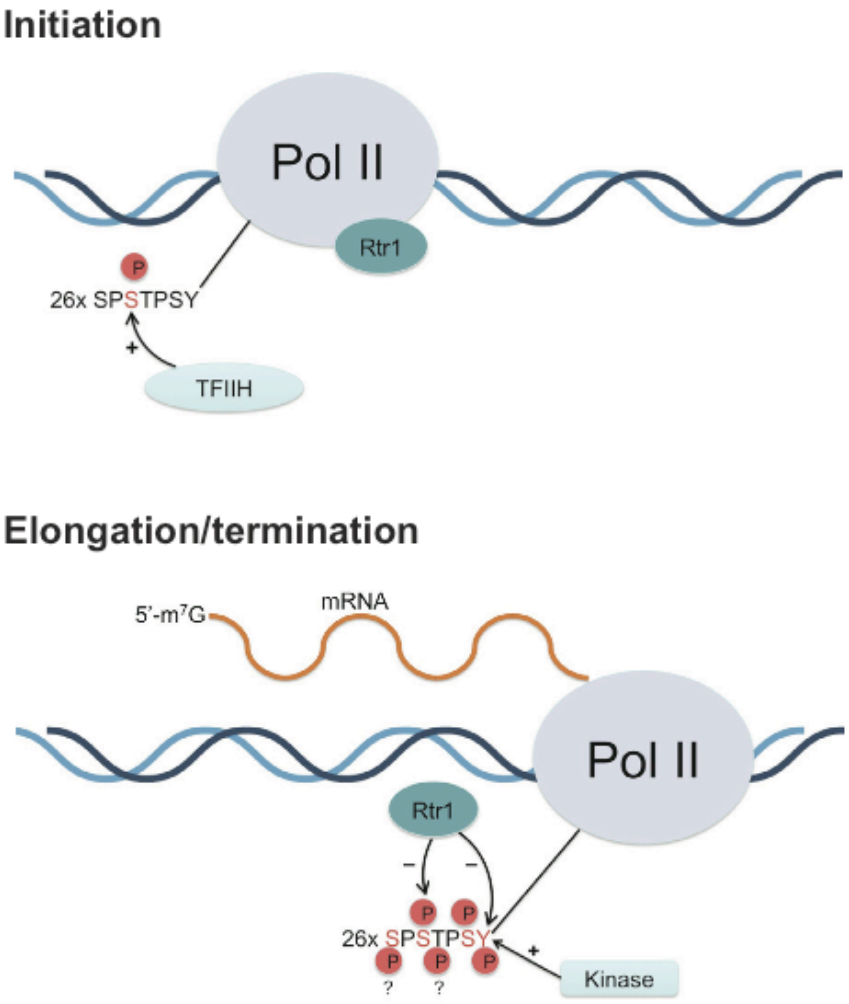

Figure 2.11 A model for Rtr1's role in the transcription cycle
The general transcription factor complex TFIIH phosphorylates Ser5 on the CTD at the start of transcription, facilitating the recruitment of mRNA capping complexes (top). As the polymerase moves into elongation and termination modes, the CTD is highly phosphorylated by multiple kinases, including an as of yet undentified Tyr1 kinase. Rtr1 is active on repeats during the elongation phase, presumably due to a signaling mechanism encoded by a phosphorylation pattern in the CTD.

Table 2.1 Crystal data collection and refinement statistics

	KlRtr1 SeMet	KlRtr1 NTD native[a]	ScRtr1 full-length
Data collection			
Space group	$P2_12_12_1$	$P2_12_12_1$	$P2_13$
Cell dimensions			
a, *b*, *c* (Å)	45.39, 103.35, 105.39	43.27, 88.92, 102.49	170.75, 170.75, 170.75
α, β, γ (°)	90, 90, 90	90, 90, 90	90, 90, 90
Resolution (Å)	50-3.19 (3.30-3.19)[b]	50-2.06 (2.10-2.06)	50-4.00 (4.07-4.00)
R_{sym} or R_{merge}	0.097 (0.620)	0.069 (0.612)	0.101 (0.540)
$I / \sigma I$	25.0 (3.9)	36.2 (3.1)	24.4 (3.9)
Completeness (%)	99.9 (100.0)	99.9 (100.0)	100.0 (100.0)
Redundancy	13.4 (12.1)	5.6 (4.8)	11.0 (11.2)
Refinement			
Resolution (Å)		44.44-2.10	
No. reflections		22581	
R_{work} / R_{free}		0.202/0.253	
No. atoms			
Protein		2,355	
Ligand/ion		2	
Water		137	
B-factors			
Protein		46.3	
Ligand/ion		35.7	
Water		64.7	
R.m.s. deviations			
Bond lengths (Å)		0.018	
Bond angles (°)		2.015	

[a] Crystals were grown from an enzymatically active protein prep. Final structure refined is that of an active protein.
[b] Values in parentheses indicate highest-resolution shell.

Table 2.2 Steady state kinetic parameters of KlRtr1 against DiFMUP (n = 3 for all experiments)

KlRtr1 construct	**K_m (µM)**	**K_{cat} (s^{-1})**
Wild-type full length	587 ± 70	0.0010 ± 0.00003
Wild-type NTD	694 ± 74	0.0015 ± 0.00005
CTR	Not detectable	Not detectable
Full length E66Q	Not detectable	Not detectable
Full length E197Q	529 ± 45	0.0016 ± 0.00004

2.5 References

Ahn, S.H., Kim, M., and Buratowski, S. (2004). Phosphorylation of Serine 2 within the RNA Polymerase II C-Terminal Domain Couples Transcription and 3′ End Processing. Mol. Cell *13*, 67–76.

Baskaran, R., Chiang, G.G., Mysliwiec, T., Kruh, G.D., and Wang, J.Y. (1997). Tyrosine phosphorylation of RNA polymerase II carboxyl-terminal domain by the Abl-related gene product. J. Biol. Chem. *272*, 18905–18909.

Brünger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.S., Kuszewski, J., Nilges, M., Pannu, N.S., et al. (1998). Crystallography & NMR system: A new software suite for macromolecular structure determination. Acta Crystallogr. D Biol. Crystallogr. *54*, 905–921.

Buratowski, S. (2003). The CTD code. Nat. Struct. Biol. *10*, 679–680.

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

Chapman, R.D., Heidemann, M., Albert, T.K., Mailhammer, R., Flatley, A., Meisterernst, M., Kremmer, E., and Eick, D. (2007). Transcribing RNA polymerase II is phosphorylated at CTD residue serine-7. Science *318*, 1780–1782.

Delaglio, F., Grzesiek, S., Vuister, G.W., Zhu, G., Pfeifer, J., and Bax, A. (1995). NMRPipe: a multidimensional spectral processing system based on UNIX pipes. J. Biomol. NMR *6*, 277–293.

Egloff, S., Szczepaniak, S.A., Dienstbier, M., Taylor, A., Knight, S., and Murphy, S. (2010). The Integrator Complex Recognizes a New Double Mark on the RNA Polymerase II Carboxyl-terminal Domain. J. Biol. Chem. *285*, 20564 –20569.

Egloff, S., Zaborowska, J., Laitem, C., Kiss, T., and Murphy, S. (2012). Ser7 phosphorylation of the CTD recruits the RPAP2 Ser5 phosphatase to snRNA genes. Mol. Cell *45*, 111–122.

Emsley, P., and Cowtan, K. (2004). Coot: model-building tools for molecular graphics. Acta Crystallogr. D Biol. Crystallogr. *60*, 2126–2132.

Forget, D., Lacombe, A.-A., Cloutier, P., Lavallée-Adam, M., Blanchette, M., and Coulombe, B. (2013). Nuclear import of RNA polymerase II is coupled with nucleocytoplasmic shuttling of the RNA polymerase II-associated protein 2. Nucleic Acids Res.

Ghosh, A., Shuman, S., and Lima, C.D. (2011). Structural insights to how mammalian capping enzyme reads the CTD code. Mol. Cell *43*, 299–310.

Gibney, P.A., Fries, T., Bailer, S.M., and Morano, K.A. (2008). Rtr1 is the Saccharomyces cerevisiae homolog of a novel family of RNA polymerase II-binding proteins. Eukaryot. Cell *7*, 938–948.

Gu, B., Eick, D., and Bensaude, O. (2013). CTD serine-2 plays a critical role in splicing and termination factor recruitment to RNA polymerase II in vivo. Nucleic Acids Res. *41*, 1591–1603.

Hausmann, S., and Shuman, S. (2002). Characterization of the CTD phosphatase Fcp1 from fission yeast. Preferential dephosphorylation of serine 2 versus serine 5. J. Biol. Chem. *277*, 21213–21220.

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

Hsin, J.-P., Sheth, A., and Manley, J.L. (2011). RNAP II CTD Phosphorylated on Threonine-4 Is Required for Histone mRNA 3′ End Processing. Science *334*, 683 –686.

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

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

Lunde, B.M., Reichow, S.L., Kim, M., Suh, H., Leeper, T.C., Yang, F., Mutschler, H., Buratowski, S., Meinhart, A., and Varani, G. (2010). Cooperative interaction of transcription termination factors with the RNA polymerase II C-terminal domain. Nat. Struct. Mol. Biol. *17*, 1195–1201.

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

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

Meinhart, A., Kamenski, T., Hoeppner, S., Baumli, S., and Cramer, P. (2005). A structural perspective of CTD function. Genes Dev. *19*, 1401–1415.

Morris, D.P., Phatnani, H.P., and Greenleaf, A.L. (1999). Phospho-carboxyl-terminal domain binding and the role of a prolyl isomerase in pre-mRNA 3'-End formation. J. Biol. Chem. *274*, 31583–31587.

Mosley, A.L., Pattenden, S.G., Carey, M., Venkatesh, S., Gilmore, J.M., Florens, L., Workman, J.L., and Washburn, M.P. (2009). Rtr1 is a CTD phosphatase that regulates

RNA polymerase II during the transition from serine 5 to serine 2 phosphorylation. Mol. Cell *34*, 168–178.

Mosley, A.L., Hunter, G.O., Sardiu, M.E., Smolle, M., Workman, J.L., Florens, L., and Washburn, M.P. (2013). Quantitative Proteomics Demonstrates That the RNA Polymerase II Subunits Rpb4 and Rpb7 Dissociate during Transcriptional Elongation. Mol. Cell. Proteomics *12*, 1530–1538.

Otwinowski, Z., and Minor, W. (1997). [20] Processing of X-ray diffraction data collected in oscillation mode. In Methods in Enzymology, (Elsevier), pp. 307–326.

Puig, O., Caspary, F., Rigaut, G., Rutz, B., Bouveret, E., Bragado-Nilsson, E., Wilm, M., and Séraphin, B. (2001). The tandem affinity purification (TAP) method: a general procedure of protein complex purification. Methods San Diego Calif *24*, 218–229.

Rodriguez, C.R., Cho, E.-J., Keogh, M.-C., Moore, C.L., Greenleaf, A.L., and Buratowski, S. (2000). Kin28, the TFIIH-Associated Carboxy-Terminal Domain Kinase, Facilitates the Recruitment of mRNA Processing Machinery to RNA Polymerase II. Mol. Cell. Biol. *20*, 104–112.

Rosen, M.K., Gardner, K.H., Willis, R.C., Parris, W.E., Pawson, T., and Kay, L.E. (1996). Selective methyl group protonation of perdeuterated proteins. J. Mol. Biol. *263*, 627–636.

Schwer, B., and Shuman, S. (2011). Deciphering the RNA Polymerase II CTD Code in Fission Yeast. Mol. Cell.

Terwilliger, T.C., and Berendzen, J. (1999). Automated MAD and MIR structure solution. Acta Crystallogr. D Biol. Crystallogr. *55*, 849–861.

Vasiljeva, L., Kim, M., Mutschler, H., Buratowski, S., and Meinhart, A. (2008). The Nrd1-Nab3-Sen1 termination complex interacts with the Ser5-phosphorylated RNA polymerase II C-terminal domain. Nat. Struct. Mol. Biol. *15*, 795–804.

Vranken, W.F., Boucher, W., Stevens, T.J., Fogh, R.H., Pajon, A., Llinas, M., Ulrich, E.L., Markley, J.L., Ionides, J., and Laue, E.D. (2005). The CCPN data model for NMR spectroscopy: development of a software pipeline. Proteins *59*, 687–696.

