Title
Characterization of the Mammalian mRNA 3' Processing Complex

Permalink
https://escholarship.org/uc/item/0m16c8r4

Author
Chan, Serena Leong

Publication Date
2014

Peer reviewed|Thesis/dissertation
Characterization of the Mammalian mRNA 3' Processing Complex

DISSEPTION

submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in Biomedical Sciences

by

Serena Leong Chan

Dissertation Committee:
Professor Yongsheng Shi, Ph.D., Chair
Professor Andrej Luptáš, Ph.D.
Professor Klemens J. Hertel, Ph.D.
Professor Marian L. Waterman, Ph.D.

2014
DEDICATION

To my wonderful,

loving

mom

for teaching me how to be a good person,
for showing me the joys of life
and
for loving me for who I am.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF FIGURES</td>
<td>v</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>vii</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>viii</td>
</tr>
<tr>
<td>CURRICULUM VITAE</td>
<td>x</td>
</tr>
<tr>
<td>ABSTRACT OF THE DISSERTATION</td>
<td>xiii</td>
</tr>
<tr>
<td><strong>Chapter 1. Introduction</strong></td>
<td>1</td>
</tr>
<tr>
<td>1.1 Pre-mRNA 3’ Processing Components Overview</td>
<td>4</td>
</tr>
<tr>
<td>1.2 Cis Elements</td>
<td>4</td>
</tr>
<tr>
<td>1.3 Trans Factors</td>
<td>7</td>
</tr>
<tr>
<td>1.4 mRNA 3’ Processing Complex Assembly and Reaction</td>
<td>23</td>
</tr>
<tr>
<td>1.5 mRNA 3’ Processing Regulation</td>
<td>27</td>
</tr>
<tr>
<td>1.6 A Look Ahead</td>
<td>31</td>
</tr>
<tr>
<td><strong>Chapter 2. Interactome Map of the mRNA 3’ Processing Machinery</strong></td>
<td>34</td>
</tr>
<tr>
<td>2.1 Summary</td>
<td>34</td>
</tr>
<tr>
<td>2.2 Introduction</td>
<td>36</td>
</tr>
<tr>
<td>2.3 Materials and Methods</td>
<td>43</td>
</tr>
<tr>
<td>2.4 Results</td>
<td>52</td>
</tr>
<tr>
<td>2.5 Conclusions and Discussion</td>
<td>72</td>
</tr>
<tr>
<td><strong>Chapter 3. CPSF 30 and WDR33 Directly Bind to AAUAAA</strong></td>
<td>76</td>
</tr>
<tr>
<td>3.1 Summary</td>
<td>76</td>
</tr>
<tr>
<td>3.2 Introduction</td>
<td>77</td>
</tr>
<tr>
<td>3.3 Materials and Methods</td>
<td>81</td>
</tr>
<tr>
<td>3.4 Results</td>
<td>94</td>
</tr>
<tr>
<td>3.5 Conclusions and Discussion</td>
<td>118</td>
</tr>
<tr>
<td><strong>Chapter 4. Regulation of CPSF-AAUAAA Interaction</strong></td>
<td>124</td>
</tr>
<tr>
<td>4.1 Summary</td>
<td>124</td>
</tr>
<tr>
<td>4.2 Introduction</td>
<td>125</td>
</tr>
<tr>
<td>4.3 Materials and Methods</td>
<td>127</td>
</tr>
<tr>
<td>4.4 Results</td>
<td>133</td>
</tr>
<tr>
<td>4.5 Conclusions and Discussion</td>
<td>146</td>
</tr>
<tr>
<td><strong>Chapter 5. Overall Conclusions and Perspectives</strong></td>
<td>151</td>
</tr>
<tr>
<td>Bibliography</td>
<td>157</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

Figure 1.1 mRNA 3’ Processing Cis Elements .......................................................... 3
Figure 1.2 mRNA 3’ Processing Trans Factors ......................................................... 8
Figure 1.3 mRNA RNA:Protein Interactions in Mammals ....................................... 13
Figure 1.4 mRNA 3’ Processing Mechanism ............................................................ 24
Figure 1.5 mRNA 3’ Processing Regulation ............................................................. 28

Figure 2.1 Immunoprecipitation Method .................................................................. 53
Figure 2.2 Immunoprecipitation Products .................................................................. 55
Figure 2.3 Summary of the Interactome Network ....................................................... 56
Figure 2.4 Ubiquitination Mechanism ...................................................................... 64
Figure 2.5 CFIm Interacts with E3 Ligases ............................................................... 65
Figure 2.6 CFIm Directly Interacts with E3 Ligases .................................................. 69
Figure 2.7 CFIm Subunits are Polyubiquitinatated by WWP2 and Itch .................... 70

Figure 3.1 Mapping Experiment 1 Reagents ........................................................... 95
Figure 3.2 Mapping Experiment 1 ........................................................................... 97
Figure 3.3 Identifying CPSF 30 as a Factor that Directly Interacts with AAUAAA ................................................................................................................................................. 100
Figure 3.4 Identifying WDR33 as Another Factor that Directly Interacts with AAUAAA ................................................................................................................................................. 101
Figure 3.5 Mapping Experiment 2 Reagents ........................................................... 102
Figure 3.6 Mapping Experiment 2 .......................................................... 103
Figure 3.7 CPSF 30 RNA-Binding Domain Identification ............................ 105
Figure 3.8 HA CPSF 30 Complex Purification .......................................... 107
Figure 3.9 Mapping CPSF 30 RNA-Binding Domains via In Vitro Assays ...... 108
Figure 3.10 Domain Specificity .................................................................. 110
Figure 3.11 Mapping CPSF 30 and WDR33 Footprints ................................ 113
Figure 3.12 PAS-SEQ .............................................................................. 115
Figure 3.13 PAS-SEQ Results ................................................................. 117
Figure 3.14 Model ................................................................................... 119

Figure 4.1 RNAs Utilized .......................................................................... 134
Figure 4.2 CPSF is Not Specific on Long Adenovirus or SVL RNAs .......... 135
Figure 4.3 CPSF Nonspecificity is Not Due to Technical Issues ............... 137
Figure 4.4 Purification of the Specificity Factor from HeLa Nuclear Extract ... 139
Figure 4.5 Chromatography can Partially Isolate the Specificity Factor ...... 141
Figure 4.6 Nonspecificity Factor is a CPSF Subunit .................................... 143
Figure 4.7 Deleting hFip1 Sequences Did Not Affect Specificity ............... 145

Figure 5.1 Summary Model ...................................................................... 153
LIST OF TABLES

Table 2.1 mRNA 3’ Processing Machinery Interactome (with other cellular processes) ........................................................................................................................................... 57
Table 2.2 FLAG CPSF 73 Mass Spectrometry Results .............................................. 58
Table 2.3 FLAG CstF 77 Mass Spectrometry Results .............................................. 59
Table 2.4 FLAG CFIm 25 Mass Spectrometry Results ........................................... 60
Table 2.5 FLAG Pcf11 Mass Spectrometry Results .............................................. 61
ACKNOWLEDGMENTS

I would like to first thank Yong for all of his advice on how to execute experiments, think like a scientist and how to present my work in a logical and appealing manner. Thanks to Yong, I can now never look at a PowerPoint presentation without noticing a font or spacing inconsistency. I would also like to thank him for his perfectionistic qualities. I am thankful to be his first graduate student.

The Shi lab has been a great lab to grow as a scientist. Everyone in there loves to help each other and encourage each other’s success. I would like to particularly thank Ling and Chengguo for their technical support. My thesis work would not be complete without them. My first friend in the Shi lab was Gabriel. He made me laugh and taught me neat technical lab skills that I still appreciate today. My best friend and sister in the lab is Esther. She did not only help me with my experiments, but she and her husband provided me a home away from home. Whenever I had a bad day, she would take time out of her hectic day to have coffee and cookies with me and let me vent. She would also leave food and gifts on my desk to cheer me up. I can never repay her for helping me remain mentally sane all these years. For future Shi lab graduate students who are reading this, don’t give up; You can do it!

I would like to thank John Wiley & Sons, Ltd. for providing me written permission to use my manuscript in Chapter 1 of this manuscript. The text of this thesis/dissertation is a reprint of the material as it appears in WIRES RNA. The coauthor listed in this publication directed and supervised research which forms the basis for the thesis/dissertation.

This document and my degree would not be possible without the generous funding from the National Institutes of Health and from the Eugene Cota-Robles Fellowship.

The friends I made in graduate school and people from my Toastmasters club are wonderful, insightful people with beautiful hearts that I hope to remain friends for many years to come.

The MMG department is the best department I have ever been in. The genuine, friendly nature of this department is incomparable and something no NIH grant can provide. The faculty members are down-to-earth people who are passionate in helping graduate students succeed. I would particularly like to thank my committee members, Dicki, Andrej and Marian for providing me with constructive feedback without scaring the living crap out of me. I would also like to thank Ming for teaching me how to answer seminar questions. Many graduate students in this department have become like siblings to me. Whether it is waiting for a seminar to start or chewing on a doughnut, seeing them usually brought a smile to my face. I would like to particularly thank Vlad for being my co-presenter in our “Elevator Talk” workshops and for being the “Gandhi of graduate students” and reminding me to remain positive. The office staff has also been a joy to work with. Their smiles every morning make me smile too. And of course, Mike. Mike
has always been there to talk to me about my troubles and I always leave his desk a bit happier than when I arrived.

The Academy of Future Scientists, Group P is my work family that I am honored to be a part of. Marina, our group leader, is the most nurturing professor I have ever met. I am so lucky to have met her and I aspire to be just like her.

I would also like to thank my elementary and middle school friends for being with me through all of my major life stages thus far and for their friendship. I hope we remain friends for many more major life stages to come. I have 5 best friends I would like to particularly thank. First of all, thanks for being my best friends; I would be lucky to only have 1 best friend but having 5 is incredibly unbelievable. Dr. Leo, thank you for understanding my crazy graduate school lifestyle. It means so much to me to be able to talk freely about my graduate career to someone who is in the same boat. Dearest Becky, thank you for letting me be my crazy self and giving me a home in college. Casey, thank you for being the best roommate ever, saving my life in Egypt and for being patient with me all these years. And of course, my two Davids. Thank you for your humor, patience, kindness and honesty with me. Your conversations have allowed me to mature and grow into the person I am today.

Lastly, I would like to thank my family. I am truly blessed to be part of a family where there is so much humor and joy. My cousins (i.e. my brothers and sisters from another mother) are hilarious people with hearts of gold. How lucky am I to grow up with you guys! My uncles and aunts have been like second parents to me, especially Uncle Luke & Auntie Della, Uncle Boone and Auntie Gini. My grandmother, with a 5-year-old personality, has been an upmost inspiration. She has taught me to always find laughter, no matter how dark the situation might be. I hope one day I can age mentally as well as her.

Dad, thank you for teaching me how to be a successful grown up and teaching me to laugh at myself.

Jeff, the best brother in the world, thank you for bringing laughter in my life and for being my brother. Life is so much easier when you have someone who knows your past, present and future and who is more than willing to tease you about it!

And finally, mom. You have been the best mom and friend any person can hope or dream for. I am so lucky to be your daughter. Thank for loving me and for shaping me to the person I am today. Everything that is good in my life, I owe to you. You are with me everyday - in my mind, in my spirit and in my heart. Thank you eternally.
CURRICULUM VITAE

Serena Leong Chan

EDUCATION

Doctor of Philosophy in Biomedical Sciences 2014
University of California Irvine, CA

Bachelor of Science in Genetics (with honors) 2008
University of California Davis, CA

RESEARCH EXPERIENCE

Graduate Research Assistant 2008-2014
University of California, Irvine

Research Technician 2008
USDA Institute of Forest Genetics

Research Intern 2006-2007
Pioneer International

Undergraduate Research Assistant 2005-2008
University of California, Davis

TEACHING EXPERIENCE

California Community College Intern 2013-2014
Orange Coast College

Vice President of Education 2013-2014
The Articulates Club, Toastmasters

“Communicating Science” Workshop Presenter 2011-2014
University of California, Irvine

Teaching Assistant 2013
University of California, Irvine
REFEREED JOURNAL PUBLICATIONS


PRESENTATIONS/POSTERS

UCI Departmental Retreat Poster, “CPSF 30 Directly Binds to AAUAAA in Mammalian mRNA 3’ Processing” 2014

UCI Departmental Seminar, “Proteomic and Functional Studies of the mRNA 3’ Processing Complex” 2014

UCI Departmental Seminar, “Proteomic and Functional Studies of the mRNA 3’ Processing Complex” 2013

UCI Departmental Seminar, “Redefining RNA-Protein Interactions in mRNA 3’ Processing” 2012

UCI Departmental Seminar, “mRNA 3’ Processing: Specificity and Regulation” 2011

HONORS

Competent Communicator Award, Toastmasters. 2013
First Runner-Up, Toastmasters Tall Tales Speech Contest 2013
First Runner-Up, Diamond Web Award, Toastmasters. 2013
Finalist, Scientific Blogging Writing Contest 2010
Eugene Cota-Robles Fellow 2008-2014
Citation Award for Excellence in Genetics 2008
Dean’s Honor List 2004-2008
Semi-finalist Oratorical Interpretation Speech Competition 2004
First Runner-Up Oratorical Interpretation Speech Competition 2004
ACADEMIC COMMITTEES

Departmental Graduate Student Representative. 2012-2013

Diverse Education Community and Doctoral Experience Student Council 2011-2012

Departmental Student Seminar Committee Member 2011-2012
mRNA 3’ processing, which typically involves an endonucleolytic cleavage followed by polyadenylation (addition of a string of adenosines), is an essential step in eukaryotic gene expression and significantly impacts many other gene expression steps such as transcription, splicing, mRNA export and translation (Zhao et al. 1999) (Chan et al. 2010) (Colgan et al. 1997) (Moore et al. 2009). Additionally, 3’ processing is needed for gene regulation. Recent studies revealed that approximately 70% of human genes produce multiple mRNA isoforms with different 3’ processing sites (Derti et al. 2012) (Hoque et al. 2013). These mRNA isoforms may encode different proteins or produce different 3’ untranslated regions. This phenomenon, called alternative polyadenylation (APA), significantly expands the coding capacity of the genome and has been increasingly recognized as a critical mechanism for eukaryotic gene regulation (Di Giammartino et al. 2011) (Shi 2012) (Proudfoot 2011) (Tian et al. 2013). In addition, aberrant mRNA 3’ processing causes a wide range of diseases, including IPEX syndrome, thalassemia and has been implicated in the development of cancer
(Danckwardt et al. 2008) (Mayr et al. 2009). Therefore, it is critical to understand both the mechanism of mRNA 3’ processing and its regulation.

mRNA 3’ processing requires specific RNA-protein and protein-protein interactions. The proteins required in mammals include four multi-subunit complexes and the poly (A) polymerase (PAP) while the main RNA sequences include the AAUAAA hexamer and U/G-rich elements (Zhao et al. 1999) (Chan et al. 2010) (Colgan et al. 1997). A central question in the mRNA 3’ processing field has been to understand how the mRNA 3’ processing sites, also called poly (A) sites, are specifically recognized and regulated. To shed light on this important question, I carried out three projects. First, comprehensive proteomic analyses of the mRNA 3’ processing machinery were performed. These were accomplished by purifying all sixteen essential mRNA 3’ processing factors by immunoprecipitation and identifying their associated proteins through high throughput mass spectrometry analyses. The results of this study not only provided a nearly comprehensive interactome map of the mRNA 3’ processing machinery, but also revealed potential new regulatory mechanisms. I have experimentally validated the association between the mRNA 3’ processing factors and some of the newly identified interacting proteins, including several ubiquitin E3 ligases, and have provided evidence that these factors regulate the stabilities of mRNA 3’ processing factors. Second, I have characterized the mechanism by which the cleavage and polyadenylation specificity factor (CPSF) specifically recognizes the AAUAAA hexamer. In contrast to the prevalent model in which the CPSF subunit, CPSF 160, recognizes the AAUAAA by itself, my data provided direct evidence that the CPSF subunits, CPSF 30 and WDR33, directly
bind to AAUAAA together. Additionally, I showed that the CPSF 30-RNA interaction is mediated by its zinc fingers two and three, which, remarkably, are directly targeted by the influenza A viral protein, NS1A, to suppress host mRNA 3’ processing (Nemeroff et al. 1998) (Twu et al. 2006). Finally, I provide evidence that the specificity of the CPSF-RNA interaction is limited and that it requires additional factor(s) for proofreading.

Together, the results from these three projects provide novel and significant insights into the fundamental mechanisms for mammalian poly(A) site recognition and regulation.
Chapter 1

Introduction

Pre-mRNA 3’ processing is an essential step in eukaryotic mRNA biogenesis, and it directly impacts many other steps in the gene expression pathway, such as transcription termination, splicing, mRNA export, translation, and mRNA turnover (Colgan et al. 1997) (Zhao et al. 1999) (Moore et al. 2009). 3’ processing is also a versatile mechanism for gene regulation (Millevoi et al. 2010). For example, the efficiency of 3’ processing, the usage of APA sites, and the length of poly (A) tails can all be modulated. Additionally, 3’ processing-mediated gene regulation can be global or transcript-specific. Accumulating EST data revealed that a significant portion of eukaryotic genes produce alternatively polyadenylated mRNAs, indicating that APA may be involved in the regulation of a large set of genes (Tian 2005) (Lutz 2008). Indeed a series of recent genomic studies showed that APA regulation is widespread and plays important roles in the immune and the neural systems, oncogenesis, stem cell
differentiation and development (Mayr et al. 2009) (Ji et al. 2009) (Sandberg et al. 2008) (Flavell et al. 2008) (Ji et al. 2009). Defects in pre-mRNA 3’ processing have been associated with a wide spectrum of human diseases (Danckwardt et al. 2008). Therefore, it is critical to understand how the pre-mRNA 3’ processing machinery functions and how its activity can be regulated.

Much progress has been made in understanding the basic molecular mechanisms and regulation of pre-mRNA 3’ processing. The core cis-elements and the majority of the basal trans-acting factors required for 3’ processing have been identified (Figure 1.1 and 1.2) (Colgan et al. 1997) (Zhao et al. 1999) (Mandel et al. 2008) (Millevoi et al. 2010). Biochemical, genetic, and structural analyses have provided insights into the functions of some 3’ processing factors (Mandel et al. 2008). A diverse array of mechanisms has been uncovered regarding how 3’ processing can be regulated (Zhao et al. 1999) (Millevoi et al. 2010). More recently, purification of the functional human pre-mRNA 3’ processing complex has allowed compositional characterization of the entire machinery and made it possible to study its structure and dynamics throughout the cleavage and polyadenylation processes (Shi et al. 2009) (Shi et al. 2009). In this introduction, I provide a broad overview of our current knowledge of the assembly and function of the pre-mRNA 3’ processing complex.
The cis elements needed for mRNA 3' processing in mammals, *Saccharomyces cerevisiae* and plants are shown above.
1.1 Pre-mRNA 3’ Processing Components Overview

Assembly of the pre-mRNA 3’ processing complex is initiated by the binding of trans-acting factors to their target cis-elements in the RNA. Unlike the spliceosome whose assembly is aided by a series of specific RNA–RNA base-pairing events, assembly of the 3’ processing complex on the pre-mRNA is directed entirely by RNA–protein interactions (Wahl et al. 2009) (Colgan et al. 1997) (Zhao et al. 1999) (Millevoi et al. 2010) (Shi et al. 2009). As will be discussed later, the cis-elements for 3’ processing are generally short and/or degenerate and the interactions of individual 3’ processing factors with RNA are usually very weak (Figure 1). Therefore, an extensive network of RNA–protein and protein–protein interactions is required for the assembly of a stable 3’ processing complex. Here the cis-elements and protein factors involved are illustrated.

1.2 Cis Elements

Assembly of the pre-mRNA 3’ processing complex is directed by multiple key cis-elements collectively called the polyadenylation signals (PAS) (Figure 1.1) (Zhao et al. 1999). Mammalian PASes can be classified into two general types, the canonical and the noncanonical (Zhao et al. 1999) (Millevoi et al. 2010) (Dickson et al. 2010). The main cis-elements in a canonical PAS include a highly conserved A(A/U)UAAA element located 10–30 nucleotides upstream of the cleavage/polyadenylation site (poly(A) site) and a more variable U/GU-rich downstream element (DSE) ~30 nucleotides downstream of the poly(A) site. These two elements determine the general location of the
poly(A) site, which is most frequently found after an adenosine residue (Chen et al. 1995). Genomic studies suggest that poly(A) site selection in vivo is fairly heterogeneous and, for any given PAS, the corresponding poly(A) sites are usually clustered in a ∼20 nucleotides window (Tian 2005) (Pauws et al. 2001). In addition to the A(A/U)UAAA and the DSE, functionally important upstream sequence elements (USE) and auxiliary downstream elements (auxDSE) have also been identified in many viral and cellular mRNAs (Zhao et al. 1999) (Millevoi et al. 2010) (Hu et al. 2005). USEs are generally U-rich while auxDSEs are mostly G-rich. These additional cis-elements help improve 3’ processing efficiency by providing binding sites for the core 3’ processing factors and/or regulatory factors.

Twenty to thirty percent of human PASes lack the A(A/U)UAAA element and it remains poorly understood how these noncanonical PASes are recognized by the 3’ processing machinery or whether they are recognized by the same 3’ processing factors that bind canonical PASes (Zarudnaya 2003) (Tian 2005). In some cases, UGUAN and related sequences have been identified as the key cis-element for directing A(A/U)UAAA-independent cleavage and polyadenylation (Venkataraman et al. 2005). Recently another type of PAS was discovered that consists of only an A-rich USE and a potent DSE (Nunes et al. 2010). Currently, however, it is not clear whether the noncanonical PASes share any common characteristics. Interestingly, noncanonical PASes are often associated with APA and may play important regulatory roles (Tian 2005).

The PASes of the budding yeast and plant pre- mRNAs are thought to be significantly different from the mammalian PASes as they lack the highly conserved A(A/U)UAAA element (Zhao et al. 1999) (Millevoi et al. 2010). Instead, yeast and plant PASes include
a degenerate A-rich element which includes AAUAAA-like sequences (called the Positioning Element (PE) in budding yeast and the Near Upstream Element (NUE) in plants), a UA-rich USE (called the Efficiency Element (EE) in budding yeast and the Far Upstream Element (FUE) in plants), and a U-rich element (URE) upstream and/or downstream of the poly(A) site (Millevoi et al. 2010) (Zhao et al. 1999). Interestingly, these cis-elements are also shared by some mammalian non-canonical PASes (Venkataraman et al. 2005) (Nunes et al. 2010) indicating that PAS sequences might have been more conserved than previously thought.

Given the lack of strong, evolutionarily conserved consensus sequence in eukaryotic PASes, RNA secondary structures have long been suspected to play a role in directing the assembly of the 3’ processing complex (Graveley et al. 1996) (Sadofsky et al. 1984). Multiple lines of evidence support this notion. First, an in vitro selection for 3’ processing enhancers enriched motifs that are related in structure rather than sequence (Graveley et al. 1996). Secondly, functionally important secondary structures have been identified in the PASes of a number of viral mRNAs (Hans et al. 2000) (Sittler et al. 1995). It remains to be seen, however, whether RNA secondary structure plays a general role in 3’ processing and if so, what specific secondary structures are required and how they are recognized.
1.3 Trans Factors

Most, if not all, basal 3’ processing factors have been identified (Figure 1.2) (Millevoi et al. 2010) (Mandel et al. 2008) (Shi et al. 2009). Mammalian 3’ processing factors include the PAP, the poly(A)-binding proteins (PABPs), the RNA polymerase II large subunit (RNAP II), and four multi-subunit protein complexes, CPSF, the cleavage stimulatory factor (CstF), cleavage factor I (CF Im) and cleavage factor II (CF IIIm) (Colgan et al. 1997) (Zhao et al. 1999) (Mandel et al. 2008). All of these factors except for PABPs are required for cleavage, but CPSF and PAP are believed to be sufficient for the subsequent polyadenylation. PABPs are not required for either cleavage or polyadenylation, but function in activating PAP and in poly(A) tail length control (Mangus et al. 2003). Despite the significant divergence in the PAS sequences between yeast and mammals, most mammalian 3’ processing factors have homologs in yeast and plants, indicating that the pre-mRNA 3’ processing machinery and the biochemical mechanism have been conserved (Zhao et al. 1999) (Mandel et al. 2008).

CPSF

CPSF is required for both cleavage and polyadenylation. It recognizes the A(A/U)UAAA element, helps to recruit other components of the 3’ processing complex, and catalyzes cleavage (Mandel et al. 2008) (Colgan et al. 1997) (Zhao et al. 1999). Originally it was shown that the purified CPSF consisted of CPSF160, 100, and 73, and in some preparations also CPSF30 (Murthy et al. 1992) (Bienroth et al. 1991). Later studies
These 5 boxes encompass the major trans factors needed for mRNA 3’ processing to occur. The top four boxes represent the multi-complexes: CPSF, CstF, CFIm and CFIIm. Within each of those boxes, the subunits of each complex are listed along with the major functions on the right. The bottom blue box lists all of the other poly (A) factors needed for the reaction.
have added symplekin, hFip1, and Wdr33 to the list of CPSF components (Shi et al. 2009) (Kaufmann et al. 2004) (Takagaki et al. 2000). Next I will discuss the roles of the individual CPSF subunits in the context of the three aforementioned critical functions of CPSF.

**CPSF in RNA Binding**

CPSF recognizes the A(A/U)UAAA element with remarkable specificity as any single mutation within the hexamer strongly represses processing (Keller et al. 1991) (Sheets et al. 1990). Surprisingly, however, none of the CPSF subunits have a canonical RNA recognition motif (RRM). CPSF160 is believed to be the major CPSF subunit responsible for recognizing the A(A/U)UAAA element based on multiple lines of evidence (Figure 1.3). For example, CPSF160 can be UV crosslinked to AAUAAA-containing RNAs, and recombinant CPSF160 preferentially binds to RNAs containing AAUAAA than those with mutant hexamers (Murthy et al. 1995) (Keller et al. 1991) (Moore et al. 1988). Two fragments within the N-terminal half of CPSF160 bear limited similarities to the RRM domain sequence and were initially suspected to mediate RNA recognition (Murthy et al. 1995). However, a later study of Cft1p, the yeast CPSF160 homolog, attributed its RNA-binding activity to a central domain (Dichtl et al. 2002). Although this region seems to be fairly well conserved from CFT1 to CPSF160, mutational analysis is needed to directly test the involvement of this segment in A(A/U)UAAA recognition by CPSF160.

