UNIVERSITY OF CALIFORNIA,
IRVINE

Identification of Novel Roles for RNA Binding Proteins in pre-mRNA Processing

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Biomedical Sciences

by

Kristianna Sarkan

Dissertation Committee:
Professor and Chancellor's Fellow Yongsheng Shi, Chair
Professor Klemens Hertel
Professor Marian Waterman

2022

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First, I would like to thank my PI, Dr. Yongsheng Shi, for his guidance and training throughout my PhD. His dedication to science is inspirational and I truly believe I would not be the scientist I am today without his mentorship. His skills with bioinformatics have also been instrumental for my project. I am very thankful that I had the opportunity to be a member of his lab.

This manuscript and my degree would not have been made possible without funding from the NIH. In addition, I have been fortunate to receive funding from the Gazzaniga Family Award as well as the School of Medicine Dean's office.

I would also like to thank my labmates. Nabila was my first lab friend and helped create a culture of collaboration and enthusiasm within all of us graduate students who joined after her. Even though she was balancing graduate school and two children, she still managed to be an incredible support both personally and scientifically. Lindsey and Liang joined the lab shortly after me so it felt like we grew together, for which I will always be grateful. Lindsey was an important collaborator on my project and was always available to discuss projects and ideas. She has also kept us all well-fed with baked goods and shared my enthusiasm for the more crafty endeavors in lab, especially the holiday decorating contest. Liang has a tenacity when it comes to experiments that I truly admire. He was also always willing to go out on our (almost) weekly boba and Trade Joe's runs on Friday nights. Although Marielle joined the lab more recently, she fit in right away. Her scientific optimism is contagious, and I also know there has been an increase in laughter since she joined the lab. Yong, Yoseop, Xiuye, and Lusong were all wonderful post-docs who were not only an abundant source of knowledge, but also always willing to help any of us experimentally. I will always be grateful for how well our lab gets along and for inside jokes like the importance of our Zodiac signs.

I am also thankful to have been a member of the MMG department, which has been incredibly welcoming and friendly. The constructive feedback and genuine interest about my projects that I have received during department seminars has been extremely beneficial in boosting my confidence as both a scientist and presenter. I also appreciate being a part of a department that prioritizes graduate student involvement and well-being. In particular, I would like to thank my thesis committee, Dr. Marian Waterman and Dr. Klemens Hertel. Their insights into my project have been extraordinarily helpful, as have their words of encouragement. I have also made so many friends within the department. Our recent departmental outings have been a welcome distraction while finishing my PhD, particularly trivia nights. I have also loved getting to know everyone's pets through our "pet pics" channel.

One of my favorite opportunities during graduate school has been ReachOut TeachOut. Working with high school students reminds me of why I love science because I remember how excited I was the first time I performed experiments that I now see as mundane after performing them on an almost-weekly basis. I have particularly enjoyed being a science fair mentor and watching all of my students grow and mature as scientists. ROTO is also full of so many committed graduate students both on the executive board and within our volunteer body, many of whom have become friends. I am especially grateful for Bahareh, Annie, and Aby, who have

been amazing co-leaders. Bahareh was my co-president and my constant support as we navigated growing a relatively-new organization. We could not have picked better successors than Annie and Aby, who not only share my love for science education but who have even in the short time they have been co-presidents already expanded and improved ROTO to an extent that I could not have imagined possible.

I would also like to thank my friends from grad school and beyond for always having my back. Amanda has been one of my closest friends throughout my entire PhD and I can't imagine going through this experience without her. Although never actually my roommate, we were together enough that people believed we were. I am grateful for having a friend who I could discuss science with at our spontaneous workday lunches outside on the patio and go on "outings" to the grocery store with, dressed up of course because when else do we get to wear our nice clothes. Karissa also spent countless hours with me on Saturdays at one of our kitchen tables practicing for presentations, working on data analysis, or just preparing for the next week. She was also always up for cooking and baking adventures or movie nights. Thank you for always being real with me and empowering me to always push for my goals. And of course Annie, my college roomie and forever one of my best friends. Thank you for showing me around St. Louis and Hershey, coming to visit me when you were in San Diego despite your crazy schedule, and being available whenever I need you.

Finally, I would like to thank my family for their infinite support and love. Matt, thank you for sharing my love of puns, not making fun of me when I send you photos of DNA concentrations that are palindromes, and always making me laugh. Being able to stay with you and Tamara during the shutdown kept my spirits up and I'll always be grateful.

Mom and Dad, thank you for always understanding and loving me. Thank you for teaching me that even the hardest of days can be fixed with chocolate, robot food, or pictures of Maggie. And thank you for always supporting my goals and encouraging me to be bold. I can't imagine having better parents.

VITA

Kristianna Sarkan

EDUCATION

PhD Candidate in Biological Sciences **September 2016-Present**
University of California, Irvine
GPA: 4.0

Bachelor of Science in General Biology **September 2011 - June 2015**
University of California, Irvine
Summa Cum Laude
GPA: 2.961

RESEARH EXPERIENCE

Graduate Research Assistant **June 2017-Present**
University of California, Irvine

Undergraduate Research Assistant **September 2013-June 2015**
University of California, Irvine

TEACHING EXPERIENCE

Adjunct Professor **August-December 2022**
Chemistry
Concordia University, Irvine

Undergraduate Teaching Assistant **March-June 2021**
Microbiology Lab and Molecular Biology
University of California, Irvine

Undergraduate Teaching Assistant **January-April 2020**
Ecology and Evolutionary Biology
University of California, Irvine

Undergraduate Teaching Assistant **September-December 2017**
Photomedicine, Biomedical Engineering
University of California, Irvine

Paraprofessional Special Education Aide **September 2015-June 2016**
Country Hills Elementary School

Learning and Academic Resource Center (LARC) Tutor **September 2014-June 2015**

University of California, Irvine

GRADUATE LEADERSHIP AND SERVICE

ReachOut TeachOut at UC Irvine Co-President University of California, Irvine	**September 2019-June 2022**
ReachOut TeachOut at UC Irvine Treasurer, UC Irvine University of California, Irvine	**May 2018-September 2019**
ReachOut TeachOut at UC Irvine Science Fair Mentor University of California, Irvine	**November 2018 – Present**
UCI School of Medicine Elevator Pitch Competition University of California, Irvine	**October 2021**
UCI School of Medicine Elevator Pitch Competition University of California, Irvine	**October 2020**
Women in STEM (WiSTEM) Mentor University of California, Irvine	**November 2020- Present**
Graduate Student Representative Department of M&MG, University of California, Irvine	**March 2022-June 2022**
Graduate Student Representative Department of M&MG, University of California, Irvine	**September 2019-June 2020**
Seminar Committee Member Department of M&MG, University of California, Irvine	**June 2018-June 2019**

AWARDS/FELLOWSHIPS

UCI School of Medicine Grad Day Poster Award	**October 2022**
RNA Society 2022 Poster Prize	**June 2022**
Associated Graduate Students Conference Grant	**September 2021**
UCI School of Medicine Travel Grant	**May 2022**
Associated Graduate Students Virtual Conference Grant	**September 2021**
School of Medicine Gazzaniga Family Award	**January 2021**
UCI School of Medicine Dean's Award	**September 2019**
Summa Cum Laude Graduate, University of California Irvine	**June 2015**
University of California, Irvine Campuswide Honors Program	**June 2015**
Honors in Biology Graduate, University of California, Irvine	**June 2015**
Excellence in Research	**June 2015**

CERTIFICATES

Division of Teaching Excellence and Innovation Course Design (UCI)
Activate to Captivate (UCI)
Improv in Teaching (UCI)

PROFESSIONAL SOCIETIES/AFFILIATIONS

RNA Society	May 2018-Present

Presentations

Sarkan, K., Soles, L., & Shi, Y. (2021, October). *The mRNA 3' processing factor CFIm participates in 3' splice site selection and regulates alternative splicing*. Poster Presentation at 5th Annual SoCal RNA Symposium

Sarkan, K., Soles, L., & Shi, Y. (2021, October). *The mRNA 3' processing factor CFIm participates in 3' splice site selection and regulates alternative splicing*. Poster Presentation at Annual UCI School of Medicine Grad Day

Sarkan, K., Soles, L., & Shi, Y. (2022, June). *The mRNA 3' processing factor CFIm participates in 3' splice site selection and regulates alternative splicing.* Poster presentation at annual RNA Society Meeting, Boulder Colorado

Sarkan, K., Soles, L., & Shi, Y. (2021, October). *The mRNA 3' processing factor CFIm participates in 3' splice site selection and regulates alternative splicing*. Poster Presentation at Annual UCI School of Medicine Grad Day

Sarkan, K., Soles, L., & Shi, Y. (2021, August). *The mRNA 3' processing factor CFIm participates in 3' splice site selection and regulates alternative splicing.* Oral presentation at biennial Cold Spring Harbor Eukaryotic mRNA Processing meeting, New York

Sarkan, K. (2021, January). *CFIm, a novel splicing regulator*. UCI Department of Microbiology and Molecular Genetics seminar

Sarkan, K. (2020, February). *Regulation and Impact of Alternative Polyadenylation*. UCI Department of Microbiology and Molecular Genetics seminar

Sarkan, K., & Shi, Y. (2019, August). *CFIm attenuates global gene expression to module cell fate decision- An "anti-myc" model*. Oral presentation at the biennial Cold Spring Harbor Eukaryotic mRNA Processing meeting, New York

Sarkan, K. (2019, April). *Insights into the Versatility of RNA Processing Regulation*. UCI RNA Club

Sarkan, K. (2019, February). *Insights into the Versatility of RNA Processing Regulation*. UCI Department of Microbiology and Molecular Genetics seminar

Publications

Vogler E, Mahavongtrakul M, Sarkan K, Catuara-Solarz S, Bohannan R, and Busciglio J. Genetic removal of synaptic Zn2+ impairs cognition, alters neurotrophic signaling and induces neuronal hyperactivity in an age-dependent manner. *eNeuro*. In submission

Sarkan, K and Shi, Y. The mRNA 3' processing factor CFIm participates in 3' splice site selection and regulates alternative splicing. In preparation

ABSTRACT OF THE DISSERTATION

Identification of Novel Roles for RNA Binding Proteins in mRNA Processing

by

Kristianna Sarkan

Doctor of Philosophy in Biomedical Sciences

University of California, Irvine, 2022

Professor and Chancellor's Fellow Yongsheng Shi, Chair

Two critical steps in mRNA processing are 3' processing and splicing, both of which diversify the human genome and create mRNA isoforms with unique regulatory properties including mRNA stability, translation efficiency or intracellular location or even create distinct proteins. In addition to being essential steps in mRNA processing, polyadenylation and splicing are highly alternatively regulated, with approximately 70% and 95% of genes producing alternative isoforms respectively. Regulation of these isoforms have important biological consequences as mis-regulation is associated with many human diseases including cancer and neurological disorders (Tian and Manley 2017; S. Chan, Choi, and Shi 2011; Q. Pan et al. 2008; Lukong et al. 2008; Y. Zhang et al. 2021).

While the core regulatory machineries for 3' processing and splicing have been identified, there is still limited ability to predict how alternative regulation will occur, making this a critical field of study. One area of particular interest for the study of alternative RNA processing is the role of RNA binding proteins (RBPs). To better understand how RNA binding proteins regulate alternative 3' processing and splicing, I carried out three projects.

First, to investigate the roles of RNA binding proteins in 3’ processing regulation, a dual luciferase reporter system was utilized for a large-scale screen of RBPs for polyadenylation site regulation. In addition to validating several known regulators of APA, we identify several novel inhibitors of polyA site selection including hnRNP A0, hnRNP G, and Musashi1. Strikingly, we also demonstrate that the SR family of proteins are polyA site position-independent repressors of polyadenylation sites, indicating that their role in polyadenylation may be unique from their role in splicing regulation where they act as position dependent activators. This screen may also be used to identify other families of RBPs with the capability to regulate polyadenylation and make predictions about other new regulators.

Next, genome-wide sequencing approaches were used to characterize the role of the core polyadenylation complex cleavage factor I (CFIm) in APA regulation. In addition to its known role in enhancing distal polyA sites, we demonstrate that CFIm promotes intronic polyadenylation, most notably within one member of each of the other core polyadenylation machinery components. We also propose a model by which CFIm regulates both 3’ UTR APA and intronic polyadenylation to modulate global protein production and therefore link 3’ processing with cell fate determination.

Finally, we interrogate the role of CFIm in splicing regulation through a combination of genome-wide sequencing and biochemical analyses. We demonstrate that CFIm is a general alternative splicing regulator that binds 3’ splice site and interacts with U2AF through RS-RS domain interactions. In addition, CFIm promotes U2AF-RNA interaction at the 3’ splice sites of CFIm activated cassette exons. Our data supports a revised model for 3’ splice site selection by U2AF: CFIm and other RNA binding proteins compete for interaction with U2AF and each regulate the alternative splicing of a specific subset of cassette exons.

CHAPTER 1

INTRODUCTION

Following transcription, pre-mRNA undergo a variety of processing events that include splicing and 3'-end processing, both of which are essential for proper gene expression as well as critical for expanding proteomic diversity (Hocine, Singer, and Grünwald 2010). 3'-end processing is the cleavage and polyadenylation of the 3'-end of pre-messenger RNA. It is linked to both transcription termination and protection from mRNA decay, making it an essential step in RNA processing. At the same time, 3'-end processing is a critical step in the regulation of gene expression as over 70% of pre-mRNAs contain more than one polyadenylation site and can be alternatively polyadenylated, leading to mRNA isoforms with differential stability, translation efficiency, and intracellular location (Tian and Manley 2017; S. Chan, Choi, and Shi 2011; Shi and Manley 2015). In addition to 3' processing, splicing is another RNA processing event essential for both constitutive and alternative gene regulation that involves the removal of non-coding introns and joining of protein-coding exons. In the case of splicing, over 90% of eukaryotic pre-mRNAs can be alternatively regulated, greatly expanding proteomic diversity (Q. Pan et al. 2008).

Regulation of both splicing and polyadenylation also have biological relevance and mis-regulation is associated with many human diseases. Alternative polyadenylation is highly tissue-specific and its accurate regulation is critical for stem cell maintenance and differentiation as well as immune system development (Lackford et al. 2014b; Justin Brumbaugh et al. 2018; Luo et al. 2013; Z. Ji et al. 2009; Kaida et al. 2010; Lee et al. 2018). Likewise, mis-regulation of alternative splicing in cancer is so common that it is considered a hallmark of cancer (Bonnal, López-Oreja, and Valcárcel 2020). As a result, it is critical to study not only how the RNA processing machineries function but also how they are regulated. This chapter will examine the mechanism, regulation and biological relevance of both splicing and mRNA 3' processing. Particular emphasis will be focused on the protein factors that regulate pre-mRNA processing. Finally, I will explore the similarities and cross-regulation between the splicing and polyadenylation machineries.

1.1 mRNA 3'-END PROCESSING MECHANISM

3'-processing is initiated by the binding of trans-acting proteins factors to key cis regulatory elements on the RNA. These trans-acting factors constitute the core polyadenylation machinery, an extensive network of protein-RNA and protein-protein interactions which are oftentimes individually weak but collectively form a stable complex. Here, these critical protein-RNA and protein-protein interactions will be described.

Cis Elements

The polyadenylation machinery assembles on a series of RNA cis elements that collectively create the polyadenylation site (PAS). PolyA sites are categorized as either canonical or non-canonical. For canonical polyA sites, the most highly conserved sequence is the A(A/U)UAAA hexamer, which is found at approximately 70-80% of polyA sites (S. Chan, Choi, and Shi 2011; Beaudoing et al. 2000; MacDonald and Redondo 2002). Removal of the A(A/U)UAAA hexamer or even single nucleotide mutations within this sequence have been shown to significantly impair cleavage efficiency, indicating the importance of this sequence.

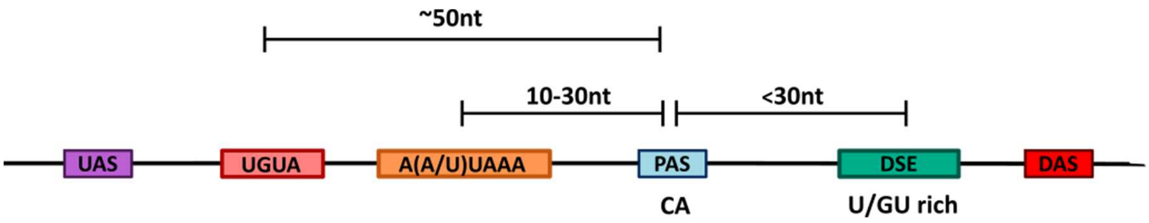


Figure 1.1 Canonical mRNA 3’-Processing Cis Elements

Schematic of cis elements important for polyA site definition in mammals.
UAS: upstream auxiliary element PAS: polyA site
DSE: downstream element DAS: downstream auxiliary element

Approximately 10-30 nucleotides downstream of the A(A/U)UAAA hexamer is the cleavage site, which is typically a C/A or U/A motif (Chen, Macdonald, and Wilusf 1995). Another key feature of a canonical PAS is the downstream element (DSE), which is more variable but is frequently U-or GU-rich, and is located approximately 30 nucleotides downstream of the cleavage site (Figure 1.1) (S. Chan, Choi, and Shi 2011). Together, the A(A/U)UAAA hexamer and the DSE define the cleavage site.