Walter, T.S., Meier, C., Assenberg, R., Au, K.-F., Ren, J., Verma, A., Nettleship, J.E., Owens, R.J., Stuart, D.I., and Grimes, J.M. (2006). Lysine Methylation as a Routine Rescue Strategy for Protein Crystallization. Structure *14*, 1617–1622.

Werner-Allen, J.W., Lee, C.-J., Liu, P., Nicely, N.I., Wang, S., Greenleaf, A.L., and Zhou, P. (2011). cis-Proline-mediated Ser(P)5 Dephosphorylation by the RNA Polymerase II C-terminal Domain Phosphatase Ssu72. J. Biol. Chem. *286*, 5717–5726.

Wilson, M., Hogstrand, C., and Maret, W. (2012). Picomolar concentrations of free zinc(II) ions regulate receptor protein-tyrosine phosphatase β activity. J. Biol. Chem. *287*, 9322–9326.

Winn, M.D., Ballard, C.C., Cowtan, K.D., Dodson, E.J., Emsley, P., Evans, P.R., Keegan, R.M., Krissinel, E.B., Leslie, A.G.W., McCoy, A., et al. (2011). Overview of the CCP4 suite and current developments. Acta Crystallogr. D Biol. Crystallogr. *67*, 235–242.

Xiang, K., Nagaike, T., Xiang, S., Kilic, T., Beh, M.M., Manley, J.L., and Tong, L. (2010). Crystal structure of the human symplekin-Ssu72-CTD phosphopeptide complex. Nature *467*, 729–733.

Xiang, K., Manley, J.L., and Tong, L. (2012). The yeast regulator of transcription protein Rtr1 lacks an active site and phosphatase activity. Nat. Commun. *3*, 946.

Zhang, M., Wang, X.J., Chen, X., Bowman, M.E., Luo, Y., Noel, J.P., Ellington, A.D., Etzkorn, F.A., and Zhang, Y. (2012). Structural and kinetic analysis of prolyl-isomerization/phosphorylation cross-talk in the CTD code. ACS Chem. Biol. *7*, 1462–1470.

Zhang, Y., Kim, Y., Genoud, N., Gao, J., Kelly, J.W., Pfaff, S.L., Gill, G.N., Dixon, J.E., and Noel, J.P. (2006). Determinants for dephosphorylation of the RNA polymerase II C-terminal domain by Scp1. Mol. Cell *24*, 759–770.

Zhang, Y., Zhang, M., and Zhang, Y. (2011). Crystal structure of Ssu72, an essential eukaryotic phosphatase specific for the C-terminal domain of RNA polymerase II, in complex with a transition state analogue. Biochem. J. *434*, 435–444.

Chapter 3 . Biochemical analysis of the CstF complex reveals selective RNA binding modulated by complex assembly[2]

3.1 Introduction

Cleavage and polyadenylation (3'-end processing) at the 3'-end of genes is the final step in the maturation of an mRNA in eukaryotes prior to packaging and export to the cytoplasm. This process is well conserved between yeast and humans and represents the final stage of quality control of the mRNA, since improperly processed transcripts are subject to nuclear retention and degradation. While 3'-end processing is a relatively simple reaction, consisting of an endonucleolytic cleavage of an RNA and subsequent polyadenylation at the newly formed and free 3'-OH, over a dozen proteins, assembled into various sub-complexes, are required to assemble to execute this reaction (reviewed in Mandel et al., 2008).

Early fractionation experiments identified three key complexes responsible for the recognition of key sequence elements (Takagaki et al., 1989). The Cleavage and Polyadenylation Factor (CPSF), comprised of five different subunits is responsible for the recognition of the universal metazoan polyadenylation signal, AAUAAA, via CPSF160 and is key in coordinating both cleavage and polyadenylation (Dichtl et al., 2002; Kaufmann et al., 2004; Murthy and Manley, 1995). Cleavage Factor I (CFI_m) recognizes a U-rich element, upstream (USE) of the polyA site, and is comprised of both a 68kDa and 25kDa subunit organized into a heterotetrameric complex (Li et al., 2011; Rüegsegger et al., 1996; Yang et al., 2010, 2011). Studies of CFI_m have suggested that

[2] The following have contributed to this work:

Hsu P, Yang W, Song J, Chen Y, Hinds T, Zheng N, Varani G

the complex can recognize a strong consensus sequence of UGUA, although it can still accommodate other U-rich RNAs. The Cleavage Stimulatory Factor (CstF) recognizes the downstream sequence element (DSE), which is heavily G/U-rich in sequence, and plays a key role in polyA site selection. Unlike the polyA site and the USE, the DSE has no clear consensus sequence (Legendre and Gautheret, 2003; Salisbury et al., 2006).

CstF is composed of three subunits of molecular weights 77kDa, 64kDa, and 50kDa (Fig. 3.1). The 77kDa subunit is highly conserved, with homologs found in all eukaryotes. Crystal structures of CstF77 show an N-terminal helical HAT (Half-a-TPR) domain that dimerizes strongly, and biochemical studies show an interaction between this domain and CPSF160, linking the two complexes together to establish a connection between the DSE and polyA site (Bai et al., 2007; Legrand et al., 2007). In addition, CstF77 has been shown to act as the primary subunit with which the other two components of CstF assemble (Moreno-Morcillo et al., 2011a; Takagaki and Manley, 2000).

CstF64 is the primary RNA binding component of the complex, with an RNA recognition motif (RRM) domain found in the N-terminal ~100 residues. Both solution and crystallographic studies have shown that CstF64 and its yeast homolog bind RNAs representative of the G/U-rich DSE with weak affinity, yet high specificity, discriminating against adenine and cytosine nucleotides (Pancevac et al., 2010; Perez Canadillas and Varani, 2003; Takagaki and Manley, 1997). CstF64's Hinge domain binds CstF77's C-terminal tail in a highly intertwined fashion, suggesting the two proteins have evolved to be stoichiometrically related to one another (Moreno-Morcillo et al., 2011a). The Hinge domain is also required for the proper assembly and nuclear import of the

CstF complex (Hockert et al., 2010). The C-terminus of CstF64 contains a small, but conserved domain that is used for the recruitment of other components of the 3' -end processing and termination complex (Qu et al., 2007). Like CstF77, CstF64 is also highly conserved, with identified homologs in all eukaryotes.

The smallest subunit, CstF50 has been identified only in multicellular eukaryotes, and currently has no known yeast homolog. CstF50 contains a predicted seven bladed WD40 domain that has been shown to interact with CstF77 (Takagaki and Manley, 2000). In addition, its N-terminus contains a small dimerization domain that mediates the homodimerization of CstF50 (Moreno-Morcillo et al., 2011b). Its biochemical role in 3'-end processing reactions is unclear, however, it has been suggested that CstF50 helps to bridge a link between the DNA damage response and 3'-end processing via its interaction with BARD1 (Edwards et al., 2008; Kleiman and Manley, 1999, 2001).

With two of the three subunits of CstF forming homodimeric associations, it has been suggested that CstF forms a dimer of heterotrimers (Bai et al., 2007; Legrand et al., 2007; Moreno-Morcillo et al., 2011b). While electron microscopy (EM) of the yeast homolog of CstF77/64 shows a tetrameric assembly (Gordon et al., 2011; Noble et al., 2004), no confirmation of the hexamerization of the CstF complex has been reported. While much has been done on the characterization of CstF64's RNA binding properties, only a single study on the RNA binding properties has been performed using natively purified complex, which may contain heterogeneities in the form of low amounts of copurifying binding partners (Takagaki and Manley, 1997). I present the *in vitro* reconstitution of a minimal CstF complex from recombinant sources, and demonstrate that CstF is a hexameric assembly, with two copies of each subunit. Using my *in vitro*

reconstitution method, I biochemically dissect CstF and show that RNA binding is mediated solely by CstF64's RRM domain, and that binding to DSE-like sequences is increased greatly upon complex formation.

3.2 Results

3.2.1 CstF can be assembled *in vitro*

While the CstF complex has been purified from native sources previously (Takagaki and Manley, 1997; Takagaki et al., 1989, 1990), to date, no *in vitro* reconstitution of the whole complex from recombinant sources has been reported. Lack of recombinant complex has hampered the analysis of how this complex recognizes the conserved G/U rich element in 3'-end processing. I began by the design of several constructs of CstF77 (see below) to test coexpression with CstF64. Due to no reports of function for the C-terminal portion of CstF64, I designed constructs using only the RRM and Hinge domains for expression, denoted as CstF64RH (Fig. 3.1). While various coexpressed constructs of CstF77 and CstF64 can be readily obtained in abundance using *E.coli* as an expression organism, production of CstF50 in *E.coli* results in complete expression in inclusion bodies (data not shown).

Thus, I used an insect cell expression strategy to obtain soluble CstF50 for use in assembling the CstF complex. In brief, CstF50 was expressed as a GST fusion in insect cells, and then the protein was bound to glutathione beads. An excess amount of purified CstF77/64 was then incubated with GST-CstF50 beads, washed extensively, and then the whole complex was eluted from beads with glutathione at once (Fig. 3.2A, lane 4). Initial purifications of the complex using full length human CstF77 revealed a co-purification of

a stoichiometric amount of an unknown protein (data not shown). Protein sequencing analysis of the band revealed that CstF77 experiences an extensive amount of degradation, particularly at its N-terminus when expressed recombinantly in *E.coli*, as shown by previous reports (Bai et al., 2007). Redesign of constructs with the N-terminal 214 residues removed resulted in highly homogenous preparations (Fig. 3.2A), devoid of noticeable degradation. Removal of this domain in CstF77 did not impair binding to CstF50, reinforcing previous work (Takagaki and Manley, 2000) that its interactions with the other components of CstF are mediated by its C-terminal tail (Fig. 3.2A, lane 3 and 4). By using a pull-down strategy to assemble the three proteins, I could ensure that all components were present in stoichiometric amounts with each other. Gel filtration analysis of purified complex also shows co-migration of all three proteins in a 1:1:1 ratio (Fig. 3.2B).

Proteolytic digestions of CstF with trypsin also provide evidence that the complex is properly assembled. When expressed as just the heterodimer, CstF77/64 is highly susceptible to proteolysis by trypsin (Fig. 3.3, lanes 1-5). Even at low concentrations of trypsin, CstF77 is cleaved to yield the HAT-C domain by itself (Fig. 3.3, lane 3). Incorporation of CstF50 into the complex increased the overall stability of the complex, likely via shielding the loop in CstF77, N-terminal to the CstF64 binding region (Fig. 3.3, compare lanes 3 and 8).

Previous solution studies of CstF64's RRM domain revealed an additional helix that lay across the canonical RNA binding surface of the RRM (Fig. 3.4, left). This helix was shown by NMR dynamics to unfold upon the addition of RNA, suggesting a mechanism that only in the presence of G/U rich RNAs, this helix would "melt" and

provide access to the binding surface (Perez Canadillas and Varani, 2003). Due to the presence of three lysines in this helix, I reasoned that if RNA binding indeed unfolded this structure it would be more susceptible to trypsin digestion. To test this hypothesis in the context of the full complex, I incubated CstF with G/U-rich RNAs (see below) and performed trypsin digestions. I observed no differences in proteolytic patterns between the free and RNA incubated complexes (Fig 3.4, compare lanes 4 and 9). This suggests that RNA binding has no effect on the folding of this helix in the assembled CstF complex.

3.2.2 CstF50 binds to CstF77 via a conserved patch found only in animals

Biochemical and structural studies have long since identified the minimal region for the CstF77/64 interaction (~ aa620-650) (Moreno-Morcillo et al., 2011a; Paulson and Tong, 2012; Takagaki and Manley, 2000). Preliminary studies also suggested that CstF50 binds to a region N-terminal to CstF64's binding site, and C-terminal to 77's HAT domain (Takagaki and Manley, 2000). I sought to further characterize the interaction between CstF77 and CstF50 by mapping the exact sequence necessary for binding.

Given that the HAT-N half of CstF77's HAT domain is dispensable for binding (Fig. 3.2A) and inclusion of this region results in heterogenous preps, I designed a series of CstF77 constructs starting at position 215-560 (HAT-C) with 10 amino acid C-terminal extensions all the way to residue 610 (Fig. 3.5A, lanes 1-6). Using GST-CstF50 as bait, we tested the ability of these CstF77 constructs to bind CstF50. As expected, CstF77 HAT-C domain alone does not bind to CstF50 appreciably above background. Constructs spanning all the way to residue 590 also lack the ability to bind 50 (Fig. 3.5A,

lanes 7-11). However, CstF77 proteins with C-termini at either 600 or 610 can bind stoichiometrically to 50; of note, there was no difference in binding strength observed between 600 and 610, suggesting that the minimal binding motif is encoded between residues 560-600 (Fig. 3.5A, lanes 12 and 13). To further validate these pull-down results, I ran the eluted complex on a gel filtration column and observed co-elution of both proteins, as well as a shift of the CstF50 peak to an earlier elution volume, consistent with complex formation (Fig. 3.5A, right).