At least two other subunits of the CPSF are involved in RNA binding, hFip1, and CPSF30 (Kaufmann et al. 2004) (Barabino et al. 1997). The C-terminal arginine-rich
domain of hFip1 binds to U-rich sequences (Kaufmann et al. 2004). The RNA-binding activity of CPSF30 is mediated by its zinc finger domain and it also prefers poly(U) sequences (Barabino et al. 1997). Since many USEs are U-rich sequences, hFip1 and CPSF30 may be involved in recognizing the USEs and other sequences, thereby providing additional RNA-protein contacts that may stabilize the specific interaction between CPSF160 and the A(A/U)UAAA element.

**Protein–Protein Interactions of CPSF**

CPSF subunits interact with one another extensively (Colgan et al. 1997) (Zhao et al. 1999) (Mandel et al. 2008). CPSF 100, 73, and symplekin have recently been suggested to form the core of CPSF complex while other subunits join the complex as peripheral factors (Sullivan et al. 2009). Furthermore, CPSF recruits other components of the 3’ processing complex through direct physical interactions. For example, CPSF can be isolated in a pre-assembled complex with CstF, and this association is mediated by multiple factors, including CPSF160, hFip1, and symplekin (Kaufmann et al. 2004) (Takagaki et al. 2000) (Murthy et al. 1995). Wdr33 may also be involved in bridging CPSF and CstF as its yeast homolog Pfs2p interacts with factors that are homologous to subunits of the CPSF and CstF complexes (Ohnacker et al. 2000). These interactions allow cooperative binding of CPSF and CstF to A(A/U)UAAA and the DSE (Colgan et al. 1997) (Zhao et al. 1999) (Mandel et al. 2008). Following cleavage, CPSF remains bound to the A(A/U)UAAA element and anchors PAP to the RNA for polyadenylation
through direct interactions between CPSF160, hFip1, and PAP (Kaufmann et al. 2004) (Murthy et al. 1995).

CPSF also interacts with a variety of regulatory proteins to modulate 3′ processing either globally or in a transcript-specific manner. For example, the influenza protein NS1 represses the 3′ processing of host cellular pre-mRNAs by inhibiting CPSF activity via direct interaction with CPSF30 (Nemeroff et al. 1998). In another example, following stress, HSF1 not only activates transcription of heat shock protein (Hsp) genes, but also promotes the 3′ processing of Hsp transcripts through its association with symplekin (Xing et al. 2004). Finally, the RRM-containing floral repressor FCA interacts with CPSF through FY, the plant homolog of Wdr33, to control flowering time at least in part by modulating the 3′ processing of specific transcripts (Simpson et al. 2003).

**CPSF in the Catalysis of Cleavage**

CPSF 73 has been identified as the endonuclease responsible for the cleavage step based on several lines of evidence (Ryan et al. 2004) (Mandel et al. 2006). First, CPSF 73 can be UV crosslinked to the cleavage site in the RNA substrate (Ryan et al. 2004). Secondly, CPSF 73 contains a metallo-beta-lactamase domain and a β-CASP domain (Mandel et al. 2006). Some members of the β-CASP protein family, such as Artemis, are known endonucleases (Callebaut et al. 2002). Thirdly, recombinant CPSF 73 shows zinc-dependent endonuclease activity (Mandel et al. 2006). It is noted that the endonuclease activity of the recombinant CPSF 73 is very weak and stimulation by other 3′ processing factor(s) may be necessary for efficient cleavage. Although CPSF
100 is structurally highly similar to CPSF 73, it lacks the zinc-binding motif and, as a result, does not possess endonuclease activity (Mandel et al. 2006). The functions of CPSF 100 remain unclear.

The endonuclease activity of CPSF 73 was also revealed by studies on metazoan histone pre-mRNA 3′ processing (Dominski et al. 2005). Directed by unique cis-elements, the 3′ processing of metazoan histone pre-mRNAs involves only a cleavage step without polyadenylation (Marzluff et al. 2008). Recent studies suggest that some CPSF subunits, including CPSF 73, 100 and symplekin, function in this process and CPSF 73 functions as the endonuclease for cleavage (Dominski et al. 2005) (Kolev et al. 2005). Intriguingly, CPSF 73 has been suggested to play additional role in histone pre-mRNA 3′ processing as the exonuclease for degradation of the downstream RNA (Yang et al. 2009).

**CstF**

CstF recognizes the DSE and is specifically required for cleavage but apparently not for polyadenylation (Colgan et al. 1997) (Zhao et al. 1999) (Mandel et al. 2008). CstF is comprised of three subunits, CstF 77, CstF 50 and either CstF 64 or its paralog CstF 64 τ (Wallace et al. 1999) (Takagaki et al. 1990). Recent studies suggest that CstF may function as a dimer in the 3′ processing complex (Bai et al. 2007) (Legrand et al. 2007).
Figure 1.3 mRNA RNA:Protein Interactions in Mammals

Here it shows which trans factor binds to which mRNA 3’ processing cis element.
CstF in RNA Binding

The RNA-binding activity of CstF is mediated by CstF 64/τ (Takagaki et al. 1990). CstF 64 contains an N-terminal RRM domain, which on its own binds with high affinity to GU-rich sequences similar to those found in natural DSEs (Figure 1.3) (Takagaki et al. 1997). The intact CstF complex, however, binds to PAS-containing pre-mRNAs very weakly as shown by both UV crosslinking and gel shift assays (Murthy et al. 1992). Stable association of CstF to DSE requires co-operative binding of CPSF to the upstream A(A/U)UAAA element (Murthy et al. 1992). Although the exact mechanism is not well understood, this type of obligatory cooperative binding may ensure that 3’ processing complex assembles only when both the A(A/U)UAAA and DSE are present and the spacing between the two elements is appropriate.

CstF 64 τ shares the same domain structure as CstF 64, but seems to have different RNA-binding specificity (Monarez et al. 2007). CstF 64 τ is highly expressed in the testis and has been proposed to mediate testis-specific poly(A) site choice (Wallace et al. 1999). CstF 64 τ knockout mice showed specific defects in spermatogenesis and male infertility, suggesting that the CstF 64 τ-mediated pre-mRNA 3’ processing plays important roles in the testis (Dass et al. 2007). Therefore, the relative abundance of CstF 64 and related proteins may be involved in regulating tissue-specific APA. Both CstF 64 and CstF 64 τ seem to be evolutionarily conserved from yeast to human (Shi et al. 2009).
Protein–Protein Interactions of CstF

CstF 77 is the scaffold holding the CstF complex together as it interacts with both CstF 64 and CstF 50 via its proline-rich domain, while CstF 64 and CstF 50 do no interact with each other (Takagaki et al. 2000) (Takagaki et al. 1994). In addition, CstF 77 also contains a so-called Half a TPR (HAT) domain. Crystal structures revealed that the CstF 77 HAT domain forms a homodimer (Bai et al. 2007) (Legrand et al. 2007). This is consistent with earlier observations that both CstF 77 and its yeast homolog RNA14p can self-associate (Takagaki et al. 2000) (Noble et al. 2004). In addition, CstF 50 may also dimerize (Takagaki et al. 2000). Together, these results provided strong evidence that CstF functions as a dimer in the 3’ processing complex.

In addition to the aforementioned RRM domain, CstF 64/τ contains a ‘hinge’ domain, a Pro/Gly-rich region, and a highly conserved C-terminal domain (CTD). Embedded within the Pro/Gly-rich region is 12 tandem copies of MEARA/G (mouse and human) or 11 copies of LEPRG (chicken) pentapeptide repeats, but the functions of these repeats as well as the entire Pro/Gly-rich domain remain unclear (Colgan et al. 1997) (Zhao et al. 1999) (Mandel et al. 2008). The CstF 64 ‘hinge’ domain mediates its interactions with CstF 77 and its CTD binds to Pcf11 and the transcription co-activator PC4 (Calvo et al. 2001) (Hockert et al. 2010) (Qu et al. 2007). As mentioned earlier, both CstF 77 and 64 interact with CPSF subunits to facilitate the co-operative binding of CPSF and CstF to pre-mRNAs.

As mentioned above, CstF 50 can self-associate and this interaction is mediated by its N-terminal region (Takagaki et al. 2000). Furthermore, CstF50 contains seven WD-40
repeats, a domain known for mediating protein–protein interactions (Colgan et al. 1997) (Zhao et al. 1999) (Mandel et al. 2008). CstF 50 binds to RNAP II CTD and this interaction has been proposed to help recruit CstF to the transcription elongation complex (McCracken et al. 1997). Following UV-induced DNA damage, CstF50 becomes associated with the BRCA1–BARD1 complex and other factors and these interactions are responsible for transiently repressing pre-mRNA 3′ processing (Kleiman et al. 2001).

**CFIm**

Similar to CstF, CFIm is involved in RNA-binding and is required only for the cleavage step (Figure 1.3) (Colgan et al. 1997) (Zhao et al. 1999) (Mandel et al. 2008). CF Im consists of CFIm 25 and one of two structurally related proteins, CFIm 59 and CF Im 68. Again similar to CstF, CF Im has recently been shown to function as a dimer (Coseno et al. 2008) (Yang et al. 2010). There are no clear homologs of CF Im in yeast. But it has been proposed that CF Im may be the functional equivalent of the yeast 3′ processing factor Hrp1, a RNA-binding factor that does not have a clear mammalian homolog (Venkataraman et al. 2005).

**CFIm in RNA Binding**

All CF Im subunits can be UV crosslinked to RNA, suggesting that they are all involved in RNA recognition (Ruegsegger et al. 1996). Recombinant CFIm25-68 binds
specifically to UGUAN motif and this interaction has recently been shown to be mediated by CFIm 25 (Figure 1.3) (Ruegsegger et al. 1996) (Yang et al. 2010). CFIm 25 contains a Nudix domain and belongs to the Nudix phosphohydrolase superfamily (Coseno et al. 2008). However CFIm 25 differs from other Nudix hydrolases at several key residues and lacks enzymatic activity (Coseno et al. 2008). Instead the CFIm 25 Nudix domain functions as a RNA-binding domain that specifically recognizes the UGUAN motif (Yang et al. 2010). CFIm 59 and 68 both have a RRM domain, but only bind to RNA in the presence of CFIm 25 (Dettwiler 2004). Currently it is unclear what RNA sequences CFIm 59 and 68 recognize and what the functional significance of their RNA-binding activities are. CF Im–RNA interactions help improve the 3’ processing efficiency at canonical PASes (Venkataraman et al. 2005) (Ruegsegger et al. 1998). For noncanonical PASes that lack the A(A/U)UAAA element but contain UGUAN motifs, CFIm can function as the primary RNA-binding factor for 3’ processing complex assembly and recruit other 3’ processing factors such as CPSF and PAP through direct interactions (Venkataraman et al. 2005). In keeping with the important roles of CFIm 25 in PAS recognition, depletion of CF Im25 by RNAi causes significant changes in the APA profile for many transcripts (Kubo et al. 2006).

Protein–Protein Interactions of CFIm

In addition to self-association, CFIm 25 also interacts with PAP and PABPN1 (Mandel et al. 2008). As mentioned earlier, both CFIm 59 and 68 have an RRM domain. Interestingly, however, the RRM of CFIm 68 is required for its interaction with CFIm 25.
(Dettwiler 2004). CFIm 59 and 68 also contain a C-terminal RS domain that is rich in arginine-serine dipeptide repeats. The RS domain is commonly found in the SR protein family, a large group of proteins best known as splicing regulators (Zhong et al. 2009). The RS domain has been shown to mediate RNA-binding as well as protein–protein interactions, including interactions with other SR proteins (Zhong et al. 2009). Indeed, CFIm 59 and 68 bind to a number of SR and SR-like proteins, and the CFIm complex has been identified as a component of the spliceosome (Dettwiler 2004) (Zhou et al. 2002) (Millevoi et al. 2006). The interactions between CFIm and splicing factors play important roles in coupling splicing and 3′ processing.

**CFIIm**

CFIIm is required only for the cleavage step, but its exact functions remain unclear (Colgan et al. 1997) (Zhao et al. 1999) (Mandel et al. 2008). CFIIm contains at least two subunits, Pcf11 and hClp1, both conserved from yeast to human. Pcf11 homologs in yeast and Drosophila are required for both pre-mRNA 3′ processing and transcription termination (Proudfoot 2004). Pcf11 binds to the RNAP II CTD in a phosphorylation-dependent manner via its CTD-interacting domain (CID domain) (Proudfoot 2004). It has been proposed that RNAP II-bound Pcf11 promotes transcription termination by dismantling the elongation complex (Zhang et al. 2005). Additionally, the yeast Pcf11p has been shown to recruit the mRNA export adaptor Yra1p to the mRNA, thereby linking pre-mRNA 3′ processing to mRNA export (Johnson et al. 2009).
Although its role in pre-mRNA 3’ processing remains unknown, hClp1 has recently been identified as an RNA 5’-kinase and it functions as a component of the endonuclease complex in tRNA splicing (Weitzer et al. 2007) (Paushkin et al. 2004). The kinase activity of hClp1 is required for maintaining the phosphorylation of the 5’-end of the 3’ exon for the subsequent ligation (Weitzer et al. 2007). It is still unclear, however, if the kinase activity of hClp1 is necessary for pre-mRNA 3’ processing. But the recombinant yeast Clp1 protein has no detectable kinase activity, indicating that the kinase activity may be unique to Clp1 homologs in higher eukaryotes and this activity may not be required for pre-mRNA 3’ processing itself (Ramirez et al. 2008).

**Poly(A) Polymerases**

PAP is perhaps the best-characterized 3’ processing factor so far (Colgan et al. 1997) (Zhao et al. 1999) (Mandel et al. 2008). A nucleotidyltransferase catalytic domain occupies the N-terminal half of PAP and is highly conserved. A RNA-binding domain is located near the middle of the protein. RNA binding by PAP alone, however, is not sequence-specific and its recruitment to 3’ processing complex requires interactions with other 3’ processing factors including CPSF and CF Im (Colgan et al. 1997) (Zhao et al. 1999) (Mandel et al. 2008). The CTD of PAP contains a bipartite nuclear localization signal (NLS) and is rich in serine and threonine residues. The CTD is a hot spot for post-translational modifications and plays important regulatory roles. First, PAP CTD is hyperphosphorylated during mitosis by the mitosis promoting factor (MPS, p34cdc2/cyclin B) and, as a result, PAP activity is strongly inhibited (Colgan et al. 1996).
Repression of PAP activity is part of the cellular mechanism for blocking protein synthesis during mitosis. Secondly, PAP CTD binds to 14-3-3ε in a phosphorylation-dependent manner and this association inhibits PAP activity and redistributes PAP to the cytoplasm (Kim et al. 2003). Thirdly, PAP CTD is acetylated by the CREB-binding protein (CBP) and acetylation inhibits its association with CFIm 25 and its nuclear localization (Shimazu et al. 2007). Finally, PAP CTD is sumoylated at multiple sites, and sumoylation stabilizes PAP, promotes its nuclear localization, but inhibits its enzymatic activity in vitro (Vethantham et al. 2008).

In addition to the canonical PAP, metazoans have at least three additional nuclear poly(A) polymerases, neo-PAP, Star-PAP and TPAP. Encoded by an intronless gene, TPAP is specifically expressed in the testis and can be found in both nucleus and the cytoplasm (Topalian et al. 2001). The functions of TPAP remain poorly understood. Neo-PAP shares the same overall domain organization with PAP and its in vitro polyadenylation activity is indistinguishable from that of the canonical PAP (Mellman et al. 2008) (Kyriakopoulou et al. 2001). Neo-PAP can be incorporated into functional human 3′ processing complex in vitro, but its in vivo functions have not been characterized (Shi et al. 2009).

Star-PAP has a very different domain structure. It contains a zinc finger domain and a RRM near its N-terminus, a split catalytic domain in the middle, and a PAP-associated domain as well as a RS domain in the C-terminal one third (Lee et al. 2000). Star-PAP was originally identified as a terminal uridylyl transferase (TUTase) specific for U6 snRNA (Trippe et al. 2006). The putative homolog of Star-PAP in fission yeast, Cid1, also has TUTase activity (Rissland et al. 2007). A later study, however, demonstrated
that Star-PAP preferentially uses ATP as substrate and mainly functions as a poly(A) polymerase (Lee et al. 2000). Star-PAP functions in a complex with the type I phosphatidylinositol 4-phosphate 5-kinases (PIPKIα) as well as subunits of the CPSF to control the 3’ processing of a subset of transcripts. Interestingly, the activity of Star-PAP is directly regulated by phosphatidylinositol-4,5-bisphosphate (PtdIns4,5P) (Lee et al. 2000). Compare to *Saccharomyces cerevisiae* which only has one nuclear PAP, the presence of multiple nuclear PAPs in metazoans may allow for more elaborate regulation of pre-mRNA 3’ processing under different conditions.

**Poly(A)-Binding Proteins**

There are at least five PABP proteins in humans, one nuclear form (PABPN1) and four cytoplasmic ones (PABPC1, 3–5) (Mangus et al. 2003) (Kuhn et al. 2004). The nuclear and the cytoplasmic PABPs have distinct domain structures. It has been shown that PABPN1 stimulates PAP activity and plays an important role in poly(A) tail length control (Bienroth et al. 1993). Following cleavage, PAP, anchored to the pre-mRNA by CPSF, initiates polyadenylation in a slow and distributive reaction (Bienroth et al. 1993). Binding of PABPN1 to the newly synthesized poly(A) tail stabilizes the polyadenylation complex and dramatically stimulates PAP activity. As a result, polyadenylation switches to a fast and processive phase. PABPN1 coats the entire length of poly(A) tail as it emerges from PAP. When the poly(A) tail reaches ~250 nucleotides in length, polyadenylation is terminated by a poorly understood but PABPN1-dependent mechanism (Bienroth et al. 1993). PABPs have also been implicated in promoting
mRNA export (Mangus et al. 2003) (Kuhn et al. 2004). In the cytoplasm, PABPs play critical roles in translation by promoting 5′ → 3′ interaction of the mRNA and stimulating initiation (Mangus et al. 2003) (Kuhn et al. 2004).

The budding yeast has only one PABP gene (PAB1), which is the ortholog of the mammalian cytoplasmic PABPs (Mangus et al. 2003) (Kuhn et al. 2004). The budding yeast does express a structurally distinct nuclear poly(A) binding protein Nab2p, which is probably the functional equivalent of PABPN1. Both Nab2p and Pab1p have been implicated in poly(A) tail length control and mRNA export (Mangus et al. 2003) (Kuhn et al. 2004).

**RNA Polymerase II C-Terminal Domain**

In addition to transcribing genes, RNAP II also plays critical roles in coordinating co-transcriptional pre-mRNA processing (Hirose et al. 2000) (Bentley 2005) (Buratowski 2003). This function is mediated by its CTD, a unique domain consisting of multiple repeats of the conserved heptapeptide YSPTSPS (McCracken et al. 1997). The number of the repeat ranges from 26 in yeast to 52 in mammals. In *in vitro*, RNAP II CTD is required for cleavage in the absence of transcription (Hirose et al. 2000). Although the exact functions of CTD in pre-mRNA 3′ processing are unclear, it may promote the assembly of a stable 3′ processing complex through its extensive interactions with other 3′ processing factors, including CPSF, CstF, and CFIm (Hirose et al. 2000) (Bentley 2005) (Buratowski 2003) (McCracken et al. 1997) (Sadowski et al. 2003).
1.4 mRNA 3’ Processing Complex Assembly and Reaction

In vivo, pre-mRNA 3’ processing factors are intimately connected with the transcription machinery and the 3’ processing complex assembles on pre-mRNAs co-transcriptionally (Hirose et al. 2000). CPSF joins the transcription machinery as early as in the preinitiation complex stage through interactions with TFIID (Dantonel et al. 1997). CPSF, CstF, and CFIm are all associated with RNAP II during transcription elongation (Hirose et al. 2000) (Bentley 2005). After transcription passes the PAS, RNAP II pauses and perhaps during this time the 3’ processing complex fully assembles on the pre-mRNA to carry out cleavage and polyadenylation (Hirose et al. 2000).

In vitro the pre-mRNA 3’ processing complex assembles in a two-step process (Figure 1.4) (Zhao et al. 1999) (Shi et al. 2009). First, CPSF, CstF and CFIm bind to the key cis-elements within PASes in a mutually stimulatory manner to form a fairly stable complex. Next CFIm and PAP join the core complex transiently to form the complete 3’ processing complex. In the latest model of the 3’ processing complex, CstF and CFIm both function as dimers. Dimerization of these factors may allow for more RNA-protein contacts which may in turn stabilize the 3’ processing complex. It is unclear whether any structural reorganization of the assembled 3’ processing complex takes place before cleavage. As mentioned earlier, the recombinant CPSF 73 has very low endonuclease
Figure 1.4 mRNA 3' Processing Mechanism

The two mRNA 3’ processing steps. The first step is cleavage followed by the second step, polyadenylation.
activity (Mandel et al. 2006). For efficient cleavage, other factor(s) within the 3’ processing complex may directly stimulate CPSF73 activity by changing its conformation and/or by positioning the pre-mRNA for optimal contact with CPSF73.

Little is known about what exactly takes place within the 3’ processing complex after cleavage. Since the polyadenylation step can be reconstituted in vitro with CPSF and PAP, it is possible that CstF, CFIm, and CFIIIm may dissociate from the complex following cleavage along with the downstream RNA (Colgan et al. 1997) (Zhao et al. 1999) (Mandel et al. 2008). These putative compositional and/or conformational changes during the cleavage/polyadenylation transition could be controlled by post-translational modification of 3’ processing factors. For example, the yeast phosphatase Glc7p is required specifically for polyadenylation, but not for cleavage and Pta1p has been identified as a substrate (He et al. 2005). A similar requirement for PP1, the human Glc7 homolog, has been demonstrated for mammalian polyadenylation (Shi et al. 2009). It has been proposed that dephosphorylating one or more 3’ processing factors by PP1 following cleavage may alter the RNA–protein and/or protein–protein interaction network to facilitate polyadenylation (Shi et al. 2009).

Following cleavage, the A(A/U)UAAA-bound CPSF anchors PAP to the pre-mRNA to carry out polyadenylation (Zhao et al. 1999) (Colgan et al. 1997). The dynamics of the 3’ processing complex during polyadenylation has been discussed earlier (see the section on Poly(A)-Binding Proteins). Meanwhile, the downstream RNA generated by cleavage is digested by the exoribonuclease Xrn2 (West et al. 2004) (Kim et al. 2004). Xrn2 is recruited to the 3’ processing complex through interactions with the multifunctional protein p54nrb/PSF and perhaps some basal 3’ processing factors as well (Kaneko et
The degradation of the downstream RNA by Xrn2 is responsible, at least in part, for transcription termination as proposed by the “torpedo” model (West et al. 2004) (Kim et al. 2004). When polyadenylation is completed, the 3′ processing complex needs to be disassembled and mature mRNAs are exported to the cytoplasm. The mechanism for 3′ processing complex disassembly remains poorly understood, but a recent study in yeast provided strong evidence that multiple mRNA export factors are required (Qu et al. 2009). For example, mutations in the mRNA export receptor Mex67 leads to retention of 3′ processing factors on RNAs and inhibition of 3′ processing. For transcripts that do get processed, their poly(A) tails are unusually long. These data indicate that mRNA export factors may play an important role in the disassembly of the 3′ processing complex (Qu et al. 2009). While the disassembly is taking place, some components of the 3′ processing complex remain associated with the polyadenylated mRNAs as part of the export-competent mRNP. For example, a recent study provided evidence that CFIm 68 remains associated with the mRNA throughout 3′ processing and serves as an adaptor protein for export (Ruepp et al. 2009). Similarly, another well-characterized mRNA export adaptor in yeast Yra1 is recruited to mRNA during 3′ processing through direct interaction with the cleavage factor Pcf11 and functions in the subsequent mRNA export (Johnson et al. 2009). Finally, the multifunctional nuclear protein nucleophosmin has been shown to bind mRNA in a polyadenylation-dependent manner, but the functional significance of this association remains unknown (Palaniswamy et al. 2006). The tight coupling between pre-mRNA 3′ processing and the mRNA export may ensure that only fully processed mRNAs are exported.
1.5 mRNA 3’ Processing Regulation

Like other steps of the gene expression pathway, pre-mRNA 3’ processing is subject to regulation (Zhao et al. 1999) (Colgan et al. 1997) (Millevoi et al. 2010). Many examples of 3’ processing regulation have been reported and the different modes of regulation can be classified into the following three general mechanisms (Figure 1.5).

**Regulation of the Basal 3' Processing Factors**

Changes in the protein levels, accessibility, or the activities of the basal 3’ processing factors can influence pre-mRNA 3’ processing and their effects can be global or transcript-specific. First, changes in the protein levels of basal 3’ processing factors can regulate 3’ processing. For example, CstF 64 is specifically up-regulated when primary B cells are induced to differentiate (Takagaki et al. 1996) (Edwalds-Gilbert et al. 1995). The limited CstF complexes in primary B cells preferentially bind to the high affinity distal poly(A) site of the IgM pre-mRNA, resulting in the production of the membrane-bound form of IgM. However, in activated B cells in which CstF is more abundant, 3’ processing shifts to a ‘first-come, first-serve’ mode and a lower affinity proximal poly(A) site is now preferentially recognized, leading to synthesis of the secreted form of IgM (Takagaki et al. 1996) (Edwalds-Gilbert et al. 1995). It is highly likely that many more transcripts are regulated in a similar fashion by the changing levels of CstF64. Additionally the protein levels of CF Im subunits show developmental stage-specific
Alternative polyadenylation mechanisms. A. When more mRNA 3’ processing subunits are present, the weaker/proximal poly (A) site is chosen over the stronger/distal poly (A) site. However, when mRNA 3’ processing subunits are scarce, the stronger poly (A) site would be preferred. B. There are factors from other gene processes that can inhibit or enhance the choice of one poly (A) site over another. C. Transcription rate affects poly (A) choice. When transcription rate is slow, the mRNA 3’ processing will occur on the first poly (A) site. However, when the transcription rate is fast, the distal poly (A) site would be preferred.
variations and may cause changes in 3’ processing efficiency as well as APA patterns (Sartini et al. 2008).

Secondly, changes in the accessibility of the basal 3’ processing factors can regulate 3’ processing. For example, regulatory viral or cellular proteins could sequester the basal 3’ processing factors and prevent them from participating in 3’ processing. As mentioned earlier, the influenza virus NS1 protein specifically interacts with CPSF 30, thereby inhibiting CPSF-RNA interaction and cleavage/polyadenylation of cellular pre-mRNAs (Nemeroff et al. 1998). As mentioned earlier, CstF50 becomes associated with the BARD1/BRCA1 complex following UV-induced DNA damage and this association leads to transient inhibition of pre-mRNA 3’ processing (Kleiman et al. 2001).