In addition to these two critical sequences, many polyA site have auxiliary elements that can influence polyA site strength. The most common downstream auxiliary element is a G-rich sequence, which can regulate cleavage and polyadenylation from hundreds of bases away (Dalziel, Nunes, and Furger 2007). Upstream of the cleavage site, the most common auxiliary sequence is a U-rich region (Neve et al. 2017). Finally, some polyA sites have a UGUA motif enriched 50 nucleotides upstream of the cleavage site (Zhu et al. 2017). Together these auxiliary elements enhance binding of the core polyadenylation machinery as well as other regulatory factors, thereby enhancing 3' processing activity (Figure 1.1).

Approximately 20-30% of polyA site are non-canonical and lack the A(A/U)UAAA hexamer (MacDonald and Redondo 2002; Beaudoing et al. 2000). The precise mechanism for PAS selection of non-canonical sites is unknown. In some cases, recognition of these A(A/U)AAA-independent polyA sites is mediated by a UGUA motif, which is consistent with the known role of UGUA as an enhancer of polyA site selection (Venkataraman, Brown, and Gilmartin 2005; Zhu et al. 2017). In addition, there have been reports of a GUKKU sequence (where K represents a G or U) approximately 50 nucleotides downstream of non-canonical PAS (Hwang et al. 2016). Other studies have demonstrated that other non-canonical sites only require the downstream element and an A-rich upstream region (Nunes et al. 2010). These diverse mechanisms for non-canonical polyA site selection indicate that non-canonical polyA sites not only have unique regulatory mechanisms from canonical polyA sites but also from each other and that no single mechanism is relevant to all non-canonical sites. Instead, non-canonical polyA sites are regulated by context-specific mechanisms and are frequently associated with alternative polyadenylation (APA), suggesting that they may play a regulatory role in polyA site selection.

Trans-Acting Factors

The core polyadenylation machinery involved in polyA site selection consists of 4 multi-subunit protein complexes: CPSF (**c**leavage and **p**olyadenylation **s**pecificity factor), CstF (**c**leavage **s**timulation **f**actor), and CFIm and CFIIm (**c**leavage **f**actor I and II) (C. R. Mandel, Bai, and Tong 2007; Tian and Manley 2017). CstF, CPSF, and CFIIm are required for cleavage, although only CPSF is believed to be required for polyadenylation (Boreikaite et al. 2022; Schmidt et al. 2022). Other proteins involved in 3'-end processing include polyA polymerase (PAP), RNA polymerase II (RNAP II), and Rbbp6, which has recently been shown to be essential for reconstitution of the 3'-processing machinery, although its exact function remains to be determined (Boreikaite et al. 2022; Schmidt et al. 2022). Here I will describe the role of each of these complexes in polyA site selection (Figure 1.2).

CPSF

The cleavage and polyadenylation specificity factor (CPSF) is required for both cleavage and polyadenylation of pre-mRNA. CPSF binds to RNA at the A(A/U)UAAA hexamer and cleaves

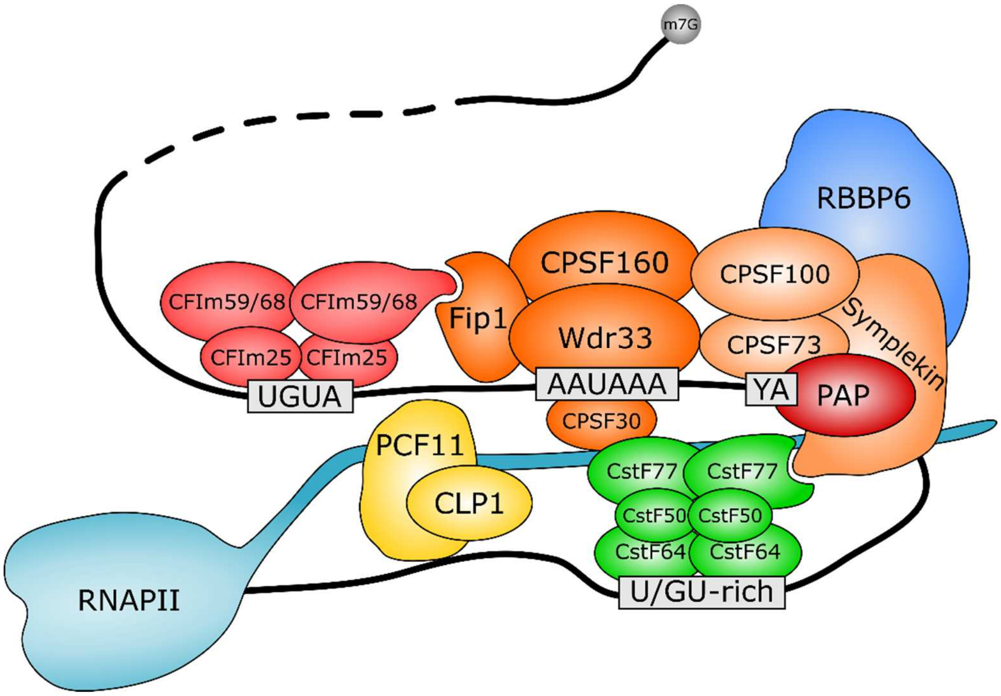


Figure 1.2 3' Processing Trans-Acting Factors

Schematic representation of trans-acting factors important for polyA site definition in mammals.
CPSF (orange) binds to the AAUAAA hexamer and physically cleavges RNA at the YA motif. CstF (green) binds to the downstream U/GU-rich region. CFIm (pink) binds the UGUA motif and enhances nearby polyadenylation sites by recruiting CPSF. CFIIm (yellow) promotes transcription termination. Rbbp6 (blue) is essential for cleavage.

CFIm: cleavage factor I	CPSF: cleavage and polyadenylation specificity factor
CFIIm: cleavage factor II	CstF: cleavage stimulation factor
RNAPII: RNA polymerase II	

RNA at the downstream C/A or U/A motif. The CPSF complex consists of 7 subunits: CPSF30, CPSF73, CPSF100, CPSF160, Fip1, Wdr33 and Symplekin (Shi et al. 2009). CPSF can be subdivided into 2 subcomplexes: the polyadenylation specificity factor (mPSF) consisting of CPSF30, CPSF160, Wdr33 and Fip1 and the cleavage factor (mCF) consisting of CPSF100, CPSF73, and Symplekin. Like its name suggests, mCF is involved in RNA cleavage while mPSF is involved in recognizing and binding polyA sites. In this section, I will discuss the roles of the different CPSF subunits in these two critical events.

CPSF in RNA Binding

The polyadenylation specificity factor (mPSF) is responsible for RNA binding. Originally, it was proposed that CPSF160 directs binding to the A(A/U)UAAA hexamer based upon several lines of evidence. First, 2 proteins of approximately 30kDA and 160 kDA were shown to directly crosslink to A(A/U)UAAA-containing RNAs. In addition, recombinant CPSF160 preferentially binds to AAUAAA containing RNAs in comparison to mutants (Murthy and Manley 1995; Keller et al. 1991). However, in 2014 it was shown that CPSF binding to the A(A/U)UAAA hexamer is in fact mediated by CPSF30 and Wdr33 (which is similar in size to CPSF160) and that the CPSF30-RNA interaction is essential for mRNA 3'-processing (Serena L. Chan et al. 2014). This finding was further validated when the structure of the complex containing Wdr33, CPSF160, CPSF30 and an AAUAAA containing RNA was resolved using cryo-electron microscopy. The polyA site was determined to be bound by both CPSF30 and Wdr33, with A4 and A5 being bound by zinc fingers 2 and 3 of CPSF30 respectively and with U3 and A6 forming Hoogsteen base pairs and contacting Wdr33. This study also confirmed that

CPSF160 is a scaffold protein that assists in Wdr33- and CPSF30-RNA binding and is not directly involved in AAUAAA recognition (Y. Sun et al. 2018) (Figure 1.2).

The fourth subunit of the polyadenylation specificity factor is Fip1. Fip1 has been shown to bind to U-rich RNA elements through its arginine-rich C-terminus (Kaufmann et al. 2004). In addition, Fip1 plays a role in alternative polyadenylation by enhancing binding to weak polyadenylation sites because it interacts with CFIm, a polyadenylation site activator that recruits CPSF to polyadenylation sites through interactions with the RE/D domain of Fip1 (Lackford et al. 2014a).

In addition to regulating polyA site selection, the mPSF complex is important for regulating interactions with other members of the core polyA machinery. For example, following RNA cleavage, CPSF remains bound to the AAUAAA hexamer. This allows CPSF and Fip1 to recruit polyA polymerase to the RNA to promote polyadenylation (Murthy and Manley 1995; Kaufmann et al. 2004). CPSF can also be isolated in a pre-assembled complex with CstF, an interaction likely mediated by CPSF160 and Fip1 as well as Symplekin (an subunit of the mCF subcomplex described below) (Y. Takagaki et al. 1990; Kaufmann et al. 2004; Murthy and Manley 1995). This interaction may allow CPSF and CstF to jointly binding to the A(A/U)UAAA hexamer and DSE to regulate PAS selection.

CPSF in RNA Cleavage

The second subcomplex of CPSF is the cleavage factor (mCF) which consists of 3 subunits: CPSF73, CPSF100, and Symplekin. Both CPSF73 and CPSF100 contain metallo-β-lactamase and β-CASP domains, which are found in other endonucleases of the β-CASP family (Corey R. Mandel et al. 2006; Callebaut et al. 2002). CPSF73 has been established as the endonuclease that cleaves RNA at the C/A cleavage site downstream of the A(A/U)UAAA hexamer because it has zinc-dependent endonucleolytic activity (Corey R. Mandel et al. 2006). The role of CPSF100 in the endonucleolytic cleavage reaction has yet to be determined. However, one interesting hypothesis is that in addition to CPSF73, CPSF100 can cleave RNA. This hypothesis is based upon the fact that while human CPSF73 has been crystallized and structurally characterized, human CPSF100 could not be crystalized and therefore the *S. cerevisiae* homolog of CPSF100 (Ydh1) was resolved instead. While the crystal structure of CPSF100 displayed no obvious access point for an RNA substrate within the active site, the crystal structure of CPSF73 showed two bound zinc ions, leading to the conclusion that CPSF73 is the endonuclease (Corey R. Mandel et al. 2006). However, several key signature residues of the metallo-β-lactamase domain of human CPSF100 are not conserved in *S. cerevisiae* although they were conserved in other organisms, from plants to vertebrates, even including *S. pombe* (Corey R. Mandel et al. 2006). This is a particularly intriguing hypothesis as it could suggest that CPSF100 and CPSF73 function as dual nucleases in regulating alternative 3' processing. Further investigation will be necessary to reveal the role of CPSF100 in 3' processing and as a potential endonuclease.

The final subunit of the mCF is Symplekin. While Symplekin is not involved in RNA cleavage, it serves as scaffold protein to recruit other factors to the core polyA machinery.

CstF

The cleavage stimulation factor (CstF) binds to the downstream element and is required for mRNA cleavage but not polyadenylation (C. R. Mandel, Bai, and Tong 2007; S. Chan, Choi, and Shi 2011; Boreikaite et al. 2022). CstF is a multi-subunit protein complex with 3 subunits:

CstF50, CstF64, and CstF77, each of which presents as a dimer (Y Takagaki and Manley 1997). CstF64 also has a closely related paralog, CstF64tau.

CstF binds to the downstream element through the RNA recognition motif (RRM) of CstF64, which by itself can bind to U or GU rich sequences similar to the downstream element (Y Takagaki and Manley 1997) (Figure 1.2). UV crosslinking studies indicate that CstF alone binds weakly to RNA. However, CstF binding increases with increasing concentrations of CPSF, indicating that stable association of CstF to the DSE requires cooperative binding of CPSF to the A(A/U)UAAA hexamer (Murthy and Manley 1992). This mechanism may ensure that the complex does not assemble unless both the AAUAA hexamer and the downstream element are present.

Unlike CPSF, which binds specifically to the consensus AAUAAA hexamer, CstF64 can bind to more diverse sequences with variable affinities, suggesting a role in alternative polyadenylation in addition to its essential role in polyA site selection (C. Yao et al. 2012). CstF64 also has a closely related paralog, CstF64tau (CstF64τ). CstF64τ is highly expressed in the testis and CstF64τ knockout has previously been shown to cause spermatogenic defects and male infertility in mice (Dass et al. 2007). It has a similar domain structure to CstF64, and the two paralogs have overlapping but distinct RNA binding specificities, indicating that tissue specific expression levels may play an important role in alternative polyadenylation. In addition, CstF64 and CstF64τ likely have related roles in polyA site selection as there are a greater magnitude of APA changes when the factors are co-depleted in comparison to depletion of CstF64 alone (C. Yao et al. 2012).

CstF77 forms a homodimer and serves as a scaffold for the CstF complex by bridging CstF50 and CstF64, which do not directly interact with each other. CstF77 interacts with CstF50 and

CstF64 through its proline rich region (Yoshio Takagaki and Manley 2000). Recently it has also been demonstrated that CstF77 also increases the affinity of the CstF64 RRM for RNA sequences in the downstream element (W. Yang et al. 2018).

In addition to its role in formation of the CstF complex, CstF77 is important for interacting with rest of the polyA machinery, specifically CPSF through multiple factors including CPSF160, Fip1, and Symplekin (Kaufmann et al. 2004; Yoshio Takagaki and Manley 2000; Murthy and Manley 1995). This supports the prevailing hypothesis that polyA site selection is mediated by cooperative binding by CPSF and CstF.

Finally, CstF50 interacts with CstF77 and, like CstF77, forms a homodimer (Yoshio Takagaki and Manley 2000). Recently it was also demonstrated that CstF50 can influence CstF-RNA interactions by recognizing G/U rich sequences of specific length and content (W. Yang et al. 2018). In addition to interacting with other subunits of CstF, CstF50 also interacts with RNA Polymerase II C-terminal domain (CTD), indicating that it may play an important role in coupling transcription and 3' processing (McCracken et al. 1997; C. R. Mandel, Bai, and Tong 2007)

CFIm

Cleavage Factor I (CFIm) is not essential for either cleavage or polyadenylation (S. Chan, Choi, and Shi 2011; C. R. Mandel, Bai, and Tong 2007; Boreikaite et al. 2022). Instead, CFIm serves as an activator of distal polyA sites and is a critical regulator of alternative polyadenylation. CFIm forms a tetramer consisting of a homodimer of the small subunit, CFIm25, and a homodimer of one of two alternative larger subunits, either CFIm59 or CFIm68 (Rüegsegger, Blank, and Keller 1998). CFIm has important roles in both RNA binding and recruitment of polyA factors.

CFIm in RNA Binding

CFIm binds to the UGUA motif, which is enriched approximately 50 nucleotides upstream of distal, or downstream, polyA sites (Zhu et al. 2017; Q. Yang, Gilmartin, and Doublié 2010; Brown and Gilmartin 2003). All three subunits of CFIm can be crosslinked to RNA, suggesting that they are all involved in RNA recognition (Rüegsegger, Beyer, and Keller 1996). The UGUA motif itself, however, is directly bound by CFIm25. CFIm25 is a member of the Nudix hydrolase superfamily of proteins and has a Nudix domain with the classic α/β/α fold. However, while the majority of proteins of this superfamily are involved in hydrolytic activity, CFIm25 lacks two critical glutamate residues, making it catalytically inactive (Coseno et al. 2008). Instead, the Nudix domain of CFIm25 directs binding to the UGUA motif (Q. Yang, Gilmartin, and Doublié 2010) (Figure 1.2).

The large subunits of CFIm can also directly crosslink to RNA, but the precise mechanism for RNA recognition by CFIm59 and CFIm68 is unknown (Rüegsegger, Beyer, and Keller 1996; G. Martin et al. 2012). While CFIm59 and CFIm68 contain an RNA recognition motif, the RNA recognition motif is in fact important for interaction with CFIm25 (Q. Yang et al. 2011). As CFIm59 and CFIm68 can bind to RNA, they may be able to fine-tune RNA binding specificity of CFIm25. It will be interesting to determine what sequences CFIm59 and CFIm68 specifically bind to as well as how they may differ from each other.

CFIm in PolyA Factor Recruitment and PolyA Site Selection

Many studies have shown that CFIm regulates alternative polyadenylation, with knockdown of CFIm causing widespread changes in polyA site usage from distal (downstream) to proximal (upstream) polyA sites (G. Martin et al. 2012; Zhu et al. 2017; Masamha et al. 2014). Recently, our lab reported that CFIm regulates APA by activating distal polyA sites. CFIm59 and CFIm68 are categorized as SR-like protein that are each composed of an N-terminal RNA recognition motif, a proline rich region, and a C-terminal RS domain. RS domains, which are common in splicing factors, are enriched for arginine serine repeats. When CFIm binds to the UGUA motif enriched upstream of distal polyA sites, it recruits CPSF to the nearby downstream AAUAA motif through interactions of the RS domain of CFIm59 of CFIm68 with the RE/D domain of Fip1, a subunit of CPSF. As a result, there is preferential usage of the distal polyA site when CFIm levels are high (Zhu et al. 2017) (Figure 1.3).