Metazoan CstF complexes have been reported to contain three components, while the yeast equivalent has only two, with no known homolog of CstF50 having been identified. Sequence analysis of CstF77 between animals and yeasts show reasonable levels of conservation in the tail beyond the HAT domain, with the highest level of conservation observed within the CstF64 binding site. A patch of highly conserved residues is also observed between aa580-600 in metazoan CstF77 proteins, while no such conservation is seen in the yeast equivalents (Fig. 3.5B boxed). The high level of conservation in this region, suggested to me that our previous pull-down results showing that constructs lacking residues 590-600 could not bind CstF50 was likely due to an incomplete binding site and not that the binding site was contained at 590-600. These results also suggest that CstF50 is a bona fide metazoan-only component of CstF and no such yeast equivalent exists.

From the above analysis, I then asked if residues N-terminal to 580 were necessary for binding. Starting with a minimal CstF50-binding competent CstF77 construct (215-600), I constructed a series of internal deletions, starting from 556-560, deleting 5 amino acids at a time all the way to 580 (Fig. 3.6A, lanes 3-7). A CstF77 (215-

560) was used in pull-down experiments as a negative control. Again, using GST-CstF50 as bait, I tested binding of each of the internal deletions for CstF50 binding. As expected, 215-600 bound CstF50 very robustly, while the HAT-C construct lacked any binding to 50 (Fig. 3.6A, lanes 9 and 10). Unexpectedly, none of the internal deletions had lost the ability to bind to CstF50 (Fig. 3.6A, gel lanes 11-15). Gel filtration analysis of this internally deleted complex also showed co-elution of both proteins. However, in this internally deleted complex, the CstF77/50 complex migrated at molecular weight consistent with that of just CstF77 alone (Fig. 3.5C, right). Molecular modeling of CstF50's WD40 domain suggests that CstF50 has a height of approximately 20Å, while having a diameter of about 50Å (data not shown). The lack of a shift in apparent size, despite complex formation, suggests that CstF50 must bind to CstF77 in an orientation that minimally increases the Stokes radius of the overall complex (see Fig. 3.6B for frame of reference for orientation). I hypothesize from this analysis CstF50 must be bound to CstF77 in an orientation that is parallel to the HAT domain.

3.2.3 CstF is a hexamer

Gel filtration analysis of CstF77/64 shows an elution profile consistent with that of a 160kDa dimer of dimers (Fig. 3.7), as previously suggested by structures of several components of CstF (Bai et al., 2007; Legrand et al., 2007; Moreno-Morcillo et al., 2011a). CstF50 elutes at volumes smaller than 100kDa, likely implying that while 50 dimerizes (Takagaki and Manley, 2000), the overall shape and conformation reflect that of a highly compact dimer.

Given the self-association of both CstF77 and CstF50, it has been suggested, but not demonstrated, that CstF assembles as a hexamer *in vivo* (Bai et al., 2007; Moreno-Morcillo et al., 2011b). Assuming two copies of each component in the complex with my domain boundaries for each subunit, my assembled complex has a theoretical molecular weight of approximately 254kDa. Oddly, when assembled as a whole, CstF runs at volumes expected for 160-170kDa globular proteins, barely larger than the CstF77/64 heterodimer (Fig. 3.7, black vs. green trace). These results suggest two possibilities: assembly of all three proteins disrupts dimeric associations (127 kDa) and results in a highly elongated trimer, or the hexameric complex forms a tight and compact structure resulting in a smaller than expected apparent molecular weight.

While gel filtration can determine molecular weights of single globular proteins with good confidence, complexes of multiple proteins may behave inconsistently on gel filtration due to their shapes. I thus turned to size exclusion chromatography multi-angle light scattering (SEC-MALS) with the help of a colleague to more accurately determine the molecular weight of the entire complex in a shape independent manner. SEC-MALS analysis of CstF using two independently purified samples showed an average molecular weight of ~253-260kDa, consistent with that of a hexameric assembly (data not shown). The difference between that of the experimental and predicted weight could partially be explained from both error of the instrument, and the possibility of contaminating RNAs bound to the complex during the purification process. All together, the data demonstrates that CstF is a highly compact hexameric assembly, with two copies of each protein in the whole complex.

3.2.4 CstF binds G/U-rich RNAs selectively with a 1:1 stoichiometry

Early SELEX experiments and bioinformatics studies showed that CstF binds G/U-rich sequences regardless of sequence (MacDonald et al., 1994; Takagaki and Manley, 1997), while still discriminating against A/C sequences. Experiments done with the CstF64 RRM show that it binds RNA with weak affinity, leading to speculation that selection of the cleavage site in 3'-end processing is defined by weak RNA interactions strengthened by cooperative interactions between protein complexes (Pancevac et al., 2010; Perez Canadillas and Varani, 2003; Takagaki and Manley, 1997). While previous biochemical and structural work have largely focused on CstF64's affinity towards G/U sequences a), little work has been done on how or if CstF64's incorporation into the full complex affects binding to RNA. To test CstF's RNA binding activity, I initially tested three G/U-rich sequences by electrophoretic mobility shift assays (EMSA) with the help of colleagues (see Table 3.1 for sequences) (Fig. 3.8). While the 10-mer and 12-mer bound only modestly, we were surprised to observe binding to the 14-mer at a K_d of between 1-10nM.

The GU12 and GU14 sequences contain a symmetric GUGU on both the 5' and 3' ends of this RNA (Table 3.1). Given the symmetry of the complex, the presence of two copies of CstF64, as well as its base specificity, we reasoned that CstF should, in principle, bind a bipartite RNA with identical G/U sequences on the 5' and 3' ends spaced by a non-binding A-spacer.

We used EMSA to test various RNAs with GUGU as the binding element, spaced apart by as few as two adenine residues. We also designed several RNAs consisting solely of the two binding sites spaced by uridines and guanines. To our surprise, CstF

bound none of the A-spaced RNAs with any appreciable affinity at the concentration range tested (1nM to 1μM) (Fig. 3.9A, all lanes). In contrast, using a G/U spacer instead, CstF showed appreciable binding for 12-mers and larger RNAs (see U_2 vs U_4), with an apparent K_d between 100-1000nM for U_4 (Fig. 3.9B, compare lanes 8 and 13). Even more strikingly, changing the central two uridines to guanines increased binding by 10-fold, with an apparent K_d of 10-100nM (Fig. 3.9B, compare lanes 4 and 14)

It was striking to see two nearly identical RNAs (U_4 vs. UGGU) display such different binding affinities based on a two base change in the center of the sequence (2U → 2G). I asked then if it was possible that rather than recognizing GUGU sequences at the 5' and 3' ends of the RNA that CstF was instead binding two separate strands of RNA via the center of the sequence. To discriminate between these binding modes (1:1 CstF:RNA, 1:2 CstF:RNA), I conducted a series of titration experiments by gel filtration, by keeping the protein concentration fixed and increasing the amount of RNA. I reasoned that when a 1:1 complex is formed, any additional RNA would result in the appearance of a free RNA peak on gel filtration.

As expected, protein:RNA ratios under 1 show no additional peak. As RNA was increased over a protein:RNA ratio of 1:1, a peak that absorbed strongly at 260nm began to appear at the volume where UGGU would elute by itself. At a protein:RNA ratio of 1:2, a very prominent peak at the RNA position is observed, suggesting that all binding sites on CstF have already been occupied (Fig. 3.10). These data establishes that CstF binds only a single strand of G/U-only RNA.

3.2.5 CstF binds RNA with both base and length specificity

To more quantitatively define CstF's affinity towards G/U rich sequences, I turned to fluorescence anisotropy (FA) binding assays using fluorescent RNAs labeled at the 5'-end. For a control, I also purified and tested CstF64's RRM domain to test against GU14. The CstF complex bound GU14 with very high affinity, with a K_d of nearly 10nM (Fig. 3.11A black trace, Table 3.2). Similarly, the 12-mer sequence, UGGU (see Table 3.1), which binds to CstF with an apparent K_d of 10-100nM by EMSA, was found by FA to bind with a dissociation constant of nearly 50nM, cross-validating my own results. In contrast with these results, binding of GU14 with CstF64's purified RRM bound with an affinity of approximately 14μM, nearly a thousand fold weaker when compared to the complex, consistent with previous results (Pancevac et al., 2010; Perez Canadillas and Varani, 2003). Both the RRM and the complex exhibited selectivity towards G/U sequences as binding experiments using A/C sequences resulted in no observable binding (Table 3.2).

Given the increase in binding after the addition of only two nucleotides (a 12-mer vs. 14-mer), I asked if RNA length played a role in CstF binding. I designed two additional RNAs that increased in length by two nucleotides at a time (Table 3.1) and tested their binding by FA. CstF's affinity towards GU16 was similar to that of GU14, in the low nanomolar range. However, to my surprise, GU18 exhibited much poorer binding characteristics, with a K_d of approximately 100nM, nearly 10-fold worse binding compared with GU14 (Fig. 3.11A green trace). These results indicate that CstF specifically optimally sequences of roughly 14-16 nucleotides in length, while losing affinity towards shorter and longer sequences.

3.2.6 CstF50 restricts the complex's affinity to shorter sequences

As seen in Fig. 3.3A binding of CstF50 can increase the overall stability of the complex *in vitro*, while *in vivo* CstF50 can bind additional factors to the 3'-end processing complex. I asked if association of CstF50 contributed to the overall binding to RNA. While CstF50 has no RNA binding domain, WD40 proteins have been known to bind to nucleic acids, including CPSF160 in the 3'-end processing complex (Dichtl et al., 2002; Kagawa et al., 2011; Murthy and Manley, 1995). To test this possibility I used purified heterodimer CstF77/64 and the complete complex in binding assays against GU14. The full complex binds with high affinity as demonstrated earlier. In parallel, binding assays using CstF77/64 showed K_d values to be about 4-fold weaker when compared with the full complex (Fig. 3.11B black vs. red trace, Table 3.2). While binding was not significantly reduced (still in the low nM range), I was surprised to see any effect on binding in the absence of CstF50. To further examine this, I used the weakest binding RNA, GU18, and tested its binding with the CstF77/64 heterodimer. Surprisingly, binding of CstF77/64 to GU18 was nearly 2-fold better compared to GU14, and when compared with full complex's affinity to GU18, nearly 4-5-fold better (Fig. 3.11B green vs blue, Table 3.2). The gain in affinity between GU14 and GU18 with respect to CstF77/64 suggests to me that avidity, due to additional binding sites on the longer RNA, may be increasing the affinity of the heterodimer.

To test to see if CstF50 contributed to RNA binding directly in the CstF complex, I assembled a complex that lacked the RRM domain of CstF64, and tested its binding to GU14. Even at the highest concentrations used (2μM), I could observe no distinguishable RNA binding from this RRM-less complex (Table 3.2). This shows that CstF64 is the

sole RNA binding component of the CstF complex, and that CstF50's role in selecting shorter RNA sequences is likely a structural one.

3.3 Discussion

Detailed biochemical analysis of how each component of the polyadenylation complex contributes to 3'-end processing has been hampered by the ability to obtain sufficient amounts of highly pure recombinant proteins expressed from heterologous organisms. Earlier studies used complexes purified from native sources and/or cancer cell lines, which might contain contaminating proteins (McDevitt et al., 1986; Murthy and Manley, 1992; Rüegsegger et al., 1996; Takagaki et al., 1989; de Vries et al., 2000). In this work, I present the *in vitro* reconstitution of the entire cleavage stimulatory factor from purified recombinant proteins and demonstrate that CstF assembles as a hexameric complex that binds G/U-rich sequences with high affinity and selectivity.