Finally, changes in the activity of the basal 3’ processing factors can regulate 3’ processing. Multiple post-translational modifications have been shown to modulate the activities of 3’ processing factors. As mentioned earlier, hyperphosphorylation of PAP during mitosis represses PAP activity and leads to general inhibition of polyadenylation (Colgan et al. 1996). Other modifications could have more profound effects. For example, sumoylation affects the stability, subcellular localization, and activities of PAP (Vethantham et al. 2008). CPSF 73 and symplekin are also sumoylated and sumoylation stimulates the assembly of the 3’ processing complexes (Vethantham et al. 2007).
Regulation of 3’ Processing Complex Assembly by Local RNA Context and Regulatory Factors

First, 3’ processing at a specific PAS can be affected by regulatory factors bound to adjacent cis-elements. For example, 3’ processing is activated by an upstream intron (Niwa et al. 1990). This stimulation is mediated by multiple interactions between 3’ processing factors and splicing factors, such as the CF Im-U2AF and CPSF-U2 snRNP interactions, and their co-operative binding to the PAS and the upstream splicing cis-elements (Millevoi et al. 2006) (Kyburz et al. 2006). Conversely, 3’ processing is strongly inhibited by a 5’ splice site located upstream of a PAS (Gunderson et al. 1998). This is due to inhibition of PAP activity by U1–70K and U1A of the U1 snRNP bound to the upstream 5’ splice site (Gunderson et al. 1998). Another example of this type of regulation is provided by the 3’ processing of the HIV pre-mRNA at the 3’ LTR. The SR protein 9G8, in a complex with CDK11 and eIF3f, binds to a specific sequence upstream of the AAUAAA hexamer. Through its interaction with CFIm bound to an adjacent UGUAN motif, 9G8 stimulates 3’ processing at the downstream PAS. Interestingly, overexpression of eIF3f or just its N-terminal 91 amino acid fragment disrupts the eIF3f-CDK11-9G8 complex and prevents efficient 3’ processing at the 3’ LTR, leading to reduced viral replication (Valente et al. 2009).

Secondly, some regulatory factors could directly compete with basal 3’ processing factors for binding to specific cis-elements within the PAS, thereby inhibiting 3’ processing. For example, both the mammalian PTB (polypyrimidine tract-binding protein) and the Drosophila Sxl (sex-lethal) proteins can compete with CstF64 for binding to the DSE (Castelo-Branco et al. 2004) (Gawande et al. 2006). The outcome of
this competition is determined by the affinities of these factors for the DSE and their relative abundance.

Regulation of Pre-mRNA 3' Processing After Complex Assembly

Although regulation of pre-mRNA 3’ processing most often occurs at the earliest step, i.e., the 3’ processing complex assembly, it should be pointed out that later steps are also subject to regulation. For example, U1A protein regulates its own expression through modulating the 3’ processing of its own transcripts (Gunderson et al. 1994). U1A proteins bind to a specific RNA structure in the 3’ UTR of its own transcripts just upstream of the PAS. The RNA-bound U1A specifically inhibits PAP activity and therefore prevents polyadenylation without affecting 3’ processing complex assembly or cleavage. This forms a negative feedback loop that helps maintain the steady state levels of U1A protein (Gunderson et al. 1994).

1.6 A Look Ahead

As illustrated above, much about mRNA 3’ processing cis elements, trans factors, mRNA 3’ processing assembly and regulation are known. However, there are still some unanswered questions. For example, how does aberrant regulation of mRNA 3’ processing lead to cancer, abnormal growth or improper stem cell differentiation (Mayr et al. 2009) (Ji et al. 2009) (Ji et al. 2009)? While there have been studies done illustrating how one mRNA 3’ processing factor can affect poly (A) site choice, a global
identification of all potential regulators of mRNA 3’ processing is needed (Lackford et al. 2014) (Edwalds-Gilbert et al. 1995) (Takagaki et al. 1996). In Chapter 2, I illustrate the identification of potential regulators of sixteen mRNA 3’ processing factors using mass spectrometry. By identifying potential regulators of mRNA 3’ processing factors, this creates the first interactome of the mRNA 3’ processing complex to help understand how mRNA 3’ processing is regulated.

While understanding the protein-protein interactions are important, understanding the RNA-protein interactions are also crucial for mRNA 3’ processing. One critical question that remains unanswered is, how does CPSF recognize AAUAAA with such high specificity when none of the CPSF subunits have a canonical RRM? Initially, only CPSF 160 has been implicated in AAUAAA recognition (Murthy et al. 1995). However, direct evidence was lacking from previous experiments. In Chapter 3, I have direct evidence that CPSF 30 directly interacts with the AAUAAA and with WDR33. The interaction between CPSF 30 and the AAUAAA was further analyzed. The CPSF 30 domains needed for AAUAAA interaction were mapped, showing that two zinc fingers are needed for AAUAAA recognition. Interestingly, the two zinc fingers needed for AAUAAA recognition are also targeted by the influenza A viral protein, NS1A, to suppress host mRNA 3’ processing (Nemeroff et al. 1998) (Twu et al. 2006). While the CPSF can recognize AAUAAA specifically via the help of CPSF 30 and WDR33, the question of whether or not the CPSF-AAUAAA interaction can be regulated remained. In Chapter 4, I show that when cis elements other than AAUAAA and G/U rich (auxiliary sequences) are present, CPSF can no longer sufficiently recognize AAUAAA. Rather, CPSF now
needs to interact with a factor present in HeLa nuclear extract to recognize AAUAAA specifically. While the identity of the HeLa factor remains unknown, several fractionation steps were taken in attempt to isolate this unknown factor. Once the unknown HeLa factor has been identified, it will provide insight into how the CPSF-AAUAAA interaction can be regulated.

By addressing the above specific questions about RNA-protein and protein-protein interactions, this manuscript provides a more in depth understanding about poly (A) site recognition and regulation. This knowledge can ultimately help us understand both mRNA 3’ processing mechanism and regulation.
Chapter 2

Interactome Map of the mRNA 3’ Processing Machinery

2.1 Summary

mRNA 3’ processing machinery is a dynamic complex that consists of many proteins outside its core complexes (Colgan et al. 1997) (Zhao et al. 1999) (Chan et al. 2010) (Shi et al. 2009). In this chapter, the regulatory functions of these peripheral proteins with respect to mRNA 3’ processing are mainly discussed. It is important to understand the functions of mRNA 3’ processing regulators because they can affect other cellular processes, mRNA 3’ processing mechanism and APA.
The regulatory proteins can affect mRNA 3’ processing through three general mechanisms. First, the regulatory proteins can regulate the basal levels of mRNA 3’ processing factors by changing its localization, stability, accessibility and activity. Second, the regulatory proteins can affect the assembly of the mRNA 3’ processing complex. And finally, the regulatory proteins can regulate mRNA 3’ processing complex after the complex has assembled. While it is known how a few mRNA 3’ processing trans factors are affected by these regulators, the central hypothesis of this chapter is that the mRNA 3’ processing complex can be regulated by other undiscovered proteins. The discovery of new regulatory proteins are important because they can affect mRNA 3’ processing mechanism and offer insights into the APA mechanism. In addition, new cellular pathways can be linked to mRNA 3’ processing. A global interactome map of the mRNA 3’ processing machinery is therefore needed to fully understand how mRNA 3’ processing is regulated. Recently, eighty-five proteins were purified with the intact and functional mRNA 3’ processing complex (Shi et al. 2009). Although new factors were found to associate with the mRNA 3’ processing machinery, the procedure did not identify a complete list of mRNA 3’ processing regulators. Therefore, the goal of this project is to produce a more complete global interactome map of the mRNA 3’ processing machinery. To accomplish this goal, I have designed the following specific aims:

Specific aim 1: Systemically identify the protein partners of each essential mRNA 3’ processing factor by immunoprecipitation and mass spectrometry and generate a comprehensive interactome map of the mRNA 3’ processing machinery.

Specific aim 2: Functionally characterize the novel interacting proteins identified in
regulating mRNA 3’ processing.

To accomplish these goals, I have immunoprecipitated all sixteen known essential mRNA 3’ processing factors and have identified their protein partners using high throughput mass spectrometry analyses. An interactome map of the mRNA 3’ processing machinery was then generated and analyzed. It was then identified via the analyses that each of the mRNA 3’ processing factors immunoprecipitated all of the subunits from the same complex. In addition, it was found that the mRNA 3’ processing complex can interact with known cellular processes such as transcription, splicing, transport, ubiquitination and the DNA damage pathway (Zhao et al. 1999) (Ingham et al. 2005) (Kleiman et al. 2001). Together, this demonstrates that this method is reliable and can produce biologically relevant information.

Novel information was also identified using this method. Specifically, novel ubiquitin E3 ligases were found to associate with each of the CFIm subunits. These interactions were verified using in vitro assays. The function of the E3 ligases was then characterized it was discovered that the E3 ligases were able to polyubiquitinate CFIm subunits. Overall, the mRNA 3’ processing machinery interactome provided new information in how mRNA 3’ processing can be regulated.

2.2 Introduction

mRNA 3’ processing requires extensive RNA:protein and protein:protein interactions for the reaction to occur (Zhao et al. 1999) (Chan et al. 2010) (Mandel et al. 2008). In
general, there are two groups of proteins that are involved in the mRNA 3' processing machinery. The first group of protein consists of the core mRNA 3' processing machinery. This consists of four multi-subunit complexes: CPSF, CstF, CFIm and CFIm. The second group of protein consists of the proteins that regulate mRNA 3' processing. This chapter focuses on those mRNA 3' processing regulators.

mRNA 3' processing regulators are important to study because they can regulate other cellular processes and can affect mRNA 3' processing and even APA. It is known that mRNA 3' processing interacts with other cellular processes such as transcription and splicing (Zhao et al. 1999) (Chan et al. 2010) (Colgan et al. 1997). For example, during transcription initiation, the CPSF complex binds to the CTD of RNA pol II (Dantonel et al. 1997). The entire CstF complex is also recruited to the CTD of RNA pol II during transcription elongation via CstF 50 (McCracken et al. 1997). mRNA 3' processing is also connected to splicing. For example, CFIm 59 and CFIm 68 contain RS domains that can interact with SR proteins crucial for splicing regulation (Zhong et al. 2009). In addition, the CFIm complex has been shown to be part of the spliceosome (Zhou et al. 2002). Although the mRNA 3' processing complex can interact with other cellular processes, mRNA 3' processing regulators can also affect the activity of these processes. For example, CstF 64 can increase transcription efficiency by binding to the transcription coactivator PC4 (Hockert et al. 2010) (Qu et al. 2007) (Calvo et al. 2001). Another example is that CDK9/cyclinT can promote the interaction of Pcf11 with the CTD of RNA pol II with its CID in order for transcription termination to occur (Proudfoot
These examples demonstrate that mRNA 3’ processing regulators can influence the activity of other cellular processes.

In addition to affecting other cellular processes, mRNA 3’ processing regulators are also important for mRNA 3’ processing. Specifically, regulators can regulate the activities, localization and accessibilities of basal mRNA 3’ processing factors. For example, regulators can decrease mRNA 3’ processing activities. The influenza viral protein, NS1A, has been shown to bind to CPSF 30 and PABPN1. By binding to these essential mRNA 3’ processing factors, NS1A blocks their interaction with the RNA and with other mRNA 3’ processing proteins. As a consequence, both cleavage and polyadenylation steps are inhibited (Chen et al. 1999) (Nemeroff et al. 1998). The activities of mRNA 3’ processing can also be affected by post-translational modifications. For example, CPSF 73 and symplekin sumolyation increases mRNA 3’ processing activity by promoting mRNA 3’ processing assembly (Vethantham et al. 2007). Another example is that during mitosis, PAP is hyperphosphorylated by cyclin B. As a consequence, PAP activity is strongly inhibited and the efficiency of mRNA 3’ processing is severely diminished (Colgan et al. 1996). In addition to phosphorylation, PAP can be sumoylated and acetylated. When PAP is sumoylated, its cellular activity decreases. In addition, the localization of PAP is altered with this modification. Specifically, PAP is localized to the cytoplasm with its sumoylation (Vethantham et al. 2008). Acetylation of PAP promotes its relocalization to the cytoplasm by inhibiting the binding of PAP to importin alpha/beta complexes (Shimazu et al. 2007). These examples demonstrate that regulators can
alter mRNA 3’ processing by affecting the activities and localizations of mRNA 3’
processing factors.

APA can also be influenced by mRNA 3’ processing regulators. APA affects mRNA
expression for 70% of human genes, producing multiple mRNA isoforms with different 3’
processing sites (Derti et al. 2012) (Hoque et al. 2013). These mRNA isoforms may
encode different proteins or have different 3’ untranslated regions. As a consequence,
APA affects many critical process such as stem cell differentiation, normal cellular
growth and disease development (Ji et al. 2009) (Lackford et al. 2014) (Shepard et al.
mechanism is largely unknown, a number of models have been proposed (Shi 2012).
One hypothesis is that if mRNA 3’ processing factors are limited, then the poly (A) site
containing the strongest match to a canonical cis element would be preferentially
chosen over a poly (A) site with a noncanonical cis element sequence. The second
hypothesis is that proteins from other gene processes can block or stimulate the binding
of mRNA 3’ processing factors to the PASs. By blocking or stimulating the activities of
these essential factors, the choice of PAS can shift from proximal to distal or from distal
to proximal. The third hypothesis is that the transcription rate determines the PAS used.
RNA-seq data from human and mouse tissues suggest that with higher transcription
rates, the proximal PASs are preferentially chosen over the distal PASs. Specifically, it
is suggested that transcription activators can promote the recruitment of mRNA 3’
processing factors, which can then promote the usage of the proximal PASs. Finally, the
fourth model is that the proximal site would be preferentially chosen over the distal site
because the proximal site is transcribed before the distal site. In *Drosophila polo*, it was demonstrated that the regulation of RNA pol II transcription rate has an affect on PAS choice. Specifically, the proximal PAS was preferred over the distal when the RNA pol II has a slower transcription rate. In contrast, when wildtype RNA pol II was used, the proximal PAS did not have an advantage (Pinto et al. 2011). These APA hypotheses illustrate how regulators can influence proteins involved in PAS choice. In addition, regulators can also affect PAS choice by affecting mRNA 3’ processing factors. For example, when CstF 64 is limited in undifferentiated B cells, the high affinity distal PAS of the IgM pre-mRNA is preferentially chosen. As a consequence, the membrane bound form of IgM is produced. However, during B cells differentiation, ample amounts of CstF 64 are present. As an affect, the proximal PAS is preferentially recognized, leading to the production of secreted IgM (Edwalds-Gilbert et al. 1995) (Takagaki et al. 1996). Similarly, the decreased expression of CFIm subunits leads to APA (Martin et al. 2012) (Kubo et al. 2006) (Kim et al. 2010) (Sartini et al. 2008). Specifically, decreased expression of CFIm 25 or CFIm 68 using siRNA can change the poly (A) site choice from distal to proximal because CFIm 25 and CFIm 68 are RNA-binding proteins (Kubo et al. 2006) (Kim et al. 2010) (Martin et al. 2012) (Ruegsegger et al. 1996). Specifically, a SELEX experiment demonstrated that CFIm25 and CFIm 68 prefer UGUAN sequences (Yang et al. 2010) (Brown et al. 2003). This allows the mRNA 3’ processing complex to recognize canonical cis elements or it can recruit the mRNA 3’ processing machinery to the RNA when the canonical AAUAAA is missing (Venkataraman et al. 2005) (Ruegsegger et al. 1996). In addition, it has been found that the CFIm complex can autoregulate CFIm 25 and CFIm 68 gene expression in male germ cells (Sartini et
al. 2008). However, it is unknown what regulates the protein expressions of CFIm 25 and CFIm 68. One possibility is that the ubiquitination pathway signals the change in protein expressions. Previously, CFIm 68 complex has been shown to interact with the Itch ubiquitin E3 ligase directly while CFIm 25 interacted with the same ligase indirectly (Ingham et al. 2005). Ubiquitin E3 ligases are crucial for ubiquitination to occur. Specifically, glycine 76 of ubiquitin is covalently bound to its targeted substrate in either a mono, multi or poly ubiquitinated manner. Depending on the ubiquitination pattern, different functions could occur. For example, ubiquitination can mediate proteasomal degradation or degradation independent pathways such as protein:protein interaction, internalization and cellular relocalization (Bernassola et al. 2008). Three crucial enzymes are involved in ubiquitination: E1, E2 and E3 ligases. Although there are hundreds of substrates prone to ubiquitination, in mammals, there are only 1 E1 ligase and 40 E2 ligases (Metzger et al. 2012). The specificity in the ubiquitin chain type and the substrate lies with the E3 ligases. In mammals, there are over 600 E3 ligases. The 600 E3 ligases can be divided into 3 families: homologous to the E6AP carboxyl terminus (HECT) family, really interesting new gene (RING) family and the U-box family (Li et al. 2008). CFIm 68 and 25 were found to associate with Itch from the HECT family. The HECT family is the smallest E3 ligase family containing only 30 members. Specifically, Itch is part of a subfamily within the HECT family that contains distinct structural domains. This subfamily typically contains a C2 domain, a HECT domain and 2-4 WW domains (Bernassola et al. 2008). Each of these domains has their own function. For example, the C2 domain is needed to interact with the intracellular membrane, the HECT domain is needed to transfer ubiquitin from E3 ligase to itself,
and the WW domains are needed to interact with adaptor proteins or target substrates. It remains to be discovered if the protein expression levels of CFIm 25 and CFIm 68 are regulated by the ubiquitination pathway.

The discovery of new regulatory proteins are important because they can affect other cellular processes, mRNA 3' processing mechanism and even APA. While it is known how a few mRNA 3' processing trans factors are affected by these regulators, the central hypothesis of this chapter is that the mRNA 3' processing complex can be regulated by other undiscovered proteins. A global interactome map of the mRNA 3' processing machinery is therefore needed to fully understand how mRNA 3' processing is regulated. Recently, eighty-five proteins were purified with the intact and functional mRNA 3' processing complex (Shi et al. 2009). Although a potential new regulator (PP1) was found to associate with the mRNA 3' processing machinery, there were limitations to the method used. First, it is unknown which mRNA 3' processing factor PP1 directly binds to. Second, the mRNA 3' processing machinery was isolated when only minimal cleavage or polyadenylation has occurred. Therefore the regulators of the cleavage and the polyadenylation steps were not identified. Third, the regulators identified were biased towards factors that positively regulated mRNA 3' processing because the mRNA 3' processing machinery was purified in its functional state. The negative regulators were consequentially not identified. Therefore, a more comprehensive interactome of mRNA 3' processing machinery is needed.

In this chapter, I systematically identified many potential mRNA 3' processing regulators. Specifically, sixteen of the major mRNA 3' processing trans factors were
immunoprecipitated. The protein partners of each major mRNA 3’ processing trans factors were then identified from mass spectrometry analyses. The interactions between each protein were then analyzed and mapped using bioinformatics. Using this method, the major and minor partners of the sixteen mRNA 3’ processing subunits were identified on a global scale. After discovering potential new mRNA 3’ processing regulators, some interactions were verified using *in vitro* assays. The functional significance of some of the interactions was then characterized. One interaction verified was the interaction between the CFIm complex and three ubiquitin E3 ligases from the same subfamily in the HECT E3 ligase class. Specifically, each of the CFIm subunits was verified to directly interact with the ubiquitin E3 ligases. The functional significance of this interaction was then characterized. Together, this chapter provides the first comprehensive interactome of the mRNA 3’ processing machinery that can significantly help us understand how regulators affect other cellular processes, mRNA 3’ processing mechanism and APA.

### 2.3 Materials and Methods

**Generating Stable Cell Lines by Overexpressing mRNA 3’ Processing Factors**

The coding sequences of CPSF 30 (NM_001081559.1), CPSF 73 (NM_016207.3), CPSF 160 (NM_013291.2), CstF 50 (NM_001033521.1), CstF 64 (NM_001325.2), CstF 64 Tau (NM_015235.2), CstF77 (NM_001326.2), CFIm 25 (NM_007006.2), CFIm 59 (NM_001136040.2), CLP1 (NM_006831.2) and Pcf11 (NM_015885.3) were cloned into FLAG-pCDNA3.1 or pCMV14-3x FLAG mammalian expression vectors. These plasmids
were then transfected into Human Embryonic Kidney 293 cells using Lipofectamine and stable transfectants were selected with neomycin/G418 (1 mg/ml) for two weeks. After selection, individual clones were selected and expanded. Western blotting analyses were then carried out with each clone to monitor the expression levels of the FLAG-tagged exogenous proteins. The clones were all screened and those that expressed the FLAG-tagged exogenous proteins near the endogenous protein levels were selected for future experiments.

**Recombinant Proteins Expression and Purification**

*Recombinant CFIm Complexes*

Recombinant CFIm complexes containing CFIm 25-59 or CFIm 25-68 were expressed using the MultiBac expression system (Berger et al. 2004) (Fitzgerald et al. 2006). MultiBac is a protein expression system in insect cells that easily allows the production of eukaryotic multi-protein complexes. What is unique about this MultiBac system is that it contains novel baculovirus transfer vectors with two important features. One feature is its multiplication module that allows for infinite gene insertions without an assembly of restriction sites. The second important property is that the transfer vectors allow improved protein expression by removal of specific viral genes. After expression of the complexes in insect cells, each recombinant complex was purified using cobalt beads. After the insect cells were sonicated, the cobalt beads were incubated with the whole-cell extract for 2 hours at 4°C with gentle rotation. The beads were then washed with the wash buffer (50 mM Tris HCl pH = 8, 0.02% NP40, 10% glycerol, 0.5 mM DTT, 0.5 mM PMSF, 500 mM NaCl) containing 5 mM imidazole. The next wash buffer contained the
same composition except it had 15 mM imidazole. Finally, the bound protein was eluted in the wash buffer containing 200 mM imidazole. The eluate was then visualized using Coomassie staining or dialyzed in Buffer D 100 (20 mM Hepes pH = 7.9, 100 mM KCl, 1 mM MgCl2, 0.2 mM EDTA, 10% glycerol, 10 mM beta mercaptoethanol).

Recombinant WWP2 Proteins
The coding sequences of full-length WWP2 and its WW domain were cloned into pGEX4T3 plasmids. The respective proteins were then expressed using *E. coli*. The transformed *E. coli* was first grown at 37°C until O.D 600 = 0.5, The protein expression was then induced by adding 0.1 mM IPTG for 12 hours at 15°C. The GST-fusion proteins were then purified with glutathione-conjugated beads following manufacturer's instructions (GE Healthcare). The eluate was then dialyzed in Buffer D 100.

Immunoprecipitation

*Endogenous mRNA 3' Processing Factors Immunoprecipitation*
To immunoprecipitate endogenous proteins, Protein A/G Agarose beads (Pierce) were covalently conjugated to the antibodies against specific target proteins (Fip1 (Bethyl. Catalog Number: A301-462A), CPSF 100 (Bethyl. Catalog Number: A301581A), Symplekin (Bethyl. A301465A), WDR33 (Bethyl. Catalog Number: BL4833) or CFIm 68 (Bethyl. A3001358A). To covalently link the antibody to the A/G beads, 10 μg of each antibody was incubated with the beads for 1 hour in room temperature with gentle rotation. The beads were then washed 3 times with 0.2 M sodium borate pH = 9. The
third wash was performed at room temperature for 30 minutes in the presence of 20 mM dimethyl pimelimidate (DMP). The beads were then washed with 0.2 M ethanolamine pH = 8 for 2 hours. The final two washes were with Buffer D 300 (20 mM Hepes pH = 7.9, 300 mM NaCl, 1 mM MgCl2, 0.2 mM EDTA, 10 mM beta mercaptoethanol) and 0.1% NP40. One milliliter of HeLa nuclear extract was then spun at 14,000 rpm for 1 minute at 4°C. The supernatant was then incubated with the beads for 2 hours at 4°C. After incubation, the supernatant was discarded and the beads were washed with Buffer D 300 + 0.1% NP40, 3 times, 10 minutes each at 4°C. The beads were then washed once with Buffer D 100 for 10 minutes at 4°C. The proteins were eluted off the beads with 0.2 M glycine pH = 3.5, 3 times, 2 minutes each at 4°C. The entire immunoprecipitation procedure was repeated two more times, utilizing 3 ml of HeLa nuclear extract total. The eluate was then gathered and precipitated. A small portion of the eluate was used for silver staining while the majority was analyzed by mass spectrometry.

**FLAG Immunoprecipitation**

For the stable cell lines overexpressing FLAG-tagged 3’ processing subunits (CPSF 73, CPSF 160, CstF 50, CstF 64, CstF 64 Tau, CstF 77, CFIm 25, CFIm 59, CLP1 and Pcf11), the FLAG immunoprecipitation protocol was used. Five to ten 15-cm plates of each respective cell line were spun down at 1.5 krpm for 5 minutes at 4°C and were resuspended with 5x the pellet volume with ice-cold Buffer A (10 mM Hepes pH = 7.9, 10 mM KCl, 1.5 mM MgCl2, 10 mM beta mercaptoethanol). Each sample was then incubated on ice for 10 min. After addition of NP40 to a final concentration of 0.5%, the
sample was spun at 4 krpm for 10 minutes at 4°C. The cytoplasmic fraction was discarded while the nuclei was resuspended with 1.5x the pellet volume with Buffer C (20 mM Hepes pH = 7.9, 420 mM NaCl, 1.5 mM MgCl2, 0.2 mM EDTA, 25 % glycerol, 10 mM beta mercaptoethanol, protease inhibitor cocktail). The nuclei were then homogenized with an 18 g needle before they were rotated for 30 minutes at 4°C. The sample was then spun at 14,000 rpm for 15 minutes at 4°C. The supernatant (nuclear extract) was then mixed with Anti-FLAG M2 Affinity beads (Sigma) for 2 hours at 4°C with gentle rotation. After incubation, the beads were spun down and the nuclear extract was removed. The beads were then washed at 4°C with Buffer D 300 + 0.1% NP40, 3x 10 minutes each and with Buffer D 100, 1x for 10 minutes. The proteins were eluted with Buffer D 100 containing 3x FLAG peptide. The elution was precipitated using acetone. A small portion of the elution was used in silver staining while the majority was analyzed by mass spectrometry.

For the FLAG-tagged CPSF 30 stable cell line, the same FLAG immunoprecipitation protocol was utilized but the proteins were immunoprecipitated from whole-cell lysate instead of from nuclear extract. The whole-cell lysate was prepared by first centrifuging at 1.5 krpm for 5 minutes at 4°C. The supernatant was discarded and the cells were resuspended in Buffer D 300 (without NP40) and sonicated. After sonication, protease inhibitor cocktail (1x) and 0.1% NP40 were added and the sample was rotated for 30 minutes at 4°C. The sample was then spun down for 10 minutes at 4°C. The whole-cell lysate was then used for FLAG immunoprecipitation as described above.
High Throughput Proteomic Mass Spectrometry

The immunoprecipitates were submitted to the Yates Lab at The Scripps Research Institute in San Diego, CA for high throughput liquid chromatography and tandem mass spectrometry (LC/MS/MS) (Motoyama et al. 2008) (Carucci et al. 2002). The immunoprecipitates were first digested with proteases to generate peptide fragments. The peptides were then separated by LC followed by fragmentation in a tandem mass spectrometer. First, the mass:charge ratios of the peptides were measured. Then the mass:charge ratios of the daughter peptides were measured. The fragmentation patterns were then used to search the genome databases for protein sequences that matched the data.