CFIIm

The final protein complex that is considered to be a member of the core polyA machinery is cleavage factor II (CFIIm), which is essential for activation of cleavage (Boreikaite et al. 2022) (Figure 1.2). CFIIm consists of two subunits: Pcf11 and Clp1. Pcf11 is a known regulator of transcription termination (Kamieniarz-Gdula et al. 2019; R. Wang et al. 2019). It interacts with the RNA polymerase II C terminal domain (Pol II CTD) and both enhances transcription termination and promotes early polyadenylation genome wide. Interestingly, Pcf11 is in sub-stoichiometric levels in comparison to other members of the polyA machinery, suggesting a mechanism to prevent premature polyadenylation (Kamieniarz-Gdula et al. 2019). In fact, Pcf11 is autoregulated through an intronic polyadenylation site that prevents formation of full-length Pcf11 transcripts (R. Wang et al. 2019).

Currently the role of Clp1 in polyadenylation is relatively unknown. As Clp1 coimmunoprecipitates both CFIm and CPSF, it was previously been hypothesized to bridge these two protein complexes (De Vries et al. 2000). However, this is likely not the case as a direct link between the CFIm59/68 subunits of CFIm and the Fip1 subunit of CPSF has since been shown (Zhu et al. 2017). It has also been suggested that ATP promotes RNA cleavage by binding to Clp1; however, this finding requires further investigation as a similar study did not find a role of Clp1 and ATP in cleavage (Schmidt et al. 2022; Boreikaite et al. 2022). If this finding is true, it would be consistent with previous findings that mutations of the ATP binding site of Clp1 interfere with binding to other polyA factors, in particular Pcf11 (Ghazy et al. 2012). In addition to a role in 3'-end processing, Clp1 is known to have RNA 5'-kinase activity and functions in tRNA splicing (Hanada et al. 2013).

Rbbp6

RB binding protein 6 (Rbbp6) was recently discovered to be essential for 3' cleavage in conjunction with CPSF, CFIIm, and CstF (Boreikaite et al. 2022; Schmidt et al. 2022) (Figure 1.2). Rbbp6 was first implicated in 3'-end processing when it was found to co-purify with the core polyA machinery (Shi et al. 2009; Di Giammartino et al. 2014). However, it was not until recently that it was determined that Rbbp6 plays a key role in stimulating cleavage. One line of evidence that supported necessity of Rbbp6 in 3' processing was that the yeast homolog of Rbbp6, Mpe1, activates the homolog of the endonuclease CPSF73, Ysh1, through interactions near the active site of Ysh1. In addition, in 2022, two independent in vitro reconstitutions of 3' endonucleolytic activity indicated that the CPSF, CstF, CFIIm, and Rbbp6 are all necessary to recapitulate efficient and specific cleavage. While Rbbp6 is not a stable component of CPSF as

Mpe1 is in yeast Cpf, it likely regulates 3’ processing activity through a similar mechanism as it interacts with CPSF73 near the active site and activates nuclease activity (Boreikaite et al. 2022; Schmidt et al. 2022).

PolyA Polymerase (PAP)

polyA polymerase (PAP) is responsible for the synthesis and addition of the polyadenosine tail to the 3’ end of premRNAs (S. Chan, Choi, and Shi 2011; Neve et al. 2017; C. R. Mandel, Bai, and Tong 2007; Tian and Manley 2017). While PAP contains an RNA binding domain, its binding is not sequence specific and it relies on interactions with other core polyA machinery components including CPSF and CFIm to be recruited to polyA sites (S. Chan, Choi, and Shi 2011; C. R. Mandel, Bai, and Tong 2007). Polyadenylation is template-independent and recent estimates suggest that the lengths of polyA tail polymerized by PAP range are relatively short at around 80-100 nucleotides in length (Legnini et al. 2019).

1.2 REGULATION AND BIOLOGICAL IMPACT OF ALTERNATIVE POLYADENYLATION

Like other RNA processing events, polyadenylation can be alternatively regulated. In fact, over 70% of pre-mRNAs contain more than one polyadenylation site and can be polyadenylated at any of these sites based upon intracellular conditions such as the availability of the core polyA machinery and other regulatory RNA binding proteins. There are two general categories of alternative polyadenylation (APA): 3’ UTR APA and intronic polyadenylation (IPA) (Figure 1.4). During 3’ UTR alternative polyadenylation, all of the polyadenylation sites lie within the

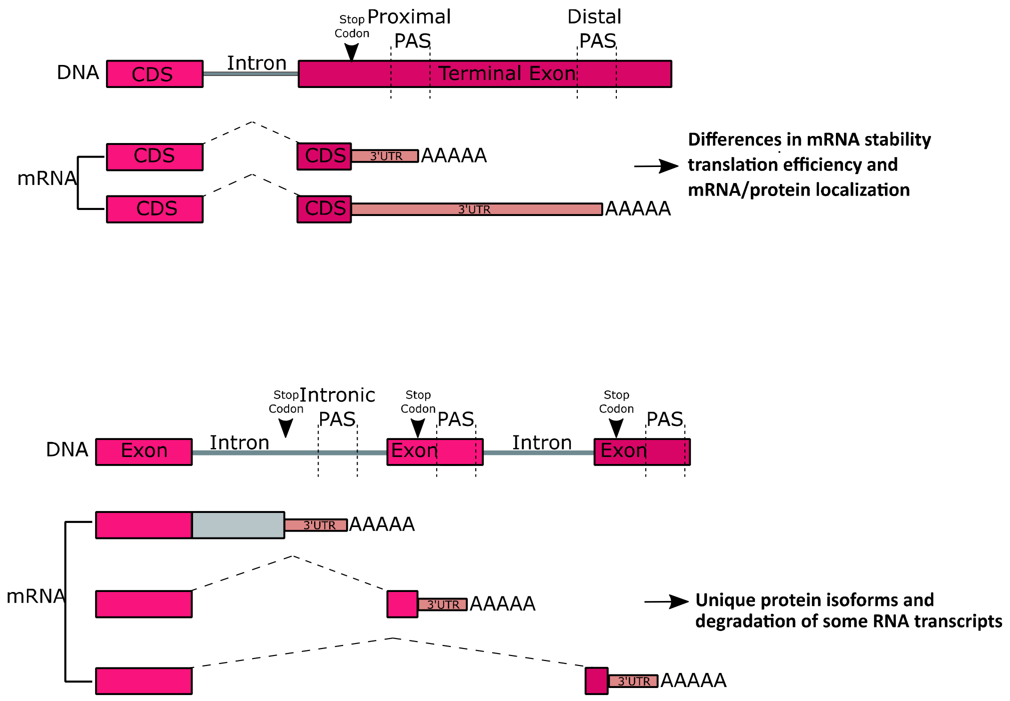


Figure 1.4 Alternative Polyadenylation

Schematic of alternative polyadenylation events. APA can either occur within the 3' UTR (top) or intronic regions (bottom). Coding regions are represented by light pink boxes, 3' UTR is represented by maroon boxes, and introns are represented by bold gray lines. Splicing events are represented by dashed gray lines.
PAS: polyA site

3’-untranslated region (3’ UTR), leading to unique mRNA isoforms with distinct mRNA stability, translation efficiency, or intracellular localization, among other properties. However, the coding region is unaffected (Tian and Manley 2017; Shi and Manley 2015). By contrast, intronic polyA sites lie within intronic regions before the end of the protein coding region, leading to transcripts that lack downstream sequences and are typically non-functional or degraded by nonsense-mediated decay (Tian and Manley 2017; Shi and Manley 2015). Here, I will elucidate mechanisms for APA regulation as well as the biological consequences of this regulation.

Mechanisms of APA

3’ UTR APA

3’ UTR APA occurs when all of the polyadenylation sites lie within the 3’ UTR following the coding sequence of the mRNA, resulting in mRNAs with different lengths of 3’ UTRs. Distal (downstream) polyA sites are oftentimes stronger than proximal (upstream) polyA sites due to a higher frequency of canonical cis elements such as the AAUAAA hexamer, the downstream U/GU rich region, and the upstream UGUA enhancer motif. By contrast, proximal polyA sites have an inherent advantage as they are transcribed first, allowing the polyA factors to begin assembly prior to the transcription of the distal polyA sites (Davis and Shi 2014). As a result, 3’ UTR APA is a balance between polyA site strength and transcription rates.

Consequences of 3’ UTR APA include mRNA stability and translation efficiency. In mammals, 3’ UTRs are highly enriched for binding sites of microRNAs, small RNA molecules that bind to complementary target mRNAs and decrease stability and translation efficiency of mRNAs (Sandberg et al. 2008; Z. Ji et al. 2009). The 3’ UTRs of mRNAs that are polyadenylated at the distal polyA site are longer and therefore have increased binding sites for microRNAs; consistently, these mRNAs have also been shown to be destabilized and have reduced translation (Sandberg et al. 2008; Hoffman et al. 2016; Fu et al. 2018). In addition to microRNA binding sites, 3' UTRs are also enriched for destabilization elements such as AU-rich elements, GU-rich elements and Puf-binding sites, which are bound by RNA binding proteins (Garneau, Wilusz, and Wilusz 2007).

A third consequence of 3' UTR APA is in mRNA and protein localization. Alternative RNA localization can be cell-type specific. For example, frequently, mRNAs isoforms with longer 3' UTRs are more often localized in the nucleus (Neve et al. 2016). 3' UTRs can also determine subcellular localization such as membrane or ER localization. One example of a gene that undergoes a change in mRNA localization is Vdac3, a voltage-dependent anion channel. When the distal polyA site of Vdac3 is used, the mRNA has reduced localization to the outer mitochondrial membrane in comparison to mRNAs in which the proximal polyA site is used. This localization is directly correlated to mitochondrial size and number, indicating that APA can regulate mitochondrial development (Arora et al. 2022). 3' UTR length can also regulate subcellular localization in the cytoplasm to promote localized translation, which is of particular relevance in neurons where isoforms with longer 3' UTRs localize to axons, dendrites, and synapses (Tushev et al. 2018).

Intronic Polyadenylation

In addition to polyA sites within the 3' UTR following the terminal exon, polyA sites can be localized in intronic regions before the end of the protein coding region of the mRNA. Polyadenylation at these sites is known as intronic polyadenylation. When intronic polyadenylation occurs, it frequently produces mRNAs that are degraded by nonsense-mediated decay although occasionally a truncated fragment may be produced if adenosines within the 3' UTR create a stop codon (Vasudevan, Peltz, and Wilusz 2002; P. Yao et al. 2012).

One role of intronic polyadenylation is in protein diversification. During B cell activation, there is a switch from usage of a polyadenylation site within the 3' UTR to an intronic polyA site in the IgM heavy chain mRNA, which shifts protein antibody production from a membrane-bound form to a secreted form. In other cases, truncated proteins produced as a result of intronic polyadenylation can function in a dominant negative role. This is seen in the case of Rbbp6, a member of the polyA machinery, which contains an IPA site that when utilized creates a truncated isoform that can compete with full length Rbbp6 for binding to the polyA machinery (Di Giammartino et al. 2014).

Finally, intronic polyadenylation can regulate protein availability. Pcf11 and CstF77, both members of the canonical polyA machinery, both contain conserved intronic polyadenylation sites. In both cases, levels of intronic polyA site usage are regulated through a negative feedback loop as high levels of the full length protein activates intronic polyA site usage (Kamieniarz-Gdula et al. 2019; R. Wang et al. 2019; Z. Pan et al. 2006). This shows that levels of core polyadenylation machinery components are highly regulated and that misregulation may influence APA.

Regulation of APA

Regulation of Canonical Polyadenylation Factors

Changes in expression levels and modifications of the canonical 3' processing factors alters alternative polyadenylation profiles. For example, a well-known case of altering polyA factor expression levels to regulate polyA site usage is during B-cell differentiation. Prior to B-cell differentiation, CstF levels are low, allowing the stronger, distal polyA site to have an advantage over the weaker, proximal polyA site. However, upon B-cell activation, CstF64 levels are upregulated, thus increasing levels of the CstF complex. Upon activation, there is now sufficient CstF available to polyadenylate at the proximal polyA site, leading to a distal to proximal shift in polyA site usage upon B-cell activation (Y Takagaki and Manley 1997). It is likely that similar mechanisms are used to regulate additional polyadenylation events.

An additional mechanism for regulating the canonical polyadenylation machinery is post-translational modification. Currently little is known about how or to what extent post-translational modifications of polyA factors regulate APA, but it remains an interesting field of study because post-translational modifications would allow for reversible regulation of polyA factors.

PolyA factors undergo a variety of modifications including phosphorylation, lysine acetylation, arginine methylation, and lysine sumoylation (Ryan and Bauer 2008). For example, PAP is hyperphosphorylated during mitosis, leading to a reduction in activity and repressed mRNA production (Colgan et al. 1996). In addition, it has been shown that hyperphosphorylation of the RS domain of CFIm59 and CFIm68 reduces its interaction with Fip1, thereby repressing its ability to regulate APA. Surprisingly, dephosphorylation of the RS domain of CFIm59 and CFIm68 did not interfere with its interaction with Fip1 (Xiao and Manley 1997; Zhu et al. 2017). This was interesting as both hyperphosphorylation and dephosphorylation of the RS domain of SR proteins that regulate splicing represses their ability to interact with U170K. One possible explanation for this difference is that while canonical SR protein RS domains are predominantly RS dipeptide rich, CFIm59 and CFIm68 are SR-like proteins with RE/D repeats in addition to

RS repeats. It is thus possible that RE/D dipeptides mimic phosphorylation, making the CFIm-Fip1 interaction less dependent on phosphorylation than the SR protein-U170K interaction (Zhu et al. 2017).

Regulation by RNA Binding Proteins

In addition to regulation by the canonical polyA machinery, a growing number of RNA binding proteins (RBPs) have been implicated in APA regulation, typically in a context specific manner. A well-known example is the Hu family of proteins, which inhibit polyA sites with U-rich elements. Interestingly, HuR, which is ubiquitously expressed, undergoes APA regulated by both itself and other neural-specific members of the Hu family including ElavL2, ElavL3, and ElavL4; this allows neurons to balance the pro-differentiation effects of the neural-specific member with the proliferative effects of HuR (Tian and Manley 2017).

Many splicing factors have also been shown to regulate APA. For example, the SR proteins SRSF3 and SRSF7 lengthen and shorten 3' UTRs respectively and SRSF10 has recently been shown to repress intronic polyadenylation (Tian and Manley 2017; Jobbins et al. 2022). In addition to SR proteins, additional members of the splicing machinery regulate polyadenylation. For example, the U2 auxiliary factor (U2AF), which promotes exon 3' splice site recognition, interacts with CFIm59 and promotes recruitment of CFIm to polyA sites. In addition, SF3b1 of the U2 snRNP interacts with CPSF and promotes efficient RNA cleavage (Millevoi et al. 2006; Kyburz et al. 2006).

Another family of splicing regulators that also impact alternative polyadenylation is the hnRNP family. hnRNP H promotes proximal polyadenylation site usage as demonstrated by iCLIP peaks near proximal sites and RNA-sequencing results indicating that there is a shift to distal polyadenylation site upon knockdown of hnRNP H. In addition, while hnRNP H and another hnRNP known as hnRNP F are closely related structurally, hnRNP H2 promotes CstF binding to polyadenylation sites whereas hnRNP F inhibits, indicating that the relative ratio of different RNA binding proteins can also impact how polyA sites are selected (Erson-Bensan and Apa 2016; Katz et al. 2010; Alkan, Martincic, and Milcarek 2006).

Functional Consequences of APA

APA and Cell Fate

A basic classification of cell fates is stem cells and differentiated cells. Stem cells are naïve, unspecialized cells that have two characteristic features: 1) the ability to self-renew indefinitely and 2) pluripotency, or the ability to differentiate into other cell types. During differentiation, stem cells change in form and become more mature, from generalized to more specialized functions (Verfaillie 2004).

Several lines of evidence suggest that APA is linked to cell fate determination. First, during mouse embryonic development, there is progressive lengthening of 3’ UTRs (Z. Ji et al. 2009). Consistently, there is also progressive shortening of 3’ UTRs during reprogramming of mouse embryonic fibroblasts (a differentiated cell type) into induced pluripotent stem cells, with analysis showing opposing changes in many genes during differentiation and reprogramming (Z. Ji and Tian 2009).

Several studies provide insight into how alternative polyadenylation and cell fate may be linked. First Fip1, a subunit of CPSF, has been shown to be critical for stem cell identity as knockdown of Fip1 in embryonic stem cells led to a widespread shift from proximal to distal polyA sites as well as loss of stem cell self-renewal and induction of differentiation. Mechanistically, Fip1 was found to promote proximal polyA sites, providing insight into how mRNA isoforms with shorter 3' UTRs may be promoted in stem cells (Lackford et al. 2014a).

By contrast, knockdown of CFIm25 in mouse embryonic fibroblasts led to a 30-fold increase in reprogramming into induced pluripotent stem cells, suggesting that CFIm is a roadblock to protect from aberrant cell division and promote a differentiated state. Consistent with the known role of CFIm in promoting distal polyadenylation sites, there was also a widespread shift from distal to proximal polyA site usage (Justin Brumbaugh et al. 2018; Zhu et al. 2017). Both the roles of Fip1 and CFIm in cell fate are consistent with mRNA isoforms with longer 3' UTRs being upregulated in differentiated cells.