In recent years several major components of the processing assembly having been characterized structurally (Bai et al., 2007; Li et al., 2011; Mandel et al., 2006; Moreno-Morcillo et al., 2011a, 2011b; Noble et al., 2007; Yang et al., 2011). In addition to high resolution structures, mass spectrometry and electron microscopy have revealed even more interactions in the 3'-end processing machinery (Shi et al., 2009). Structures of CstF77's and CstF50's dimerization domains have been solved, showing that dimerization of these components in the complex have extensive interfaces, suggesting that the entire CstF complex functionally assembles as a dimer of trimers (Bai et al., 2007; Moreno-Morcillo et al., 2011b). Studies on the yeast versions of CstF also suggest a dimeric assembly of the proteins, implying that these interfaces are evolutionarily

conserved (Gordon et al., 2011; Noble et al., 2004). I show in my work that CstF behaves indeed as a highly compact dimer of trimers, as gel filtration elution shows a far smaller complex than expected when compared to molecular weight standards on my columns. CstF's smaller than expected migration has been observed previously from natively purified complexes, where people noticed the complex migrating at a molecular weight of about 200kDa. My light scattering data show conclusively, that, despite the irregular migration on a column, CstF's molecular weight is consistent with that of a hexameric assembly, confirming previous speculations based on structural data.

By using GST pull-down assays, I demonstrate that CstF50 binds to CstF77 through a patch of conserved residues just before the CstF64 binding site (Fig. 3.5). CstF has three known components, while the yeast version has only two components, with no known CstF50 equivalent. An earlier report suggested that the protein Pfs2, also a WD40 repeat protein that could bind Rna14 (yeast CstF77), could potentially be the CstF50 homolog in yeast (Ohnacker et al., 2000). Sequence analysis of the C-terminus of CstF77, particularly near the CstF50 binding site, shows moderate levels of conservation between yeasts and metazoans just before the CstF50 binding region. However, conservation with yeasts drop off dramatically at the CstF50 binding site, while species with known CstF50 homologs all show high levels of sequence identity. My sequence analysis suggests that CstF50 has no equivalent in yeasts and that its function evolved only in metazoans. This is not unexpected given that the overall 3'-end processing architecture in yeasts differ in a number of aspects with the metazoan machine.

Additional pull-down assays demonstrate that the entire segment C-terminal to the HAT domain, and N-terminal to CstF50's binding site is completely dispensable for

complex formation. While this by itself is not remarkable given that WD40 propellers often interact with other proteins through short peptide sequences (Davis et al., 2005; Jennings et al., 2006; Zhang et al., 2012), my gel filtration analysis with these internal deletion complexes show that CstF50 minimally shifts the CstF77 peak once the loop has been deleted, suggesting that CstF50 inserts itself parallel to the CstF77HAT dimer (Fig. 3.6B), negating any major increase to the overall radius of the complex. A perpendicular insertion, given that WD40 propellers can be nearly 50Å in diameter, is more likely to increase the overall radius of the complex and result in a shift on gel filtration. These results are in agreement with my gel filtration data with regards to the assembly of the full complex, compared to the CstF77/64 heterodimer, showing that incorporation of CstF50 minimally increases the apparent molecular weight on columns (Fig. 3.7).

Previous work, including structures of CstF64, have shown that CstF has a preference for binding to guanine and uridine rich sequences, typically found downstream of the polyA site in 3'-UTRs (McDevitt et al., 1986; Perez Canadillas and Varani, 2003). As previously reported, in my control experiments with purified CstF64 RRM, I observe only modest binding to a G/U-rich sequence while maintaining exquisite selectivity against sequences consisting exclusively of A/C residues. To my surprise, however, the full complex binds to RNA 1000-fold more strongly. While I identified this sequence serendipitously, other G/U-rich RNAs I tested also bound with nanomolar affinity, albeit more weakly compared to GU14 (Fig. 3.8). The binding constants I observe are consistent with those seen for natively purified complexes from cells (Takagaki and Manley, 1997), reinforcing that my *in vitro* reconstituted complex represents a native-like state. I surmise that due to the predicted anti-parallel symmetry of the complex, CstF's

1:1 stoichiometry with RNA, and the presence of two possible binding sites on the sequence used, that the bound RNA must adopt a U-shaped conformation in order to allow both RRMs of CstF64 access to the 5' and 3' binding sites. How the central linker between the sites plays a role in binding is unclear (U_4 much weaker than UGGU), and will require structural studies in order to elucidate the contribution of this element to boosting affinity.

My binding experiments reveal a previously unknown role of CstF50 in the complex for binding RNA. RNA binding assays performed both in the presence and absence of CstF50 show a difference in binding affinities towards identical RNAs. The full complex displayed very strong affinity (Table 3.2) towards the GU14 sequence, while only slightly longer RNAs (GU18) displayed nearly 10-fold weaker K_d values. In contrast, binding to GU14 by CstF77/64 showed a binding constant approximately 4-fold weaker compared to the full complex. Even more strikingly, binding to the GU18 sequence by the heterodimer showed comparable binding as to the GU14 RNA. These data, together with the observation that CstF64's RRM is the only direct RNA binding component in the complex, suggests that CstF50 plays a role in restricting the complex to bind only to G/U-rich sequences of 12-16 nucleotides in length. These results indicating that CstF has a preference to RNA lengths are in agreement with previous bioinformatics analysis done on downstream sequence elements, which showed that DSEs are typically 12-15 nucleotides in length (Salisbury et al., 2006). I speculate that CstF50 dependent restriction of RNA length plays a role in polyA site selection, where DSEs of 12-16 nucleotides are favored over longer G/U-sequences in strong constitutively utilized polyA sites. While the differences in binding affinity in my assays are not drastic (4-10 fold)

between the full complex and the heterodimer, small changes in sequences in the DSE have been correlated with “weak” versus “strong” polyadenylation sites (Legendre and Gautheret, 2003). I hypothesize that the presence of CstF50 can help bias 3’-end processing towards constitutive polyA sites. A genome wide study on polyA site usage in CstF50 depleted cells will be necessary to see if there is indeed an influence by CstF50 on polyA site selection.

My work illustrates that CstF assembles as a tight and compact hexameric complex. The entire complex then binds to G/U sequences with very high affinity, compared to CstF64’s RRM alone, and protein-protein interactions within the CstF complex assist in the selection of downstream sequence elements. Further work using purified recombinant proteins will be needed to understand the interactions between CstF and CPSF, and how their association helps to define the cleavage and polyadenylation site.

3.4 Materials and methods

3.4.1 Purification and assembly of CstF

Sequences coding for human CstF77 and CstF64 were cloned into a modified pRSF-Duet-1 (Novagen) vector with a Protein G B1 domain (GB1) inserted between the NcoI and BamHI sites to facilitate expression and solubility of expressed complexes. Sequences coding for CstF50 were cloned into a modified pFastBac (Life Technologies) vector with an N-terminal GST tag to facilitate purification. Identity of all clones were verified by DNA sequencing.

CstF77/64 clones were transformed into BL21 (DE3) *E.coli* cells. Transformants were then grown in LB media at 37^{o}C until OD600 = 0.6. Cells were then induced with 0.2mM IPTG and expressed overnight at 18^{o}C. The next morning, cells were harvested by centrifugation, and resuspended in buffer A (50mM HEPES pH7.5, 200mM NaCl, 30mM imidiazole, 5mM βME). Cells were then lysed by sonication, and then clarified by high-speed centrifugation. Soluble material was purified by nickel affinity chromatography, using buffer B for elution (A + 500mM imidazole). TEV protease was added to the purified material to remove the His-GB1 tag, and this mixture was then dialyzed against buffer C (20mM HEPES pH7.5, 5mM βME).

Following dialysis, dialyzed material was further purified by anion exchange chromatography by loading onto a Q HP column (GE Healthcare) equilibrated in buffer C. Bound material was eluted by a linear gradient against buffer D (C + 1M NaCl). Fractions containing pure material were concentrated and loaded onto a Superdex 200 10/300 GL (GE Healthcare) equilibrated in storage buffer (20mM HEPES pH7.5, 200mM NaCl, 5mM DTT). Purified heterodimer was finally concentrated to ~1mg/mL and flash frozen for complex reconstitution.

Standard methods in Sf9 cells were used for production/expansion of baculoviruses expressing CstF50. Viruses at the P3 or P4 stage of expansion were used to infect monolayer HighFive insect cells. Cells were harvested 72-96 hours post infection, then resuspended in PBS. Lysis was done by sonication, and material was clarified by high-speed centrifugation. Lysates were then loaded onto a glutathione sepharose (GE Healthcare). Bound material was then washed extensively with PBS, and finally with storage buffer. Following washes, purified CstF77/64 (5-10mg total material) was

incubated with the resin for one hour at 4°C. The column was then washed extensively with storage buffer, and the bound material was eluted with storage buffer + 10mM reduced glutathione. TEV protease was then added to cleave the GST tag from CstF50 overnight. Samples were concentrated the next morning and loaded onto a Superdex 200 10/300 GL equilibrated in storage buffer. Fractions containing stoichiometric amounts of all three subunits were pooled, concentrated to 5-10mg/mL, and flash frozen for subsequent use.

3.4.2 GST pull-down assays

All CstF77 constructs were expressed in *E.coli* as a His_6GB1 fusion and purified by single step nickel affinity chromatography. GST-CstF50 was expressed and purified as described above. 2 mg of each CstF77 construct was incubated with GST-CstF50 beads for one hour at 4°C in storage buffer and washed extensively with storage buffer following incubation. For gel analysis, samples were eluted from beads by the addition of 2x SDS-PAGE loading dye, then heated at 95°C for 10 minutes. Samples were resolved by 10% SDS-PAGE gels and proteins were detected by Coomassie blue staining.

For native gel filtration analysis, complexes were eluted from beads with 10mM reduced glutathione, and tags were cleaved overnight by TEV protease. Samples were then loaded to a Superdex 200 10/300 GL equilibrated in storage buffer.

3.4.3 Limited proteolysis

10 μg of CstF was used per 10μL reaction. Samples were incubated with increasing amounts of bovine trypsin (Sigma-Aldrich) in storage buffer on ice for one hour.

Digestions were then quenched by the addition of 2x SDS-PAGE loading dye and heated at 95^{o}C for 10 minutes. For digestions in the presence of RNA, CstF was incubated with GU14 at a protein:RNA ratio of 1:1.1 on ice for 15 minutes and digested with trypsin.

3.4.4 RNA binding assays

All RNAs used for binding experiments were purchased as RNase-free HPLC purified oligos from IDT DNA. Oligonucleotides were resuspended in DNase/RNase-free TE buffer to a stock concentration of 100μM and kept frozen at -20^{o}C until needed.

For EMSA binding experiments, RNAs (see Table 3.1 for sequences) were radiolabelled at the 5′ end with [γ-^{32}P] ATP (PerkinElmer) by T4 polynucleotide kinase (Takara). Oligonucleotides were then purified from excess radiolabel and kinase using a NAP10 column (GE Healthcare). Labeled RNAs were then used in binding reactions with CstF at a final concentration of 50pM in EMSA binding buffer (20mM Tris pH7.0, 100mM NaCl, 1mM EDTA pH8.0, 2mM DTT) on ice for 30 minutes. Reactions were resolved on a 6% native acrylamide gel. Gels were vacuum dried and visualized on a phosphor-screen.

For fluorescence anisotropy experiments, RNAs were modified with a 5'-end Cy5 fluorophore (absorbance maxima = 650nm, emission maxima = 670nm) and used at a concentration of 0.5μM in FA binding buffer (20mM HEPES pH7.5, 200mM NaCl, 5mM DTT, 1mM EDTA) for binding experiments with CstF. Fluorescence anisotropy values were converted to RNA fraction bound by the equation:

$$FB = \frac{FA_{obs} - FA_{min}}{FA_{max} - FA_{min}}$$

Where FB is the fraction of RNA bound by CstF, FA_{obs} is the anisotropy at any protein concentration point, FA_{min} is the anisotropy at no protein added, and FA_{max} is the maximum value of anisotropy obtained in the experiment.

The fraction of RNA bound was then plotted against protein concentration. K_d's were determined by fitting to the equation:

$$FB = K_d + [RNA] + [P] - \sqrt{\frac{(K_d + [RNA] + [P])^2 - (4[RNA][P])}{2[RNA]}}$$

Where FB is the fraction bound, K_d is the dissociation constant (to be determined), [RNA] is the concentration of RNA used in the experiment, and [P] is the protein concentration used.