Bioinformatic Analysis of the Mass Spectrometry Results

To analyze the mass spectrometry data, a program called “significance analysis of interactome” (SAINT) was used (Choi et al. 2011). SAINT is a program that allows probabilistic scoring of affinity purification-mass spectrometry data to derive bona fide protein-protein interactions. First, the mass spectrometry data was formatted as required by the SAINT program. Then the strengths between the bait and the prey were measured using the sequence coverage from the mass spectrometry results. Cytoscape was then used to visualize the interactions obtained from SAINT. All of the interactions mapped with Cytoscape had a probability of 0.95 and sequence coverages greater than 10%. After the Cytoscape analyses, all of the proteins identified were compared to the negative control (FLAG immunoprecipitation). The proteins commonly identified in the
control FLAG immunoprecipitate were then filtered, preserving only the proteins that unique interacted with each mRNA 3’ processing subunit.

**CFIm – WWP2 Interaction Assays**

*FLAG CFIm 25 Immunoprecipitation*

Whole-cell lysate was generated from two 15-cm plates of FLAG CFIm 25 stable cell line. The whole-cell lysate from FLAG-CFIm 25 stable cell line was then immunoprecipitated with Anti-FLAG M2 Affinity beads similar to the protocol for the FLAG-CPSF 30 stable cell line described above. Differently, here the eluate was resolved on a 10% SDS-PAGE gel and analyzed by Western blotting analyses.

*Endogenous CFIm 25 Immunoprecipitation*

To determine if endogenous CFIm complex interacts with WWP2, 4 µg of actin or CFIm 25 (Santa Cruz, Catalog number: 2203C3) antibody was covalently conjugated to Protein A/G beads as described above. Four 15-cm plates of HeLa cells were then spun down at 1.5 krpm for 5 minutes. The supernatant was discarded. Buffer 200 (20 mM Hepes pH = 7.9, 200 mM NaCl, 1 mM MgCl2, 0.2 mM EDTA, 10 mM beta mercaptoethanol) was then added to the cells with protease inhibitor cocktail (1x). The cells were then sonicated. After sonication, NP40 was added to a final concentration of 0.1%. The sample was then rotated in 4°C for 15 minutes before spun down at 14,000 rpm for 10 minutes. The supernatant was then incubated with the actin or CFIm 25 beads for 12 hours with gentle rotation at 4°C. The beads were then spun down and the
supernatant was discarded. The beads were washed with Buffer 200 with 0.1% NP40 for 3x, 10 minutes each at 4°C. Finally, the beads were washed with Buffer D 100 for 1x, 10 minutes each. The proteins were eluted from the beads with glycine and precipitated with acetone. Finally the eluate was resolved on a 10% SDS-PAGE gel and analyzed by Western blotting analyses.

Protein Pulldown Assays with Recombinant CFIm

GST WWP2 and GST fused to the WW domain were incubated with recombinant CFIm complexes to determine if they interacted with the E3 ligase directly. Four microgram of the GST proteins were first bound to glutathione-conjugated beads for 30 minutes at room temperature. After binding, the supernatant was discarded and the beads were washed 3x with Buffer D 300. The final wash was discarded. Five percent of filtered BSA was then incubated with the beads for 1 hour at 4°C. After incubation, the filtered BSA was discarded and the beads were washed with Buffer D 300. The final wash was discarded. Half a microgram of each respective recombinant CFIm complexes were then incubated with the glutathione-conjugated beads for 2 hours at 4°C. After incubation, the supernatant was discarded and the beads were washed with Buffer D 300 + 0.1% NP40 for three times and then one time with Buffer D 100. Each wash was 10 minutes. The final wash was discarded. The protein was then eluted with 2x SDS loading buffer and Western blotting analyses were performed.
Pulldown Assays with In vitro Translated Proteins

Two microgram of FLAG cDNA3.1 plasmids containing the coding sequences of FLAG CFIm 25, 59 or 68 were translated in vitro using the TNT T7 Quick Coupled Transcription/Translation System (Promega) in the presence of 134 micro curies of L-[\(^{35}\)S] methionine. After in vitro translation, the radioactive in vitro translated proteins were incubated with the glutathione-conjugated beads bound to 4 µg of GST WWP2 for 12 hours at 4°C. After incubation, the supernatant was discarded. The beads were then washed with Buffer D 300 +0.1% NP40 3x and then once with Buffer D 100. Each wash was 10 minutes. After the washes, the final wash was discarded. The sample was then eluted with SDS loading buffer and loaded onto a 10% SDS PAGE gel. The radioactive in vitro translated proteins were imaged using film.

In Vitro Ubiquitination Assay

Ten microgram of each plasmid (His Ubiquitin pQCXIP, FLAG CFIm 59 pcDNA3.1, WWP2 pCDNA3, Myc Itch pRK5, Myc Itch Mutant (C830A) pRK5) were transfected into 293T cells using the calcium phosphate method (Kingston et al. 2001). Forty-eight hours after transfection, 50 µM of Mg132 was added to the cells. The cells were then incubated for another hour at 37°C before the cells were harvested. The cells were spun down and resuspended in Buffer 1 (8 M Urea, 300 mM NaCl, 0.5% NP40, 50 mM sodium phosphate, 50 mM Tris pH = 8, 1 mM PMSF). The cells were then sonicated. After sonication, the samples were spun down at 14,000 rpm for 10 minutes at room temperature. The whole-cell lysate was then incubated with cobalt beads for 6 hours in room temperature. After binding, the whole-cell lysate was discarded. The beads were
washed 3x, 10 minutes each with the same Buffer 1 as above except with 10 mM imidazole. The last wash was Buffer 1 containing 10 mM imidazole similar to above except it did not contain 8 M urea. The proteins were then eluted with 2x SDS loading buffer. Finally, the samples were resolved on a 10% SDS PAGE gel and analyzed by Western blotting analyses.

2.4 Results

Purification of mRNA 3’ Processing Subunits

The mRNA 3’ processing interactome was mapped to answer outstanding questions about the APA mechanism. To generate the interactome, I immunoprecipitated sixteen known essential mRNA 3’ processing subunits from mammalian cell lysates and identified their binding partners with mass spectrometry (Figure 2.1). For eleven mRNA 3’ processing subunits (CPSF 30, CPSF 73, CPSF 160, CstF 50, CstF 64, CstF 64 τ, CstF 77, CFIm 25, CFIm 59, CLP1, and Pcf11), their coding sequences were cloned into FLAG-pCDNA3.1 or pCMV14-3x FLAG mammalian expression vectors. Stable cell lines in Human Embryonic Kidney 293 cell lines overexpressing each FLAG-tagged subunit were then generated. Nuclear extract was prepared from all of the stable cell lines except for the stable cell line overexpressing FLAG CPSF 30; whole-cell lysate was generated from the FLAG CPSF 30 stable cell line. The nuclear extracts and the whole-cell lysate were used in immunoprecipitation with anti-FLAG antibodies. For the other five major mRNA 3’ processing subunits (Fip1, CPSF 100, Symplekin, WDR33 and CFIm 68), the respective endogenous proteins were immunoprecipitated from HeLa nuclear extract using protein A/G beads. The protein A/G beads were covalently-
Figure 2.1 Immunoprecipitation Method

Depicted here is the method of which the mass spectrometry samples were generated. Each of the endogenous or FLAG-tagged mRNA 3' processing subunit was immunoprecipitated from HeLa or 293 nuclear or whole cell extract. The immunoprecipitate were then sent to mass spectrometry to identify its protein partners.
conjugated to the respective antibodies. Using this sensitive method, I was able to immunoprecipitate the targeted subunit, other mRNA 3’ processing factors from the same multi-subunit complex, proteins from other mRNA 3’ processing mutli-subunit complexes and even proteins from other gene regulatory processes (Figure 2.2, Figure 2.3).

**Verifying the Immunoprecipitation-Mass Spectrometry Procedure**

To test the validity of this method, the mass spectrometry results were first used to determine if factors from transcription, splicing and translation pathways were immunoprecipitated with mRNA 3’ processing factors (Zhao et al. 1999) (Chan et al. 2010). Indeed each of the mRNA 3’ processing complexes were able to immunoprecipitate factors from other cellular processes (Table 2.1). In addition, each mRNA 3’ processing factor can immunoprecipitate subunits from the same complex as expected. As shown in Tables 2.2-2.5, CPSF, CstF, CFIm and CFIIm subunits can interact with not only all of the subunits within the same complex, but also interact with subunits from other complexes. For example, FLAG CPSF 73 can immunoprecipitate the other six core members of the CPSF complex: CPSF 160, WDR33, Symplekin, CPSF 100, hFip1 and CPSF 30. In addition, it can also immunoprecipitate CstF subunits. This agrees with previous understanding that the CPSF complex interacts with the CstF complex (Table 2.2) (Murthy et al. 1992) (Takagaki et al. 2000). (Kaufmann et al. 2004) (Murthy et al. 1995). The reciprocate is also true. FLAG CstF 77 mass spectrometry results shows that the CstF complex can immunoprecipitate CPSF subunits. In addition, CstF 77 can immunoprecipitate all of the other three CstF
The four mRNA 3' processing multi-subunit complexes are shown on the left of each red box. Under the name of each complex are representative silver stains of each complex. The red arrows in each silver stain point to each of the subunits within those complexes. Within the red boxes on the right are the IP products of each of the subunits within the corresponding complexes. Each IP product was visualized by silver stain.
Figure 2.3 Summary of the Interactome Network

The mass spectrometry data was analyzed using SAINT software and the data was visualized using Cytoscape (see materials and methods).

A. Interactome Map depicting the interactions between how each multi subunit complex in mRNA 3’ processing interact with each other. The red circle highlights each complex while each of the subunits within the complexes are represented by a green circle. The tighter the interaction between the subunits found via mass spectrometry, the darker the line connecting them.

B. Interactome Map depicting the interactions between how each mRNA 3’ processing factor interacts with other gene regulatory pathways. Each green circle represents a mRNA 3’ processing subunit while each pink circle represents a protein from another gene regulatory pathway. The tighter the interaction between the proteins found via mass spectrometry, the darker the line connecting them.
Table 2.1 mRNA 3' Processing Machinery Interactome (with other cellular processes)

Partial list of the how each of the complexes interact with proteins from other cellular processes as identified from mass spectrometry.

<table>
<thead>
<tr>
<th>mRNA 3' processing factor</th>
<th>Splicing</th>
<th>Translation</th>
<th>Sequence count</th>
<th>Sequence coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>PABPN1</td>
<td>2</td>
<td>22%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spen b subunit 9</td>
<td>2</td>
<td>10.7%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spen C</td>
<td>4</td>
<td>7.4%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RNKExoribonuclease II</td>
<td>2</td>
<td>3.6%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Splicing factor Prp8</td>
<td>3</td>
<td>10.0%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>U1 small nuclear RNA-polymerase II</td>
<td>3</td>
<td>6.0%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Splicing factor U1</td>
<td>2</td>
<td>2.0%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Splicing factor U2</td>
<td>2</td>
<td>2.0%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Splicing factor U2, interacting protein</td>
<td>2</td>
<td>2.0%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>U3A poly(A) polymerase</td>
<td>2</td>
<td>2.0%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ku-86</td>
<td>2</td>
<td>2.0%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ku-86</td>
<td>2</td>
<td>2.0%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ku-86</td>
<td>2</td>
<td>2.0%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ku-86</td>
<td>2</td>
<td>2.0%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Splicing factor 3' spliceosome**

- Splicing factor Prp8
- U1 small nuclear RNA-polymerase II
- Splicing factor U1
- Splicing factor U2
- Splicing factor U2, interacting protein
- U3A poly(A) polymerase
- Ku-86
- Ku-86
- Ku-86
- Ku-86

**Ku-86**

- Splicing factor Prp8
- U1 small nuclear RNA-polymerase II
- Splicing factor U1
- Splicing factor U2
- Splicing factor U2, interacting protein
- U3A poly(A) polymerase
- Ku-86
- Ku-86
- Ku-86
- Ku-86

**Ku-86**

- Splicing factor Prp8
- U1 small nuclear RNA-polymerase II
- Splicing factor U1
- Splicing factor U2
- Splicing factor U2, interacting protein
- U3A poly(A) polymerase
- Ku-86
- Ku-86
- Ku-86
- Ku-86

**Ku-86**

- Splicing factor Prp8
- U1 small nuclear RNA-polymerase II
- Splicing factor U1
- Splicing factor U2
- Splicing factor U2, interacting protein
- U3A poly(A) polymerase
- Ku-86
- Ku-86
- Ku-86
- Ku-86

**Ku-86**

- Splicing factor Prp8
- U1 small nuclear RNA-polymerase II
- Splicing factor U1
- Splicing factor U2
- Splicing factor U2, interacting protein
- U3A poly(A) polymerase
- Ku-86
- Ku-86
- Ku-86
- Ku-86

**Ku-86**

- Splicing factor Prp8
- U1 small nuclear RNA-polymerase II
- Splicing factor U1
- Splicing factor U2
- Splicing factor U2, interacting protein
- U3A poly(A) polymerase
- Ku-86
- Ku-86
- Ku-86
- Ku-86

**Ku-86**

- Splicing factor Prp8
- U1 small nuclear RNA-polymerase II
- Splicing factor U1
- Splicing factor U2
- Splicing factor U2, interacting protein
- U3A poly(A) polymerase
- Ku-86
- Ku-86
- Ku-86
- Ku-86
Table 2.2 FLAG CPSF 73 Mass Spectrometry Results

Shown here is a chart of some proteins IPed with FLAG CPSF 73 from HEK293 cells along with the sequence count and the sequence coverage. The red color highlights subunits from the CPSF complex while the green highlights the factors from the CstF complex.
Table 2.3 FLAG CstF 77 (HEK 293)

<table>
<thead>
<tr>
<th>Name</th>
<th>Sequence Count</th>
<th>Sequence Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>CstF 50</td>
<td>48</td>
<td>73.5%</td>
</tr>
<tr>
<td>CstF 77</td>
<td>118</td>
<td>62.2%</td>
</tr>
<tr>
<td>CstF 64 tau</td>
<td>50</td>
<td>56.3%</td>
</tr>
<tr>
<td>CstF 64</td>
<td>45</td>
<td>50.6%</td>
</tr>
<tr>
<td>CPSF 100</td>
<td>24</td>
<td>38%</td>
</tr>
<tr>
<td>Fip-like 1</td>
<td>13</td>
<td>30.8%</td>
</tr>
<tr>
<td>CPSF 160</td>
<td>22</td>
<td>25.2%</td>
</tr>
<tr>
<td>CPSF 73</td>
<td>9</td>
<td>21.1%</td>
</tr>
<tr>
<td>CPSF 30</td>
<td>2</td>
<td>16.4%</td>
</tr>
<tr>
<td>WDR33</td>
<td>10</td>
<td>14.4%</td>
</tr>
</tbody>
</table>

Table 2.3 FLAG CstF 77 Mass Spectrometry Results

Shown above is a chart of some proteins IPed with FLAG CstF 77 from HEK293 cells along with the sequence count and the sequence coverage. The green color highlights subunits from the CstF complex while the red highlights the factors from the CPSF complex.
Table 2.4 FLAG CFIm 25 Mass Spectrometry Results

<table>
<thead>
<tr>
<th>Name</th>
<th>Sequence Count</th>
<th>Sequence Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>CFIm 25</td>
<td>68</td>
<td>71.4%</td>
</tr>
<tr>
<td>CFIm 59</td>
<td>56</td>
<td>53.9%</td>
</tr>
<tr>
<td>CFIm 68</td>
<td>99</td>
<td>52.6%</td>
</tr>
<tr>
<td>CstF 64 tau</td>
<td>2</td>
<td>6.8%</td>
</tr>
</tbody>
</table>

Shown above is a chart of some proteins IPed with FLAG CFIm 25 from HEK293 cells along with the sequence count and the sequence coverage. The purple color highlights subunits from the CFIm complex while the green highlights the factors from the CstF complex.
Table 2.5 FLAG Pcf11 Mass Spectrometry Results

Shown above is a chart of some proteins IPed with FLAG Pcf11 from HEK293 cells along with the sequence count and the sequence coverage. The grey color highlights subunits from the CFIm complex while the purple and green highlights the factors from the CFIm and the CstF complexes, respectively.
subunits (Table 2.3). Similarly, FLAG CFIm 25 can also immunoprecipitate the other two CFIm subunits and FLAG Pcf11 can immunoprecipitate Clp1 from the CFIm complex (Tables 2.4 and 2.5, respectively). Together, this demonstrates that immunoprecipitation together with mass spectrometry can reveal valid, biological interactions.

**Interactome Map of the mRNA 3’ Processing Machinery**

To identify new mRNA 3’ processing regulators, an interactome of the mRNA 3’ processing machinery was physically mapped based on the mass spectrometry results (Figure 2.3). To analyze the mass spectrometry data, the SAINT program was used (Choi et al. 2011). SAINT is a program that allows probabilistic scoring of affinity purification-mass spectrometry data to derive bona fide protein-protein interactions. Cytoscape was then used to visualize the interactions obtained from SAINT. All of the interactions mapped with Cytoscape had a probability of 0.95 and sequence coverages greater than 10%. From the bioinformatics analyses, two maps were drawn. One map focused on the interactions within mRNA 3’ processing complex (Figure 2.3.A). From that map, the interactions between the multi-subunit complexes were visualized. As expected, many interactions exist between the multi-subunit complexes (Zhao et al. 1999) (Mandel et al. 2008). On the other hand, the interactions between the mRNA 3’ processing complex and other gene regulatory pathways (such as splicing, transcription, replication, and translation) were also mapped (Figure 2.3.B). From that map, two groups of proteins were identified. One group of proteins was commonly bound by most of the mRNA 3’ processing subunits. That group of proteins are exemplified by the pink dots in the middle of the map that are linked to multiple mRNA 3’
processing subunits. In contrast, the other group of proteins interacts with specific mRNA 3’ processing subunits. That group of proteins are exemplified by the pink dots located on the periphery of the map. From these two interactome maps, new regulators of mRNA 3’ processing complex can be identified.

**Novel Post-Translational Modifications Identified**

Previously, it has been shown that some mRNA 3’ processing subunits can be modified by post-translational modifications such as phosphorylation, sumoylation, and acetylation. These modifications can affect mRNA 3’ processing efficiency, protein:protein interactions, protein localization and enzymatic activity (Colgan et al. 1997) (Kim et al. 2003) (Shimazu et al. 2007) (Vethantham et al. 2007) (Vethantham et al. 2008). In addition, ubiquitination has been shown to affect mRNA 3’ processing (Ingham et al. 2005). Specifically, Ingham et. al., showed that CFIm 68 can bind to the ubiquitin E3 ligase, AIP4/Itch. Its functional significance was also explored. Ingham et. al., showed that AIP4/Itch can polyubiquitinate CFIm 68. Interestingly, my mass spectrometry results showed that AIP4/Itch and 2 other ubiquitin E3 ligases from the same HECT sub-family can also interact with the CFIm complex. The other 2 ubiquitin E3 ligases identified by my mass spectrometry results were WWP1 and WWP2 (Figure 2.4). Based on the sequence coverage (ranging from 6.9% to 39.3%), the interaction between the CFIm complex and the 3 ubiquitin E3 ligases are strong. Furthermore, AIP4/Itch, WWP1 and WWP2 are from the same E3 ligase family. Together, this
Figure 2.4 Ubiquitination Mechanism

The ubiquitination mechanism is depicted above. Using mass spectrometry and bioinformatics analyses, three ubiquitin E3 ligases were bound to associate with the CFIm complex. The three ubiquitin E3 ligases were WWP1, WWP2 and ITCH.
Figure 2.5 CFIm Interacts with E3 ligases

A. 293 cell line stably expressing FLAG CFIm 25 was IPed using FLAG beads. The sample was resolved on a 10% SDS PAGE. A Western was then performed with WWP2, WWP1, CFIm 59 and CFIm 25 antibodies. B. Endogenous CFIm 25 was immunoprecipitated from HeLa whole cell extract. The sample was resolved on a 10% SDS PAGE. A Western was then performed with WWP2, WWP1, CFIm 59 and CFIm 25 antibodies.
suggests that the interaction between the CFIm complex and these 3 E3 ligases are biologically significant. However, *in vitro* assays are still needed to verify the mass spectrometry results.

**Novel Post-Translational Modifications Verified**

Using the whole-cell lysate of the FLAG CFIm 25 stable cell line, FLAG immunoprecipitation was performed. As a negative control, FLAG immunoprecipitation was also performed from whole-cell lysate of untransfected 293 cells (Figure 2.5.A). After FLAG immunoprecipitation, the eluate was used for Western blot analyses. The data showed that FLAG CFIm 25 was immunoprecipitated along with another CFIm complex subunit, CFIm 59, as expected. In addition, the analyses illustrated that WWP1 and WWP2 were also immunoprecipitated with FLAG CFIm 25. In contrast, CFIm 25, CFIm 59, WWP1 and WWP2 were not immunoprecipitated from untransfected 293 cells. This demonstrates that FLAG-CFIm 25 interacts with WWP1 and 2 in a specifically.

Next, the endogenous CFIm 25 was immunoprecipitated to determine if the interaction between CFIm and WWP1 and WWP2 is dependent on the FLAG tag. The endogenous CFIm 25 was immunoprecipitated from HeLa whole-cell extract with CFIm 25 antibodies bound to protein A/G beads. The immunoprecipitated product was then resolved on a 10% SDS PAGE gel and used for Western blotting analyses (Figure 2.5.B). The data demonstrated that beads covalently conjugated to CFIm 25 antibodies can specifically immunoprecipitate CFIm 25, CFIm 59 as well as WWP1 and 2. In contrast, actin
antibody was not able to immunoprecipitate CFIm 25, CFIm 59, WWP1 or WWP2. This confirms that CFIm 25 can specifically interact with WWP1 and WWP2 E3 ligases independent of the FLAG tag. Together, the data from Figures 2.5.A and 2.5.B show that the CFIm complex interacts with WWP1 and 2.

Although Figures 2.5.A and 2.5.B show that the CFIm complex interacts with WWP1 and WWP2, it is unclear whether this interaction is indirect or direct. To address this question, recombinant His-tagged CFIm complexes were generated using the baculovirus expression system (Figure 2.6.A left). From the baculovirus expression system, two CFIm complexes were made; one version contains His CFIm 25 and CFIm 59 while the other contains His CFIm 25 and CFIm 68. Two recombinant CFIm complexes were generated to determine if there were differences between the structurally-similar CFIm 59 and CFIm 68 subunits. In addition, recombinant GST-fusion proteins were made (Figure 2.6.A right). The GST tag was fused to full-length WWP2 or just the WW domain of WWP2 to determine if the WW domain alone is sufficient in CFIm binding. To identify if the CFIm complexes directly interact with WWP2, the recombinant His-CFIm complexes were incubated with recombinant GST-WWP2 Full or GST-WWP2 Domain in a GST pulldown assay (Figure 2.6.B). First, the GST-WWP2 Full and GST-WWP2 Domain proteins were fused to glutathione-conjugated beads. The respective recombinant CFIm complexes were then incubated with the GST-WWP2 Full or GST-WWP2 Domain beads in a GST pull-down assay. After the pulldown, the eluate was used for Western blot analyses. From the analyses, both GST-WWP2 Full and GST-WWP2 Domain pulled down recombinant His CFIm25-59 complex as well as the
recombinant His CFIm 25-68. This interaction was specific because GST alone was not able to pull down either recombinant CFIm complexes. These data show that the CFIm complex can interact with WWP1 and WWP2 directly. Finally, the CFIm subunits in direct contact with WWP2 were identified. To determine which CFIm subunit is in direct contact with full-length WWP2, a GST-pulldown was performed. Specifically, GST-WWP2 Full was bound to glutathione-conjugated beads. Radioactive, in vitro translated CFIm 25, CFIm 59 or CFIm 68 was then added to the beads. After the GST-pulldown, the eluate was resolved on a 10% SDS PAGE gel and the radioactive in vitro translated CFIm proteins were visualized using film. Figure 2.6.C shows that all three in vitro translated CFIm subunits interacted to GST-WWP2 specifically. Together, these data verified that all three CFIm subunits can interact with WWP2 directly and specifically.

**Functional Significance Characterization**

After confirming the interaction of the CFIm complex with the ubiquitination pathway, the functional significance of this interaction was determined. An in vivo ubiquitination assay was used to determine if the CFIm subunits are ubiquitinated (Figure 2.7). Mammalian expression vectors containing the coding regions of His-tagged ubiquitin, FLAG-tagged CFIm 59, FLAG-tagged CFIm 68, WWP2, Myc tagged Itch or the catalytically inactive Itch (Myc-Itch MT) were transfected into 293 T cells. The cells were then grown for 48 hours. After two days, 50 μM of Mg132 was added to the cells. The cells were then incubated at 37°C for an additional hour. The cells were then harvested. Whole-cell lysates were generated from the transfected cells and used in a pulldown assay under
Figure 2.6 CFIm Directly Interacts with E3 Ligases

A. Left: Coomassie stain of recombinant His CFIm 25-59 and His CFIm 25-68 purified from the baculovirus system. Right: Coomassie stain of recombinant GST fusion proteins from E. coli. B. GST pulldown of recombinant His CFIm 25-59 or His CFIm 25-68 with the respective recombinant GST-tagged proteins. After the pulldown, the samples were resolved on a 10% SDS PAGE gel and probed with CFIm 25 and either CFIm 59 or CFIm 68. C. GST-pulldown with GST or GST-WWP2 full length with radioactive in vitro translated CFIm 25, CFIm 59 or CFIm 68 proteins. After the GST pulldown, the samples were resolved on a 10% SDS PAGE gel and the radioactive in
vitro translated proteins were imaged using film.

<table>
<thead>
<tr>
<th></th>
<th>His-Ub</th>
<th>FLAG</th>
<th>CFIm 59</th>
<th>WWP2</th>
<th>Myc-Itch</th>
<th>Myc-Itch MT</th>
<th>Mg132</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transient Transfection</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+ Mg132</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Whole cell lysate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>His pulldown</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>His-Ub</th>
<th>FLAG</th>
<th>CFIm 59</th>
<th>WWP2</th>
<th>Myc-Itch</th>
<th>Myc-Itch MT</th>
<th>Mg132</th>
</tr>
</thead>
</table>
| **Left:** Experimental procedures. **Center and Right:** Transient transfection was performed by transfecting plasmids expressing His-Ubiquitin (His-Ub), FLAG CFIm 59, WWP2, Myc-Itch and Myc-Itch mutant (Myc-Itch MT) into 293T cells. Forty eight hours after transfection, 50 uM of Mg132 was added to the cells. The samples were then harvested and pulled down using cobalt beads under denaturing conditions (8 M urea). Finally, the samples were resolved on a 10% SDS PAGE gel and a Western was performed with either CFIm 59 or CFIm 68 antibody.