APA and Cancer

Like other RNA processing events including RNA splicing, mis-regulation of alternative polyadenylation is associated with many human diseases including cancer. Several studies indicate that there is a widespread shift from distal to proximal polyA site usage during cellular transformation and carcinogenesis, consistent with the knowledge that more proliferative cells have mRNA isoforms with shorter 3' UTRs (Mayr and Bartel 2009; Lin et al. 2012; Sandberg et al. 2008). In addition, cancer cell lines exhibit 3' UTR shortening of proto-oncogenes that increases both mRNA stability and protein levels, although the proto-oncogene itself is not genetically altered (Mayr and Bartel 2009). Importantly, these studies also identified mRNAs that undergo 3'UTR lengthening during cancer progression, suggesting that while there is a general trend towards proximal polyA site usage in cancer cells, there are exceptions to this generalization (Mayr and Bartel 2009).

Cancer-relevant APA changes are not limited to 3'-UTR APA as there are also changes in intronic polyadenylation. For example, in chronic lymphocytic leukemia, there is aberrant production of truncated mRNAs for tumor suppressor genes caused by enhanced intronic polyadenylation. These truncated proteins either lacked their tumor-suppressive functions or were even oncogenic (Lee et al. 2018). In addition, intronic polyadenylation is also relevant within BRCA tumors, which are frequently caused by loss of function mutations with homologous repair (HR) genes. Interestingly CDK12 is one of the 22 genes common in BRCA tumors and is the only gene not involved in HR. Recently it was discovered that CDK12 globally suppresses intronic polyadenylation, particularly within genes critical for homologous repair upon DNA damage in human tumors. (Dubbury, Boutz, and Sharp 2018).

The precise mechanism that regulates the cancer-specific APA profile is relatively unknown. Several studies have identified different polyA factors as being dysregulated in cancer, suggesting that the APA changes may not have a universal but are in fact cancer-type specific. CFIm25 was shown to be downregulated in glioblastoma, leading to both 3' UTR shortening and increased proliferation (Masamha et al. 2014). Consistently, CFIm is also downregulated in lung cancer and regulates APA of oncogenes (Huang et al. 2018). In addition to CFIm, CPSF73 was recently discovered to be the target of the anti-cancer compound JTE-607, suggesting that CPSF73 may also play a role in carcinogenesis in acute myeloid leukemia and Ewig's sarcoma (Ross et al. 2020). Other polyA factors that have been linked to cancer progression include Pcf11 and CstF64 (Ogorodnikov et al. 2018; Hwang et al. 2016).

1.3 PRE-mRNA SPLICING

In addition to polyadenylation, eukaryotic pre-mRNAs undergo additional processing events that also include splicing, or the exclusion of non-coding introns and the joining together of protein-coding exons. Like polyadenylation, splicing can also be alternatively regulated, which is critical for diversification of the genome. As both splicing and 3' processing are key regulatory events in gene expression, it is crucial not only understand their regulation but also their relationship to each other. In this section, I will examine the mechanism of splicing, splicing regulation, and its link to polyadenylation.

Splicing Mechanism

Splicing, or the removal of non-coding introns and joining together of protein coding exons, is mediated by the interaction of cis regulatory elements and the spliceosome, a complex of the U1, U2, U4, U5 and U6 small nuclear proteins (snRNPs) and its associated proteins. Each snRNP is composed of small nuclear RNAs and associated proteins (Black 2003; Wilkinson, Charenton, and Nagai 2020).

Critical sequences for intron definition include the 5' splice site, the 3' splice site and the branch point site (Figure 1.5). The 5' splice site is located at the exon-intron boundary. It is highly conserved and consists of the sequence AG/GURAGU (where R represents a purine [A or G] and / represents the exon-intron boundary). The 3' splice site is located at the intron-exon boundary. While less conserved than the 5' splice site, it consists of 3 elements: the YAG sequence directly preceding the intron-exon boundary, the branch point sequence with the consensus sequence YNYURAY (where Y represents a pyrimidine [C or U]), and the polypyrimidine tract characterized by a high frequency of pyrimidines (C or U) (Black 2003; Wilkinson, Charenton, and Nagai 2020; Moore, Query, and Sharp 1993).

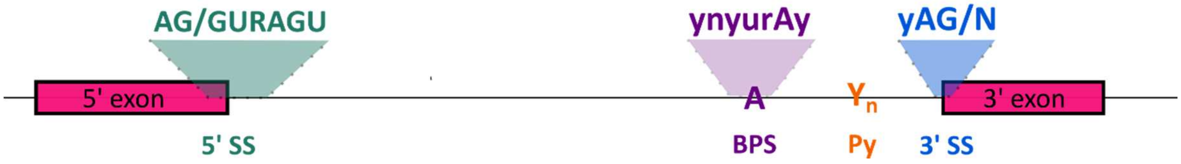


Figure 1.5 Cis Elements for Splice Site Selection

Schematic of 5’ splice site (exon/intron) and 3’ splice site (intron/exon) junctions and associated cis elements. Y is a pyrimidine (C or U), R is a purine (A or G), N is any nucleotide, and / is junction between intronic and exonic regions. Exons are represented by boxes and introns are represented by a thin solid line.

5’ss: 5’ splice site	BPS: branch point site
Py: polypyrimidine tract	3’ss: 3’ splice site

Spliceosome assembly is initiated by the direct base pairing of the U1 snRNP to the 5' splice site and the binding of the U2 auxiliary factor (U2AF) to the 3' splice site. U2AF consists of 2 subunits: U2AF35 which binds to the YAG sequence at the 3' splice site and U2AF65 which binds to the polypyrimidine tract, frequently a series of T and C. This binding of the U1 snRNP and U2AF forms the E complex, which is the only ATP-independent step of spliceosome assembly. Additionally, Sf1 binds to the branch point. The next step in spliceosome assembly is the formation of the pre-splicesome or A complex when Sf1 is displaced and the U2 snRNP is recruited to the branch point. Subsequently, the tri-snRNP complex consisting of U4-U5-U6 is recruited and the mRNA is incorporated into the fully assembled pre-spliceosome during pre-catalytic B complex formation. Structural rearrangements lead to the disassociation of the U1 and U4 snRNPs to form the catalytic B complex, which is followed by formation of C complex following further structural rearrangements and the association of the U6 snRNP with both U2 and the 5' splice site. Once the catalytic B complex is formed, the first transesterification reaction of splicing occurs. During this reaction, the 2'OH of the branch point adenosine attacks the 5' phosphate of the 5' splice site; this causes the 5' end of the intron to be ligated to the branch site and produces two reaction intermediates: a detached 5' exon and the lariat intermediate consisting of the intron-3'exon fragment. During the second transesterification reaction, the free 3'OH of the free 5' exon attacks the 5' phosphate of the 3' exon. The exons are now ligated together and the intron is released in the lariat conformation (Black 2003; Moore, Query, and Sharp 1993; Wilkinson, Charenton, and Nagai 2020) (Figure 1.6).

Alternative Splicing

Approximately 95% of eukaryotic mRNAs undergo alternative splicing, greatly expanding the diversity of the human genome (Q. Pan et al. 2008). Four general categories of alternative

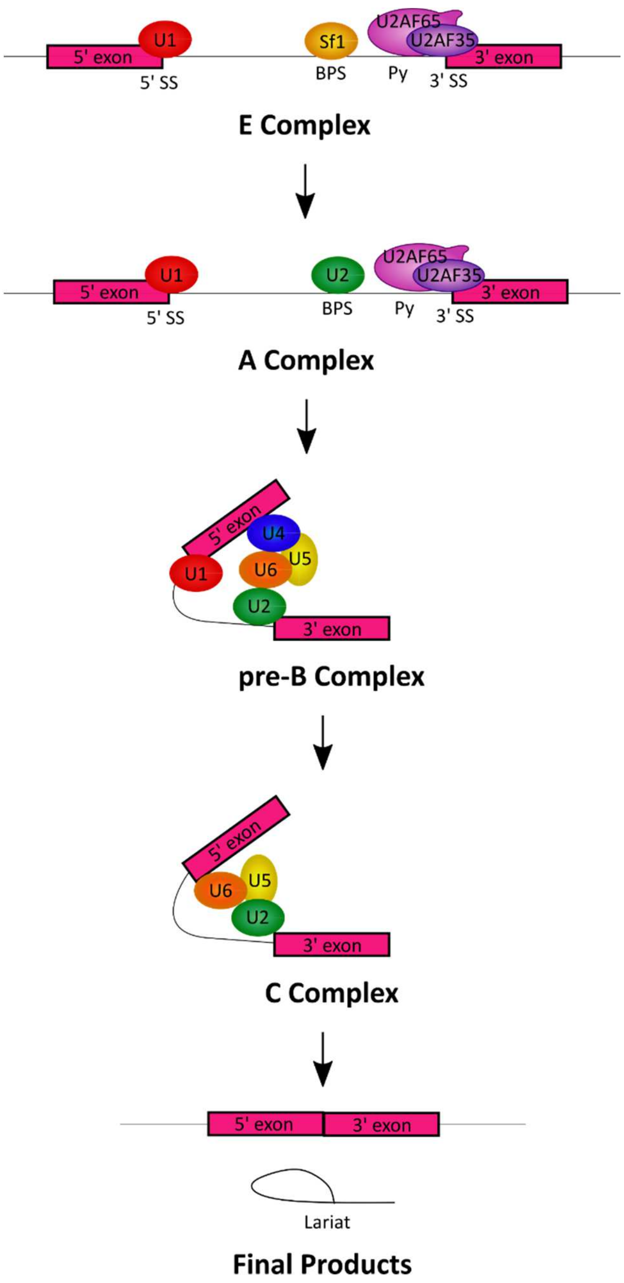


Figure 1.5 Spliceosome Assembly

Splicing is catalyzed by the spliceosome which consists of the U1, U2, U4, U5 and U6 snRNPs. Spliceosome components assemble on the pre-mRNA in a stepwise manner and transitions between E, A, pre-B, and C complexes. The mature mRNA contains exons that have been joined together and intron is released as a lariat byproduct.

splicing include cassette exons (exon is alternatively included or excluded), intron retention (intron is either included or excluded), and alternative 5' or 3' splice sites (altering the 3' boundary of the upstream exon or the 5' boundary of the downstream exon respectively) (Figure 1.7). Alternatively spliced exons frequently have weaker 5' and 3' splice sites than constitutive exons and are therefore less efficiently recognized by the spliceosome (Baek and Green 2005; Garg and Green 2007; Zheng, Xiang-Dong, and Gribskov 2005). As a result, alternatively spliced exons rely on other regulatory factors to promote proper exon inclusion or exclusion. In this section, I will highlight the RNA cis elements and protein factors involved in alternative splicing regulation as well as its biological significance.

Alternative Splicing Regulation

Many variables influence whether an alternative splicing event will occur. First, 5' and 3' splice sites have variable strengths based upon their ability to be recognized by U1 and U2AF respectively. 5' splice site recognition is mediated by direct base-pairing of the U1 snRNA to splice site. As a result, 5' splice sites with high complementarity to the U1 snRNA are strong 5' splice sites, whereas 5' splice sites with lower complementarity are weaker (Roca, Sachidanandam, and Krainer 2005). 3' splice sites, by contrast, are defined by three sequence features: the branch point sequence, the polypyrimidine tract, and the intron/exon junction. Of these sequences, the polypyrimidine tract is the most variable, and therefore 3' splice site strength is determined by the ability of U2AF65 to recognize the polypyrimidine tract. Factors that influence polypyrimidine tract strength include length and composition, with strong 3' splice sites being defined by longer, contiguous polypyrimidine tracts with a high enrichment of uridine (Coolidge, Seely, and Patton 1997; Hertel 2008). Exons with strong 5' and 3' splice sites are

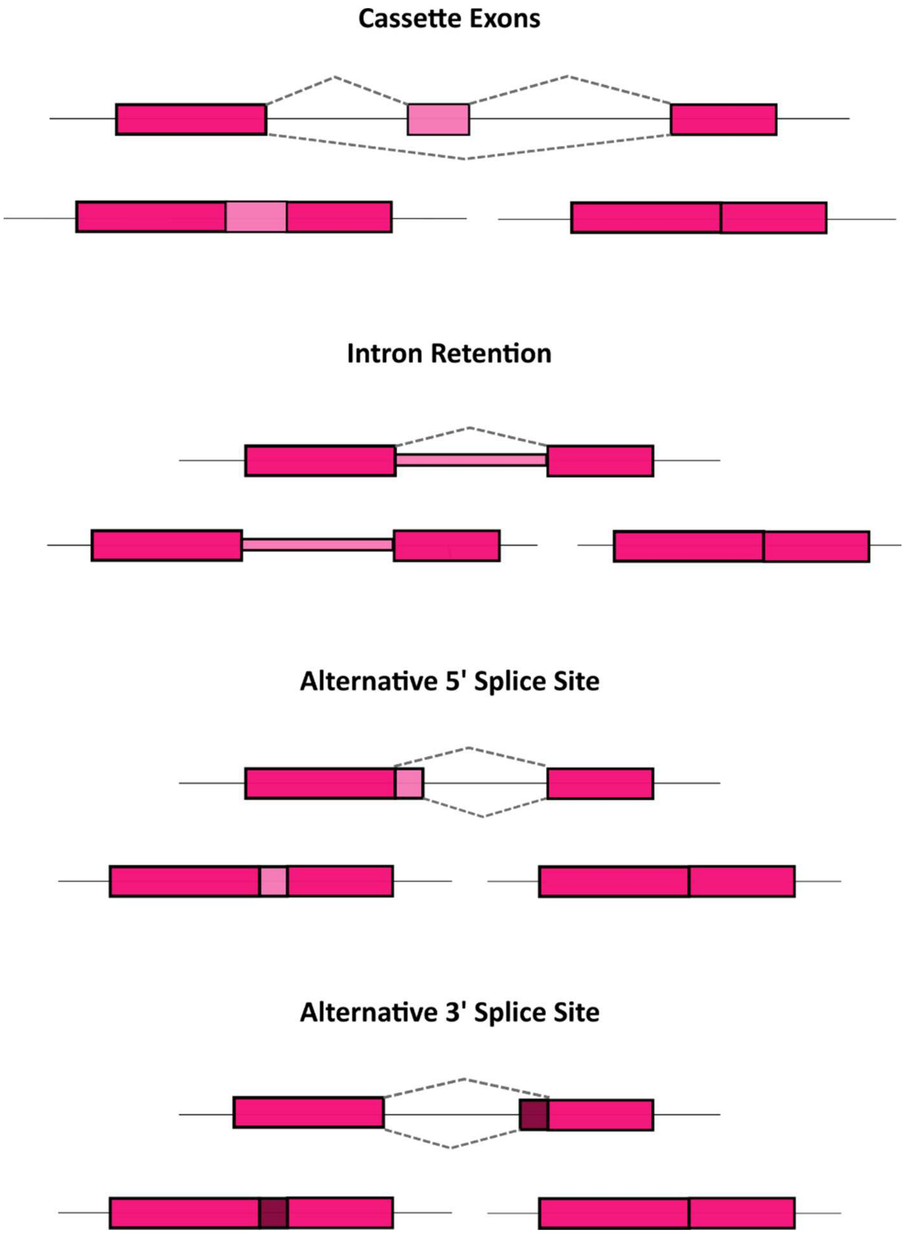


Figure 1.7 Alternative Splicing

Schematic of alternative splicing events. Constitutive exons are represented by dark pink boxes, alternative events are represented by either light pink or maroon boxes, and introns are represented by solid lines. Alternative splicing events are represented by dashed gray lines.

more frequently constitutive exons while exons with weaker 5' and/or 3' splice sites are more frequently skipped (Zheng, Xiang-Dong, and Gribskov 2005).

Additionally, splicing regulatory elements (SREs) can modulate alternative splicing by promoting or inhibiting recruitment of members of the splicing machinery. Broadly, SREs can be divided into two categories: splicing enhancers and splicing inhibitors. Both enhancers and inhibitors can be either exonic (exonic splicing enhancers [ESE] or exonic splicing silencers [ESS]) or intronic (intronic splicing enhancers [ISE] or intronic splicing silencers [ISS]).

SREs are recognized by two main families of trans-acting protein factors: SR proteins and hnRNPs. SR proteins are characterized by an N-terminal RNA recognition motif and a C-terminal RS domain which is enriched with arginine serine repeats. There are twelve canonical SR proteins that are designated SRSF1-12. There are also numerous SR like proteins that also regulate alternative or constitutive splicing and share an RS domain but either lack or have a different RNA binding domain or do not have the ability to complement splicing reactions (Busch and Hertel 2012). SR proteins activate splicing when bound to ESEs by recruiting U2AF and U1 to 3' and 5' splice sites respectively (Dvinge 2018; J. C. Long and Caceres 2009; Graveley, Hertel, and Maniatis 2001). By contrast, hnRNPs are characterized by one or more RNA binding domains as well as unstructured regions that are frequently engaged in protein-protein interactions. The RNA binding domain can consist of an RNA recognition motif, a quasi-RRM, a KH domain, or a RGG box, with the RRM being most common. There are over 20 major hnRNPs as well as minor hnRNPs which sometimes lack RNA binding capabilities but regulate major hnRNPs (Geuens, Bouhy, and Timmerman 2016). Over half of alternative splicing events are regulated by more than one hnRNP, which can act both synergistically and antagonistically (Dvinge 2018; Huelga et al. 2012).