3.4.5 Size exclusion chromatography multi-angle light scattering

The SEC-MALS system consisted of a P900 HPLC pump (GE), a UV-2077 detector (Jasco), a Tri Star Mini Dawn light scattering instrument (Wyatt), and an Opti Lab T-Rex refractive index instrument (Wyatt). Approximately 200µg of purified CstF was injected into a Superdex 200 (10/300GL) gel filtration column and eluted at 0.5 ml/min in a buffer containing 20mM HEPES pH7.5, 250mM NaCl, 0.05% NaN_3. The specific refractive index of CstF was assumed to be 0.186 ml/g. Data collection and analysis was performed with Astra 6 software (Wyatt). Total molecular mass of the complex was determined with Astra6 software using protein analysis. Both peak overlap and peak broadening were corrected with Astra 6 software. The SEC-MALS system was pre-calibrated with BSA.

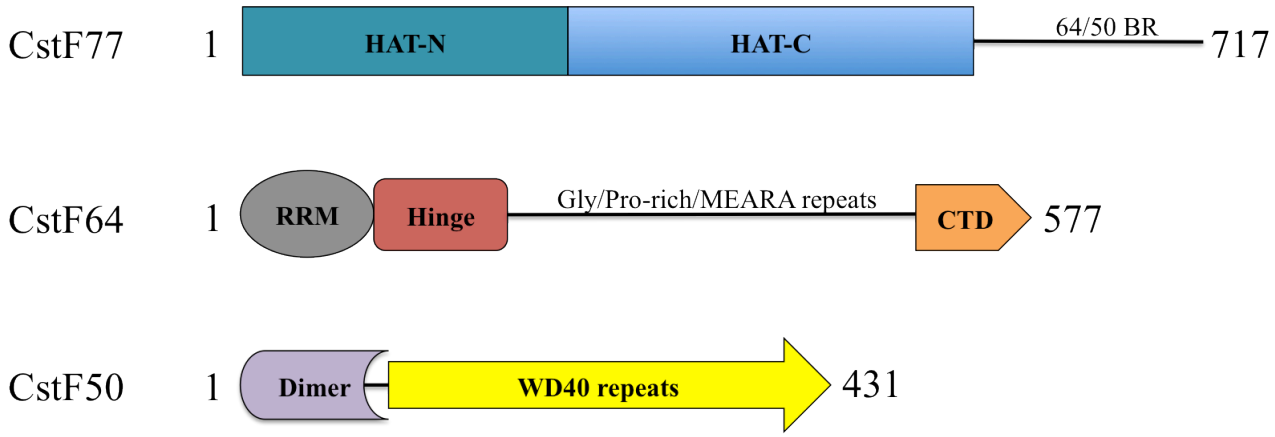

Figure 3.1 Domain breakdown of individual CstF subunits
All numbering and domain breakdowns are based on the human subunits of CstF. CstF77's HAT domain comprises the first 550 amino acids, with a break between HAT-N and HAT-C at residue 215. The remaining 167 residues lack a well-defined domain and are used for binding CstF64 and 50. CstF64 contains an RRM domain within its first 95 residues, and then followed by a unique Hinge domain (aa100-200) that interacts with CstF77. A large region of low complexity sequence is located following Hinge. The CTD of CstF64 is a novel structure that interacts with other members of the 3'-end processing complex. CstF50 has a unique homo-dimerization domain in its N-terminal 70 amino acids, followed by a short linker. The WD40 propeller domain is located between residues 100-431 and is used for binding CstF77.

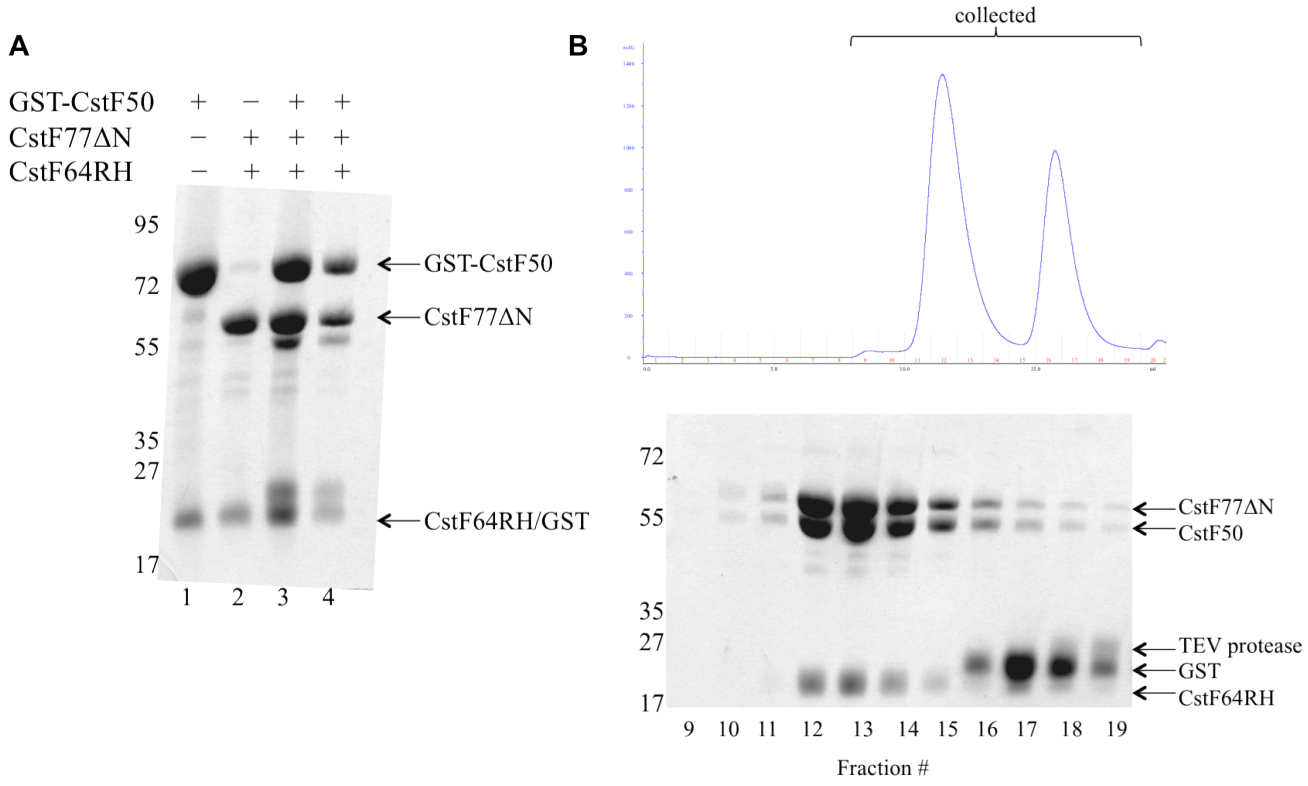

Figure 3.2 *In vitro* assembly of the CstF complex

A) Affinity purification of CstF on glutathione-sepharose beads. Lanes 1 and 2 represent input of GST-CstF50 and purified CstF77/64 respectively. Lane 3 is the bound ternary complex on glutathione-sepharose after washes. Lane 4 is the co-elution of all three proteins from beads. Molecular weight markers are indicated on the left side of the gel, and positions of the CstF subunits are indicated by arrows.

B) Gel filtration of the CstF complex on a Superdex 200 column (top). Collected fractions were resolved on SDS-PAGE. Molecular weight markers are indicated on the left side of the gel and positions of all proteins resolved are indicated by arrows on the right side of the gel.

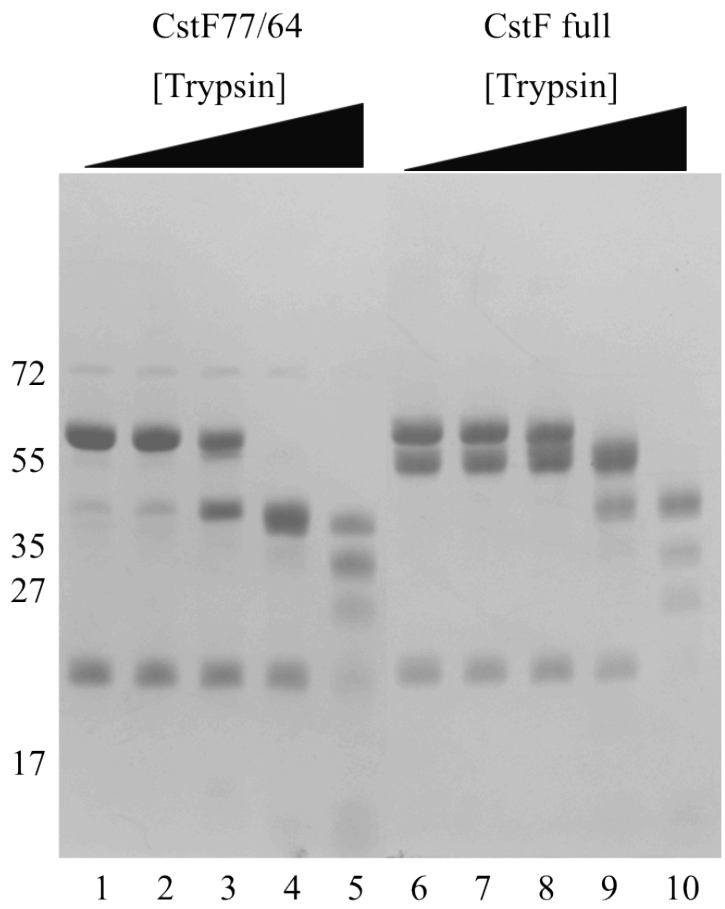

Figure 3.3 CstF50 increases the overall stability of the CstF complex *in vitro*
10μg of protein were digested by trypsin for one hour, then resolved on SDS-PAGE, and visualized by Coomassie blue staining. Lanes 1 and 6 represent input samples with no addition of trypsin.

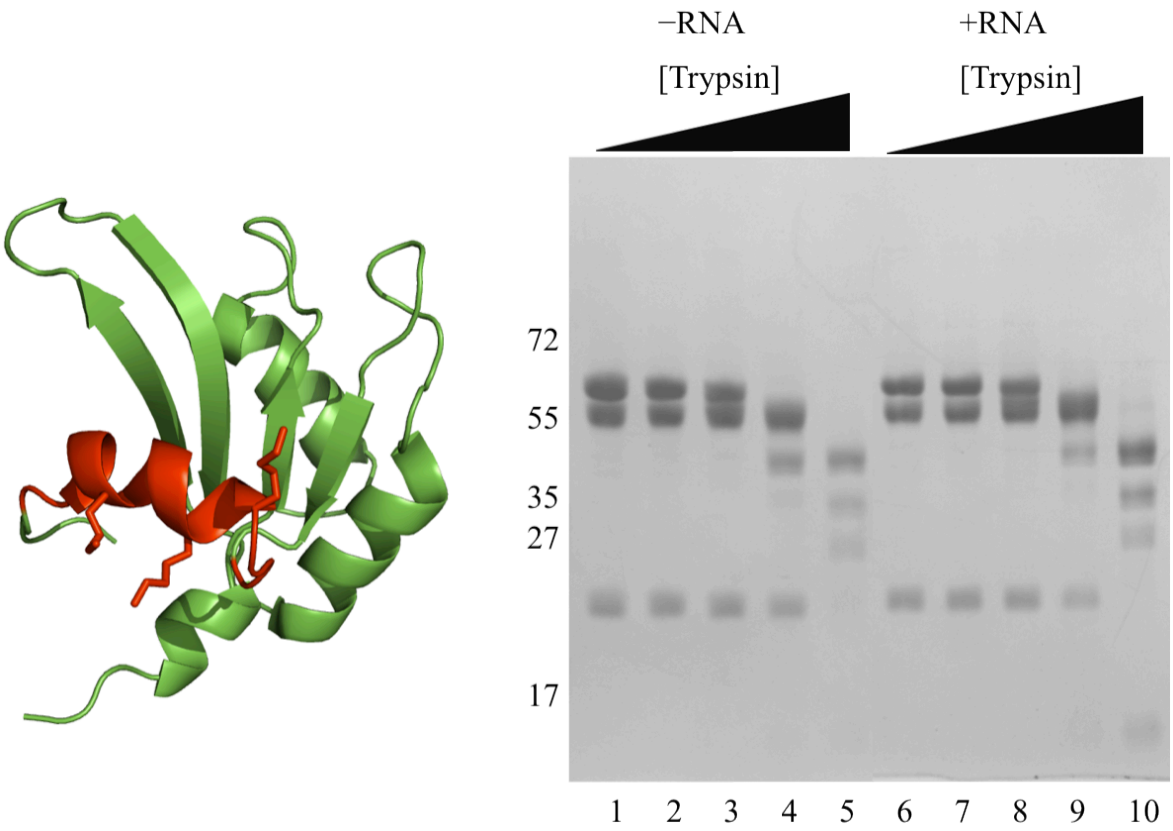

Figure 3.4 RNA binding does not unfold the C-terminal helix of CstF64
Solution structure of CstF64's RRM showed a C-terminal helix that lay across the canonical RNA binding surface, hypothesized to unfold upon RNA binding (left, red). Potential trypsin vulnerable lysines are shown as sticks. Trypsin digests of 10μg of CstF complex in the presence and absence of RNA (right). Lanes 1 and 6 represent input samples without trypsin.