**Figure 2.7 CFIm Subunits are Polyubiquitinated by WWP2 and Itch**

IB: CFIm 59

IB: CFIm 66
denaturing conditions with cobalt beads. The pulldown assay was denaturing because the buffers contained 8 M urea. The eluate was then resolved on a 10% SDS PAGE gel for Western blotting analyses with antibodies targeting CFIm 59 and CFIm 68. The data demonstrated that there was a lack of ubiquitination when only exogenous His ubiquitin or His ubiquitin with FLAG CFIm 68 were transfected into 293 T cells. Ubiquitination is identified as a smear that resolves higher than the nonubiquitinated product. However, ubiquitination was present when exogenous His ubiquitin, FLAG CFIm 68 and WWP2 were simultaneously transfected into 293 T cells. This demonstrates that WWP2 can polyubiquitinate CFIm 68 in an in vivo ubiquitination assay. In addition, I incubated exogenous His Ubiquitin, FLAG CFIm 68 with Myc Itch to replicate previous findings that Itch can polyubiquitinate CFIm 68 (Ingham et al. 2005). Identically, CFIm 68 was polyubiquitinated by Itch. This polyubiquitination is specific because the catalytically inactive form of Itch could not polyubiquitinate CFIm 68 (Figure 2.7 right). A similar pattern was shown with CFIm 59. When His ubiquitin, FLAG CFIm 59 and WWP2 were transfected into the cells, polyubiquitination was also seen. However, in contrast to CFIm 68, CFIm 59 can be polyubiquitinated in the presence of endogenous E3 ligases; polyubiquitination is readily apparent when only FLAG CFIm 59 and His ubiquitin are transfected into the cells (Figure 2.7 left). This is in contrast to CFIm 68 in that endogenous E3 ligases are not capable of polyubiquitinating CFIm 68. Rather, an exogenous source of E3 ligase is needed for polyubiquitination to occur (Figure 2.7 right). Together this confirms that CFIm 59 and CFIm 68 can interact with WWP2. As a result of that interaction, polyubiquitination occurs. This suggests that ubiquitination may play a major role in the CFIm complex.
2.5 Conclusions and Discussion

In this chapter, a broad approach was taken to discover novel regulators of mRNA 3’ processing. Specifically, sixteen of the major mRNA 3’ processing subunits were immunoprecipitated. Their binding partners were then identified by mass spectrometry. From the mass spectrometry results, the mRNA 3’ processing interactome was mapped. This method was valid and revealed biologically relevant interactions such as between the CFIm complex and three ubiquitin E3 ligases: Itch/AIP4, WWP1 and WWP2. The interaction between the CFIm complex and the ubiquitin E3 ligases were then verified using in vitro assays. My data confirmed that the endogenous CFIm complex can interact with WWP1 and WWP2 and that this interaction was a direct interaction between the CFIm complex and WWP2. Furthermore, each of the CFIm subunits can interact with WWP2. The functional significance of the interaction between CFIm and ubiquitin E3 ligases was also characterized. Using in vivo ubiquitination assay, it was demonstrated that CFIm 59 and CFIm 68 can be polyubiquitinated. These results demonstrate the validity of the approach combining immunoprecipitation and mass spectrometry in finding new regulators of mRNA 3’ processing.

Polyubiquitination of the CFIm complex may play a role in APA. Previously it was shown that depleting CFIm 25 or CFIm 68 using siRNA can change the poly (A) site choice from distal to proximal (Kubo et al. 2006) (Kim et al. 2010) (Martin et al. 2012). Perhaps the polyubiquitination of CFIm 25 and CFIm 68 can lead to the degradation of these proteins and therefore shift in favor of the proximal PAS. Interestingly, not all of the
CFIm subunits can affect APA. CFIm 59, in contrast to CFIm 25 and CFIm 68, does not seem to affect APA (Gruber et al. 2012) (Zhao et al. 1999) (Mandel et al. 2008). This suggests that CFIm 59 and CFIm 68 can affect APA differently. The reasons why two structurally similar proteins in the same complex can affect APA differently is still uncertain. However, one possible reason might have emerged from my in vivo ubiquitination experiment. From my in vivo ubiquitination assay, CFIm 59 and CFIm 68 have different polyubiquitination patterns. Specifically, CFIm 59 can be polyubiquitinated by endogenous AIP4/Itch, WWP1 or WWP2 E3 ligases (and possibly other undiscovered ubiquitin E3 ligases) while CFIm 68 cannot. This result suggests that the CFIm 59 can more easily be degraded by the proteasomal pathway than CFIm 68. The different polyubiquitination patterns of CFIm 59 and CFIm 68 can create different APA patterns. CFIm 59 can be degraded so quickly that the amount of CFIm 59 present under normal conditions is limited. With a scarce amount of CFIm 59, the distal PAS will be preferentially chosen over the proximal and weaker PAS. When siRNA targeting CFIm 59 is introduced into the system, this would decrease the levels of CFIm 59 even further and the distal PAS would be preferred even more. Therefore, the proximal PAS would not be chosen and APA would not occur.

While polyubiquitination usually signals a substrate for degradation, some polyubiquitination signals a change in protein localization (Metzger et al. 2012). Previously it has been shown that CFIm 68 can shuttle between the nucleus and the cytoplasm to export mRNA (Ruepp et al. 2009). Perhaps the polyubiquitination of CFIm 68 is the signal to accomplish this task. Future experiments are needed to determine if
CFIm 59 and CFIm 68 polyubiquitination leads to their degradation, a change in localization or other functions. In addition, the difference in polyubiquitination pattern of CFIm 59 and CFIm 68 needs to be further explored.

The immunoprecipitation combined with mass spectrometry experiment also shed some light into the role of CFIIIm in mRNA 3’ processing. The role of CFIIIm in mRNA 3’ processing is still largely a mystery (Zhao et al. 1999). From immunoprecipitating Pcf11 and the mass spectrometry results, my data shows that Pcf11 can bind to isoforms 1 and 2 of Bcl-2. Their interactions were strong considering that the sequence coverage was 32.2% for isoform 1 and 30.3% for isoform 2. The role of Bcl-2 is to regulate cellular apoptosis. Because Bcl-2 regulates apoptosis, perhaps the role of CFIIIm in mRNA 3’ processing is to signal non-polyadenylated, premature RNA for degradation. Another possibility is the Bcl-2 is a signal for cells under stressful conditions to stop mRNA 3’ processing from occurring. Further analysis needs to be done to verify the interaction of CFIIIm with Bcl-2 as well as its functional significance.

In addition to characterizing possible interaction of Pcf11 with Bcl-2, more analyses need to be done to harvest other potential regulators. One interesting interaction seen from the mass spectrometry analyses is the interaction between CstF 50 and p53. According to the mass spectrometry data, CstF 50 interacts with p53 isoform A, B, C, D, E and F with 12.5%, 14.9%, 10.10%, 10.3%, 12% and 11.7% sequence coverage, respectively. This suggests that CstF 50 interacts with 6 out of the possible 12 different p53 isoforms, implying that this interaction is biologically significant. This interaction
correlates with previous data showing that CstF 50 can interact with another tumor suppressor, BARD1, in preventing premature polyadenylated mRNA from expressing under cellular damage conditions (Cevher et al. 2010). Thus, it would be interesting to determine if p53 can prevent premature polyadenylated mRNA from expression by interacting with CstF 50.

In conclusion, this chapter demonstrated and validated the use of immunoprecipitation combined with mass spectrometry to identify new regulators of mRNA 3’ processing. While mass spectrometry can lead to many false positives with its highly sensitive method, there are also many authentic interactions waiting to be discovered such as the interaction of CstF 50 with p53 and Pcf11 with Bcl-2.
Chapter 3

CPSF 30 and WDR33 Directly Bind to AAUAAA

3.1 Summary

Protein factors and cis elements are essential for mRNA 3’ processing to occur. The cis elements required for mammalian mRNA 3’ processing include the cleavage site, AAUAAA, U/GU-rich sequences and auxiliary elements located both up- and downstream of AAUAAA. The most conserved and invariable cis element is the AAUAAA. The AAUAAA is so invariable that a mutation of the U to any other nucleotide will abolish both the cleavage and the polyadenylation steps *in vivo* and *in vitro*. It is known that the CPSF complex recognizes the AAUAAA element. Previously, researchers thought that only the 160 kilodalton (kDa) CPSF subunit, CPSF 160, is needed to recognize AAUAAA. However, CPSF 160 does not have a canonical RRM
and direct evidence demonstrating the role of CPSF 160 in AAUAAA recognition is lacking. In this chapter, the central hypothesis is that other CPSF subunits are required to recognize AAUAAA. To address this hypothesis, two specific aims were developed. Specific aim 1: Identify the CPSF subunits that directly bind to AAUAAA. Specific aim 2: Characterize the CPSF30-AAUAAA interaction in detail.

Here I present data that both CPSF 30 and WDR33 directly bind to AAUAAA. The interaction between CPSF 30 and the AAUAAA was further studied. Specifically, zinc fingers 2 and 3 of CPSF 30 were shown to be responsible for RNA binding, for the entire CPSF complex to function and possibly to specifically recognize AAUAAA. Interestingly, these two zinc fingers are blocked by the influenza A viral protein, NS1A, to inhibit host mRNA 3' processing. In addition, the footprint of CPSF 30 and WDR33 were mapped on the RNA and global analyses of the role of CPSF 30 in polyadenylation were performed. Together, this chapter provides new insight in how AAUAAA is recognized for mRNA 3' processing to occur.

3.2 Introduction

mRNA 3’ processing is an essential process in eukaryotic gene expression that requires an extensive network of protein-protein and RNA-protein interactions. In Chapter 2, the interactome of the mRNA 3’ processing machinery was discussed. Here, the RNA:protein interactions of mRNA 3’ processing are explored. The three core cis elements needed for mammalian mRNA 3' processing are the cleavage site, the highly variable U/GU-rich element (located approximately 30 nucleotides downstream of the

The CPSF complex has been shown to recognize AAUAAA (Bienroth et al. 1991) (Gilmartin et al. 1989) (Gilmartin et al. 1991) (Keller et al. 1991) (Moore et al. 1988) (Murthy et al. 1992). Specifically, CPSF 160 is thought to recognize AAUAAA (Keller et al. 1991) (Murthy et al. 1995) (Moore et al. 1988) (Gilmartin et al. 1991). However, there are still some unanswered questions about the details of this interaction. First, it is uncertain how CPSF 160 can recognize AAUAAA when it does not have a canonical RRM and the mammalian CPSF 160 cannot bind to RNA by itself in a UV crosslinking assay (Murthy et al. 1995). Second, although Murthy and Manley demonstrated that CPSF 160 can recognize AAUAAA-containing RNA with slight specificity in an in vitro RNA pulldown assay, their data showed that CPSF 160 can also recognize AACAAA-containing RNA (Murthy et al. 1995). Lastly, it is still uncertain if CPSF 160 or other CPSF subunits bind directly at AAUAAA because body-radiolabeled or end-labeled RNAs were used in previous UV crosslinking experiments with partially purified CPSF.
complexes (Gilmartin et al. 1989) (Gilmartin et al. 1991) (Keller et al. 1991) (Moore et al. 1988) (Murthy et al. 1995). Therefore the CPSF subunits in direct contact to AAUAAA are unknown. The central hypothesis in this chapter is that other CPSF subunits are required to recognize AAUAAA specifically.

Similar to CPSF 160, CPSF 30 is a bona-fide subunit of the CPSF complex that is needed for mRNA 3’ processing (Barabino et al. 1997) (Bienroth et al. 1991). However there are a few differences between CPSF 160 and CPSF 30. First, CPSF 30 has been demonstrated extensively that it is an RNA-binding protein (Addepalli et al. 2007) (Bai et al. 1998) (Barabino et al. 1997) (Delaney et al. 2006) (Tacahashi et al. 2003) (Barabino et al. 2000). Second, NS1A can inhibit host’s mRNA 3’ processing by blocking the interaction of CPSF 30 with the RNA (Nemeroff et al. 1998) (van de Sandt et al. 2012). Specifically, CPSF 30 interacts with NS1A via its zinc fingers (Twu et al. 2006). CPSF 30 domain structure consists of three to five zinc fingers (C-X7-C-X5-C-X3-H) and possibly a zinc knuckle (C-X2-C-X4-H-X4-C), depending on the homolog (Barabino et al. 1997). In addition to interacting with NS1A, the zinc fingers of CPSF 30 are also needed to interact with other CPSF subunits and the RNA (Addepalli et al. 2007) (Barabino et al. 1997) (Barabino et al. 2000) (Tacahashi et al. 2003) (Hunt et al. 2008) (Rao et al. 2009). Although there have been many studies done to understand the interaction between CPSF 30 and the RNA, there are still discrepancies about the role of CPSF 30 role in RNA recognition. For example, it has been demonstrated that the zinc knuckle of mammalian CPSF 30 is needed for poly U binding (Barabino et al. 1997). However, the S. cerevisiae homolog doesn’t have a zinc knuckle and the protein
is thought to bind either near the cleavage site or the U-rich elements via its zinc fingers two and four (Barabino et al. 2000) (Tacahashi et al. 2003). Similarly, in Arabidopsis, zinc fingers two and four are also needed for RNA binding. However, zinc finger 3 can also bind to RNA (Addepalli et al. 2007). Therefore, the role of CPSF 30 in RNA recognition needs to be better understood.

WDR33 is another bona-fide CPSF subunit (Ohnacker et al. 2000) (Shi et al. 2009). Similar to CPSF 160, WDR33 is a large CPSF subunit (146 kDa). In contrast to CPSF 160, however, WDR33 contains repetitive WD40 domains that are required for protein-protein interactions such as between the CPSF and the CstF complexes (Ohnacker et al. 2000). In addition, WDR33 is also needed to control flowering time (Simpson et al. 2003). However, it remains largely unknown if WDR33 can bind to RNA via its WD40 domain. Therefore, the roles of CPSF 30 and WDR33 in RNA binding need to be better understood.

In this chapter, I provide evidence that CPSF 30 and WDR33 directly recognize AAUAAA. The interaction between CPSF 30 and the AAUAAA was further characterized. Zinc fingers two and three of CPSF 30 are needed for RNA interaction, for the entire CPSF complex to function, and possibly for AAUAAA recognition. A global analyses was done to understand the role of CPSF 30 in mRNA 3’ processing. Together, this study provides a more complete understanding of how the CPSF complex recognizes AAUAAA.
3.3 Materials and Methods

RNA substrates

*Splint Ligated, 46 nucleotides RNAs*

RNAs were site specifically radiolabeled at one position using a method called splint ligation. The splint ligated RNAs utilized in this chapter were derived from the commonly utilized substrate, L3. Successful splint ligation of the RNA was achieved by first purchasing chemically synthesized RNA oligonucleotides. For each splint ligated RNA, two chemically synthesized RNA oligonucleotides were purchased. The RNA oligonucleotide that contains the 5’ sequences was called the 5’ RNA oligonucleotide. Respectively, the RNA oligonucleotide that contains the 3’ sequences was called the 3’ RNA oligonucleotide. The 3’ RNA oligonucleotide was first radiolabeled at their 5’ end with $[\gamma-^{32}\text{P}]$ ATP and Optikinase using the standard protocol from Affymetrix. The radiolabeled 3’ RNA oligonucleotide was then ligated to the respective 5’ RNA oligonucleotide with the aid of T4 DNA ligase and a DNA bridge oligonucleotide that base paired to both the 5’ RNA oligonucleotide and the 3’ RNA oligonucleotide. After ligation, the DNA bridge oligonucleotide was degraded by DNase before the ligated RNA product was resolved on an 8% urea gel. The correct sized band (46 nucleotides) was then eluted from the gel. After gel purification, the splint ligated RNAs were then quantified using a scintillation counter. The sequences of the splint RNAs utilized in this chapter are the following:
RNA 1: 5’ AUGUACUAGGAGACACUUUCAAU*AAAGGCAAAUGUUUUUAUUUGUA
RNA 2: 5’ AUGUACUAGGAGACACUUUCAAU*AACGGCAAAUGUUUUUAUUUGUA
RNA 3: 5’ AUGUACUAGG*AGACACUUUCAAUAAAGGCAAAUGUUUUUAUUUGUA
RNA 4: 5’ AUGUACUAGGAGACACUUUCAAUAAAGGCAAAUGUUUUUAUUUGUA
RNA 5: 5’ AUGUACUAGGAGACACUUUCAAUAAAGGCAAA*UGUUUUUAUUUGUA

(46 nucleotides)

* represents the radiolabeled position.

**Bold** nucleotides represents wildtype or mutant hexamer

**5’ Radiolabeled, 18 nucleotide RNAs**

Besides L3, another commonly utilized substrate was used in this chapter. That substrate is called SVL. An 18 nucleotide RNA used in this chapter was derived from the SVL sequence. This 18 nucleotide SVL sequence was chemically-synthesized and radiolabeled at its 5’ end with \([\gamma^{-32}P] ATP\) and Optikinase using the standard protocol from Affymetrix. The 5’ radiolabeled, 18 nucleotide RNAs utilized in this chapter are the following:

WT: 5’ *CUGCAUAAAACAAGUUAA
MT: 5’ *CUGCAAGAAAACAAGUUAA

(18 nucleotides)

* represents the radiolabeled position.

**Bold** nucleotides represents the wildtype or mutant hexamer
3' Radiolabeled, 39 nucleotides RNAs

The 3' radiolabeled RNAs used in this chapter were derived from SVL and generated by *in vitro* transcription without radioactivity. The transcripts were then resolved on an 8% urea gel. The correct sized transcripts (39 nucleotides) were gel purified from 8% urea gel using UV shadow. The purified RNAs were then 3' radiolabeled with T4 RNA ligase and [5'-32P] PcP. The 3' radiolabeled RNAs utilized in this chapter are the following:

M1: GGGUAACCAUUAUAACCUGCAAUAAACAACUUAAACAACA*
M1-MT: GGGUAACCAUUAUAACCUGCAAUAAACAACUUAAACAACA*
M2: GGGUAACCAUUAUAACCUCGCAUAACCAACUUAAACAACA*
M3: GGGUAACCAUUAUAACCUCGCAUAACCAAGUUAACAACA*
M4: GGGUAACCAUUAUAACCUCGCAUAACCAACUUAAAGAACA*

* (39 nucleotides)

* represents the radiolabeled position

**Bold nucleotides represents the wildtype or mutant hexamer**

**Clones**

Coding regions encoding for mammalian full-length CPSF 30 isoform 2 (Full), without zinc finger 1 (-ZF1, missing amino acids 41 through 59), without zinc finger 2 (-ZF2, missing amino acids 68 to 86), without zinc finger 3 (-ZF3, missing amino acids 96 to 114), without zinc finger 4 (-ZF4, missing amino acids 124 to 140), without zinc finger 5 (-ZF5, missing amino acids 148 to 166) or without zinc knuckle (-ZK, missing amino acids 220 to 233) were cloned into FLAG-pcDNA3.1 vector. The coding regions for
some CPSF 30 variants (Full, -ZF2, -ZF3, -ZK) were then cloned into HA pQCXIP vectors.

Clones that exchanged the sequences encoding mammalian CPSF 30 zinc finger 2 and 3 for *Saccharomyces cerevisiae*’s zinc finger 2 or zinc finger 3, respectively, were also made. Specifically, the mammalian zinc finger 2 (amino acids 68-86) was switched with *Saccharomyces cerevisiae*’s zinc finger 2 (amino acids 67 to 85) while the mammalian zinc finger 3 (amino acids 96 to 114) was switched with *Saccharomyces cerevisiae*’s zinc finger 3 (amino acids 95 to 114). They are referred to as CPSF 30 Y2 and Y3, respectively. CPSF 30 Y2 and Y3 were cloned into HA pQCXIP vectors.

**Protein Complexes**

*FLAG CPSF 30-Containing Complexes*

293T cells were transiently transfected with the FLAG pcDNA3.1 vectors containing the coding sequences for CPSF 30 Full, CPSF 30 -ZF1, CPSF 30 -ZF2, CPSF 30 -ZF3, CPSF 30 -ZF4, CPSF 30 -ZF5, or CPSF 30 –ZK using calcium phosphate (Kingston et al. 2001). After 48 hours, the transfected cells were harvested and whole-cell extracts were generated using standard procedures. The whole-cell extracts were then incubated with Anti-FLAG M2 Affinity beads (Sigma-Aldrich) for 2 hours at 4° C. After incubation, the supernatant was discarded and the beads were washed with Buffer D 300 with 0.1% NP40, 3 times, 10 minutes each with gentle rotation at 4° C. Buffer D 100 was then used as the last wash. After the washes, the FLAG CPSF 30 complexes were
eluted from the beads with chemically synthesized FLAG peptides from Genescript in Buffer D 100. The eluates were then used in vitro assays.

The CPSF Complex

HEK 293 cells was transfected with 3x FLAG CPSF 73 pCMV14 using Lipofectamine. The stable transfectants were selected with neomycin/G418 (1 mg/ml) for two weeks. After selection, individual clones were picked and expanded. Western blotting analyses were then carried out with each clone to monitor the expression levels of the FLAG-tagged exogenous proteins. The clones were all screened and those that expressed the FLAG-tagged exogenous proteins near the endogenous protein levels were selected for further experiments. After the stable transfectants were selected, the cells were expanded and nuclear extract was generated using standard procedures. The nuclear extract was then incubated with Anti-FLAG M2 Affinity beads for 2 hours in 4°C. After incubation, the supernatant was discarded. The beads were then washed and eluted under identical conditions as the FLAG CPSF 30 complexes mentioned above. The eluted FLAG CPSF 73 complex was then used for in vitro assays.

HA CPSF 30-Containing Complexes

HA CPSF 30 (Full, -ZF2 or -ZF3) pQCXIP vectors were transfected with Lipofectamine 2000 into 293 GP2 packaging cell line in the presence of the envelope protein, pVSV-G. The supernatant containing the retroviruses were collected 48 hours later and filtered. The sterile retroviruses were then utilized to infect FLAG CPSF 73 stable cell lines (see above) in the presence of 8 µg/ml polybrene. Twenty four hours later, the infection was
repeated. The FLAG CPSF 73 cell lines were then placed under puromycin (1 µg/µl) and G418 (0.5 mg/ml) selection for at least two weeks. After selection, individual clones were picked and expanded. Western blotting analyses were then carried out with each clone to monitor the expression levels of the HA-tagged exogenous proteins. The clones were all screened and those that expressed the HA-tagged exogenous proteins near the endogenous protein levels were selected for further experiments.

To purify the complexes, HA CPSF 30 Full complex was purified using the FLAG immunoprecipitation protocol as described above. Meanwhile the nuclear extract from HA CPSF 30 complex without ZF2 or ZF3 was first used in a pull down assay with 4 µg of GST-NS1A at 4° C overnight. After incubation, the flow through was immunoprecipitated with FLAG beads (see above protocol).

**Recombinant Proteins**

*6x His PAP*

Sequences encoding PAP was cloned into 6x His pEAK. The plasmid was then transformed into *E. coli* and grown until the density at OD 600 = 0.5. The *E. coli* was then induced by adding 0.5 mM IPTG for 18 hours at 15° C. The *E. coli* was then pelleted and resuspended in a buffer containing 20 mM Tris HCl, pH = 7.4, 100 mM NaCl, 0.05% NP40, 5 mM imidazole and 0.5 mM PMSF. Afterwards, the sample was sonicated. The sample was then spun down and the supernatant was incubated with cobalt beads overnight at 4° C with gentle rotation. After incubation, the supernatant was discarded and the beads were washed with the same buffer as above except the
buffer contained high salt (500 mM NaCl). The beads were washed 3x, 5 minutes each with gentle rotation. Then the beads were washed with the same high salt buffer except this time it contained 15 mM imidazole. Finally, the protein was eluted off the beads with the high salt buffer containing 200 mM imidazole. The eluate was dialyzed into Buffer D 100 without MgCl2 and used in in vitro polyadenylation assays.

**GST-NS1A**

Pgex-4T2 plasmid containing the coding sequences of NS1A (CY121129.1) was kindly provided as a gift from Dr. Diana Lee Noah. After transforming the plasmid into competent BL21 E. coli, the cells were grown to OD 600 = 0.5. IPTG was then added to a final concentration of 0.4 mM before the cells were grown for 15 hours at 18° C. The cells were then pelleted and resuspended in ice-cold lysis buffer (20 mM Tris HCl pH = 7.5, 300 mM NaCl, 0.5 mM EDTA, 5% glycerol, 1 mM DTT, 1x protease inhibitor cocktail). Sonication was then performed to lyse the cells. NP40 was then added to a final concentration of 0.5% before the lysate was rotated in 4° C for 30 minutes. Lysate was then spun in 4° C at 14,000 rpm for 10 minutes. The supernatant was then incubated with glutathione beads in 4°C with gentle rotation for 12 hours. After incubation, the supernatant was discarded. The beads were washed with lysis buffer with 0.5% NP40, 3x, each time 10 minutes in 4° C with gentle rotation. The final wash was with lysis buffer but without NP40. GST-NS1A was then eluted with elution buffer (100 mM Tris HCl pH 7.5, 100 mM NaCl, 0.5 mM EDTA, 5% glycerol, 20 mM glutathione). Finally, the eluate was dialyzed in Buffer D 100 in 4° C for 3 hours.
**FLAG CPSF 30**

Recombinant FLAG CPSF 30 was made using the baculovirus expression insect cell system (Invitrogen) using standard procedures. After infection, the insect cells were harvested and the whole-cell lysate was made using standard procedures. Recombinant FLAG CPSF 30 was then purified using FLAG beads as described above.

**His CPSF 160**

Recombinant His CPSF 160 was made using the baculovirus expression insect cell system (Invitrogen) using standard procedures. After infection, whole-cell lysate was made using standard procedures. His pulldown was then performed. Cobalt beads were washed with Buffer D 100 (without EDTA) with 0.01% NP40 and incubated with His CPSF 160 whole-cell lysate at 4°C for 12 hours. After incubation, the supernatant was discarded and the beads were washed with Buffer D 100 (without EDTA) with 0.01% NP40, 3x, 10 minutes each at 4°C. The proteins were then eluted with Buffer D 100 (no EDTA) containing 200 mM imidazole for 2 hours at 4°C. The eluate was then dialyzed in Buffer D 100 overnight at 4°C before it was used in *in vitro* assays.