Historically, SR proteins were considered activators of splicing and hnRNPs were repressors. However, it was more recently demonstrated that both protein families have position dependent effects with SR proteins activating splicing when bound to ESEs and inhibiting splicing when bound to ISSs and hnRNPs have the converse effect: activating splicing when bound to ISEs and inhibiting splicing when bound to ESSs (Erkelenz et al. 2013). This highlights that the context of SR protein or hnRNP binding is highly relevant to its role in splicing regulation.

Biological Relevance of Alternative Splicing

Like other RNA processing events, regulation of alternative splicing is critical, and mis-regulation can cause numerous disease phenotypes including various forms of cancer, retinitis pigmentosa, and neurological disorders including Alzheimer's disease (Scotti and Swanson 2015).

Splicing is of particular relevance in cancer. Analysis of 8000 tumors over 32 cancer types revealed thousands of alternatively spliced variants and cancer-specific markers and neo-antigens. Over 119 cancer driver mutations have been discovered within splicing factors or regulators and over 70% of splicing factors change expression levels during tumorigenesis, within cancer subtypes showing independent signatures in both alternative splicing events and mis-regulated factors (Bonnal, López-Oreja, and Valcárcel 2020). Splicing-relevant mutations frequently increase cell proliferation, induce angiogenesis, enhance invasion and/or metastasis, or alter energy metabolism. They can also either activate oncogenes or inhibit tumor suppressors. The most frequently mutated splicing regulatory proteins include SF3B1 (component of the U2 snRNP), U2AF35, and SRSF2. These mutations are frequently missense mutations and resemble oncogenes (Bejar 2016).

Interestingly, it has been hypothesized that while splicing alterations can be advantageous to tumors, this comes at the cost of decreased splicing fidelity and efficiency. This can be exploited in chemotherapy treatments as further perturbations such as mutations or inhibitors can cause selective cytotoxicity against cancer cells. For example, cancer driver mutations are usually heterozygous and mutually exclusive and co-mutation of SRSF2 with Sf3b1 causes cell death. In addition, Myc driven cancers are especially vulnerable to the depletion of splicing inhibitors or splicing inhibitors, likely due to increased burden on the spliceosome caused by activation of gene expression induced by Myc (Bonnal, López-Oreja, and Valcárcel 2020). This suggests that while dysregulation of splicing can enhance tumor progression, it is also an important therapeutic target.

Links Between Polyadenylation and Splicing

Several lines of evidence suggest that splicing and polyadenylation are not in fact isolated events but are in fact tightly coupled. First, affinity purifications of the spliceosome have identified members of the core polyadenylation machinery as co-purifying including subunits of CFIm, CPSF, and CstF (Zhou et al. 2002; Rappsilber et al. 2002). Consistently, purification of the polyadenylation machinery also co-purified members of the splicing machinery including spliceosomal components and U2AF (Shi et al. 2009). Together, this indicates that the splicing and polyadenylation machineries are tightly linked with each other. One possible explanation for this phenomenon is that both polyadenylation and splicing are co-transcriptional RNA processing events, suggesting that they may occur simultaneously (McCracken et al. 1997; Zeng and Berget 2000; Bittencourt and Auboeuf 2012). As both splicing and polyadenylation are co-transcriptional and their machineries interact, it seems likely that they may be able to cross-regulate.

Polyadenylation and splicing factors both regulate terminal exon recognition. The exon definition model of splicing states that splicing occurs when U1 recognizes the downstream 5' splice site of the exon and U2AF recognizes the upstream 3' splice site. However, first and terminal exons each lack one of these features: the upstream 3' splice site and the downstream 5' splice site respectively (Berget 1995). Previously, it has been shown that terminal exon recognition is regulated by U2AF binding at the 3' splice site and interactions between the polyadenylation and splicing machineries, specifically CFIm or PAP with U2AF (Millevoi et al. 2006; Cooke, Hans, and Alwine 1999; Vagner, Vagner, and Mattaj 2000). In addition, terminal exon recognition is also likely regulated by interactions of CPSF with SF3B (which is a part of the U2 snRNP) as the presence of CPSF was necessary for efficient splicing activity and, likewise, mutation of the U2 snRNP impaired cleavage activity in addition to splicing (Kyburz et al. 2006). While it has been demonstrated that splicing and polyadenylation are linked for terminal exon recognition and processing, little is known about how the polyadenylation machinery may affect upstream alternative splicing events or vice versa, making this an interesting area of research as it is a potentially novel way to diversify the genome.

A final well-known example of splicing regulators regulating alternative polyadenylation is U1 snRNA-mediated inhibition of intronic polyA sites, also known as telescripting. Telescripting is caused by U1 snRNA binding within introns and involves formation of complexes that are distinct from the U1-snRNP (Kaida et al. 2010). This highlights that the splicing machinery and polyadenylation machineries may be in competition for regulation of intronic polyadenylation sites.

1.4 SUMMARY

Polyadenylation and splicing are two highly-regulated RNA processing events that require the precise coordination of numerous cis-regulatory elements and trans-acting factors. Changes to this regulation can be detrimental and cause aberrant disease phenotypes. This chapter has outlined what is currently known about their regulatory mechanisms as well as the outcomes when these regulatory mechanisms are altered. However, there are still many unanswered questions. This work will provide insight into the interconnected relationship between RNA processing events, particularly splicing and polyadenylation. Specifically, RNA binding proteins with well-established functions in one particular RNA processing event will be re-evaluated to identify previously uncharacterized roles in other RNA processing events.

Chapter 2 utilizes a dual-luciferase reporter to screen RNA binding proteins for novel roles in alternative polyadenylation regulation. Many of the RBPs utilized within the screen have previously characterized roles in other RNA processing events including splicing, transcription, RNA export, and mRNA stability, highlighting the often-overlapping roles of RNA binding proteins in RNA processing. As predicted, we identified several novel regulators of APA, particularly polyA site inhibitors. Interestingly, we also establish that the SR family of proteins that activate splicing in fact inhibit polyadenylation.

Chapter 3 narrows in on a new mechanism for the regulation of one specific type of polyadenylation event: intronic polyadenylation. Through genome-wide sequencing and biochemical analyses, it was discovered that the core polyadenylation machinery complex component cleavage factor I (CFIm) promotes intronic polyadenylation site selection in a variety of genes in addition to its canonical role in enhancing distal polyA sites. In fact, CFIm regulates the intronic polyadenylation of one subunit of each other member of the core polyadenylation machinery, indicating that it may be a master regulator of 3’ processing. By promoting both distal and intronic polyA sites, CFIm also regulates global protein abundance, which may play a role in linking 3’ processing to cell fate determination.

Finally, Chapter 4 explores another new role for CFIm, in this case in linking 3’ processing and splicing. In addition to being a 3’ processing factor, CFIm is determined to be a global alternative splicing regulator at exons upstream of the terminal exon. CFIm interacts with U2AF and binds to 3’ splice sites, thereby activating a subset of 3’ splice sites. In addition to proposing a new role for CFIm, our finding also presents an intriguing model for 3’ splice site recognition in which a variety of RNA binding proteins can compete to complex with U2AF to activate unique 3’ splice sites.

CHAPTER 2

RNA BINDING PROTEINS HAVE NOVEL ROLES IN REGULATING ALTERNATIVE POLYADENYLATION

2.1 SUMMARY

Over 70% of eukaryotic pre-mRNA contain more than one polyadenylation site, allowing them to be alternatively regulated. As mis-regulated alternative polyadenylation has been associated with many human diseases including neurological disorders and cancers, it is critical to understand the mechanisms involved in polyA site selection. However, despite the importance of understanding the mechanisms involved in APA, there is still limited ability to predict polyA site selection based upon our knowledge of the core polyadenylation machinery itself, suggesting that additional mechanisms may be involved.

One factor that contributes to our inability to predict how alternative polyadenylation will occur is that although the polyadenylation machinery has been characterized, it is in fact highly dynamic and includes many proteins outside of its core components, suggesting that by analyzing RNA binding proteins, we may be able to better understand how polyA sites are selected (Shi et al. 2009). The human genome encodes over 1500 RNA binding proteins (RBPs) (Dominguez et al. 2018). Given the abundance of RBPs, it is likely that there are additional proteins that regulate APA, particularly in context-specific manners including cell type. Among the potential candidates for polyadenylation regulators are splicing regulatory (SR) proteins because the polyadenylation factor cleavage factor I (CFIm) is structurally similar to SR

proteins; in addition, CFIm and SR proteins use similar mechanisms to regulate polyA site and splice site selection respectively (Zhu et al. 2017).

To identify novel APA regulators, we screened splicing regulatory proteins and other RBPs for polyadenylation regulatory activity using a dual luciferase reporter. In addition to recapitulating known regulators of alternative polyadenylation such as hnRNP H, we identified several RBPs with likely roles in polyA site selection that have not previously been identified in humans including several hnRNPs as well as Musashi1. We also identify SR proteins as a novel class of proteins that inhibit polyadenylation, suggesting cross-regulation between polyadenylation and splicing. Finally, we establish that for many RBPs, regulation of APA is position independent, with similar effects when RBPs binding upstream and downstream of the cleavage site within our reporter. Our study not only identifies novel polyadenylation regulators, but also provides insight into the types of RNA binding proteins that regulate APA and therefore allows us to make predictions about other related proteins or RBP subtypes that may have additional roles in APA.

2.2 INTRODUCTION

mRNA 3' processing requires a wide range of both protein-RNA and protein-protein interactions to occur. Many of these protein factors are subunits of the core polyadenylation machinery consisting of CPSF, CstF, CFIm, and CFIIm. These protein complexes have been extensively studied to better understand how they select polyadenylation sites. Understanding how polyA sites are selected is especially important in the context of alternative polyadenylation, which occurs on the approximately 70% of pre-mRNAs that contain more than one polyadenylation site (S. Chan, Choi, and Shi 2011). Mis-regulation of alternative polyadenylation is linked to many human diseases including neurological disorders and cancer, so it is critical to understand the mechanisms involved in polyadenylation site selection. However, studies have shown that in cases of alternative polyadenylation, only 70% of the dominant cleavage sites have the canonical polyadenylation signal A(A/U)UAAA. In addition, while binding of CstF and CFIm are considered the most predictive of dominant polyA sites, CstF binding only successfully predicts the dominant site 50% of the time and CFIm binding has an even lower predictive rate at only 39-42%. While joint binding of CstF and CFIm has a higher prediction accuracy than either protein complex alone at 60%, 40% of cleavage sites still cannot be predicted by binding of the core polyadenylation machinery, suggesting that other mechanisms are involved (G. Martin et al. 2012).

One additional mechanism may include regulation by RNA binding proteins (RBPs) that are not members of the core polyadenylation machinery. Recent studies estimate that the human genome encodes over 1500 RNA binding proteins, which is approximately 7.5% of the human genome (Dominguez et al. 2018; Weiße et al. 2020). RBPs are a broad family of proteins that have one or more RNA-binding domain. These RNA binding domains can include RNA recognition motifs (RRMs), K-homology domains, RGG boxes, and zinc fingers, among others. Of these, RNA recognition motifs are the most common (Oliveira et al. 2017). Given the general abundance of RBPs, it seems likely that there may be additional members of the RBP family that regulate alternative polyadenylation. Additionally, 3' UTRs are enriched for binding sites of RNA binding proteins, in particular AU-rich elements which are bound by trans-acting factors that can have diverse roles including regulating mRNA stability and translation efficiency. There are at least 10 RBPs known to bind AU-rich elements; for example, KHSRP destabilizes mRNAs while HuR stabilizes (Mayr 2019). As 3' UTRs are enriched for binding sites of RBPs, it seems likely that they may be able to regulate APA as well.

There are a variety of models by which RBPs may regulate APA. For example, an RBP could bind to a cis regulatory element on pre-mRNA and recruit the core polyA machinery to a nearby site, leading to polyA site activation (Figure 2.1A). Alternatively, an RNA binding protein might compete for binding to or near a polyadenylation site, leading to decreased polyadenylation site usage (Figure 2.1B). At the same time, many proteins have charged amino acids including arginine, histidine, and lysine (positive) or aspartic acid and glutamic acid (negative) that when clustered together in the amino acid sequence create a net charge on a protein or domain. RBPs with opposing charges to members of the core polyadenylation machinery may activate polyadenylation through recruitment (Figure 2.1C) whereas RBPs with similar charges may interfere with machinery binding through repulsion (Figure 2.1D). While there are no reported cases of attraction or repulsion affecting 3' processing machinery binding, it has been shown in prostate cancer that different subcategories of transcription factors can inhibit tandem binding by electrostatic repulsion of positively charged residues (Madison et al. 2018).

Supporting the role for RBPs in APA, in 2009, an affinity purification and proteomic characterization of the polyadenylation machinery was performed to identify the proteins involved in 3' processing. In addition to identifying all known core polyadenylation machinery members expect for Clp1 (a subunit of CFIIm), the study revealed that the 3' processing machinery consists of approximately 85 proteins, many of which were not members of the core machinery. Other associated proteins have a wide range of functions including DNA damage response proteins, transcription factors, and splicing factors. Interestingly, two of these proteins, Wdr33 and Rbbp6, were later determined to be critical members of the core polyadenylation

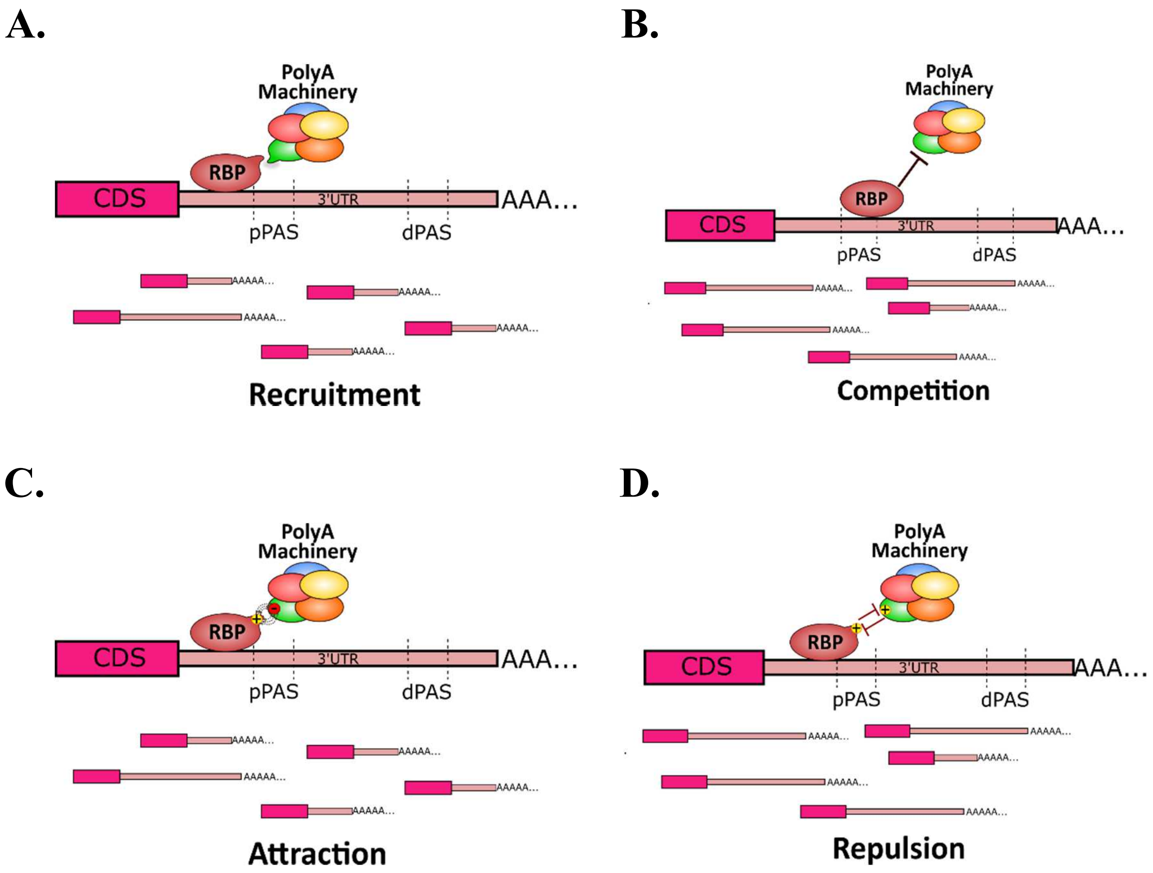


Figure 2.1 Potential Models for RBP Regulation of APA

A. **Recruitment**: When an RBP interacts with a member of the core 3' processing machinery and binds to a region nearby a polyadenylation site, it recruits the polyadenylation machinery and activates the polyadenylation site.
B. **Competition**: If an RBP binding sequence overlaps with the polyadenylation site, it may compete with the 3' processing machinery for binding and inhibit the polyadenylation site.
C. **Attraction**: When a RBP or RBP domain has an opposing charge to the core polyadenylation machinery, it can attract the polyadenylation machinery and enhance binding.
D. **Repulsion**: When an RBP or RBP domain has a similar charge to a member of the 3' processing machinery, it can repel the machinery and inhibit 3' processing.

pPAS: Proximal polyA site dPAS: distal polyA site machinery, with Wdr33 being a subunit of CPSF and Rbbp6 associating with CPSF and being essential for RNA cleavage (Shi et al. 2009; Boreikaite et al. 2022; Schmidt et al. 2022). This indicates that the affinity purification was able to identify novel polyadenylation factors. Additionally, it suggests that other associated proteins may have real roles in 3' processing, despite having known roles in other RNA processing events.