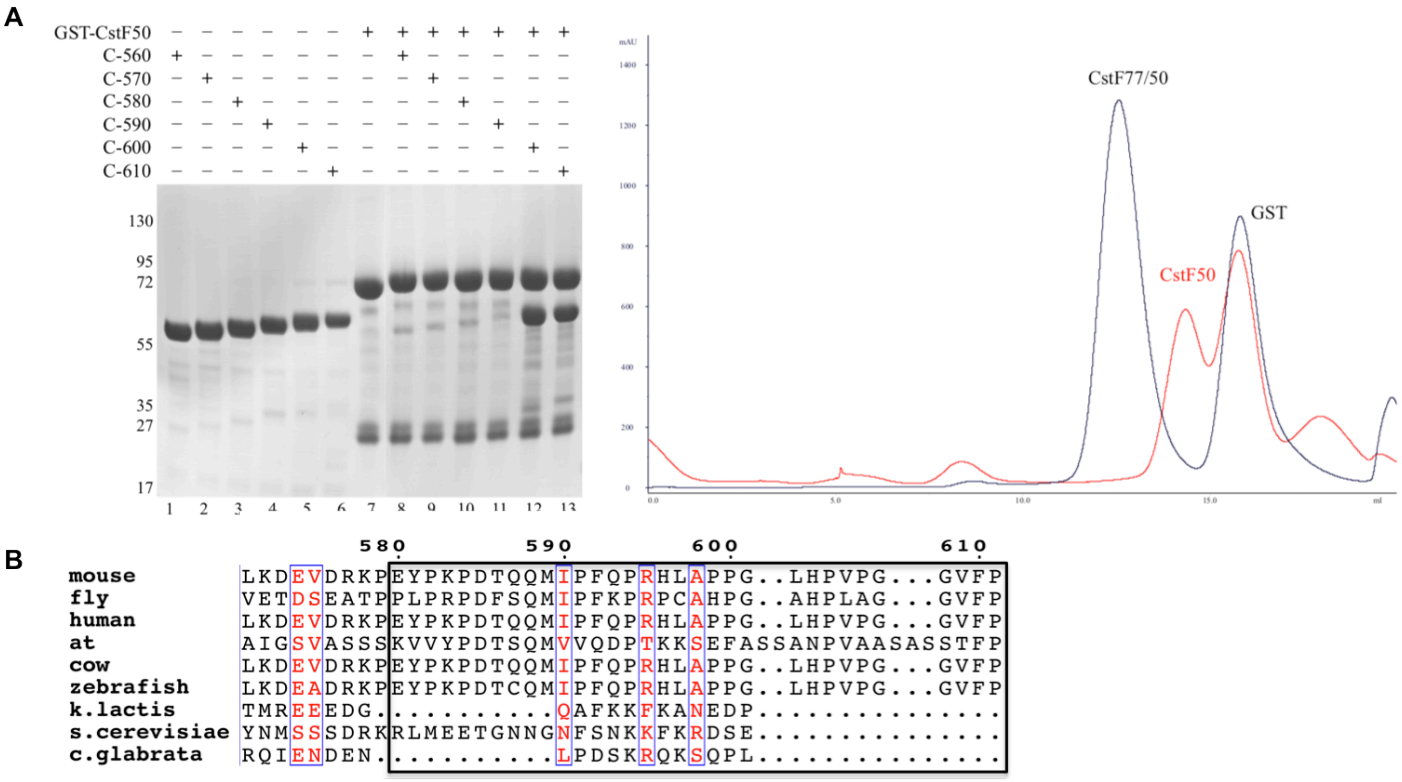

Figure 3.5 CstF50 binds to CstF77 via a conserved patch of residues

A) GST pull-down assay using numerous C-terminal extensions of CstF77 as prey (left). C-# denotes the last residue in the construct.Gel filtration on Superdex 200 (right) of a minimal CstF77/50 complex (black), and CstF50 alone (red).

B) Sequence alignment of CstF77 across various organisms. The putative CstF50 binding site is boxed.

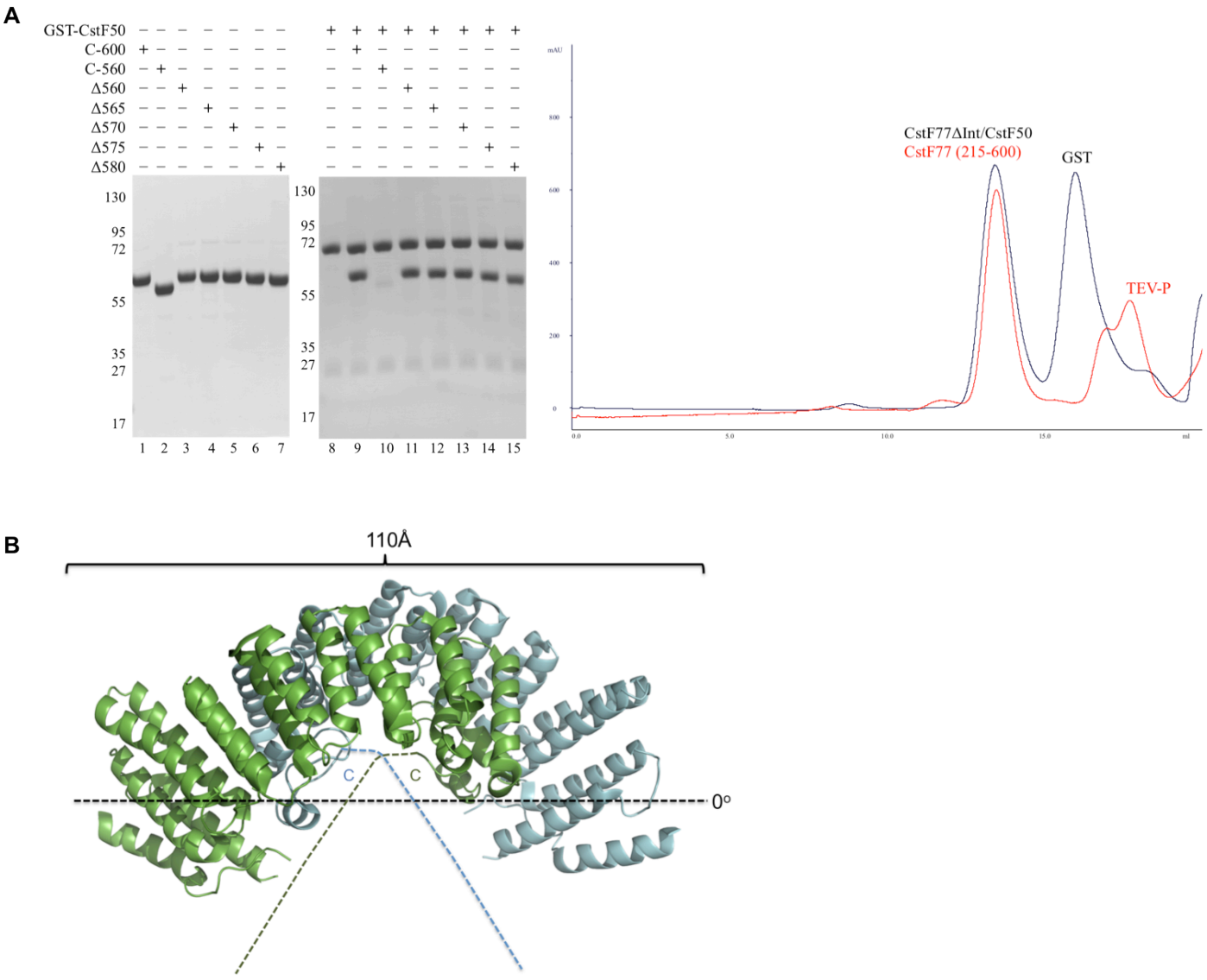

Figure 3.6 CstF50 binds CstF77 in an orientation that minimally increases the size of the complex

A) GST pull-down assay using numerous internal deletions of CstF77, starting from residue 556 (left). C-600 and C-560 are used as positive and negative controls respectively. Gel filtration on Superdex 200 (right) of a maximum internal deleted complex of CstF77/50 (black) and CstF77C-600 alone (red).

B) Crystal structure of CstF77's HAT-C domain (PDB 2OOE) showing the overall length of the dimer. The position of the last residue seen is labeled with a C for each monomer and the colored dotted lines represent the C-terminal tail used for binding CstF50/64. The black dotted line shows the frame of reference used in the discussion (see manuscript).

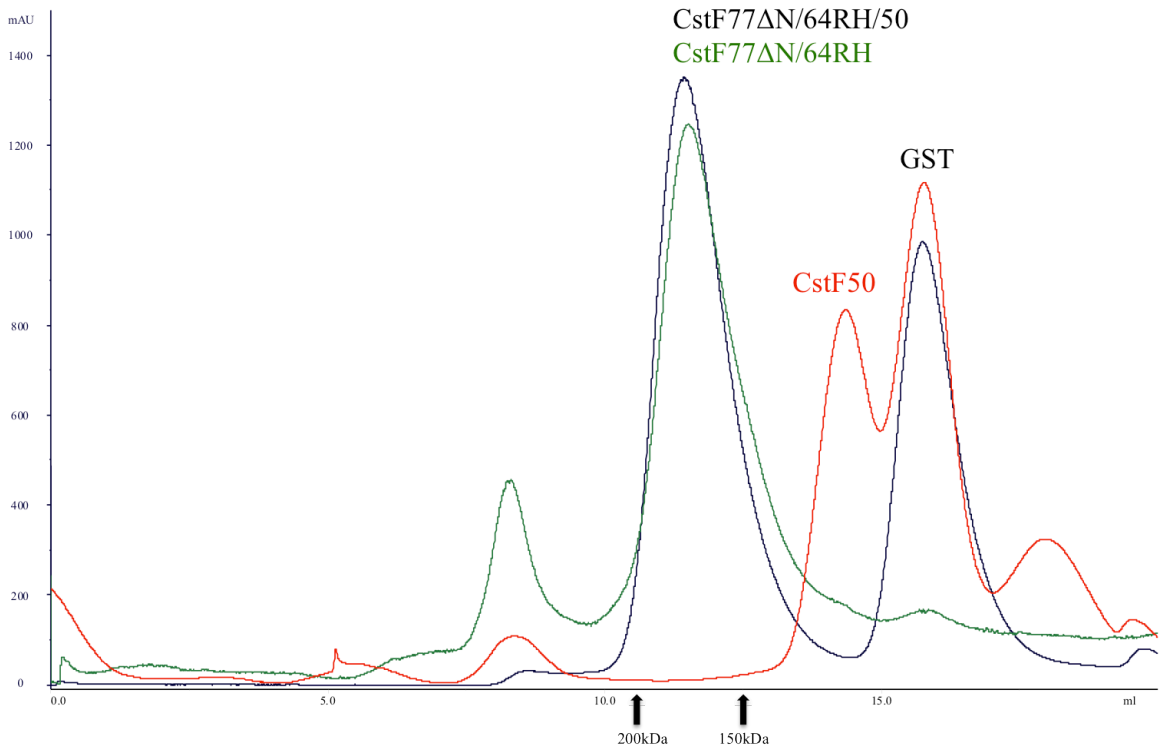

Figure 3.7 CstF is a compact hexamer
Gel filtration on Superdex 200 of CstF (black), CstF77/64 (green), and CstF50 (red). The GST tag from the purification of CstF50 and the CstF complex acts as an internal control. The void peak at 8mL also acts as an internal control between runs. The arrows indicate the positions where the globular weight standards for 200kDa and 150kDa run.