**In Vitro Assays**

*UV Crosslink Assay*

The respective RNAs (0.7 fmol) and proteins (68 fmol) were incubated in the presence of 1.6 mM ATP, 0.1 mg/ml tRNA, 0.02 mM CP, 0.2 mg/ml BSA, 10 mM Hepes, 50 mM NaCl or KCl, 0.5 mM MgCl2, 0.1 mM EDTA, 5% glycerol and 10 mM beta mercaptoethanol for 10 minutes at 30°C. The samples were then subjected to UV
crosslinking for 10 minutes at 254 nm (240,000 microjoules). Afterwards, the samples were either left on ice or incubated with the 20 units of RNase T1 or 60 units of RNase 1 for 5 minutes in 37° C. The proteins were then resolved on a 10% SDS PAGE gel. After the samples have been resolved, the gels were fixed in a solution containing 40% methanol and 10% acetic acid before imaged by phosphorimager.

**Immunoprecipitation Under Denaturing Conditions**

Large scale nuclear extract was heated in the presence of 0.5% SDS and 1.5% Triton X100 in Buffer IP 100 (100 mM NaCl, 2 mM MgCl2, 50 mM Tris pH 7.6, 0.05% NP40, 0.5 mM DTT) for 3 minutes at boiling temperatures. The denatured nuclear extract was then diluted 10 folds with Buffer IP 100 and incubated with protein A/G beads conjugated to either 1 µg of actin, beta-actin, CPSF 30, CPSF 160 or WDR33 for 12 hours at 4° C. After incubation, the supernatant was discarded. The beads were then washed twice with Buffer IP 100 containing 1 M urea for 10 minutes each at 4° C. The final wash was with Buffer IP 100 without urea. SDS loading buffer was then added to the beads. The eluted RNA and peptides were resolved on a 10% SDS-PAGE gel and Western blotting analyses were then performed.

**UV Crosslink and Immunoprecipitation Under Denaturing Conditions**

After CPSF was UV crosslinked to the RNA (see protocol above), 0.5% SDS, 1.5% Triton X100 in Buffer IP 100 was added to the sample and boiled for 3 minutes. The samples were then immunoprecipitated under denaturing conditions as described above.
**In Vitro Polyadenylation Assay**

The polyadenylation assay reaction is similar to the UV crosslinking assay (see above) except it contained 0.8 mM ATP and 20 ng of PAP. After the reaction was assembled, the samples were incubated for 1-2 hours at 30° C. After incubation, the samples were treated with proteinase K for 15 minutes at 37° C. Phenol chloroform was then added to the sample to denature proteins. Afterwards, the RNA was ethanol extracted and resolved on an 8% urea gel. The data was then imaged by phosphorimager.

**RNase Digestion Assay**

CPSF was UV crosslinked to RNA1 and digested with RNase 1 or T1 similar to the above UV crosslinking protocol. Difference is, the samples were then treated with proteinase K to denature the protein. The RNA was ethanol extracted and resolved on a 20% urea before imaged by phosphorimager.

**Luciferase Assay**

HeLa stable cell lines expressing HA CPSF 30 Full, HA CPSF 30 –ZF2, HA CPSF 30 –ZF3, HA CPSF 30 –ZK, HA CPSF 30 Y2 or HA CPSF 30 Y3 were transfected with either AAUAAA or AAAAAA-containing pPassport. The transfected cells were lysed 24 hours later with 100 µl of Passive Lysis Buffer from the Dual-Luciferase Reporter Assay (Promega). The lysates were then incubated for 15 minutes in room temperature with gentle rotation. Fifteen microliters of the lysates were then added to 75 µl of Luciferase Assay Reagent. The firefly signal was then quantified by a luminometer. Seventy five
microliter of Stop & Glo Buffer was added to the sample before the renilla signal was quantified. The renilla to firefly ratio was then normalized to the cell lines expressing HA CPSF 30 Full. The data was graphically represented.

**Poly (A) Site Sequencing (PAS-SEQ)**

*Library Generation*

Total RNA was isolated from HeLa cell lines overexpressing HA CPSF 30 Full, HA CPSF 30-ZF3, HA CPSF 30 Y3, FLAG NS1A, siCPSF 30 or siCPSF 160 using standard procedures. Ten microgram of total RNA was fragmented at 70°C for 10 minutes using the fragmentation reagent from Ambion. The fragmentation was then stopped using Stop Buffer. RNA was then ethanol precipitated. The pelleted RNA was then resuspended in water before it was reverse transcribed using Superscript III and PASSEQ7-2 RT primer. After reverse transcription, the samples were treated with RNase H. cDNA was then gel purified from unincorporated RT primer with an 8% urea gel. Afterwards, the purified cDNA was circularized using Circligase II ssDNA ligase (Epicentre) overnight at 60°C. During the circularization, a hairpin structure was formed, generating a now cleavable BamHI site. After BamHI digestion, the circularized cDNA is now linear. The linear cDNA was then amplified with PE1.0 and PE2.0 index primers (see below for sequences). All of the samples were amplified with PE1.0. However, each sample was amplified with a unique PE2.0 index primer that contained its own barcode. HeLa was amplified with PE2-index1, HA CPSF 30 Full was amplified with PE2-index2, HA CPSF 30 (-ZF3) was amplified with PE2-index3, HA CPSF 30 (yeast ZF3) was amplified with PE2-index4, siCPSF 30 was amplified with PE2-index5,
siCPSF 160 was amplified with PE2-index6 and FLAG NS1A was amplified with PE2-index7. After PCR amplification, the products were sent to UCI Genomics High Throughput Facility where single read, 100 base pair sequences were read from the 5’ end of the cDNA into the poly (A) tract.

List of Primers used

RT primer:

5’[phos]NNNNAGATCGGAAGAGCGTCTCTGATcggatcattaggatccgaGACGTGTGCTCTTCGATCTTTTTTTTTTTTTTTTTTTTTTTT[V-Q]
where V-Q represents any DNA nucleotide except for T

PCR primers:

1) PE1.0

5’AATGATACGGCGACCACCAGATCTACACTCTTTCCCTACACGACGCTCTTTCCGA
TCT

2) PE2-index1

5’CAAGCAGAAGACGCGATACGAGATCGTGATGGACTGGAGGAGTTCAGACGCTCTTCCGATC

3) PE2-index2

5’CAAGCAGAAGACGCGATACGAGATACATCGGTGACTGGAGGAGTTCAGACGCTCTTCCGATC

4) PE2-index3

5’CAAGCAGAAGACGCGATACGAGATGGCTAAATGGACTGGAGGAGTTCAGACGCTCTTCCGATC

5) PE2-index4

5’CAAGCAGAAGACGCGATACGAGATTGGTCAATGGACTGGAGGAGTTCAGACGCTCTTCCGATC
6) PE2-index5
5’CAAGCAGAAGACGGCATACGAGATCACTGTGTGACTGGAGTTCAGACGTGTGCT
CTTCCGATC

7) PE2-index6
5’CAAGCAGAAGACGGCATACGAGATATTGGCGTGGAGTTCAGACGTGTGCT
CTTCCGATC

8) PE2-index7
5’CAAGCAGAAGACGGCATACGAGATGATCTGGACTGGAGTTCAGACGTGTGCT
CTTCCGATC

The underlined nucleotides represent a unique bar code that is used to identify each sample.

**PAS-SEQ Analysis**

First each sample was separated based on their unique 6 nucleotides bar code. Then only the sequences with 15 continuous As or more were used for isolated for further analysis. After trimming, if the remaining length was greater than 15 nucleotides, the sequence would be mapped to the genome using a program called Bowtie. During the mapping, at most 2 mismatches are allowed. Only uniquely mapped reads are kept for further analysis. Sequences that had the same barcode, same length, and mapped to the same genome location were treated as PCR duplicates. Those samples were merged together and treated as one.

Sequences that primed internally were then removed. If the 3’ end of mapped reads were immediately followed by 6 continues As or 7 As were present within the next 10
nucleotides, those reads were removed. For the remaining reads, they were normalized to each other by dividing the original read count by a factor. The factor is the total number of reads in the sample divided by 1,000,000.

To analyze the APA of each sample, pairwise comparison between each sample and the HeLa control was done. The P-value of each individual poly (A) site was calculated using Fisher-Exact test. For each gene, the proximal poly (A) and the distal poly (A) with the smallest p-values were selected for further analysis. A false discovery rate threshold of 0.05 was applied to adjust the p-value threshold. If a gene has both a proximal and a distal poly (A) site with p-values smaller than the adjusted p-value threshold, and the difference between the proximal and distal is greater than 0.2, that gene was considered to have a significant APA change. The APA genes were further classified into two categories: genes that shifted significantly from proximal to distal (P2D) and genes that shifted significantly from distal to proximal (D2P).

3.4 Results

Determining the CPSF Subunits in Direct Contact with AAUAAA on L3 Substrates

The CPSF complex utilized in my experiments was first immunoprecipitate with FLAG conjugated beads from FLAG CPSF 73 stable cell line. After FLAG immunoprecipitation, the CPSF complex was visualized by silver staining (Figure 3.1.A). Based on previous mass spectrometry results, the CPSF complex was not only highly purified, but it also contained all of the known CPSF subunits (Shi et al. 2009). To determine which CPSF subunit binds directly to AAUAAA, splint RNAs site-specifically
Figure 3.1 Mapping Experiment 1 Reagents

A. CPSF complex purified by FLAG immunoprecipitation from 293 cell lines stably overexpressing FLAG CPSF 73. The complex, containing all of the known CPSF subunits, is visualized using silver staining. B. Each adenovirus-derived splint ligated RNAs were uniquely site specifically radiolabeled at one position as depicted by the red astricks. C. In vitro polyadenylation assay was performed by incubating 0.7 fmol of the respective splint RNAs in B. with 68 fmol of CPSF and 20 ng of PAP for 1 hour at 30 °C. After the incubation, RNAs were extracted and the samples were resolved on a 8% urea gel.
radiolabeled between the third and fourth position of the AAUAAA was used (RNA 1, Figure 3.1.B). A splint RNA that was site-specifically radiolabeled between the third and fourth position of the mutant hexamer, AAUAAC, was used as a negative control (RNA 2, Figure 3.1.B). Additional splint RNAs were identical to RNA 1 but the site-specifically radiolabeled positions were different (RNA 3,4,5, Figure 3.1.B). Afterwards, the splint RNAs were incubated with the CPSF complex to determine if the CPSF complex was functional on them. An in vitro polyadenylation assay was performed by incubating CPSF and recombinant PAP with the corresponding splint RNAs. If the CPSF complex was functional, then the splint RNAs containing AAUAAA (RNA 1,4,5) would be polyadenylated. In contrast, the RNA containing the mutant AAUAAC hexamer should not be polyadenylated (RNA 2). The data showed that as expected, RNA 1,4, and 5 were polyadenylated but not RNA 2 (Figure 3.1.C). This demonstrates that our CPSF can specifically recognize and function on the splint RNAs.

To determine which CPSF subunit is in direct contact to AAUAAA, an UV crosslinking assay was performed with the splint RNAs. The CPSF complex was first UV crosslinked to RNA 1. The UV crosslinked product was then resolved on a 10% SDS-PAGE gel (Figure 3.2.A left). The data showed that two proteins (approximately 40 and 160 kDa) UV crosslinked to RNA 1 (Figure 3.2.A right). From the result so far, the data shows that a 40 kDa and a 160 kDa protein can bind near or at the AAUAAA hexamer. To determine which protein binds directly at AAUAAA, the UV crosslinked product was treated with RNases. Two RNases were used in this experiment: RNase T1 and
RNase1. RNase T1 is an endoribonuclease that hydrolyzes 3’ of guanines while RNase

Figure 3.2 Mapping Experiment 1

A. Left: Experimental scheme Right: UV crosslinking experiment where 68 fmol of CPSF was incubated with 0.7 fmol of either RNA 1 or RNA 2. After UV crosslinking, the samples were treated with 20 units of RNase T1 or 60 units of RNase 1. The samples were then resolved on a 10% SDS PAGE gel. Phosphorimager was used to visualize the image. B. RNase digestion assay where 68 fmol of CPSF was incubated with 0.7 fmol of RNA1. After UV crosslinking and RNase digestion with either 20 units of RNase T1 or 60 units of RNase 1, the RNA was extracted and resolved on an 20% urea gel. Phosphorimager was then used to visualize the image.
1 is an endoribonuclease that hydrolyzes 3’ of every ribonucleotide. The specificity of RNase T1 and RNase 1 were first experimentally determined (Figure 3.2.B). From the experiment, it shows that RNase T1 can hydrolyze 3’ of guanines while RNase 1 can hydrolyze 3’ of every ribonucleotides as expected. Therefore, the protein in direct contact with AAUAAA can be accurately determined. To determine which CPSF subunits directly bind to AAUAAA, after UV crosslinking the CPSF complex to the splint RNAs, RNase T1 or 1 treatment was performed. Two proteins, approximately 30 and 160 kDa remain bound to AAUAAA after RNase treatment. In contrast, no proteins UV crosslinked to RNA 2. This suggests that a 30 kDa and a 160 kDa protein bind directly at AAUAAA specifically together.

**CPSF 30 and WDR33 Directly Bind to AAUAAA**

To determine the identity of the 30 kDa in direct contact with the AAUAAA, immunoprecipitation under denaturing conditions was performed after UV crosslinking. First, the denaturing procedure was tested to provide evidence that the procedure worked (Figure 3.3.A). Indeed, the denaturing immunoprecipitation protocol worked in that only the CPSF 30 protein was immunoprecipitation with the CPSF 30 antibody. After determining that the denaturing procedure works, UV crosslinking followed by immunoprecipitation under denaturing conditions were performed with either actin or CPSF 30 antibody (Figure 3.3.B). The data shows that the CPSF 30 antibody can indeed immunoprecipitate the 30 kDa protein. This illustrates that CPSF 30 directly recognize AAUAAA.
To identify the 160 kDa protein, UV crosslinking assay was again performed. After UV crosslinking, the samples were treated with RNase T1. Immunoprecipitation under denaturing conditions was then performed with either beta actin, CPSF 30, WDR33 or CPSF 160 antibody. The data shows that the WDR33 antibody, but not the CPSF 160 antibody, can immunoprecipitate the 160 kDa protein that is in direct contact with the AAUAAA (Figure 3.4). In addition, CPSF 30 was again confirmed to directly interact with the AAUAAA. Contrary to previous data, this suggests that WDR33 and CPSF 30 directly interact with AAUAAA.

**CPSF 30 is a Major Factor in AAUAAA Recognition on SVL Substrates**

To further confirm that CPSF 30 and WDR33 bind directly to AAUAAA, I utilized another commonly used substrate, SVL. I took a fragment of the SVL substrate and mutated the substrate such that it contained guanosines only at the selected positions. The RNAs were then 3’ radiolabeled with [5’–³²P] pCp (Figure 3.5.A). The guanosines were positioned such that when the RNAs were digested with RNase T1, unique 3’ sized fragments remained (Figure 3.6.A). The fingerprints of the CPSF subunits can then be mapped. The CPSF complex was first incubated with each of the RNAs, UV crosslinked and completely digested with RNase T1. As expected, CPSF only bound to RNA sequences containing AAUAAA but not AAUAAC (Figure 3.5.B). The experiment was then repeated but with the addition of M2, M3 and M4. After RNase digestion, the samples were split in half; one-half was resolved on a 10% SDS-PAGE gel while the other half was resolved on a 6% SDS-PAGE gel (Figure 3.6.B). Different percentage
Figure 3.3 Identifying CPSF 30 as a Factor that Directly Interacts with AAUAAA

A. Denature immunoprecipitation assay was tested with HeLa nuclear extract to determine if the procedure worked. HeLa nuclear extract was boiled in the presence of 0.5% SDS and 1.5% Triton X-100. The sample was then diluted 10 folds with Buffer IP 100. Immunoprecipitation was performed with either actin or CPSF 30 antibody. B. Left: Method scheme. Right: CPSF (68 fmol) was UV crosslinked to RNA 1 (0.7 fmol) and denature immunoprecipitation was performed with either actin or CPSF 30 antibody to determine the identity of the bottom band that UV crosslinked to RNA1.
Figure 3.4 Identifying WDR33 as another Factor that Directly Interacts with AAUAAA

CPSF (133-267 fmol) was UV crosslinked to RNA 1 (27-54 fmol). After UV crosslinking, each sample was digested with 40 units of RNase T1. The samples were then subjected to denature immunoprecipitation with either beta actin, CPSF 30, WDR33 or CPSF 160 antibody.
Figure 3.5 Mapping Experiment 2 Reagents

A. Five 3’ radiolabeled RNAs derived from SVL are depicted. The green lettering depicts the positions where RNase T1 will digest the RNA while the purple letter highlights the mutant AAUAAA hexamer. B. Left: Method scheme Right: UV crosslinking assay where M1 or M1 MT was UV crosslinked with 68 fmol of CPSF. After UV crosslinking, samples were digested with or without 20 units of RNase T1. Samples were then resolved on a 10% SDS PAGE gel and imaged using phosphorimager.
A. Each of the RNAs (M1, M2, M3, M4, M1 MT, 67 fmol) was UV crosslinked with 68 fmol of CPSF and then digested with 120 units of RNase T1. The samples were then resolved on a 20% urea gel to determine if the digestion sizes were correct. B. CPSF (68 fmol) was UV crosslinked with each of the respective RNAs (67 fmol each) and digested with 120 units of RNase T1 before resolved on either a 10% (left) or 6% SDS PAGE (right) gel. C. Quantification of the CPSF 30 (Figure B, left) and WDR33 (Figure B, right) signal was illustrated in a bar graph.
gels were utilized to ensure that both the large and small proteins were clearly resolved. The data showed that CPSF 30 bound to M1, both before and after RNase T1 digestion. This result was identical to the result achieved in Figure 3.5.B. However, after RNase T1 digestion, CPSF 30 cannot bind to M2, M3 or M4 as efficiently as to M1 (Figure 3.6.B left). This suggests that CPSF 30 binds to the AAUAAA hexamer and slightly upstream of it but not downstream of the AAUAAA. This result agrees well with the mapping experiment with adenovirus (Figure 3.2). In contrast, WDR33 was able to efficiently bind to M1, M2 and M3 with similar efficiency after RNase T1 digestion (Figure 3.6.B right). This suggests that while WDR33 can bind to the AAUAAA hexamer and slightly upstream of it, WDR33 can also bind downstream of the hexamer while CPSF 30 cannot. The difference in UV crosslinking signal was quantified in Figure 3.6.C. Together, this mapping experiment with the SVL substrate along with the mapping experiment with the adenovirus substrate suggests that CPSF 30 and WDR33 can both directly interact with the AAUAAA.

**Zinc Fingers Two and Three of CPSF 30 are Needed for RNA Binding**

After determining that CPSF 30 is a major factor in AAUAAA recognition, this suggests that CPSF 30 recognizes AAUAAA with either its zinc fingers or zinc knuckle (Figure 3.7.A). To identify the zinc finger or zinc knuckle domain responsible for AAUAAA recognition, plasmids containing CPSF 30 coding sequences lacking either zinc finger one, two, three, four, five or the zinc knuckle were made and transiently expressed in 293T cells. The complexes containing each of the respective CPSF 30 mutants were purified and utilized to perform UV crosslink with a short, eighteen nucleotides fragment
Figure 3.7 CPSF 30 RNA-Binding Domain Identification

A. Pictured is mammalian CPSF 30 structural domain from yeast to mammals. It contains 5 conserved zinc fingers (green) and at least one variable zinc knuckle (yellow). B. Left: FLAG cpsf 30 with its respective zinc fingers or zinc knuckle deletions were transiently transfected into 293T cells using calcium phosphate. After 48 hours, whole cell extract was made and each sample was immunoprecipitated with FLAG beads. The complexes were resolved on a 10% SDS PAGE gel and visualized by silver stain. Right: The immunoprecipitates from B. were incubated with 230 fmol of 5' radiolabeled 18 nt WT RNA derived from SVL. The samples were then UV crosslinked and resolved on a 10% SDS PAGE gel. The image was visualized by phosphorimager. C. The transient transfected CPSF complexes from B. were resolved on a 10% SDS PAGE gel. A Western was then performed to determine the protein:protein interactions within the CPSF complex.
derived from SVL. The UV crosslinking experiment showed that without zinc fingers two and three and possibly zinc knuckle, CPSF 30 can no longer bind to the RNA (Figures 3.7.B). This inability to bind to the RNA was not due to the lost of protein-protein interactions within the CPSF complex (Figure 3.7.C). While there does seem to be a decrease in RNA binding when the zinc knuckle is missing, the FLAG CPSF 30 (-ZK) complex (purified from stable cell line overexpressing FLAG CPSF 30 (-ZK)) was purified and used in a UV crosslinking assay. The data showed that that complex was still able to bind to RNA (data not shown). Therefore, zinc fingers 2 and 3 of CPSF 30 seem to contribute significantly to RNA binding.

**Zinc Fingers Two and Three of CPSF 30 are Needed for the CPSF Complex to Function**

To further analyze the roles of zinc finger two and three in RNA recognition, CPSF 30 complexes containing zinc fingers two and three mutations were made (Figure 3.8 and 3.9). Specifically, complexes overexpressing FLAG CPSF 73 and HA CPSF 30 with either its zinc finger two or zinc finger three deleted were made (Figure 3.8.A). To purify the complex containing HA CPSF 30 Full, FLAG immunoprecipitation was performed after nuclear extract was made (Figure 3.8.C top). For the complexes containing HA CPSF 30 without zinc finger two or three, NS1A was utilized to remove the endogenous CPSF 30. Previously, it has been shown that NS1A can bind to CPSF 30 via its zinc fingers two and three to block CPSF 30, and therefore the entire mRNA 3' processing
machinery, from processing cellular mRNAs (Twu et al. 2006). To accomplish this, first

A CPSF 30 WT/MT Domain Diagram

B Recombinant GST-NS1A

C Complex Purification Flow Scheme

D HA CPSF 30 Complexes after Purification

Figure 3.8 HA CPSF 30 Complex Purification

A. Diagram of mammalian CPSF 30 WT and MT domain structures tested. B. Purified recombinant GST-NS1A as shown in Coomassie Stain. C. Purification scheme of HA CPSF 30 Full, -ZF2 or -ZF3 from cell lines stably expressing FLAG CPSF 73 with the respective HA CPSF 30. D. The HA CPSF 30 Full, -ZF2 or -ZF3 complexes after purification. Left: Western Right: Quantitative Western
A UV Crosslink

HA CPSF 30 Full-ZF2-ZF3

Quantification

B Polyadenylation

HA CPSF 30 IN Full -ZF2-ZF3 PAP - + + +

Quantification

Figure 3.9 Mapping CPSF 30 RNA-Binding Domains via In Vitro Assays

A. *Left:* HA CPSF 30 Full, -ZF2 or –ZF3 complexes were incubated with 1000 fmol of 18 nt WT RNA derived from SVL in a UV crosslinking assay without RNase digestion. *Right:* Quantification of the UV crosslinked signal in A. Red represents CPSF 30 while blue represents WDR33. Standard deviations are also shown.

B. *Left:* In vitro polyadenylation assay with HA CPSF 30 Full, -ZF2 or –ZF3 complexes and PAP with 36 fmol of 18 nt WT RA derived from SVL. The lower band represents unprocessed RNA while the slower migrating smear represents polyadenylation RNA. *Right:* Quantification of the polyadenylated signal.
Recombinant GST-NS1A was made (Figure 3.8.B). Then GST-NS1A was incubated with nuclear extract made from cell lines overexpressing HA CPSF 30 without zinc finger two or three (Figure 3.8.C bottom). After the GST-NS1A pulldown, the supernatant was used in FLAG immunoprecipitation. After purification, the protein composition of each complex was visualized by performing Quantitative Western blotting analyses (Figure 3.8.D). The data demonstrated that regardless if zinc finger two or three was deleted, it did not affect the protein-protein interactions within the CPSF complex. Therefore, any difference seen in CPSF:RNA interaction is not due to the lack of CPSF subunits present. After the purification, the complexes were then used in UV crosslinking and in vitro polyadenylation assays. As expected, CPSF complexes containing HA CPSF 30 Full can bind (Figure 3.9.A) and polyadenylate (Figure 3.9.B) AAUAAA-containing RNA. CPSF complexes containing HA CPSF 30 without zinc finger two or three, however, were not able to bind nor polyadenylate AAUAAA-containing RNA as efficiently. Together, these experiments demonstrate that zinc fingers two and three are needed not only for AAUAAA recognition, but are also needed for the entire CPSF complex to function.

**CPSF 30 Zinc Fingers Two and Three May Be Needed for AAUAAA Recognition**

After determining that zinc fingers two and three of CPSF 30 are needed for RNA recognition, it was hypothesized that perhaps zinc fingers two and three are also needed for CPSF 30 to distinguish AAUAAA from mutant hexamers like AAAAAA. To test this hypothesis, the dual-reporter pPassport vector was used in a Luciferase assay (Figure 3.10.A). Two versions of this pPASPORT were made: one contained AAUAAA
Figure 3.10 Domain Specificity

A. Diagram of the pPassport vector. B. Diagram of mammalian CPSF 30 with its zinc fingers 2 and 3 swapped with yeast zinc fingers 2 and 3, respectively. C. Luciferase assay result. Each of the respective stable cell lines expressing HA CPSF 30 were transfected with the pPassport vector containing wildtype SVL (left) or mutant SVL (right).
at the multiple cloning site while another vector contained AAAAAA. Each construct was transiently transfected into HeLa cell lines that stably expressed either HA CPSF 30 Full, HA CPSF 30 –ZF2, HA CPSF 30 -ZF3 or HA CPSF 30 –ZK. In addition, each construct was transfected into cell lines that stably expressed HA CPSF 30 that contain the yeast ZF2 or ZF3 rather than the mammalian ZF2 or ZF3, respectively (Figure 3.10.B). The logic is if yeast uses A-rich sequences (rather than AAUAAA) as a signal for polyadenylation, then perhaps switching the mammalian ZF2/ZF3 for yeast ZF2/ZF3, respectively, can determine the CPSF 30 domain needed for AAUAAA recognition. After transfecting the wildtype and mutant pPASPORT into the stable cell lines, the cells were harvested and the renilla and firefly signals were measured using a luminometer. The results were graphed in Figure 3.10.C. Two possibilities can occur. One: if a strong polyadenylation signal is in the multiple cloning site or the mRNA 3’ processing complex is recruited efficiently to the PAS, then cleavage and polyadenylation will occur efficiently. As a result, the renilla signal would be high compared to the firefly signal. Two: if a weak polyadenylation site is in the multiple cloning site or if the mRNA 3’ processing complex is recruited inefficiently to the PAS, then the polyadenylation site 3’ of the firefly would be alternatively used. Consequentially, this would lead to a lower renilla to firefly ratio.

When the wildtype pPASPORT was transfected into the HA CPSF 30 mutant stable cell lines, the renilla to firefly ratio was lower than if wildtype pPASPORT was transfected into HA CPSF 30 Full cell line. This suggests that mutating the CPSF 30 domains can jeopardize CPSF binding to the RNA. In contrast, when pPASPORT containing the
mutant AAAAAA hexamer was transfected into the same HeLa stable cell lines, the
renilla to firefly ratio was different. The largest difference was seen when the mutant
pPASPORT was transfected into cell lines overexpressing HA CPSF 30 without zinc
finger 3 and with the yeast zinc fingers two and three. Specifically, the renilla to firefly
ratio is larger compared to the HeLa negative controls. This suggests that without
mammalian zinc fingers two and three, CPSF 30 can now polyadenylate the mutant
SVL RNA more efficiently than full-length CPSF 30. Together, this experiment provides
evidence that zinc fingers two and three are needed to specifically distinguish AAUAAA
from its close mutants.