There are also several individual reports of RBPs with roles in polyadenylation site selection. For example, hnRNP H2 has been reporter to activate proximal polyadenylation sites by recruiting CstF whereas the closely related hnRNP F inhibits polyadenylation (Katz et al. 2010; Erson-Bensan and Apa 2016; Alkan, Martincic, and Milcarek 2006). In other cases, RBPs regulation of APA can be position dependent such as in the case of Muscleblind1 (Mbnl1). Mbnl1 activates polyadenylation when bound upstream of the polyadenylation site by recruiting members of the core polyadenylation machinery but inhibits when its binding site overlaps with the A(A/U)UAAA hexamer (Erson-Bensan and Apa 2016; Batra et al. 2014).

Another family of RNA binding proteins that may regulate APA is the SR family of proteins. SR proteins are a family of 12 proteins that all contain at least 1 RNA recognition motif and a C-terminal RS domain (Figure 2.2A). SR proteins have roles in both constitutive and alternative splicing. During constitutive splicing, SR proteins promote pairing of 5' and 3' splice sites across introns (Shepard and Hertel 2009). In addition, to regulate alternative splicing, SR proteins bind to 2 different regions: exons and introns with position dependent effects. Within exons, SR proteins bind to exonic splicing enhancers and promote exon inclusion (Cho et al. 2011; Saha and Ghosh 2022; J. Y. Wu and Maniatis 1993; Graveley, Hertel, and Maniatis 2001). Conversely, SR proteins also bind to intronic splicing silencers thereby repress alternative splicing (Erkelenz et al. 2013). SR proteins likely to regulate APA because they use similar

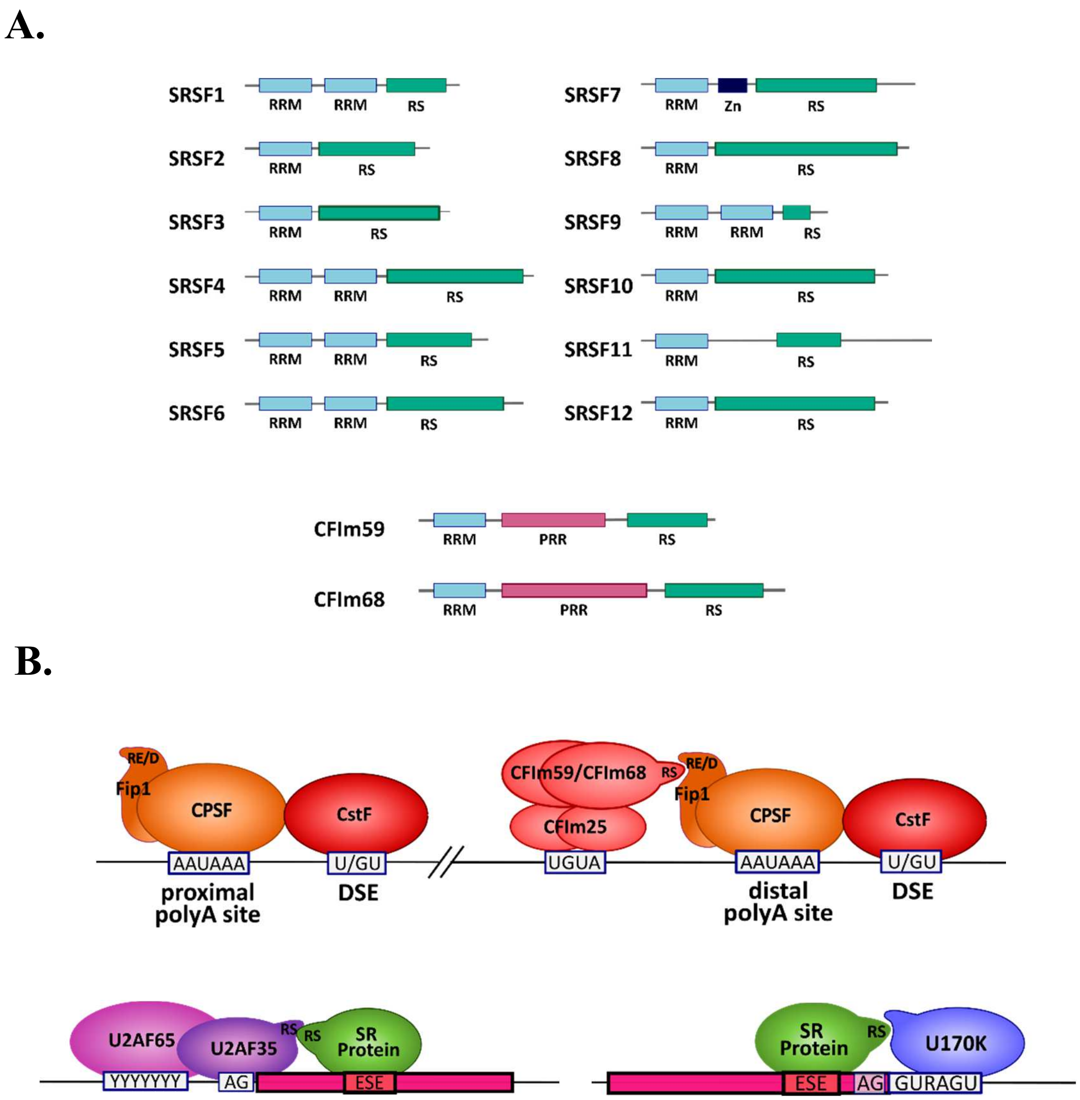


Figure 2.2 CFIm and SR Proteins Use Similar Regulatory Mechanisms

A. Schematic comparing domain of canonical SR proteins and SR-like proteins CFIm59 and CFIm68.
B. Model comparing regulation of polyadenylation by CFIm and splicing by SR proteins. The RS domain of CFIm59 and CFIm68 recruits CPSF to nearby polyadenylation sites by interacting with the RE/D domain of Fip1. Similarly, SR proteins bind to exonic splicing enhancers and recruit U2AF or U170K to nearby 3’ splice site or 5’ splice sites respectively by interacting with the RS domain of U2AF35 or U170K.

mechanisms to regulate splicing as CFIm uses to regulate polyadenylation. CFIm59 and CFIm69, the large subunits of CFIm, are categorized as SR-like proteins because they share in common several key functional domains with SR proteins including the RRM and RS domains (Figure 2.2A). CFIm is an activator of distal polyA sites that binds to the UGUA enhancer motif enriched upstream of distal polyA sites through the Nudix domain of the small subunit, CFIm25. CFIm then recruits the rest of the core polyadenylation machinery to the A(A/U)UAAA motif approximately 50 nucleotides downstream through interactions of its RS domain with the RE/D domain of Fip1, a subunit of the cleavage and polyadenylation specificity factor (CPSF). This is strikingly similar to the role that SR proteins play in splicing regulation because SR proteins activate splicing when bound to exonic splicing enhancers by recruiting U2AF to the 3' splice site and U170K to the 5' splice site of the exon through RS-RS interactions (Cho et al. 2011; Saha and Ghosh 2022; J. Y. Wu and Maniatis 1993; Graveley, Hertel, and Maniatis 2001). As both CFIm and SR proteins use their RS domains to regulate polyadenylation and splicing respectively, we hypothesized that SR proteins may play a role in polyadenylation site selection (Figure 2.2B).

Additionally, individual SR proteins have previously been shown to regulate APA. For example, knockdown of SRSF3 promotes 3' UTR shortening and cellular senescence. Mechanistically, SRSF3 binds stronger to proximal polyA sites than distal polyA sites, making it likely that SRSF3 inhibits proximal polyA sites (T. Shen et al. 2019). Additionally, SRSF7 has been found to promote distal polyA sites by recruiting Fip1, a mechanism similar to that which we propose is universal for SR proteins (Schwich et al. 2021). As SRSF3 and SRSF7 seem to have opposing roles, it will be important to gain further knowledge about the SR family in polyadenylation to establish whether all proteins regulate APA and if this is true, whether they all have similar functions.

In this chapter, I utilize a dual luciferase reporter to perform a large-scale screen to characterize RNA binding proteins with novel roles in polyA site selection. In addition to confirming the regulatory capability of RBPs such as hnRNP H, our screen identifies RBPs with no previous known role in APA including Musashi1 (Msi1) and several hnRNPs. We also establish SR proteins as a novel family of proteins with an inhibitory effect on APA. Finally, as most RBPs screened within our study have similar effects when bound upstream and downstream of the cleavage site, we propose that in the majority of cases, the role of RBPs on polyA site selection is position independent. While previous studies have reported individual cases of RBPs that regulate APA, there has never previously been an unbiased and comprehensive effort to identify both the types of RBPs that regulate APA but also to compare their regulatory capacity with one another. As a result, our screen provides critical insight into how APA is regulated.

2.3 RESULTS

Dual-Luciferase Reporter Screen Identifies To Test for Regulation of APA by RBPs

To investigate the role of RBPs in polyadenylation regulating, approximately 60 RBPs of interest were first identified for the screen (Table 2.1). RBPs came from several categories. Due to the interest in understanding whether splicing regulatory proteins also regulate polyadenylation, all 12 canonical SR proteins and 21 canonical hnRNPs were included. In addition, several RBPs with either known or predicted roles in APA were selected. For example, ElavL1 (HuR) has previously been identified to autoregulate its own expression by inhibiting binding of CstF to proximal polyA sites and thereby causing 3' UTR lengthening. This proximal to distal shift

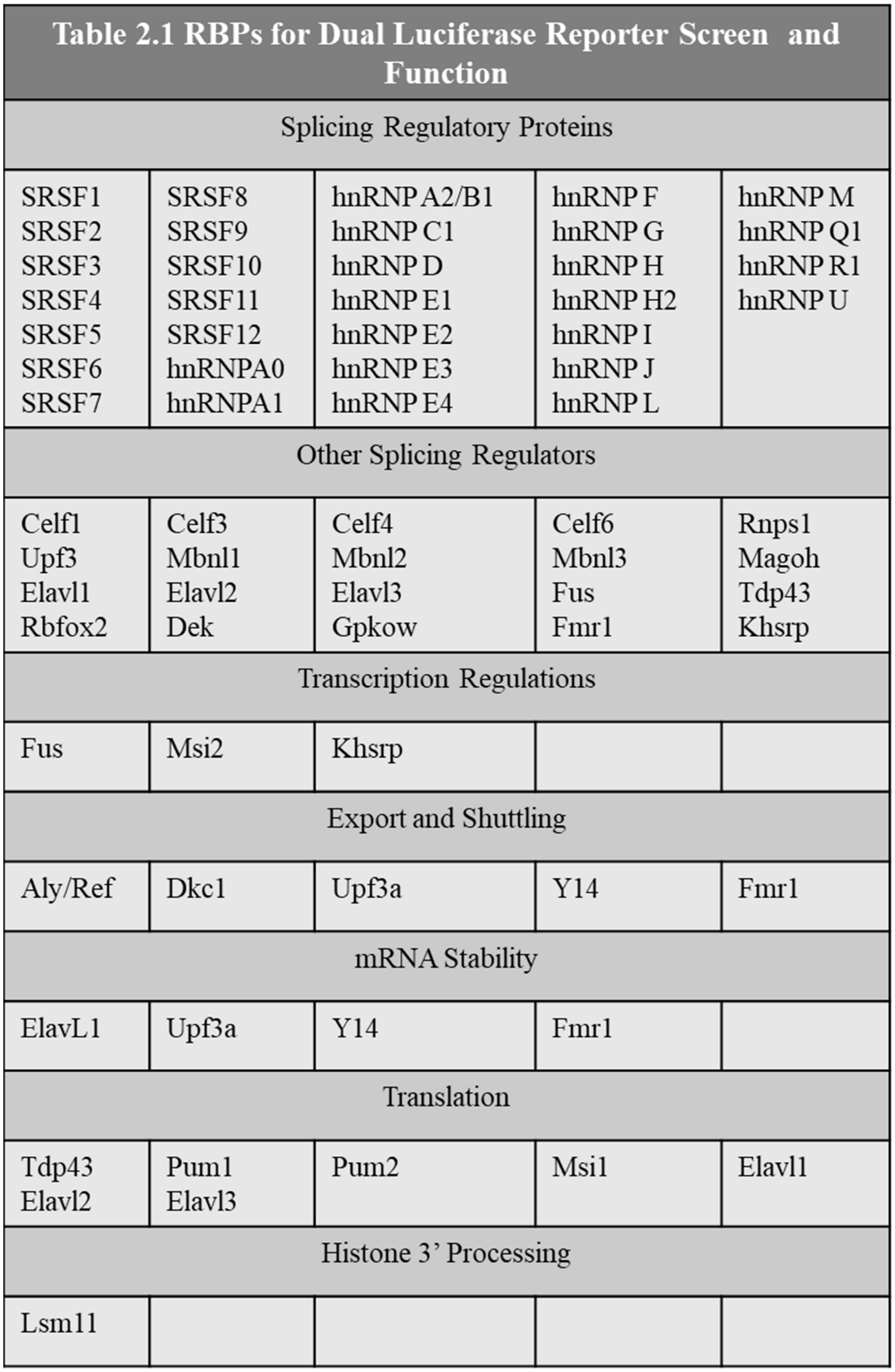


Table 2.1 RBPs for Dual Luciferase Reporter Screen and Function

Splicing Regulatory Proteins				
SRSF1 SRSF2 SRSF3 SRSF4 SRSF5 SRSF6 SRSF7	SRSF8 SRSF9 SRSF10 SRSF11 SRSF12 hnRNPA0 hnRNPA1	hnRNP A2/B1 hnRNP C1 hnRNP D hnRNP E1 hnRNP E2 hnRNP E3 hnRNP E4	hnRNP F hnRNP G hnRNP H hnRNP H2 hnRNP I hnRNP J hnRNP L	hnRNP M hnRNP Q1 hnRNP R1 hnRNP U
Other Splicing Regulators				
Celf1 Upf3 Elavl1 Rbfox2	Celf3 Mbnl1 Elavl2 Dek	Celf4 Mbnl2 Elavl3 Gpkow	Celf6 Mbnl3 Fus Fmr1	Rnps1 Magoh Tdp43 Khsrp
Transcription Regulations				
Fus	Msi2	Khsrp		
Export and Shuttling				
Aly/Ref	Dkc1	Upf3a	Y14	Fmr1
mRNA Stability				
ElavL1	Upf3a	Y14	Fmr1	
Translation				
Tdp43 Elavl2	Pum1 Elavl3	Pum2	Msi1	Elavl1
Histone 3' Processing				
Lsm11				

causes production of an mRNA that contains an AU-rich element that destabilizes ElavL1 mRNA, thus creating a negative feedback loop (Dai, Zhang, and Makeyev 2012). Because of the known role of ElavL1 in APA, ElavL1 and the closely related, predominantly neuronal ElavL2, 3, and 4 were all included within the screen. Finally, RBPs that bind near polyadenylation sites were included as well because it was hypothesized that they are likely APA regulators. To identify these proteins, publicly available ENCODE eCLIP data was examined to identify RBPs that are enriched for binding from 200 nucleotides upstream to 200 nucleotides downstream of polyA sites genome-wide (Figure 2.3B). eCLIP is a modified version of iCLIP (individual nucleotide resolution cross-linking immunoprecipitation) and is a technique that utilizes UV-irradiation of cells to crosslink RNA binding proteins to RNAs. Proteins of interest are then immunoprecipitated and RNA is released and sequenced (Van Nostrand et al. 2016) (Figure 2.3A). Analysis of the eCLIP data revealed that members of the core polyadenylation machinery were enriched for binding near polyadenylation sites, with CstF enriched between 50-100 nucleotides downstream of the AAUAAA hexamer and CFIm enriched approximately 50 nucleotides upstream (Figure 2.3C). As both CstF and CFIm binding profiles are consistent with their known roles in polyA site selection, this suggests that eCLIP data may accurately predict 3' processing regulators. Additional RBPs that were enriched for binding near polyA sites and therefore included within the screen include Fmr1, Khsrp, Lsm11, and GPKOW (Figure 2.3A). Notably, Fmr1 is the most enriched RBP, even more so that the polyA factor CstF64τ (a paralog of CstF64), which is the second-most enriched protein. Fmr1 is enriched for binding downstream of the polyA site. Both Khsrp and Lsm11 are enriched upstream of the AAUAAA hexamer. Khsrp has maximal enrichment approximately 100-50 nucleotides upstream whereas