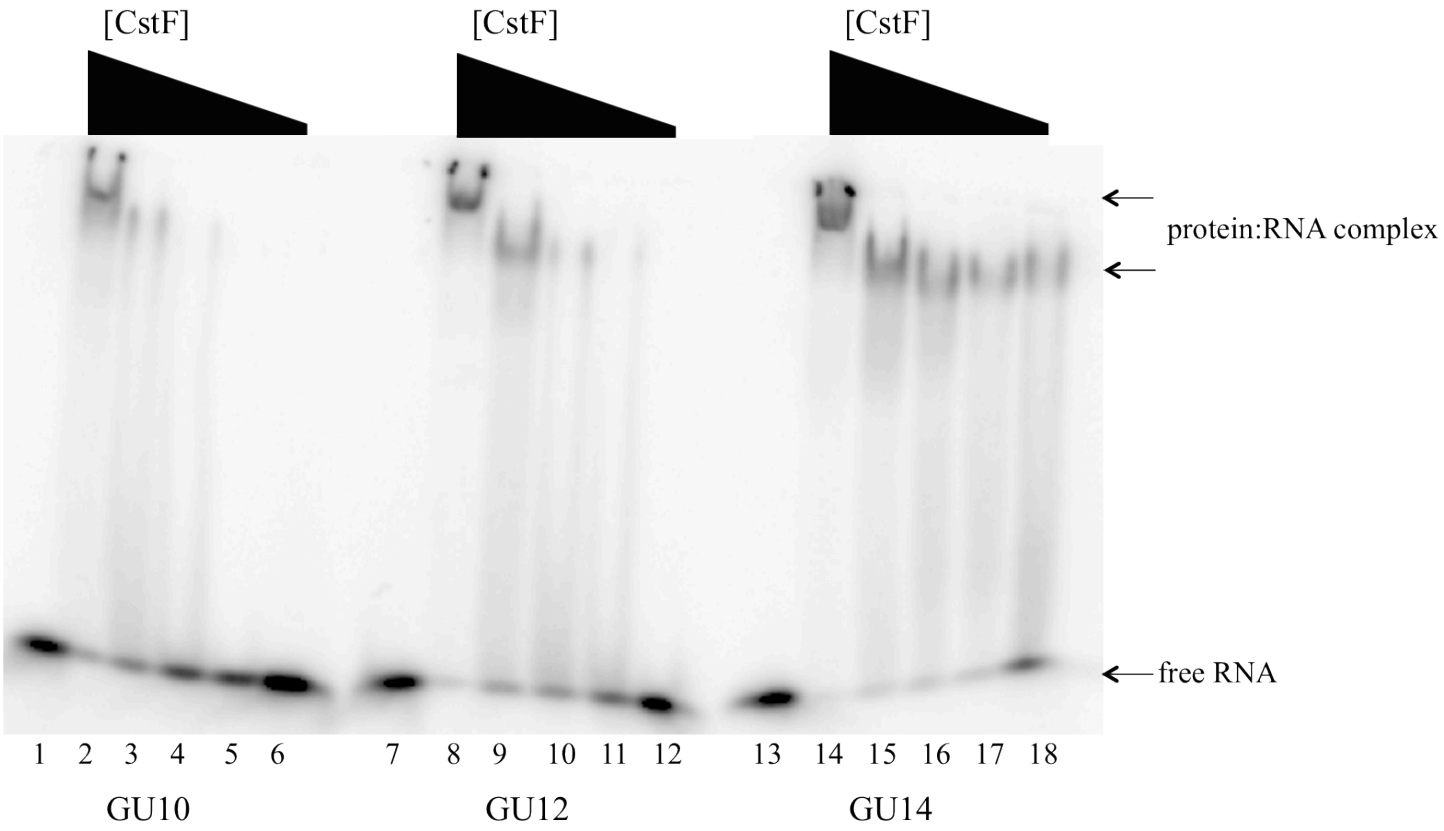

Figure 3.8 EMSA identification of RNA binding by CstF
EMSA of CstF with three distinct GU-rich RNA sequences. Formation of protein:RNA complexes are indicated by arrows. Sequence details are shown in Table 3.1. CstF was used at final concentrations of 1nM, 10nM, 100nM, 1μM, and 10μM.

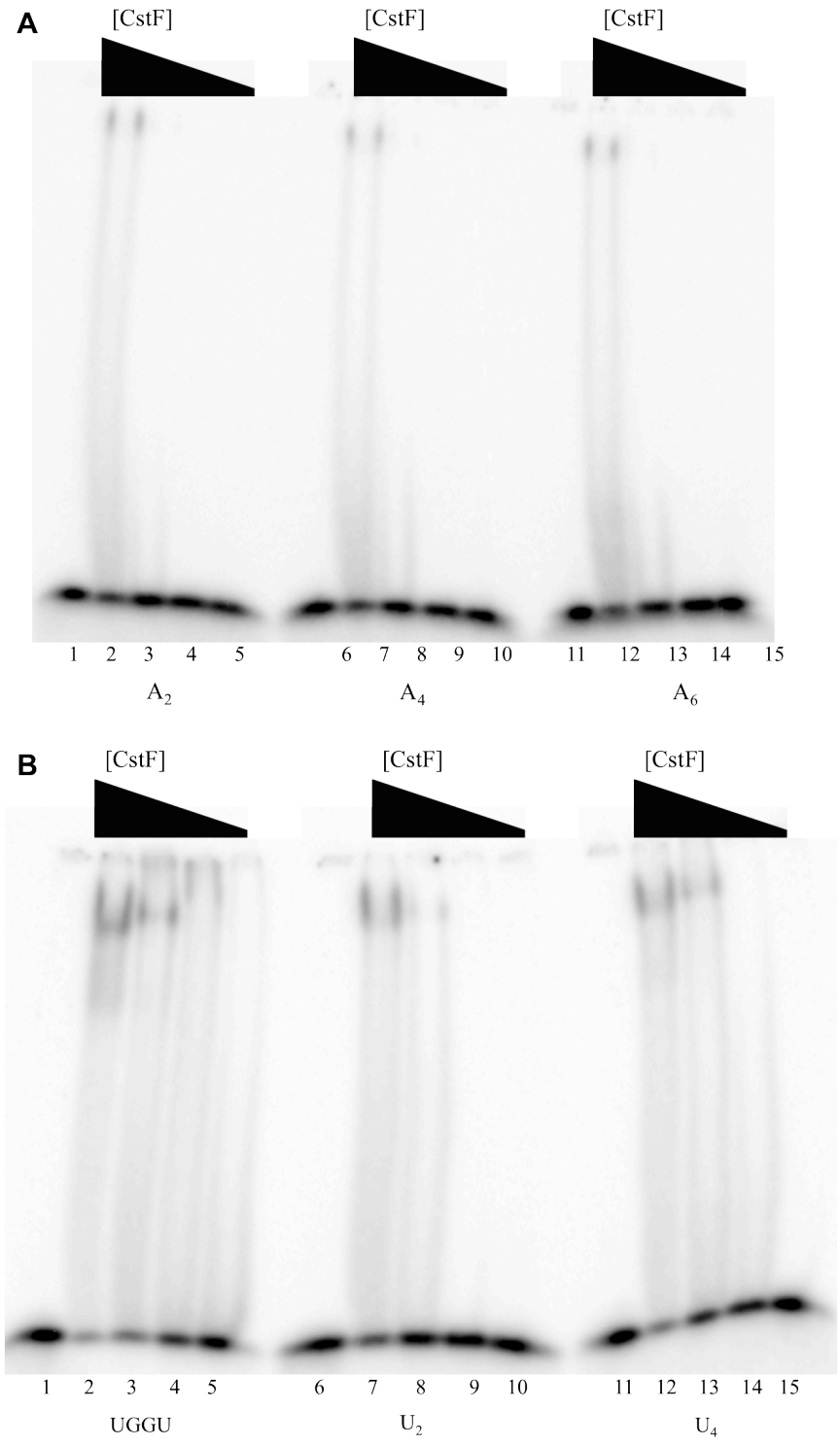

Figure 3.9 CstF binds to G/U-rich only sequences

A) EMSA of CstF against GUGU(A_n)GUGU sequences, where n is the number of adenine nucleotides between GUGU motifs. CstF was used at 1nM-1μM.

B) EMSA of CstF against GUGU(G/U)GUGU sequences. Nucleotides spacing the GUGU motifs are denoted underneath each RNA. CstF was used at 1nM-1μM.

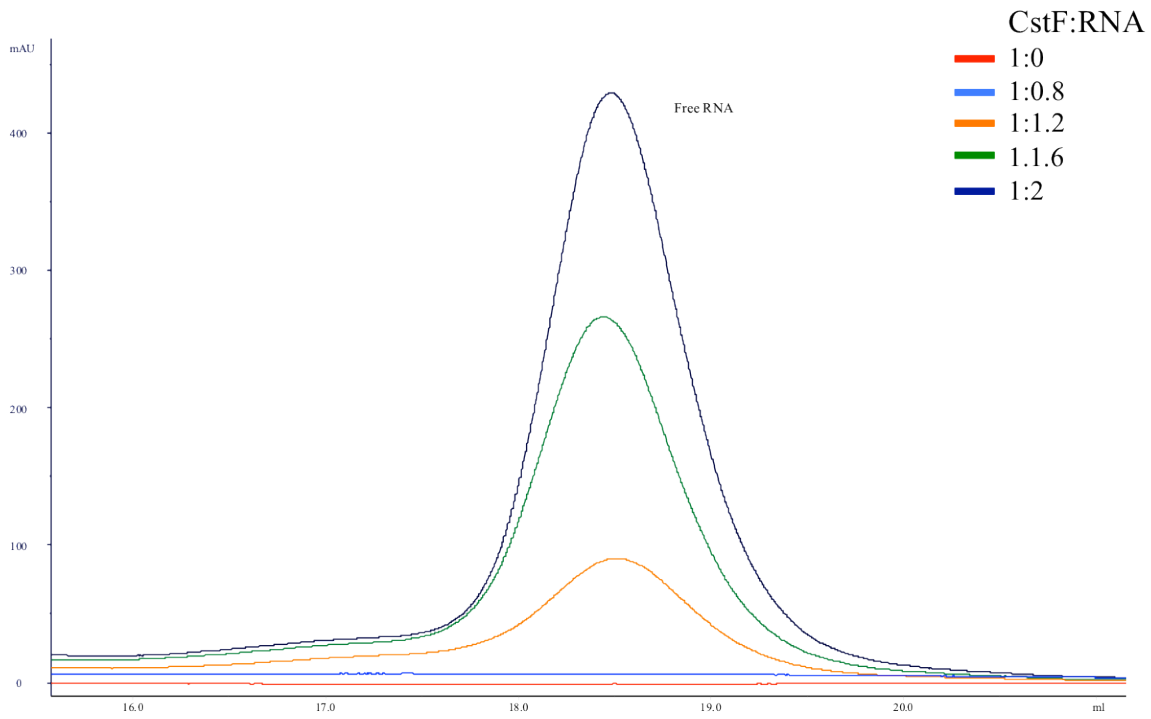

Figure 3.10 CstF binds to target RNAs with a 1:1 stoichiometry
Gel filtration chromatogram observing absorbance at 260nm of CstF titrated with increasing UGGU RNA. The elution volume of the free RNA peak is observed at ~18.5mL.