Mapping the Footprints of CPSF 30, WDR33 and hFip1

In addition to mapping which CPSF subunit directly binds to AAUAAA, a footprint
experiment was done to determine which CPSF subunits bound up- and downstream of
the AAUAAA. To accomplish this, CPSF was UV crosslinked to RNA three, four or five.
RNA three, four or five were site-specifically radiolabeled at unique positions. For
example, RNA three was site-specifically radiolabeled thirteen nucleotides upstream of
the AAUAAA. After CPSF was UV crosslinking to RNA three, the sample was treated
with RNase 1. After digestion, only one protein, approximately 70 kDa, remain bound to
RNA three (Figure 3.11.A). Based on the size, it seems likely that this 70 kDa protein is
another CPSF subunit, hFip1. When RNA four (site-specifically radiolabeled seven
nucleotides upstream of the AAUAAA) was used in a similar experiment, the proteins
that remain bound to the RNA after RNase 1 treatment were identical to the proteins
that remain bound to RNA one. This suggests that CPSF 30 and WDR33 do not only
Figure 3.11 Mapping CPSF 30 and WDR33 Footprint

A. CPSF 30 and WDR33 footprints were mapped using a UV crosslinking assay. Site-specifically radiolabeled RNAs 1, 2, 3, 4 and 5 (0.7 fmol each) were incubated with 68 fmol of CPSF complex. The samples were then UV crosslinked and digested with 60 units of RNase 1. The samples were then resolved on a 10% SDS PAGE gel and imaged using phosphorimager. B. The results from A is illustrated.
bind to the AAUAAA specifically, but they also bind seven nucleotides upstream of the AAUAAA as well. RNA five (site-specifically radiolabeled nine nucleotides downstream of the AAUAAA) was then used in a similar experiment. Nine nucleotides downstream of the AAUAAA, WDR33 remain bound to RNA five after RNase 1 digestion. Together, this experiment mapped which protein bound directly at the AAUAAA and to the nearby cis elements. Specifically, this experiment demonstrated that CPSF 30 and WDR33 bind directly at the AAUAAA and seven nucleotides upstream of the AAUAAA. WDR33 (and not CPSF 30) also binds nine nucleotides downstream of the AAUAAA while hFip1 binds thirteen nucleotides upstream of the AAUAAA (Figure 3.11.B).

**PAS-SEQ Analyses**

Total RNA were extracted from HeLa cell lines overexpressing HA CPSF 30 Full, HA CPSF 30-ZF3, HA CPSF 30 Y3, FLAG NS1A, siCPSF 30 or siCPSF 160 using standard procedures (Figure 3.12.A). Libraries were then generated from these total RNA in preparation for a deep sequencing-based method called PAS-SEQ. PAS-SEQ is a protocol developed by the Shi lab to analyze polyadenylation on a global scale (Shepard et al. 2011). After generation of the libraries, they were sent to UCI Genomics High Throughput Facility for sequencing. Afterward, the results were analyzed as described in the Materials and Methods section of Chapter 3. From the analyses, some of the samples had only a handful of genes that displayed a significant APA shift. For example, the library generated from the HA CPSF 30-ZF3 cell line only had five genes with a P2D shift and four with a D2P APA shift (Figure 3.13.A) while FLAG NS1A
Figure 3.12 PAS-SEQ

A. The PAS-SEQ procedure is illustrated. B. Two of the samples sent to PAS-SEQ were libraries made from cell lines transiently expressing siRNA that targeted CPSF 160 or CPSF 30. The Western above shows the efficiency of the knock downs.
sample only had one gene with a P2D shift and two with a D2P shifts (data not shown). Similarly, HA CPSF 30 Y3 had nine genes with a P2D shift and eleven with a D2P shift (Figure 3.13.A). However the PAS-SEQ libraries generated from siCPSF 30 and siCPSF 160 cell lines showed many more significant APA shifts (Figure 3.12.B). Specifically, siCPSF 30 had seventy-five genes with a P2D shift and twenty-six genes with a D2P shift. siCPSF 160 had even more APA shifts; two hundred and seventeen genes had a P2D shift and three hundred and thirteen genes had a D2P shift. One possibility of why there were so few APA shifts in the HA CPSF 30 (-ZF3), HA CPSF 30 Y3 and NS1A samples is because those cell lines still contained the endogenous CPSF 30. With the endogenous CPSF 30 still present, the effects of the CPSF 30 mutants or NS1A could be masked.

Of the significant APA shifts, one particular result that was interesting was that without CPSF 30, APA occurs on the calmodulin gene. Calmodulin is a calcium binding messenger protein. Specifically, without siCPSF 30, the PAS of CALM2 (encoding calmodulin) shifts significantly from distal to proximal. This is interesting because it has been shown previously that ability of CPSF 30 to bind to RNA is regulated by calmodulin (Delaney et al. 2006). Specifically, in the presence of calcium, calmodulin inhibited CPSF 30 from binding to the RNA. It is interesting that not only can calmodulin affect CPSF 30 activity, but CPSF 30 can also affect calmodulin expression. Further analyses are needed to verify this PAS-SEQ data.
Figure 3.13 PAS-SEQ Results

Significant APA changes from the PAS-SEQ analyses are illustrated in the scatter plots here. Blue dots represent P2D shifts while red dots represent D2P shifts. A. Left HA CPSF 30 (-ZF3). Right HA CPSF 30 Y3. B. Left siCPSF 30. Right siCPSF 160
A sample of particular interest was the library generated from HA CPSF 30 Y3 cell line. Because this CPSF 30 contains the yeast ZF3 instead of the mammalian ZF3, it is hypothesized that with the yeast ZF3, now CPSF 30 can no longer recognize AAUAAA. Other samples such as HA CPSF 30 –ZF3 might generate similar results but it is more likely that the structure of CPSF 30 remains intact with the yeast ZF3 swap rather than with the ZF3 deletion. From the HA CPSF 30 Y3 PAS-SEQ data, only 20% of the genes with significant APA shifts occurred on genes that did not contain AAUAAA. *In vitro* experiments need to be done to determine the validity of these results.

### 3.5 Conclusions and Discussion

In this chapter, it is demonstrated that CPSF 30 and WDR33 directly bind to the AAUAAA (Figure 3.14). The footprints of CPSF 30 and WDR33 were similar regardless if the RNA used in the mapping experiments was derived from adenovirus or from SVL. Specifically, CPSF 30 bind directly AAUAAA and seven nucleotides upstream of the AAUAAA. Meanwhile, WDR33 is also needed to directly interact with AAUAAA by binding to the AAUAAA, seven nucleotides upstream of the AAUAAA and nine nucleotides downstream of the AAUAAA. The interaction between CPSF 30 and the AAUAAA was further analyzed. Since it has been unknown thus far that CPSF 30 contributes to AAUAAA recognition directly, the domains necessary for CPSF 30 to interact with the RNA were mapped. Zinc fingers two and three are needed for CPSF 30 to directly interact with AAUAAA, for the entire CPSF complex to function and to possibly even distinguish the wildtype hexamer from a mutant hexamer. This data
Figure 3.14 Model

Summary Model. Both CPSF 30 and WDR33 bind directly to AAUAAA while hFip1 binds somewhere nearby. CPSF 30 binds to AAUAAA via its zinc fingers 2 and 3.
detailing the importance of zinc fingers two and three in AAUAAA recognition agrees
nicely with previous data showing that NS1A inhibits cellular polyadenylation by binding
to specifically zinc fingers two and three (Twu et al. 2006). Since the zinc fingers two
and three have such an important role in mRNA 3’ processing, it makes logical sense
that NS1A would target those structural domains. Although CPSF 30 did not affect APA
significantly, this chapter provided a novel, detail understanding and appreciation for the
roles of CPSF 30 and WDR33 in AAUAAA recognition.

It is surprising that CPSF 30 and WDR33 have not been implicated as AAUAAA binding
proteins earlier. One possibility for the slow development is that in the past it was a
discrepancy whether or not CPSF 30 was a bona fide CPSF subunit (Bienroth et al.
1991) (Murthy et al. 1992). Similarly, it was unknown until fairly recently that WDR33
was a component of the mammalian mRNA 3’ processing complex (Shi et al. 2009). In
addition, the similar sizes of WDR33 (146 kDa) and CPSF 160 (160 kDa) probably led
to the slow discovery of the role of WDR33 in AAUAAA recognition.

However, there are currently still some unanswered questions. First is, does the zinc
knuckle domain of CPSF 30 play a role in RNA binding? In the UV crosslinking
experiment with transient transfected FLAG CPF 30 (-ZK), the zinc knuckle does seem
to be needed for AAUAAA binding. In addition, the luciferase assay shows that when
pPassport containing the wildtype hexamer is transfected into HA CPSF 30 (-ZK) cell
line, renila:firefly signal is low. This suggests that the mRNA 3’ processing complex
cannot be recruited to the RNA as well without zinc knuckle. However, when the
complex containing FLAG CPSF 30 (-ZK) was made, the complex can bind to the RNA (data not shown). One explanation for this discrepancy is that the zinc knuckle behaves differently when it is in the cytoplasm compared to when it is in the nucleus. During transient transfection, FLAG CPSF 30 (-ZK) is mostly expressed in the cytoplasm. In the cytoplasm there are not many other mRNA 3’ processing factors relative to the nucleus. Thus, FLAG CPSF 30 (-ZK) will not have as many protein:protein interactions in the cytoplasm and the inability to bind to RNA cannot be compensated by another RNA:binding protein. Contrary, when stable cell lines were made, now FLAG CPSF 30 (-ZK) is expressed mostly in the nucleus. In the nucleus, there are many mRNA 3’ processing factors for FLAG CPSF 30 (-ZK) to interact with and thus the RNA binding can be compensated by another RNA binding protein that can also bind to FLAG CPSF 30 (-ZK). Based on the domain structure of CPSF 30, it makes sense that its zinc knuckle can contribute in AAUAAA recognition. The zinc knuckle is not a conserved domain. The species that do have zinc knuckles are mammals, *Mus musculus* and *D. melanogaster*. The species that do not have zinc knuckles are *A. thaliana* and *S. cerevisiae*. Interestingly, *A. thaliana* and *S. cerevisiae* do not strictly polyadenylate AAUAAA-containing RNA; many other AAUAAA-variants are naturally used in *A. thaliana* and *S. cerevisiae* as poly (A) cis elements (Graber et al. 1999). On the other hand, mammals, *Mus musculus* and *D. melanogaster* have a higher percentage of RNAs in their cells that contain AAUAAA. This correlative-based hypothesis can be tested by doing PAS-SEQ with a normal *S. cerevisiae* strain compared to a *S. cerevisiae* strain that overexpressed the zinc knuckle of FLAG CPSF 30. If the zinc knuckle does play a role in AAUAAA recognition, than the strain containing the zinc
knuckle can now polyadenylate AAUAAA-containing RNA at a higher frequency than the normal strain.

Another question that needs to be addressed is why are so few genes altered globally by CPSF 30. One possibility is that although mutant CPSF 30 were overexpressed in the cell lines, the endogenous CPSF 30 might mask the APA changes. In that situation, one future experiment is to generate PAS-SEQ libraries from cell lines that lack the endogenous CPSF 30 but overexpresses the mutant CPSF 30. This might generate a larger APA affect.

Ultimately, the most important question is: what components are needed to reconstitute specificity? Reconstituted proteins were utilized to address this question. With recombinant CPSF 30 by itself, CPSF 30 was not able to distinguish AAUAAA from AAGAAAA. However, CPSF 30 can distinguish an RNA containing CCCCCC from AAUAAA (data not shown). Reconstitution with recombinant CPSF 30 and CPSF 160 was also done. However, specificity was not reconstituted with these proteins (data not shown). The next experiment would be to make recombinant WDR33 and perform a UV crosslinking experiment with recombinant CPSF 30 and WDR33 together. This should determine if WDR33 and CPSF 30 can together recognize AAUAAA specifically. If specificity can still not be regenerated, perhaps a third CPSF subunit is need. The third factor might be CPSF 160 or hFip1 because although they don't bind directly to AAUAAA, they do bind nearby as shown by the mapping experiments above. Perhaps the protein:protein interactions of CPSF 160 and hFip1 are needed for specific AAUAAA
recognition by CPSF 30 and WDR33. Another possibility is that other non-RNA binding CPSF subunits such as Symplekin, CPSF 100 and CPSF 73 are needed to help position CPSF 30 and WDR33. The third possibility is that perhaps mRNA 3’ processing is more similar to splicing than previously thought. Perhaps, in addition to RNA:protein interactions, RNA:RNA interactions are also crucial for the reaction to occur. One piece of data suggests that this might be the case (data not shown). Within the purified CPSF complex, an RNA strand (roughly 75 nucleotides) was found. While this is preliminary data, this brings excitement that perhaps CPSF 30 and WDR33 need RNA as a guide in AAUAAA recognition.

In the next chapter, we will see that AAUAAA recognition can be a bit more challenging for the CPSF complex to distinguish when other cis elements are present in the near vicinity. In order for mammalian mRNA 3’ processing to still occur specifically, AAUAAA recognition needs to be regulated by an unidentified protein present in the nuclear extract.
Chapter 4

Regulation of CPSF-AAUAAA Interaction

4.1 Summary

AAUAAA is needed for both the cleavage and the polyadenylation step to occur. The CPSF complex, specifically CPSF 30 and WDR33, recognizes AAUAAA with such high specificity that a mutation of the U to any other nucleotide leads to mRNA 3’ processing deficiency \textit{in vivo} and \textit{in vitro}. However, it is unknown whether this important interaction can be regulated. In this chapter, it is shown that when auxiliary cis elements are present, CPSF can no longer recognize the AAUAAA specifically. Therefore, a nonspecificity factor is responsible for the lack of specificity. The central hypothesis of this chapter is that the nonspecificity factor is a CPSF subunit and a specificity factor is
present to regulate the nonspecificity factor. To address this hypothesis, two specific aims were developed.

Specific aim 1: Identify and Characterize the Nonspecificity Factor.
Specific aim 2: Identify and Characterize the Specificity Factor.

In this chapter, correlative evidence is provided showing that the nonspecificity factor is hFip1. The specificity factor on the other hand is a factor present in HeLa nuclear extract. Fractionation procedures were attempted to identify the specificity factor. While the identity of the nonspecificity and specificity factor were not conclusively determined, this chapter provides initial understanding in how the CPSF-AAUAAA interaction can be regulated.

4.2 Introduction

mRNA 3’ processing requires an extensive amount of protein:protein and RNA:protein interactions (Colgan et al. 1997) (Chan et al. 2010) (Zhao et al. 1999). The core cis elements required for mRNA 3’ processing to occur is the AAUAAA element and the U/GU rich element located downstream of the cleavage site. The AAUAAA element is the most crucial element because it is needed for both steps of the reaction such that a mutation of the hexamer could abolish mRNA 3’ processing (Connelly et al. 1988) (Fitzgerald et al. 1981) (Montell et al. 1983, Manley et al. 1985) (Zarkower et al. 1986) (Conway et al. 1987) (Sheets et al. 1990) (Wilusz et al. 1989). Previously, researchers demonstrated that partially purified or highly purified CPSF is responsible in recognizing AAUAAA specifically (Bienroth et al. 1991) (Murthy et al. 1992) (Keller et al. 1991)
A variety of RNAs and assays were used to test the specificity of the CPSF complex. For example, Bienroth, Wahle et al. in 1991 used an adenovirus RNA that terminated one nucleotide upstream of the natural cleavage site and used AAGAAA as their mutant. They determined that their highly purified CPSF was specific by competition UV crosslinking and gel shift assays. Similarly, they used a shorter SVL RNA and their experiment showed that CPSF can bind and function specifically on the AAUAAA-containing RNA but not the mutant (AAGAAA) RNA (Keller et al. 1991). Contrary, Murthy and Manley in 1992 were not able to see any RNA binding with highly purified CPSF complex. They did, however, see specific in vitro polyadenylation with a RNA derived from SVL. Together, these data show that the interaction between CPSF and the AAUAAA needs to be further studied.

The CPSF complex I purified is different from previous researchers’ in two ways. First the purification procedure is more efficient. Previously, researchers purified the CPSF complex using chromatography, which can be tedious and irreproducible (Gilmartin et al. 1989) (Gilmartin et al. 1991) (Bienroth et al. 1991) (Murthy et al. 1992). Instead of using chromatography to purify the CPSF complex, affinity immunoprecipitation was used here that allowed the purification to be higher in efficiency and reproducibility. The second difference between the CPSF complex used here and the CPSF complexes purified by other laboratories is that the CPSF complex here contains all of the known CPSF subunits (Shi et al. 2009). Using the CPSF complex, I have already redefined mRNA 3’ processing mechanism by discovering that CPSF 30 and WDR33 directly recognize AAUAAA (see Chapter 3). In this chapter, it is shown that the interaction
between CPSF and the AAUAAA is no longer specific when auxiliary cis elements are present. The central hypothesis is that one of the CPSF subunits is a nonspecificity factor and that there is a specificity factor that regulates this nonspecificity factor. Here, hFip1 is suggested to be the nonspecificity factor. Furthermore, biochemical characteristics of the specificity factor are identified. Together, this provides for the first time, evidence that the CPSF-AAUAAA interaction can be regulated.

4.3 Materials and Methods

RNA Substrates

*Long RNAs*

Two populations of long RNAs are made in this chapter. One population was derived from SVL while the other was derived from adenovirus RNA. Both sequences were cloned into the pGem plasmid. SVL-pGem was restriction digested with the DraI restriction enzyme while the L3-pGem was restriction digested with the Ava1 restriction enzyme. The linear DNAs were then used in an *in vitro* transcription assay with [³²P-UTP] to body radiolabel the newly transcribed RNAs. After the run-off *in vitro* transcription, each of the long RNAs were gel purified and quantified using a scintillation counter. The long RNAs used in this chapter have the following sequences:

SVL:

5'AGACA*UGA*UAAGA*UACA*U*UGA*UGAG*U*UGGACAAACCACAAC*UAGAA*
UGCAG*UGAAAAAAA*UGC*U*U*UA*U*UG*UGAAAA*U*U*UG*UGA*UGC*UA*U*UGC*U*UA*U*UG
C*U*U*UA*U*U*UG*UAACCA*U*UA*UAAGC*UGC*AA*UAAAACAAG*U*UAACAAACAC
AA*U*UGCA*U*UCA*U*U*U*UA*UG*U*U*U*UCAGGC*U*UCAGGGGAGG*UG*UGGGA
GG*U*U*U*U*U*U

(191 nucleotides)

* represents the radiolabeled position

**Bold** nucleotides represent the wildtype of mutant hexamer

L3:

5’*U*UC*U*U*U*U*UG*UCAC*U*UGAAAAACA*UG*UAAAAA*UAA*UG*UAC*UGGAG
ACAC*U*UCAA*UAAAAGGCAA*UG*U*U*U*UA*U*U*UG*UACAC*UC

(84 nucleotides)

* represents the radiolabeled position

**Bold** nucleotides represent the wildtype of mutant hexamer

L3 (-U rich sequences):

5’GAAAAACA*UG*UAAAAA*UAA*UG*UAC*UAGGAGACAC*U*U*UCAA*UAAAAGGCA
AAACAC*UC

(56 nucleotides)

* represents the radiolabeled position

**Bold** nucleotides represent the wildtype of mutant hexamer
Short RNAs

The short eighteen nucleotides SVL RNAs made in this chapter were identical to the ones used in Chapter 3. These RNAs were made by radiolabeling the 5' end of the chemically-synthesized RNA oligos with $[\gamma^{-32}P]$ ATP and Optikinase using the standard protocol from Affymetrix. The 5' radiolabeled, eighteen nucleotides RNAs utilized in this chapter have the following sequences:

18 nucleotides WT: 5' *CUGCAAUAAAACAAGUUAA
18 nucleotides MT: 5' *CUGCAGAAAACAAGUUAA

* represents the radiolabeled position.

**Bold** nucleotides represents the wildtype or mutant hexamer

Protein Complexes

The CPSF Complex

The CPSF complex described here is identical to the one described in Chapter 3. HEK 293 cells was transfected with 3x FLAG CPSF 73 pCMV14 using Lipofectamine. The stable transfectants were selected with neomycin/G418 (1 mg/ml) for two weeks. After selection, individual clones were picked and expanded. Western blotting analyses were then carried out with each clone to monitor the expression levels of the FLAG-tagged exogenous proteins. The clones were all screened and those that expressed the FLAG-tagged exogenous proteins near the endogenous protein levels were selected for further experiments. After the stable tranfectants were selected, the cells were expanded and nuclear extract was generated using standard procedures. The nuclear
extract was then incubated with Anti-FLAG M2 Affinity beads for 2 hours in 4° C. After incubation, the supernatant was discarded. The beads were then washed and eluted under identical conditions as mentioned above. The eluted FLAG CPSF 73 complex was then used for *in vitro* assays.

*Endogenous WDR33 Containing Complexes*

Protein A/G beads were conjugated to 5 µg of anti-WDR33. The beads were then washed with Buffer D 100 before incubated with large scale nuclear extract for 2 hours in 4° C. The beads were then washed with Buffer D 300 for 5 minutes, Buffer D 100 for 5 minutes and finally, Buffer D 50 (same composition as Buffer D 300 except it has 50 mM NaCl) for 5 minutes.

*In Vitro Assays*

*Gel Shift Assay*

Ninety six nanograms of CPSF were incubated with 70 fmol of RNA in the presence of 1.6 mM ATP, 0.1 mg/ml tRNA, 0.02 M CP, 0.2 mg/ml BSA, 0.01% NP40, RNase OUT, 10 mM Hepes, 50 mM NaCl, 0.5 mM MgCl2, 0.1 mM EDTA, 5% glycerol for 10 minutes at 30° C. The sample was then loaded onto a 4% native polyacrylamide gel and imaged with a phosphorimager.

*UV Crosslinking*

Ninety six to a hundred and fifty nanograms of CPSF was incubated with 70-600 fmol of RNA in the presence of 0.02 M CP, 1-1.6 mM ATP, 0.1 mg/ml tRNA, 0.2 mg/ml BSA, 6-
10 mM Hepes, 30-50 mM NaCl, 0.3-0.5 mM MgCl₂, 0.06-0.1 mM EDTA, 3-5% glycerol, 5 mM beta mercaptoethanol. The reaction was incubated at 30° C at 10 minutes followed by UV crosslinking at 254 nm for 10 minutes (240,000 microjoules). The UV crosslinked sample was then treated with or without RNase A/T mix for 30 minutes at 37° C. The sample was finally loaded on a 10% SDS PAGE gel and imaged with a phosphorimager.

In Vitro Polyadenylation

HeLa nuclear extract or 50-100 ng of CPSF and 10-20 ng of PAP were incubated in the presence of 70-80 fmol of RNA, 0.8-1 mM ATP, 0.1 mg/ml tRNA, 0.02 M CP, 6-10 mM Hepes, 30-50 mM NaCl, 0.44-1 mM MgCl₂, 0.06-0.1 mM EDTA, 3-5% glycerol and 5 mM beta mercaptoethanol for 1.5 hours at 30° C. The samples were then treated with proteinase K and phenol chloroform. The RNA was ethanol precipitated before resolved on a 8% urea gel. The image was visualized using phosphorimager.

In Vitro Polyadenylation on beads

The beads with the respective CPSF complex were incubated with 1 mM ATP, 0.1 mg/ml tRNA, 0.02 M CP, 2.5% SA PVA, 8.8 mM Hepes, 44 mM NaCl, 1 mM MgCl₂, 0.08 mM EDTA, 4.4% glycerol, 100 ng PAP, RNA for 1.5 hours at 30° C in a thermomixer. The reactions were mixed periodically at 1000 rpm.
Protein Fractionation

Ammonium Sulfate Fraction

Large scale nuclear extract was heated at 49 °C for 15 minutes. The heated nuclear extract was then fractionated using ammonium sulfate. Different percentages of ammonium sulfate were added to the heated nuclear extract, fractionating the heated nuclear extract into the corresponding fractions (0-20%, 20-40%, 40-60%, > 60%).