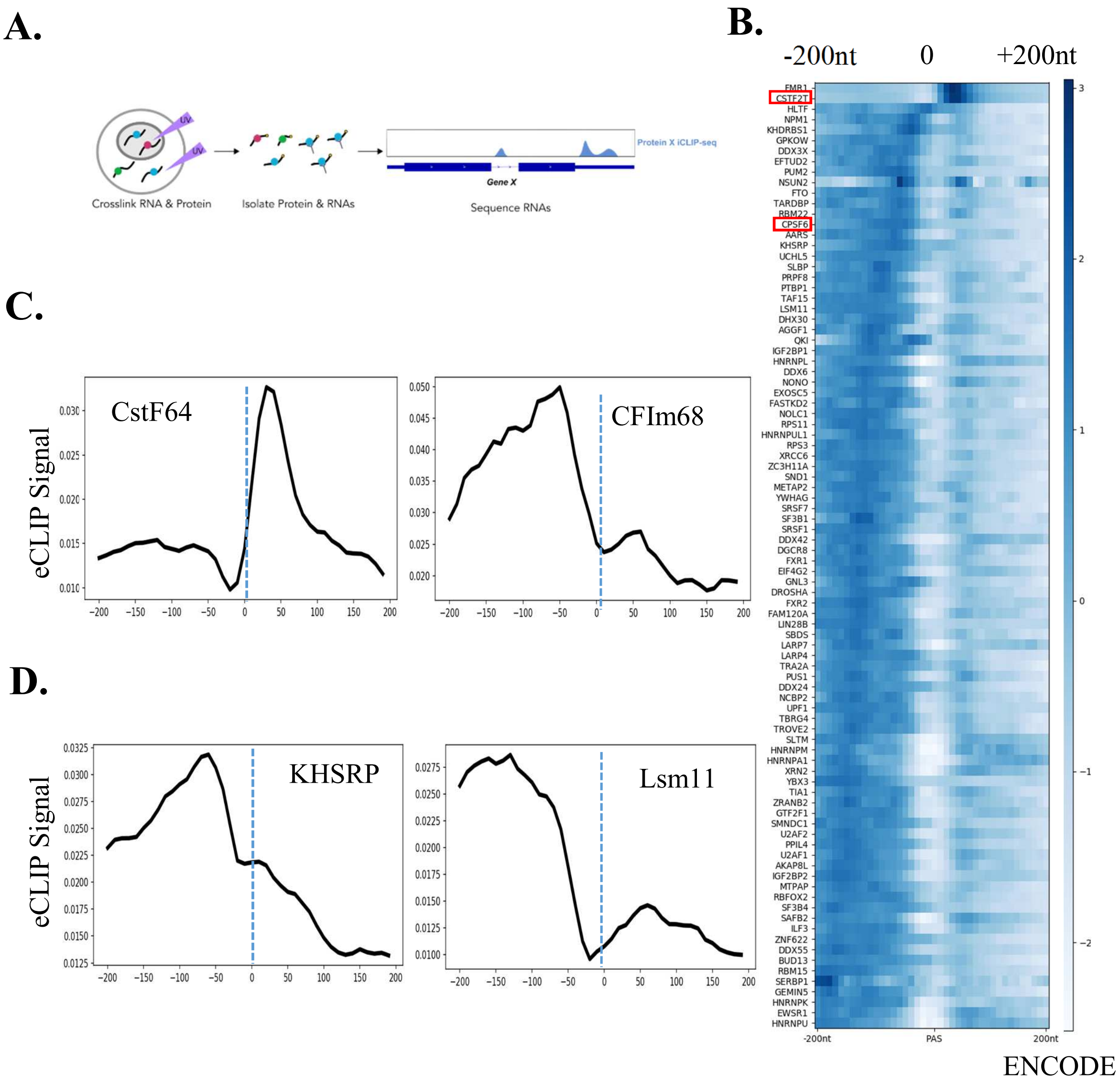


Figure 2.3 Enrichment of RBPs Near Polyadenylation Sites

A. Schematic depicting CLIP protocol. RNA is UV-crosslinked to protein followed by immunoprecipitation of RPB of interest. RNAs bound by protein are subsequently sequenced and aligned to reference genome.
B. eCLIP signal for 90 RNA binding proteins from 200 nucleotides upstream of the A(A/U)UAAA hexamer to 200 nucleotides downstream genome-wide. RBPs are ranked by enrichment. PolyA factors are indicated with a red box.
C. eCLIP signal for core polyadenylation machinery components CstF64 (left) and CFIm68 (right) from 200 nucleotides upstream to 200 nucleotides downstream of the AAUAAA hexamer genome-wide.
D. eCLIP signal for KHSRP (left) and Lsm11 (right) from 200 nucleotides upstream to 200 nucleotides downstream of the AAUAAA hexamer genome-wide.

Lsm11 has a broader enrichment from approximately 200-100 nucleotides upstream of the AAUAAA hexamer (Figure 2.3C)

After determining the RBPs that would be tested, a dual-luciferase reporter was generated from the dual luciferase reporter pPASPORT (Lackford et al. 2014a). The pPASPORT reporter contains Renilla and Firefly luciferase co-expressed on a bicistronic mRNA. Both luciferases can be translated because an IRES element between Renilla and Firefly luciferase allows for translation of Firefly luciferase. Firefly luciferase is followed by the globin polyadenylation site, a strong polyA site (Figure 2.4). This strategy allows for the approximation of the strength of the polyA site following Renilla luciferase. If the polyadenylation site is strong, cleavage and polyadenylation are efficient so only Renilla luciferase is expressed. However, if the polyA site is weaker, there will be transcription readthrough and both Renilla and Firefly luciferase are expressed. Polyadenylation site strength is thus measured as the ratio of Renilla to Firefly luciferase.

The polyA site following Renilla luciferase is the L3 viral polyA site with two BoxB hairpins nearby. Due to the strong interaction between λN peptides and BoxB hairpins, cotransfection of the reporter with λN-tagged RBPs effectively tethers RBPs to the Renilla polyA site, allowing for evaluation of the RBP's effects on polyA site selection (Figure 2.4). Activation of polyadenylation is represented as an increase in the ratio of Renilla to Firefly luciferase in comparison to a reporter-only control because there is increased cleavage and polyadenylation at the first cleavage site. In contrast, inhibition of polyadenylation decreases this ratio as there is inefficient cleavage and polyadenylation, leading to increased transcription readthrough and increased expression of Firefly luciferase.

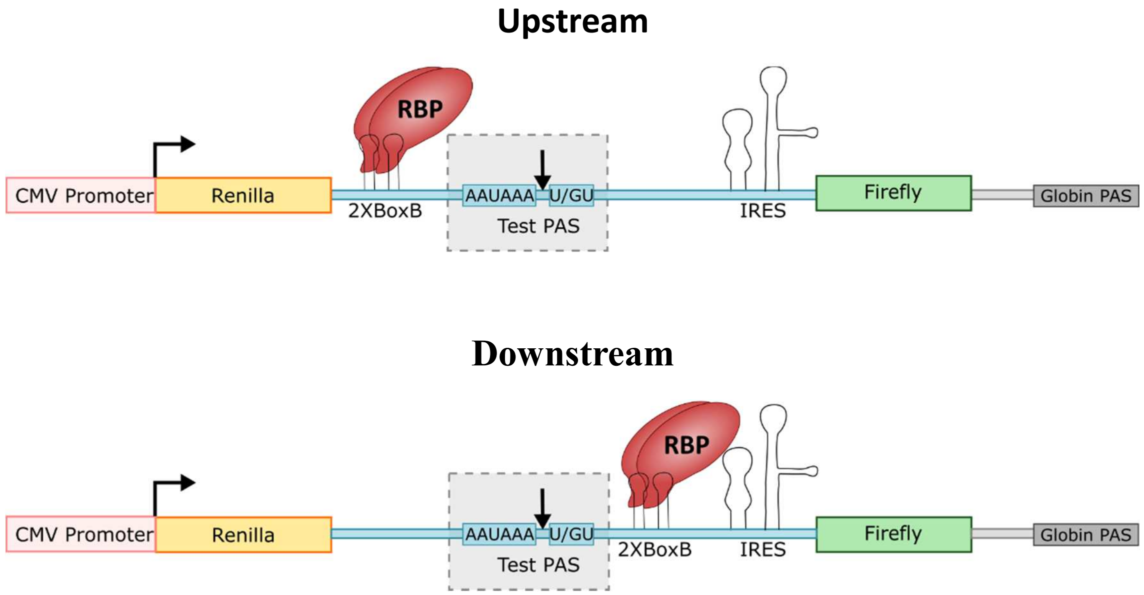


Figure 2.4 Dual Luciferase Reporter for Monitoring APA Regulations by RNA Binding Protein

Schematic of pPASPORT dual luciferase reporter for APA analysis. Renilla and Firefly luciferase are co-expressed on a bi-cistronic mRNA in which an IRES element allows for translation of Firefly luciferase. Renilla luciferase is followed by the test polyA site, which contains 2 BoxB hairpins either upstream (top) or downstream (bottom) of the cleavage site of the L3 viral polyA site. Reporter is co-transfected with λN-tagged RBPs , which bind to the boxB hairpin. Activation of polyadenylation is expressed as an increase in the ratio of Renilla to Firefly luciferase in comparison to a vector only control whereas inhibition is expressed as a decrease in this ratio.

It has been previously shown that RBPs have position-dependent effects on polyadenylation. For example, Mbnl1 activates polyadenylation when upstream of the polyadenylation site but inhibits when its binding site overlaps with the A(A/U)UAAA hexamer (Batra et al. 2014). In addition, one of the purposes of the screen was to examine the roles of the SR family of proteins on APA. As SR proteins have been shown to have a position dependent effect on splicing, it was predicted that their effect on APA may also be position dependent (Erkelenz et al. 2013). To establish the position dependence of APA regulation, two different dual luciferase reporters were generated, with the BoxB hairpins upstream and downstream of the polyadenylation site. In the upstream position, the boxB hairpins were added at the location of the two UGUA motifs that are found within the L3 viral polyA site. In the downstream position, the two BoxB hairpins are located downstream of the U/GU rich region where CstF binds (Figure 2.4).

Reporters were cotransfected with λN tagged RBPs and the ratio of Renilla to Firefly luciferase was calculated relative to a reporter-only control (Figure 2.5). As a positive control, CFIm25 was transfected as it has previously been to activate polyadenylation 6-fold when bound upstream of the polyadenylation site (Zhu et al. 2017). All overexpressed proteins were verified by western blotting with an antibody specific to FLAG tag, as all proteins were tagged with an N-terminal FLAG tag downstream of the λN tag (Figure 2.6).

The reporter recapitulated the role of several RBPs with known roles in activating polyA sites. First, hnRNP H was found to activate polyadenylation site selection (fold change 1.9 and 2.5 upstream and downstream of the AAUAAA hexamer respectively), consistent with studies that found that hnRNP H promotes proximal polyA sites by recruiting CstF64 to nearby polyadenylation sites (Katz et al. 2010). hnRNP F, which is closely related to hnRNP H, was also observed to activate 3’ processing (fold change 2.4 upstream and 2.9 downstream).

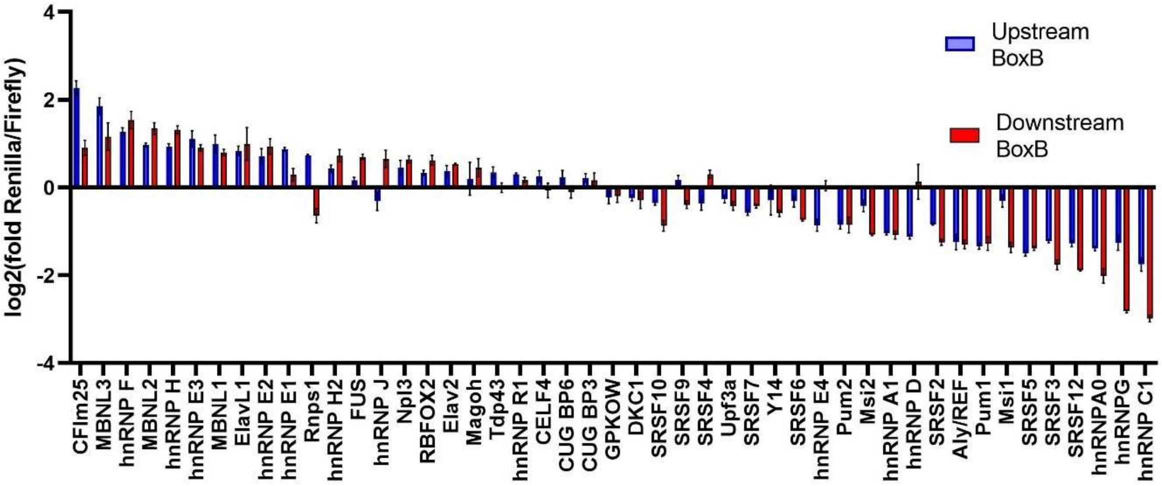
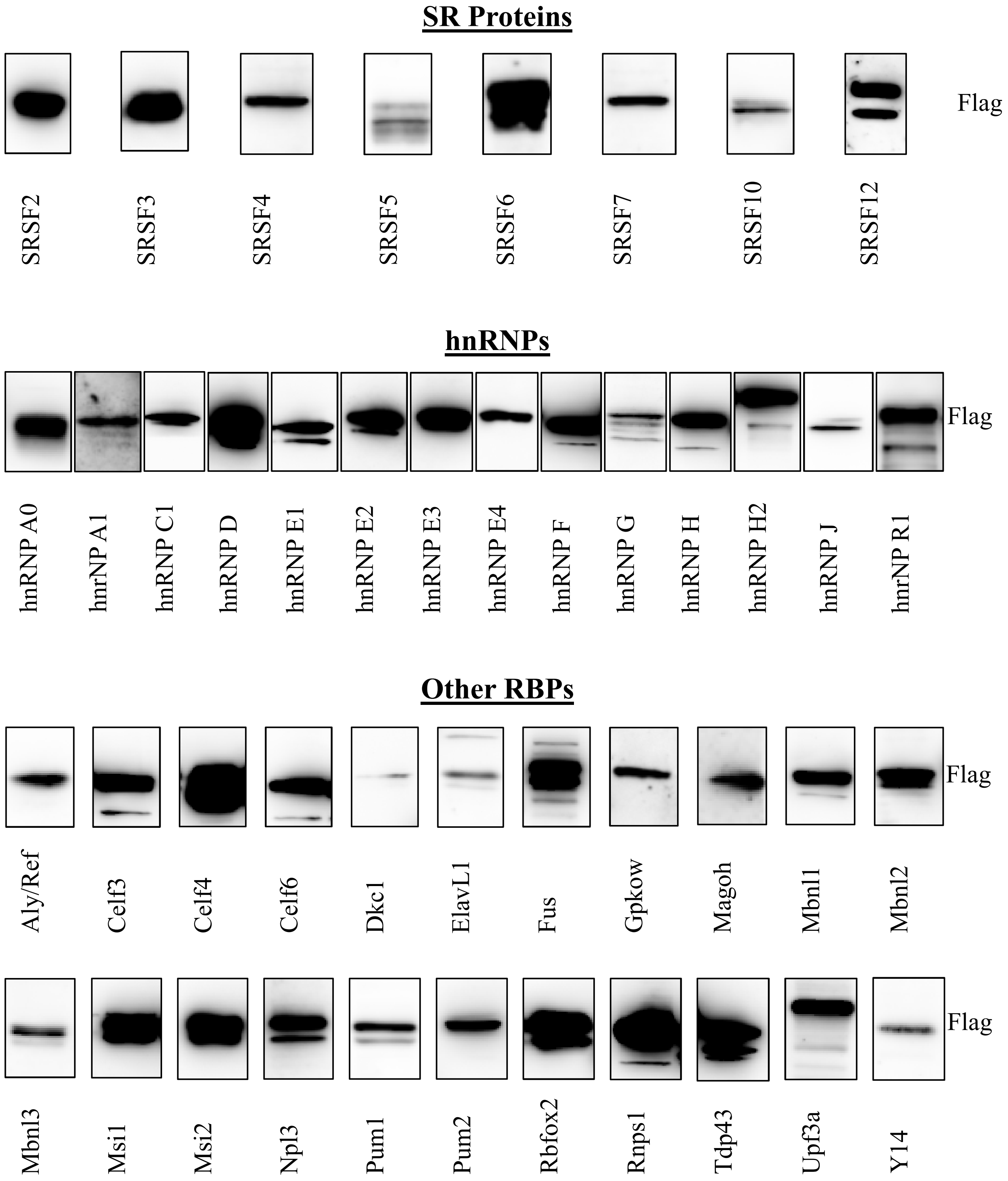


Figure 2.5 Tethering of RBPs to Dual Luciferase APA Reporter

47 unique RNA binding proteins were tethered using to the pPASPORT APA reporter using the interaction of BoxB hairpins and the λN peptide. RBPs were tethered both upstream (blue) and downstream (red) of the cleavage site of the L3 viral polyA site. Activation is shown is represented as an increase in the ratio of Renilla to Firefly luciferase whereas inhibition is a decrease in the ratio.



SR Proteins

hnRNPs

Other RBPs

Figure 2.6 Expression of λN-flag RBPs

Western blot analysis of expression of λN-flag-tagged RBPs. RBPs are co-transfected into 293T cells with dual luciferase reporter for polyA site strength analysis. All blots are shown with an anti-flag antibody.