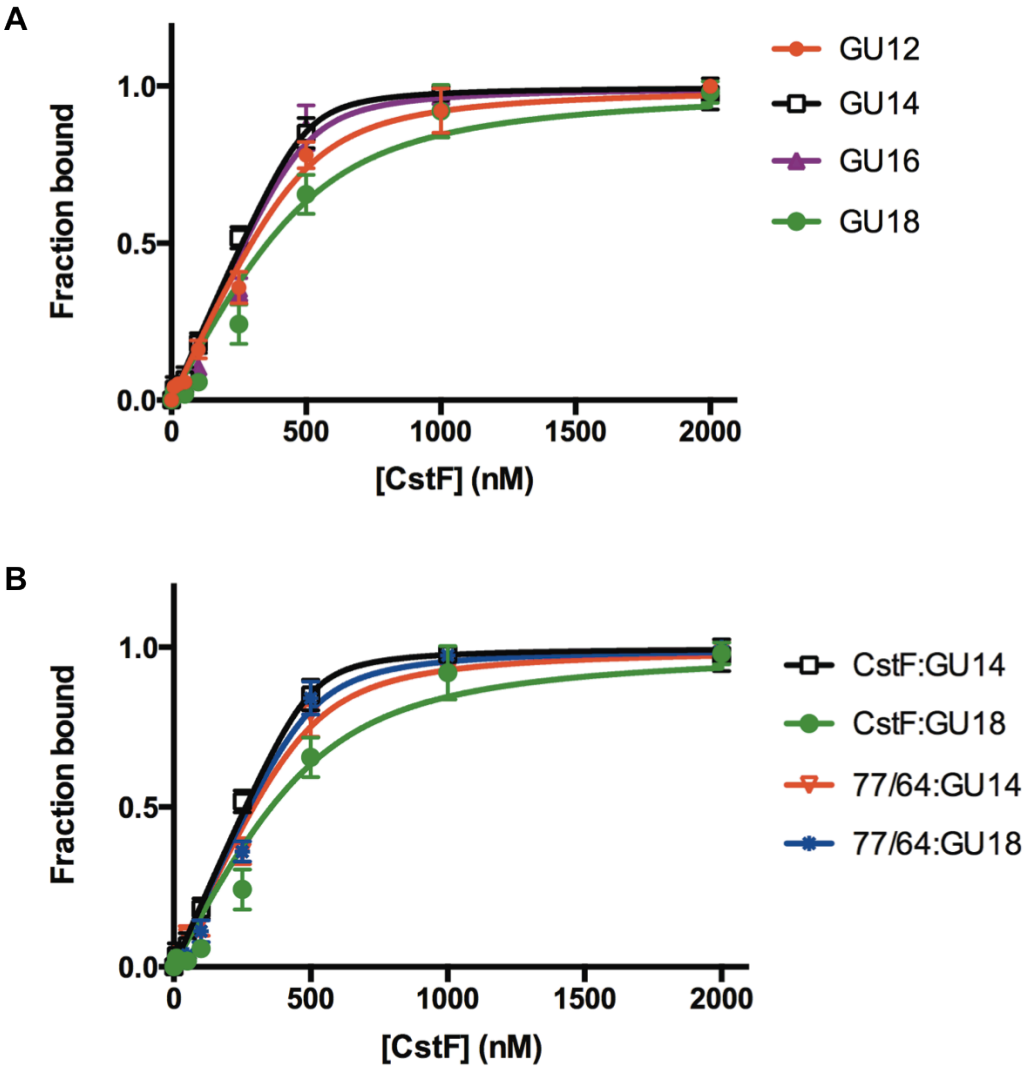

Figure 3.11 CstF binds G/U-rich RNAs length specifically

A) Fluorescence anisotropy binding curves of CstF against GU RNAs of increasing lengths. Sequence details and binding constants are given in Tables 3.1 and 3.2 respectively.

B) Fluorescence anisotropy binding curves of CstF and CstF77/64 against GU14 and GU18. Binding constants are given in Table 3.2.

Table 3.1 RNA sequences used in this study

RNA	Sequence
GU10	GUGUGUGUUG
GU12	UGUGUUUGUGUG
GU14	GUGUGUUGGUGUGU
GU16	GUGUGUUGGUGUGUGU
GU18	GUGUGUGUUGGUGUGUGU
U_2	GUGUUUGUGU
U_4	GUGUUUUUGUGU
UGGU	GUGUUGGUGUGU
A_2	GUGUAAGUGU
A_4	GUGUAAAAGUGU
A_6	GUGUAAAAAAGUGU
AC14	CACACAACCACACA

Table 3.2 Protein:RNA complex dissociation constants
(n = 3 for all FA experiments)

Protein construct	**RNA**	**K_d (EMSA/FA)**
CstF 77ΔN/64RH/50	GU10	100-1000nM (EMSA)
CstF 77ΔN/64RH/50	GU12	10-100nM (EMSA)
CstF 77ΔN/64RH/50	GU14	8.0nM ± 3.6nM (FA)
CstF 77ΔN/64RH/50	GU16	20.6nM ± 7.4nM (FA)
CstF 77ΔN/64RH/50	GU18	106.0nM ± 22.2nM (FA)
CstF 77ΔN/64RH/50	U2	100-1000nM (EMSA)
CstF 77ΔN/64RH/50	U4	10-100nM (EMSA)
CstF 77ΔN/64RH/50	UGGU	46.9nM ± 8.8nM (FA)
CstF 77ΔN/64RH/50	A2	>1μM (EMSA)
CstF 77ΔN/64RH/50	A4	>1μM (EMSA)
CstF 77ΔN/64RH/50	A6	>1μM (EMSA)
CstF 77ΔN/64H/50	GU14	No binding (FA)
CstF 77ΔN/64RH/50	AC14	No binding (FA)
CstF64 RRM	GU14	13.7μM ± 3.1μM (FA)
CstF64 RRM	AC14	No binding (FA)
CstF 77ΔN/64RH	GU14	41.3nM ± 7.1nM (FA)
CstF 77ΔN/64RH	GU18	24.6nM ± 7.5nM (FA)

3.5 References

Bai, Y., Auperin, T.C., Chou, C.-Y., Chang, G.-G., Manley, J.L., and Tong, L. (2007). Crystal structure of murine CstF-77: dimeric association and implications for polyadenylation of mRNA precursors. Mol. Cell *25*, 863–875.

Davis, T.L., Bonacci, T.M., Sprang, S.R., and Smrcka, A.V. (2005). Structural and molecular characterization of a preferred protein interaction surface on G protein beta gamma subunits. Biochemistry (Mosc.) *44*, 10593–10604.

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

Edwards, R.A., Lee, M.S., Tsutakawa, S.E., Williams, R.S., Nazeer, I., Kleiman, F.E., Tainer, J.A., and Glover, J.N.M. (2008). The BARD1 C-terminal domain structure and interactions with polyadenylation factor CstF-50. Biochemistry (Mosc.) *47*, 11446–11456.

Gordon, J.M.B., Shikov, S., Kuehner, J.N., Liriano, M., Lee, E., Stafford, W., Poulsen, M.B., Harrison, C., Moore, C., and Bohm, A. (2011). Reconstitution of CF IA from Overexpressed Subunits Reveals Stoichiometry and Provides Insights into Molecular Topology. Biochemistry (Mosc.).

Hockert, J.A., Yeh, H.-J., and MacDonald, C.C. (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*, 695–704.

Jennings, B.H., Pickles, L.M., Wainwright, S.M., Roe, S.M., Pearl, L.H., and Ish-Horowicz, D. (2006). Molecular recognition of transcriptional repressor motifs by the WD domain of the Groucho/TLE corepressor. Mol. Cell *22*, 645–655.

Kagawa, W., Sagawa, T., Niki, H., and Kurumizaka, H. (2011). Structural basis for the DNA-binding activity of the bacterial β-propeller protein YncE. Acta Crystallogr. D Biol. Crystallogr. *67*.

Kaufmann, I., Martin, G., Friedlein, A., Langen, H., and Keller, W. (2004). Human Fip1 is a subunit of CPSF that binds to U-rich RNA elements and stimulates poly(A) polymerase. EMBO J. *23*, 616–626.

Kleiman, F.E., and Manley, J.L. (1999). Functional interaction of BRCA1-associated BARD1 with polyadenylation factor CstF-50. Science *285*, 1576–1579.

Kleiman, F.E., and Manley, J.L. (2001). The BARD1-CstF-50 interaction links mRNA 3' end formation to DNA damage and tumor suppression. Cell *104*, 743–753.

Legendre, M., and Gautheret, D. (2003). Sequence determinants in human polyadenylation site selection. BMC Genomics *4*, 7.

Legrand, P., Pinaud, N., Minvielle-Sébastia, L., and Fribourg, S. (2007). The structure of the CstF-77 homodimer provides insights into CstF assembly. Nucleic Acids Res. *35*, 4515–4522.

Li, H., Tong, S., Li, X., Shi, H., Ying, Z., Gao, Y., Ge, H., Niu, L., and Teng, M. (2011). Structural basis of pre-mRNA recognition by the human cleavage factor Im complex. Cell Res. *21*, 1039–1051.

MacDonald, C.C., Wilusz, J., and Shenk, T. (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*, 6647–6654.

Mandel, C.R., Kaneko, S., Zhang, H., Gebauer, D., Vethantham, V., Manley, J.L., and Tong, L. (2006). Polyadenylation factor CPSF-73 is the pre-mRNA 3'-end-processing endonuclease. Nature *444*, 953–956.

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

McDevitt, M.A., Hart, R.P., Wong, W.W., and Nevins, J.R. (1986). Sequences capable of restoring poly(A) site function define two distinct downstream elements. EMBO J. *5*, 2907–2913.

Moreno-Morcillo, M., Minvielle-Sébastia, L., Fribourg, S., and Mackereth, C.D. (2011a). Locked Tether Formation by Cooperative Folding of Rna14p Monkeytail and Rna15p Hinge Domains in the Yeast CF IA Complex. Structure *19*, 534–545.

Moreno-Morcillo, M., Minvielle-Sébastia, L., Mackereth, C., and Fribourg, S. (2011b). Hexameric architecture of CstF supported by CstF-50 homodimerization domain structure. RNA New York N *17*, 412–418.

Murthy, K.G., and Manley, J.L. (1992). Characterization of the multisubunit cleavage-polyadenylation specificity factor from calf thymus. J. Biol. Chem. *267*, 14804–14811.

Murthy, K.G., and Manley, J.L. (1995). The 160-kD subunit of human cleavage-polyadenylation specificity factor coordinates pre-mRNA 3'-end formation. Genes Dev. *9*, 2672–2683.

Noble, C.G., Walker, P.A., Calder, L.J., and Taylor, I.A. (2004). Rna14-Rna15 assembly mediates the RNA-binding capability of Saccharomyces cerevisiae cleavage factor IA. Nucleic Acids Res. *32*, 3364–3375.

Noble, C.G., Beuth, B., and Taylor, I.A. (2007). Structure of a nucleotide-bound Clp1-Pcf11 polyadenylation factor. Nucleic Acids Res. *35*, 87–99.

Ohnacker, M., Barabino, S.M.L., Preker, P.J., and Keller, W. (2000). The WD-repeat protein Pfs2p bridges two essential factors within the yeast pre-mRNA 3|[prime]|-end-processing complex. EMBO J. *19*, 37–47.

Pancevac, C., Goldstone, D.C., Ramos, A., and Taylor, I.A. (2010). Structure of the Rna15 RRM-RNA complex reveals the molecular basis of GU specificity in transcriptional 3'-end processing factors. Nucleic Acids Res.

Paulson, A.R., and Tong, L. (2012). Crystal structure of the Rna14-Rna15 complex. RNA New York N *18*, 1154–1162.

Perez Canadillas, J.M., and Varani, G. (2003). Recognition of GU-rich polyadenylation regulatory elements by human CstF-64 protein. EMBO J *22*, 2821–2830.

Qu, X., Perez-Canadillas, J.-M., Agrawal, S., De Baecke, J., Cheng, H., Varani, G., and Moore, C. (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*, 2101–2115.

Rüegsegger, U., Beyer, K., and Keller, W. (1996). Purification and characterization of human cleavage factor Im involved in the 3' end processing of messenger RNA precursors. J. Biol. Chem. *271*, 6107–6113.

Salisbury, J., Hutchison, K.W., and Graber, J.H. (2006). A multispecies comparison of the metazoan 3'-processing downstream elements and the CstF-64 RNA recognition motif. BMC Genomics *7*, 55.

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

Takagaki, Y., and Manley, J.L. (1997). RNA recognition by the human polyadenylation factor CstF. Mol. Cell. Biol. *17*, 3907–3914.

Takagaki, Y., and Manley, J.L. (2000). Complex protein interactions within the human polyadenylation machinery identify a novel component. Mol. Cell. Biol. *20*, 1515–1525.

Takagaki, Y., Ryner, L.C., and Manley, J.L. (1989). Four factors are required for 3'-end cleavage of pre-mRNAs. Genes Dev. *3*, 1711–1724.

Takagaki, Y., Manley, J.L., MacDonald, C.C., Wilusz, J., and Shenk, T. (1990). A multisubunit factor, CstF, is required for polyadenylation of mammalian pre-mRNAs. Genes Dev. *4*, 2112–2120.

De Vries, H., Rüegsegger, U., Hübner, W., Friedlein, A., Langen, H., and Keller, W. (2000). Human pre-mRNA cleavage factor II(m) contains homologs of yeast proteins and bridges two other cleavage factors. EMBO J. *19*, 5895–5904.

Yang, Q., Gilmartin, G.M., and Doublié, S. (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*, 10062–10067.

Yang, Q., Coseno, M., Gilmartin, G.M., and Doublié, S. (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. Struct. Lond. Engl. 1993 *19*, 368–377.

Zhang, P., Lee, H., Brunzelle, J.S., and Couture, J.-F. (2012). The plasticity of WDR5 peptide-binding cleft enables the binding of the SET1 family of histone methyltransferases. Nucleic Acids Res. *40*, 4237–4246.