Ion Exchange Chromatography

Large scale nuclear extract was dialyzed such that the final concentration was 50 mM Tris HCl pH = 7, 50 mM NaCl, 1 mM MgCl2, 0.2 mM EDTA, 0.5 mM PMSF, 0.5 mM DTT. Afterwards, the nuclear extract was heated at 49 °C for 15 minutes before it was centrifuged. The supernatant was then filtered through 0.2 µm filter. The heated nuclear extract was fractionated using IEX Q FF or the IEX SP FF Sepharose column. The sample was bound to the beads using a binding buffer with the composition of 50 mM Tris pH = 7, 50 mM NaCl, 1 mM MgCl2, 0.2 mM EDTA. The samples were then eluted from the beads with the same composition as the binding buffer above except it contained 1 M NaCl. The fractions were then dialyzed into Buffer D 100 before use in in vitro assay.
4.4 Results

CPSF Cannot Recognize AAUAAA When Long Substrates are Used

In Chapter 3, it was shown that CPSF purified from FLAG CPSF 73 stable cell lines was specific when splint RNAs were used (Figure 3.2.A). However, here the data shows that CPSF behaves differently in the presence of long RNAs (Figure 4.1.A). When long SVL RNA was incubated with CPSF in a UV crosslinking assay, the data shows that CPSF can bind to long AAUAAA-containing SVL as well as to long SVLs containing AAAAAA, AACAAA or AAGAAA (Figure 4.2.A). Based on the UV crosslinking pattern, four proteins bound to wildtype or mutant long SVL. Based on the sizes of the bands, the proteins mostly likely correspond to WDR33, CPSF 100, hFip1 and CPSF 30. When the CPSF complex was incubated with long adenovirus RNA in a gel shift assay, again, the experiment shows that CPSF cannot distinguish between the wildtype and the mutant hexamer (Figure 4.2.B). The next question was, can CPSF function specifically on an RNA containing the wildtype hexamer? To determine this, CPSF and PAP were incubated with long adenovirus RNA containing AAUAAA or AAGAAA hexamer in an in vitro polyadenylation assay (Figure 4.2.C). Similar to the UV crosslinking and the gel shift assay, no difference was seen between the interaction of the CPSF complex with the wildtype or the mutant hexamer containing RNA. When in vitro polyadenylation was expanded to include other mutants including AAAAAA, AACAAA and no hexamer altogether, polyadenylation was still seen with CPSF and PAP (Figure 4.2.D). Together, this demonstrates that although CPSF can specifically recognize AAUAAA when short RNAs (such as splint RNAs) are used, CPSF can no longer recognize AAUAAA by itself when longer substrates are used.
Figure 4.1 RNAs Utilized

A. Illustration of the long RNAs used in this project. Two long RNAs were used – one was derived from SVL and one was derived from adenovirus. Both RNAs contained the AAUAAA hexamer (or the mutant as described in each figure) and the cleavage site (CA). Each of the long RNAs used were body radio-labeled. B. The short RNA used in this project was derived from adenovirus. It contains the AAUAAA (or the corresponding hexamer mutant as described in each figure). The short RNA was 5’
Figure 4.2 CPSF is Not Specific on Long Adenovirus or SVL RNAs

A. Wildtype long SVL RNA and its respective hexamer mutants (AAAAAAA, AACAAA, AAGAAAAA) were incubated with CPSF in a UV crosslinking assay. The samples were loaded on a 10% SDS PAGE after UV crosslinking and RNase A/T1 digestion. B. Wildtype long adenovirus RNA and the corresponding mutant containing AAGAAA were incubated with CPSF and loaded onto a 4% non-denaturing polyacrylamide gel in a gel shift assay. C. Wildtype long adenovirus RNA and the corresponding mutant containing AAGAAAAA were incubated with CPSF and/or PAP in an in vitro polyadenylation assay. D. Wildtype long adenovirus RNA and the corresponding mutants (AAAAAAA, AACAAA, AAGAAAAA, or lacking the entire AAUAAA) were incubated with CPSF and/or PAP in an in vitro polyadenylation assay in the presence of 0.005 mg/ml BSA and 0.01% NP40.
**CPSF Nonspecificity is Not Due to Technical Issues**

CPSF nonspecifically binding to long RNAs was surprising. One possible explanation is that there were technical issues that caused this nonspecificity. More specifically, perhaps the CPSF was purified under conditions that were not stringent enough; perhaps the nonspecificity was purified with the CPSF complex. Therefore, higher amounts of salt were utilized to purify CPSF without the nonspecificity factor. The CPSF was then used in an *in vitro* polyadenylation assay with long adenovirus RNA containing the wildtype of mutant hexamer (Figure 4.3.B). However, the CPSF purified with higher amounts of salt did not specifically polyadenylate the wildtype long RNA. Another possibility is that during the CPSF complex purification, crucial post-translational modifications required for the CPSF complex to specifically recognize AAUAAA were interrupted. To test that hypothesis, the CPSF complex was purified in the presence of phosphatase inhibitor to preserve possible phosphorylations needed for specificity to occur (Figure 4.3.A). The CPSF complex was then used in an *in vitro* polyadenylation assay with recombinant PAP to determine if specificity can occur. However, the *in vitro* polyadenylation assay showed that specificity can still not be achieved. The third possibility is that since the CPSF complex was purified using FLAG beads targeting the stably overexpressed FLAG CPSF 73 subunit, perhaps the FLAG tag is somehow causing the nonspecificity. The CPSF complex was thus immunoprecipitated using an antibody targeting the endogenous WDR33. The complex was then used in an *in vitro* polyadenylation assay with recombinant PAP with wildtype and mutant hexamer containing long RNAs (Figure 4.3.C). However, specific interactions between CPSF and
Figure 4.3 CPSF Nonspecificity is Not Due to Technical Issues

A. CPSF was treated with phosphatase inhibitor and incubated with PAP and long adenovirus RNA (MT = AAGAAAA) in an in vitro polyadenylation assay on beads. B. CPSF was purified with 300, 500 or 1000 mM NaCl from stable cell lines expressing FLAG CPSF 73. The complexes were incubated with PAP and long adenovirus RNA (MT = AAGAAAA) in an in vitro polyadenylation assay on beads. C. CPSF was immunoprecipitated using antibody targeting WDR33. That CPSF and PAP were then incubated with long adenovirus (MT = AAGAAAA) in an in vitro polyadenylation assay on beads with 0.005 mg/ml BSA and 0.01% NP40.
the AAUAAA still did not occur. Together, these experiments demonstrate that the nonspecificity is not due a technical error.

**HeLa Nuclear Extract Contains the Specificity Factor**

Interestingly, when HeLa nuclear extract was incubated with RNA containing wildtype or the mutant hexamer in an *in vitro* polyadenylation assay, specific polyadenylation was achieved (Figure 4.4.A). This suggests that while the CPSF complex contains a factor that is responsible for the nonspecific recognition of the AAUAAA, there is a regulatory factor present in the HeLa nuclear extract that can alleviate the nonspecificity. The goal was then to identify this regulatory factor in the HeLa nuclear extract by fractionation. The first fractionation step was to heat the nuclear extract. After heating the nuclear extract, the polyadenylation activity is significantly decreased because the heat destroyed crucial enzyme activities (figure not shown). The heated nuclear extract was then added to CPSF and PAP. Surprisingly, specificity can now be restored (Figure 4.4.B). This suggests that the specificity factor in the HeLa nuclear extract is insensitive to heat. The heated HeLa nuclear extract was then further fractionated using different percentage of ammonium sulfates. The heated HeLa nuclear extract was fractionated into 0-20%, 20-40%, 40-60% and > 60% ammonium sulfate fractions. These fractions were then added to CPSF and PAP in an *in vitro* polyadenylation assay (Figure 4.4.C). One fraction (20-40%) was distinctively specific when added to CPSF and PAP. This suggests that the specificity protein can be precipitated with 20-40% ammonium sulfate. What is also interesting is that with the 20-40% ammonium sulfate fraction, CPSF and PAP, the polyadenylation pattern is unusual; the poly (A) tail generated was less
Figure 4.4 Purification of the Specificity Factor from HeLa Nuclear Extract

A. Nuclear extract was added to long adenovirus RNA that either contained the wildtype hexamer or AAAAA, AACAAA, AAGAAA or missing the entire hexamer in an in vitro polyadenylation assay. B. Different microliters of nuclear extract heated at 49°C for 15 minutes were added to wildtype or mutant (AAGAAA) long adenovirus RNA in an in vitro polyadenylation assay with CPSF, PAP, 0.01% NP40 and RNase OUT. C. Nuclear extract heated at 49°C for 15 minutes were fractionated using ammonium sulfate salt such that the final concentrations were 0-20%, 20-40%, 40-60% or >60% ammonium sulfate. Each of the fractions were then added to CPSF and PAP in an in vitro polyadenylation assay with 0.01% NP40. RNase OUT, wildtype or mutant (AAGAAA) long adenovirus RNA.
efficient compared to normal polyadenylation as shown by the ladder-like pattern. This suggests that the specificity factor might also affect polyadenylation efficiency in addition to specificity. Next, the specificity factor was further purified by performing chromatography. An ion exchange chromatography with IEX Q Fast Flow Sepharose beads was used. This column is a strong anion exchanger. Thus, any protein that binds to this column will have a net negative charge on its surface while any protein that is in the flow through will have a net positive charge on its surface. After separation of the heat nuclear extract with IEX Q Fast Flow Sepharose beads, the proteins that were bound to the column or were in the flow through were incubated with CPSF and PAP in an *in vitro* polyadenylation assay with long adenovirus RNA (Figure 4.5.A). From the *in vitro* polyadenylation assay, the specificity factor was present in the flow through fraction. This suggests that the specificity factor has an overall positive charge on its protein structure since it did not bind to the IEX Q Fast Flow Sepharose beads. Next a column consisted of IEX SP Fast Flow Sepharose beads was used. This column, in contrast to the IEX Q column, is a strong cation exchanger. Thus, any proteins that bound to these beads would have a net positive charge on its surface. However, if the proteins cannot bind to the beads, then those proteins will have an overall negative charge on its surface. After separation of the heat nuclear extract with IEX SP Fast Flow Sepharose beads, the proteins that were bound to the column or were in the flow through were added to CPSF and PAP in an *in vitro* polyadenylation assay with long adenovirus RNA (Figure 4.5.B). This time, the specificity factor remained bound to the beads suggesting again that the specificity factor has an overall positive charge on its surface.
Figure 4.5 Chromatography Can Partially Isolate the Specificity Factor

A. Nuclear extracts heated at 49 °C for 15 minutes were fractionated using IEX Q FF column. The flow through (FT) and proteins bound to the beads after wash (B) were added to CPSF and PAP in an in vitro polyadenylation assay with long adenovirus wildtype or mutant (AAGAAA) RNA with 0.01% NP40 and RNase OUT. B. Nuclear extract heated at 49 °C for 15 minutes were fractionated using IEX SP FF column. The flow through and proteins bound to the beads were added to CPSF and PAP in an in vitro polyadenylation assay with 0.01% NP40, RNase OUT, long adenovirus wildtype or mutant (AAGAAA) RNA.
Together, some key biochemical characteristics about the specificity factor were discovered. First, the specificity factor is present in HeLa nuclear extract. Second, the factor is insensitive to heat, suggesting that perhaps it does not contain enzymatic capabilities. Third, this specificity factor can be found in a fraction that contains 20-40% ammonium sulfate. Fourth, the surface of the specificity factor has an overall positive charge. While the identity of this specificity factor is still unknown and more experiments need to be done, the experiments above provided a good starting point in identifying the specificity factor.

The CPSF Complex is Specific on Short RNAs

As shown in Chapter 3, the CPSF complex can be specific if the RNA used is short, such as with the splint RNAs. Here I took a short piece of RNA derived from adenovirus (Figure 4.1.B) and used it in various in vitro assays to determine if the CPSF complex can now recognize this short RNA. First, a gel shift assay was performed (Figure 4.6.A). This experiment demonstrated that CPSF can now recognize and bind the short adenovirus RNA specifically. In addition, the CPSF complex and PAP can also function specifically on wildtype short adenovirus RNA in an in vitro polyadenylation assay (Figure 4.6.B). Together this suggests that while CPSF cannot recognize the AAUAAA element when long RNAs are used, CPSF can specifically recognize the AAUAAA hexamer when short RNAs are used.
Figure 4.6 Nonspecificity Factor is a CPSF Subunit

A. CPSF was incubated with short RNA in a gel shift assay. The short RNAs used contained either the wildtype hexamer or AAGAAA. B. CPSF and PAP were incubated in an in vitro polyadenylation assay with wildtype or mutant (AAGAAA) short RNA in the presence of 0.2 mg/ml BSA, 0.01% NP40 and RNase OUT. C. Left: CPSF was incubated with short wildtype or mutant (AAGAAA) RNA in a UV crosslinking assay with RNase OUT and 0.01% NP40. Right: Long wildtype and mutant adenovirus RNA were UV crosslinked to CPSF as shown previously.
hFip1 Might be the Nonspecific CPSF Factor

Lastly, a UV crosslinking assay was performed to observe the proteins responsible for specific interaction between CPSF and the AAUAAA (Figure 4.6.C left). The CPSF UV crosslinked to the wildtype short adenovirus RNA specifically with the help of two proteins. Those two proteins are WDR33 and CPSF 30. This UV crosslinking pattern was different from when CPSF was UV crosslinked to long RNAs (Figure 4.6.C right). When the CPSF complex was UV crosslinked to the long RNA, four proteins bound to the RNA. Similar to the UV crosslinking pattern with the short RNA, WDR33 and CPSF 30 bound to the RNA. However, now with the long RNA, two extra proteins were present. Those proteins are approximately 100 kDa and 70 kDa. Based on the size of these proteins, it seems likely that these proteins are two other CPSF subunits, CPSF 100 and hFip1, respectively. Since CPSF 100 and hFip1 are bound to the RNA when CPSF is not specific (but not bound when CPSF is specific), this suggests that CPSF 100 or hFip1 is the nonspecificity factor when auxiliary cis elements are present.

While not much is known about CPSF 100, binding preferences of hFip1 are known. It was previously shown that hFip1 preferentially binds to U-rich sequences (Kaufmann et al. 2004). Therefore, one hypothesis is that if the U-rich sequences are removed, perhaps now the CPSF complex can specifically recognize AAUAAA in the presence of auxiliary cis elements. To test this hypothesis, two regions of U-rich sequences were deleted from the long adenovirus RNA (Figure 4.7.A). CPSF and PAP was then used in an in vitro polyadenylation assay with this RNA (Figure 4.7.B). The data shows that regardless what RNA was used, CPSF and PAP can polyadenylate both wildtype and
Figure 4.7 Deleting hFip1 Sequences Did Not Affect Specificity

A. Top: Illustration of long adenovirus RNA. Bottom: Illustration of long adenovirus RNA without 2 U-rich regions. The first region contains the sequence 5’UUCUUUUUGUCACUU and the second region contains the sequence 5’UGUUUUUAUUUGU. B. CPSF and PAP were incubated with wildtype or mutant (AAGAAA) short, long adenovirus or long adenovirus (-U rich sequences) in an in vitro polyadenylation assay with 0.2 mg/ml BSA, 0.01% NP40 and RNase OUT.
mutant RNA. This suggests that either hFip1 is not the nonspecificity factor or the binding site of hFip1 has not been completely eliminated.

4.5 Conclusions and Discussion

In Chapter 3 and 4, evidence was provided that our highly-purified CPSF can recognize the AAUAAA specifically when short RNAs, such as the eighteen nucleotide SVL or splint RNAs, were used. Interestingly, the same highly-purified CPSF can no longer recognize AAUAAA specifically when longer RNAs were used. From the UV crosslinking experiments comparing the binding patterns of CPSF complex with the shorter RNA versus the longer RNA suggested that hFip1 and/or CPSF 100 might contribute to the nonspecificity. However, there is at least one factor in the HeLa nuclear extract that can regulate this nonspecificity. Although the identify of the specificity factor is now known, some biochemical characteristics of this factor has been discovered. First, the specificity factor is not affected by heat at 49 °C (although this temperature did abolish PAP activity in HeLa cells). This suggests that the specificity factor is probably not an enzyme. Second, it seems to have a relatively strong positive charge at the surface of the protein as shown by the ammonium sulfate fractionation and the chromatography experiments.

The identity of the nonspecificity factor is still unknown. There are, however, speculations of its identity. Comparing the UV crosslinked pattern of the CPSF complex on short RNAs versus on the long RNAs, two proteins bound to the long RNA but not to
the short RNA. Based on their sizes, those two proteins are most likely CPSF 100 and hFip1. However, the identity of these proteins need to be determined by immunoprecipitation under denaturing conditions. This led to the conclusion that either or both of these proteins are indirectly or directly responsible for the nonspecificity. Not a lot is known about the function of CPSF 100 except for its highly similar structural domain to the endonuclease of mRNA 3’ processing, CPSF 73. In addition, it is known that CPSF 100 is a core factor of the mRNA 3’ processing complex and has a role in histone polyadenylation (Mandel et al. 2006) (Jenny et al. 1994) (Dominski et al. 2005) (Marzluff et al. 2008). It is possible that CPSF 100 plays a role in the nonspecificity. However, the current hypothesis is that hFip1 is the nonspecificity factor. It is hypothesized that the nonspecificity factor can interact with the RNA and have the capacity to recruit PAP to the RNA nonspecifically. The current knowledge of hFip1 fits these criterions. hFip1 has been shown to be one of the three RNA-binding subunit of the CPSF complex while CPSF 100 is not known to be an RNA-binding protein (Zhao et al. 1999) (Chan et al. 2010) (Colgan et al. 1997). Specifically, it has been shown that hFip1 prefers to bind to U-rich sequences (Kaufmann et al. 2004). In addition, it has been demonstrated that hFip1 can polyadenylate wildtype hexamer containing RNAs with PAP. This suggests that hFip1 cannot only bind to the RNA, but it has the ability to recruit PAP to the RNA for mRNA 3’ processing. Interestingly, hFip1 has already been shown to nonspecifically polyadenylate mutant RNA; it has been shown that hFip1 can recruit PAP to AAGAAA-containing RNAs in addition to AAUAAA-containing RNAs. All together, this provide evidence that hFip1 is directly involved in the nonspecificity of the CPSF complex.
While the role of hFip1 as the nonspecificity factor was tested by deleting its U rich binding sequences, the CPSF complex containing hFip1 can still polyadenylate these RNA nonspecifically. The next experiment is to make recombinant hFip1 and incubate with PAP to replicate the nonspecific functions mentioned in the above in an *in vitro* polyadenylation assay results. If the results can be duplicated, then hFip1 can be knocked down in the FLAG CPSF 73 stable cell line. The CPSF complex can then be IPed with FLAG beads. Consequentially, that CPSF complex will contain minimal hFip1. That CPSF complex can then be used in gel shift, UV crosslinking or *in vitro* polyadenylation assays to determine if specificity is achieved without hFip1. Ideally, the CPSF complex can be reconstituted to not include hFip1 and used in *in vitro* assays to test for specificity.

The model is that with short RNAs, hFip1 cannot bind to the RNA because it lacks enough U rich sequences. Because hFip1 cannot bind to the RNA anymore, CPSF 30 and WDR33 can now recruit PAP specifically to AAUAAA-containing RNAs. However, when longer RNAs are present, and therefore more cis elements are available for binding, hFip1 now can bind to U-rich sequences and recruit PAP the RNA independent of CPSF 30 and WDR33.

The nonspecific function of hFip1 can be regulated such that ultimately only AAUAAA will be polyadenylated. Although many natural hexamer mutations exist in the cell, if hFip1 can polyadenylate RNA independent of AAUAAA, many diseases can occur.
At least one protein present in HeLa nuclear extract can regulate the nonspecific function of hFip1. In addition to not being an enzyme and having an overall positive charge on its surface, it seems like this specificity factor plays a role in polyadenylation efficiency. This HeLa factor seems to have a role in polyadenylation efficiency because when the 20-40% ammonium sulfate was used in HeLa nuclear extract fractionation, the fraction that polyadenylated the wildtype RNA specifically had a distinct pausing polyadenylation pattern (Figure 4.4.C). In relation to hFip1, this suggests that perhaps hFip1 was no longer present in the 20-40% fraction. Because hFip1 can recruit PAP for polyadenylation, perhaps the pausing polyadenylation pattern was caused by less efficient recruitment of PAP without hFip1 present. Another possibility is that a poly (A) binding protein is responsible for specific polyadenylation in the presence of hFip1. The role of poly (A) binding proteins in polyadenylation consists of increasing poly (A) tail elongation efficiency, preventing and initiating mRNA degradation and transport of the mRNA from the nucleus to the cytoplasm for translation (Eliseeva et al. 2013). There are two broad classes of poly (A) binding proteins; one that is predominately cytoplasmic and one that is predominately nuclear. To affect polyadenylation, the poly (A) binding protein/specificity factor should be present in the nucleus. One poly (A) binding protein, PABPN1, is present in the nucleus. The roles of PABPN1 specifically are to increase poly (A) tail elongation efficiency, prevent mRNA degradation and act as a molecular ruler to determine when to terminate polyadenylation. Perhaps an abundant of PABPN1 was present in the 20-40% fraction,
causing a change in the polyadenylation pattern. A poly (A) binding protein as the specificity factor makes sense in relations to hFip1 because since hFip1 can bind to PAP, hFip1 might also can interact with the poly (A) binding protein that associates with PAP (Kaufmann et al. 2004). The specificity factor cannot be PAP itself because PAP has catalytic activity and would have been inactivated in the presence of heat. However, the specificity factor is heat insensitive. To experimentally determine if PAPN1 is the specificity factor, the direct experiment is to add recombinant PABPN1 to CPSF and PAP in an in vitro polyadenylation assay. If PABPN1 is the specificity factor, then CPSF and PAP can now polyadenylate only the wildtype RNA but not the mutant RNA.

Together, the data suggests that hFip1 is the CPSF factor that binds to longer RNAs nonspecifically. PABPN1, on the other hand, might be the specificity factor that regulates nonspecific functions of hFip1. This mechanism can regulate the interaction between CPSF 30 and WDR33 with AAUAAA. This extra regulatory step can increase mRNA 3’ processing specificity and make sure that aberrant mRNA 3’ processing is limited.
Chapter 5

Overall Conclusions and Perspectives

mRNA 3’ processing is not merely a two-step reaction. As shown in this manuscript, this process is highly regulated. While the general mRNA 3’ processing mechanism is known, the data in this manuscript suggests that perhaps the mRNA 3’ processing mechanism and regulation needs to be redefined (Figure 5.1). From the current understanding, CPSF 160 cannot recognize AAUAAA by itself (Murthy et al. 1995). In Chapter 3, evidence is provided that CPSF 30 and WDR33 directly recognize the AAUAAA together. This interaction was verified with two commonly used substrates in the mRNA 3’ processing field, SVL and adenovirus. The interaction between CPSF 30 and the AAUAAA makes sense because it contains five zinc finger domains and 1 zinc knuckle domain. In Chapter 3, the detailed interaction between CPSF 30 and the AAUAAA is explored. Specifically, zinc fingers 2 and 3 were found to be responsible for
the AAUAAA interaction, mRNA 3’ processing to function and possibly recognize AAUAAA. Coincidentally, NS1A binds to CPSF 30 specifically via zinc fingers 2 and 3 to inhibit cellular mRNA 3’ processing in influenza A affected cells (Twu et al. 2006). This interaction between NS1A and the zinc fingers 2 and 3 of CPSF 30 reinforces the hypothesis that CPSF 30 binds to the most important mRNA 3’ processing cis element. This interaction is also interesting in that it identifies mRNA 3’ processing machinery as a possible therapeutic agent (Twu et al. 2006). The zinc knuckle might also play a role in RNA binding. Following this interaction, the next step is to generate recombinant CPSF 30 along with its mutant versions for a UV crosslinking assay.

Ultimately, specificity needs to be reconstituted to determine what is required for specificity to occur. Recombinant WDR33 and CPSF 30 could be incubated together to determine if specificity can be reconstituted. If specificity is not achieved, another possibility is that mRNA 3’ processing needs RNA:RNA interaction in addition to RNA:protein and protein:protein interaction for the reaction to occur. This hypothesis was entertained previously by researchers but the idea was ultimately rejected (Bienroth et al. 1991) (Christofori et al. 1988) (Murthy et al. 1992) (Takagaki et al. 1989). However, it can be demonstrated that our purified CPSF complex does indeed contain an RNA that is approximately 70 kDa (data not shown). This suggests that perhaps a snRNA is needed for CPSF 30 and WDR33 to fully achieve specificity.

In Chapter 4, I provided evidence that the CPSF-AAUAAA interaction can be regulated. It is shown that while CPSF can recognize the AAUAAA element when short RNAs are
Figure 5.1 Summary Model

Summary of the findings from this manuscript.
used, CPSF cannot recognize AAUAAA when longer RNAs are used. This nonspecificity is most likely due to the interaction between hFip1 and U-rich sequences.

It is hypothesized that hFip1 is the nonspecificity factor because it is one of the two proteins that can UV crosslink to the RNA when the CPSF was not specific and it cannot UV crosslink to the short RNA when CPSF was specific. CPSF 100 was similar to hFip1 in that regard. However, hFip1 is more likely the nonspecificity factor because hFip1 has the ability to bind to RNA and recruit PAP nonspecifically (Kaufmann et al. 2004). However, more experiments need to be done to verify that hFip1 is indeed the nonspecificity factor.

In the HeLa nuclear extract, there is a factor that is needed to regulate the interaction between CPSF with the RNA to regulate specific polyadenylation. While the identity of this specificity factor is still unknown, a few biochemical characteristics were determined. First, the specificity factor is not an enzyme because it is insensitive to heat. Second, it has a net positive charge on its surface. Third, it has a role in polyadenylation efficiency. The current hypothesis is that this specificity factor is a PABPN1 because it plays a role in polyadenylation efficiency and it does not have a catalytic domain like PAP (Chan et al. 2010). In addition, since hFip1 can recruit PAP nonspecifically, this suggests that PABP should also interact with hFip1 and PAP. This suggests that the specificity factor should directly interact with the nonspecificity factor to regulate its function. To directly test if PABPN1 is the specificity factor, recombinant PABPN1 should be added to CPSF and PAP in an in vitro polyadenylation assay to test if the polyadenylation is now specific. hFip1 can now be used to test if it is the
nonspecificity factor by titrating it in the assay mentioned. If hFip1 is the nonspecificity factor, then CPSF would no longer be specific with additional hFip1. This extra level of regulation in recognizing the AAUAAA element displays the elegance of the mRNA 3’ processing machinery.

Finally, the entire mRNA 3’ processing complex can be regulated. To find potential novel regulators of mRNA 3' processing on a global level, immunoprecipitation and mass spectrometry were performed (Chapter 2). From those experiments, the interactome of mRNA 3’ processing machinery was mapped. From the results, the interactions between the CFIm subunits and E3 ligases were verified. The data also showed that at least two of the CFIm subunits can be polyubiquitinated. CFIm 68 was able to be polyubiquitinated by Itch E3 ligase, agreeing with previous results (Ingham et al. 2005). In addition, it was shown here that CFIm 68 can also be polyubiquitinated by WWP2. CFIm 59 can also be polyubiquitinated but different from CFIm 68, the endogenous E3 ligases were sufficient in recognizing CFIm 59. This suggests that CFIm 59 is more likely the target of the proteasomal degradation pathway than CFIm 68. This result demonstrates that perhaps CFIm 59 and CFIm 68 are not as similar as previously thought, although they both have the similar structural domains. Perhaps the different degradation pattern is a regulatory mechanism to alter the interaction between CFIm and the RNA by degrading one of the larger subunits and not the other. Future experiments include to verify the functional significance of why CFIm 59 and CFIm 68 are polyubiquitinated are needed.
Besides degradation, another function of polyubiquitination is to change protein localization. It has been shown that CFIm 59 and CFIm 68 shuttle between the nucleus and the cytoplasm (Ruepp et al. 2009). Perhaps polyubiquitination is the signal needed for the shuttling of CFIm 59 and CFIm 68. Immunohistochemistry is needed to visualize the localizations of CFIm 59 and CFIm 68 in the presence or absence of WWP2. Besides finding ubiquitination as another mode of regulation for mRNA 3' processing, other novel regulators could potentially be found. For example, the mass spectrometry data showed that CstF 50 is associated with p53. It is interesting to understand how mRNA 3' processing regulates cancer. However, the problem with the mass spectrometry results is that many false positives could be present. Separation of the false positives from the biologically significant interactions would be a crucial task. One possibility to decipher the biologically significant interactions from the false positives is to immunoprecipitate each subunit three times. By doing triplicates, a degree of significance can be generated, making the search for biological significance easier.

Overall, the work shown in this manuscript has made an impact in both the mRNA 3' processing mechanism as well as its regulation. The detailed mechanism of how the CPSF complex recognizes AAUAAA and how its interaction is regulated has been redefined. In addition, novel methods of how mRNA 3' processing can be regulated has been discovered. Together, this work has provided a deeper understanding of an essential process that hopefully one day would be useful in a therapeutic setting.
Bibliography


Edwalds-Gilbert, G. and C. Milcarek (1995). "Regulation of Poly(A) site Use During Mouse B-Cell Development Involves a Change in the Binding of a General