Previously, hnRNP F has had conflicting roles in alternative polyadenylation. While it was first identified to inhibit polyadenylation by repressing CstF64 binding in mouse B cells, it was later found to promote polyadenylation sites within the 3' UTR of the MeCP gene (Veraldi et al. 2001; Newnham et al. 2010). Our data supports the role of hnRNP F as an activator, although this result does not eliminate the possibility that it may have context-specific effects such as cell type. Finally, the hnRNP E family (also known as the PCPB family) was found to be activators of polyA site selection (hnRNP E1 fold change 1.8 upstream and 1.2 downstream; hnRNP E2 fold change 1.6 upstream and 1.9 downstream; hnRNP E3 fold change 2.2 upstream and 1.9 downstream), consistent with reports that the PCBP proteins are global enhancers of 3' processing (Figure 2.5) (X. Ji et al. 2013).

Some polyadenylation site repressors recapitulate previous studies as well. hnRNP C was found to be the strongest inhibitor of polyadenylation (fold change 0.3 upstream and 0.1 downstream), consistent with previous studies that have found that hnRNP C regulates cancer specific APA and also binds to the U-rich region downstream of simian virus 40 late polyadenylation signal (Fischl et al. 2019; Wilusz1 and Shenk2 1990) (Figure 2.5). Overall, this suggests that our reporter screen is an effective method to identify RBPs that regulate polyadenylation site selection and therefore may also be able to identify novel APA regulators as well.

Tethering assay data suggests that generally there are more strong inhibitors of polyadenylation than activators. Mbnl3, the strongest activator, causes an approximately 4-fold increase in the ratio of Renilla to Firefly luciferase, similar to the activation effect of CFIm25. Besides Mbnl3, only hnRNP F, Mbnl2, and hnRNP F and hnRNP E3 have fold changes above 2. By contrast, the repressors hnRNP C1 and hnRNP G both have an approximately 8-fold affect, and 11 other factors decrease the ratio to Renilla to Firefly luciferase more than 2-fold (Figure 2.5).

Reporter Screen Identifies Novel APA Regulators

In addition to validating previous studies that identified roles for RBPS in polyA site selection, our reporter screen also identifies several novel regulators of polyA site selection. First, all tested SR proteins reduce the ratio of Renilla to Firefly luciferase and therefore repress polyA site selection (Figure 2.5). SRSF12, SRSF3, and SRSF5 have the greatest inhibitory effect on polyadenylation with fold changes of 0.4, 0.4, and 0.3 tethered upstream and 0.3, 0.3, and 0.4 tethered downstream of the polyA site respectively. While only 9 of the 12 SR proteins were tested within this study, the consistent repressive role suggests that the SR family as a role inhibit polyadenylation site usage.

In addition, despite having no previously identified role in APA, hRNP G and hnRNP A0 had the second and third largest inhibitory effects on polyA site usage respectively. hnRNP G (RBMX) is an alternative splicing regulator that preferentially binds the sequence CC(A/C) (Heinrich et al. 2009). hnRNP G induced a fold change of 0.4 when tethered upstream of the polyA site and a fold change of 0.1 tethered downstream of the polyadenylation site (Figure 2.5). While hnRNP G has no known role in polyA site selection, a GWAS study that used a computational model to identify the effects of genetic variation on alternative 3' UTRs found that single nucleotide polymorphisms (SNPs) that alter RBMX binding sites are associated with APA changes (Mariella et al. 2019), supporting the finding of this study. One method to characterize the mechanism for regulation of APA by RBPs such as hnRNP G is to determine the members of the core polyadenylation machinery that it interacts with. Previously, flag-immunoprecipitation or endogenous immunoprecipitations of core polyA factors followed by mass spectrometry was performed to identify proteins that interact with each polyA factors. hnRNP G was found to interact with the core polyA machinery members CPSF100 and CFIm25, which supports a role for hnRNP G in 3' processing although further functional studies will be necessary to elucidate the precise mechanism (Table 2.2).

Tethering hnRNP A0 also repressed polyA site selection, with a fold change of 0.4 tethered upstream of the polyA site and a fold change of 0.25 when tethered downstream of the polyA site. Notably, another member of the hnRNP A family, hnRNP A1, also causes a fold change of 0.5 when tethered both upstream and downstream of the polyA site (Figure 2.5). The hnRNP A family, while also includes hnRNP A2/B1 and hnRNP A3, has never previously been shown to regulate APA but Nab4p, a yeast protein structurally related to hnRNP A1, was previously shown to repress cleavage at cryptic polyA sites in a concentration dependent manner (Krecic and Swanson 1999; Minvielle-Sebastia et al. 1998). As both hnRNP A0 and hnRNP A1 were found to inhibit polyadenylation, it is possible that they use a similar mechanism as the yeast homolog. Also consistent with the hnRNP A family of proteins regulating 3' processing, flag immunoprecipitation of core 3' processing factors indicates that hnRNP A1 interacts with members of subunits of CstF (CstF64, CstF64tau and CstF77) as well as CPSF (CPSF160, Symplekin and Wdr33) (Table 2.2). Similarly, hnRNP A2/B1 interacts with components of CstF (CstF64), CFIm (CFIm68 and CFIm25), and CPSF (CPSF160 and Symplekin) and hnRNP A3 interacts with CFIm25 and CPSF160 (Table 2.2).

A final example of a RBP that may have a novel role in 3' processing is Musashi1 (Msi1). Msi1 is a neural-progenitor marker that is important for brain development. It is also a marker for

Table 2.2 RBPs Co-Immunoprecipitated by Core Polyadenylation Factors (data from Serena Chan)

CPSF					
CPSF30					
hnRNP F	hnRNP K				
CPSF73					
SRSF1	SRSF2	SRSF7	hnRNP C	hnRNP F	hnRNP H1
hnRNP H2	hnRNP L	hnRNP M	hnRNP R	hnRNP U	hnRNP R
Magoh					
CPSF100					
SRSF1	SRSF2	hnRNP E1	hnRNP E2	hnRNP E3	hnRNP F
hnRNP G	hnRNP K	hnRNP L	hnRNP U		
CPSF160					
SRSF1	SRSF3	SRSF4	SRSF5	hnRNP A1	hnRNP A2/B1
hnRNP A3	hnRNP D	hnRNP E1	hnRNP E2	hnRNP E3	hnRNP F
hnRNP G	hnRNP H1	hnRNP H2	hnRNP K	hnRNP L	hnRNP M
hnRNP R	hnRNP U	ElavL1	Fus	Magoh	Tdp43
Fip1					
hnRNP K					
Symplekin					
SRSF2	hnRNP A1	hnRNP A2/B2	hnRNP C	hnRNP K	hnRNP R
Aly/Ref					
Wdr33					
SRSF2	hnRNP A1	hnRNP E1	hnRNP H1	hnRNP K	hnRNP L
hnRNP U	Fus	Tdp43	Aly/Ref		

Table 2.2 (continued) RBPs Co-Immunoprecipitated by Core Polyadenylation Factors (data from Serena Chan)					
CFIm					
CFIm25					
SRSF1	SRSF3	SRSF4	SRSF5	SRSF7	SRSF9
SRSF10	hnRNP A2/B1	hnRNP A3	hnRNP C	hnRNP D	hnRNP F
hnRNP G	hnRNP H1	hnRNP I	hnRNP K	hnRNP L	hnRNP M
hnRNP Q	hnRNP R	hnRNP U	ElavL1	Fus	Magoh
CFIm59					
hnRNP C	hnRNP F	hnRNP H	hnRNP H2	Fus	
CFIm68					
SRSF2	SRSF11	hnRNP A2/B1	hnRNP F	hnRNP H1	hnRNP H2
hnRNP H3	hnRNP K	hnRNP R			
CFIIm					
Pcf11					
SRSF1 hnRNP A0 hnRNP E2 hnRNP H3 ElavL1 Aly/Ref	SRSF3 hnRNP A2/B1 hnRNP E3 hnRNP K Fus	SRSF4 hnRNP A3 hnRNP F hnRNP L Rnps1	SRSF7 hnRNP C hnRNP G hnRNP M Tdp43	SRSF9 hnRNP D hnRNP H1 hnRNP R Celf1	SRSF9 hnRNP E1 hnRNP H2 hnRNP U Gpkow
Clp1					
SRSF3	hnRNP L				

Table 2.2 (continued) RBPs Co-Immunoprecipitated by Core Polyadenylation Factors (data from Serena Chan)

CstF					
CstF50					
hnRNP H1	hnRNP L				
CstF64					
hnRNP A1 hnRNP K Aly/Ref	hnRNP A2/B1 hnRNP L	hnRNP C hnRNP L	hnRNP D hnRNP R	hnRNP H1 hnRNP U	hnRNP H2 Fus
CstF64tau					
hnRNP A1 hnRNP R	hnRNP C hnRNP U	hnRNP E1 Fus	hnRNP K	hnRNP L	hnRNP M
CstF77					
hnRNP A1 hnRNP M	hnRNP C hnRNP U	hnRNP F	hnRNP H1	hnRNP H2	hnRNP K

some adult stem cells and is upregulated in tumors including breast, prostate, lung, and brain (Forouzanfar et al. 2020). In our screen, Msi1 is identified as a repressor of polyadenylation, with fold inhibition of 0.8 when tethered upstream and log2 fold inhibition of 0.4 when tethered downstream (Figure 2.5). One possible explanation for the role of Musashi1 in 3' processing is that an interactome study identified that the *C. elegans* ortholog for Musashi1 interacts with the homolog of CstF64 and CstF77. Strikingly, the same screen also identified Wdr33 as interacting with subunits of CPSF (CPSF73, CPSF160, Fip1L1, and CPSF30) as well as PAP, and CstF77. In 2006, Wdr33 was not known to be a polyadenylation factor but it is now known to be a subunit of CPSF, indicating that Msi1 may be a true APA regulator as well (Gandhi et al. 2006). Msi1 was not identified as interacting with any core polyA machinery components, but this does not suggest the Msi1 is not a true interactor because immunoprecipitations were performed in either 293T cells or from HeLa cell nuclear extract, neither of which express Msi1.

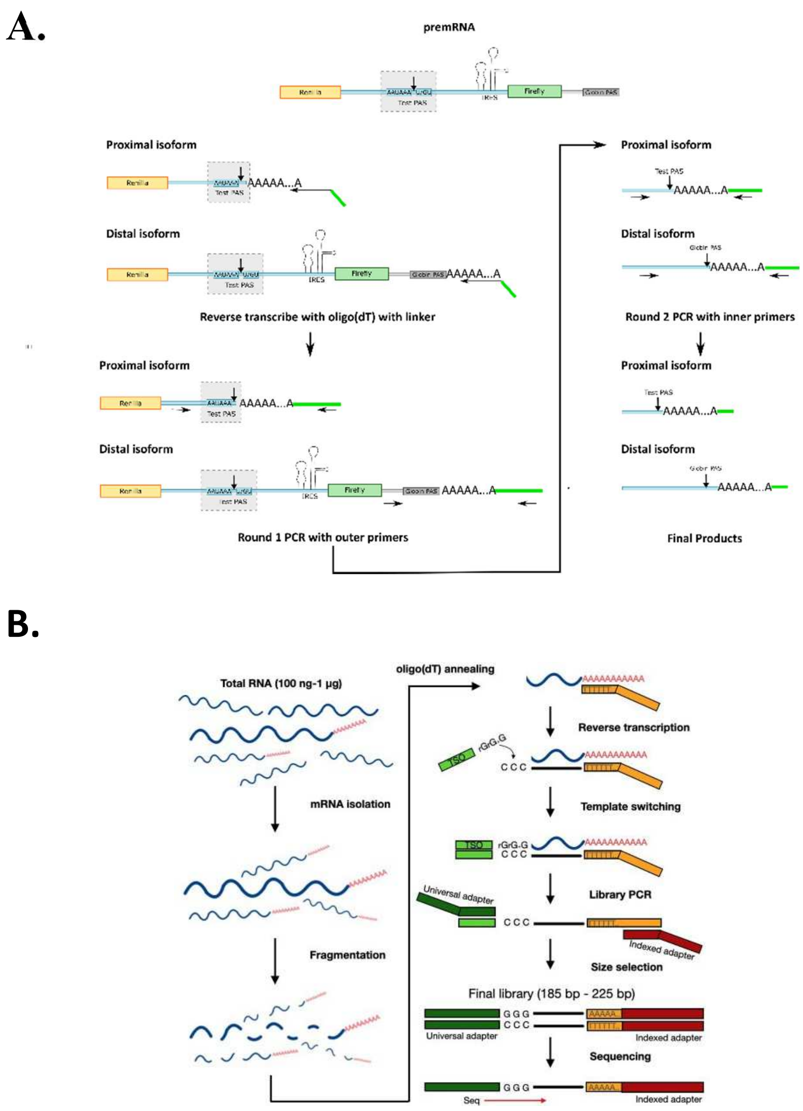
One important consideration, however, is that Msi1 is a known translation repressor. Musashi1 binds to the 3'UTR of Numb and represses translation by competing with polyA binding protein for interacting with eIF4G; as a result, Msi1 inhibits ribosome assembly (Kawahara et al. 2008). Currently, our screen cannot differentiate between changes in APA and changes in translation because the readout is luminescence, which is determined by the protein levels of Renilla and Firefly luciferase. If Musashi1 represses translation of Renilla luciferase when tethered to the L3 polyA site, there would also be a decrease in the ratio of Renilla to Firefly luciferase. While currently untested, one way to differentiate between the two would be to analyze the mRNA itself through 3' RACE. 3' RACE is a cDNA amplification technique utilized for polyA site analysis. cDNA is synthesized by using a reverse primer oligo(dT) for the polyA tail that has a linker at the 3' end. cDNA is subsequently used for two rounds of PCR using two forward primers in the same PCR reaction that bind to either the shared region upstream of the Renilla luciferase polyA site or a Firefly luciferase polyA site specific region along with a reverse primer that binds to the linker. The ratio of Renilla to Firefly is used to compare distal polyA site usage to proximal polyA site usage and therefore the effect of Musashi1 on polyA site selection at the mRNA level (Figure 2.7A).

Most RBPs Have Position-Independent Effects on 3' Processing

It has previously been shown for 3' end processing as well as other RNA processing events such as splicing that the effects of RBPs on RNA processing are position dependent. As a result, the dual luciferase reporter was designed with the 2xBoxB hairpins both upstream and downstream of the cleavage site to test any position dependent effects (Figure 2.4B)

Generally, we found that tethered RBPs have similar effects, whether it be activation or activation and inhibition, when tethered upstream and downstream of the polyA site, although the magnitude of change sometimes differs. For example, while hnRNP C1 and hnRNP G inhibit polyadenylation site usage from both positions, they both have stronger inhibitory effects when tethered downstream (fold change of 0.125 for both) than upstream (fold change of 0.3 and 0.4) (Figure 2.5).

Most strikingly, all members of the SR protein family have position independent repressive effects on polyA site selection. In splicing, SR proteins activate alternative cassette exon inclusion when bound to exonic sequences and inhibit exon inclusion when bound to intronic



Yoon et al (2021) Methods Enzymol.

Figure 2.7 APA Analysis at the mRNA Level by 3’ RACE and PAS-Seq

A. Schematic demonstrating procedure to use 3’ RACE to verify APA changes induced by tethering of RBP. While initial studies analyzed APA events by measuring changes in the ratio of Renilla to Firefly luminescence, 3’ RACE allows for analysis at the mRNA level.
B. Schematic of polyadenylation site sequencing (PAS seq), a method to map polyadenylation sites globally and quantify the relative usage of alternative sites.

sequences (Erkelenz et al. 2013). This suggests that the mechanism regulating SR protein-mediated regulation of APA may be different from the mechanism governing SR-mediated splicing regulation. hnRNPs have also been shown to have position dependent effects with the opposite role of SR proteins: inhibiting exon inclusion from exonic positions and activating exon inclusion from intronic positions. However, like SR proteins, most hnRNPs have similar results tethered upstream and downstream of the polyA site, as exemplified by hnRNP F which activates both upstream and downstream of the polyA site (fold change of 2.5 and 2.8 respectively) and hnRNP C1 when inhibits both upstream and downstream of the polyA site (fold change of 0.3 and 0.1 respectively) (Figure 2.5). As a result, it is unlikely that SR proteins and hnRNPs have the same global position dependent effects on alternative polyadenylation as they do on alternative splicing.

There are, however, some notable exceptions of RBPs that do have position dependent effects. The first is fused in sarcoma (Fus), which has a previously characterized role in position dependent APA. Fus is an RNA binding protein whose mutation is associated with amyotropic lateral sclerosis (ALS). It has previously been implicated in transcription, RNA splicing, and mRNA 3' processing. ChIP-seq analysis revealed that Fus promotes pausing RNA Pol II pausing upstream of APA sites that are upregulated upon Fus knockdown and downstream of APA sites downregulated upon Fus knockdown. This suggests that Fus-mediated APA regulation is both co-transcriptional and position dependent. This likely occurs because when Fus is bound downstream of polyA sites, it promotes CPSF160 binding. By contrast, when Fus binds upstream of the polyA site, it induces Pol II stalling and reduced gene expression (Masuda, Takeda, and Ohno 2016; Masuda et al. 2015). This finding is recapitulated within our tethering assay as there is a stronger activation effect when Fus is tethered downstream of the polyA site (fold change 1.6) than upstream, where the effect is minimal (log2 fold change 1.1) (Figure 2.5). This suggests that while few position-dependent effects on APA were identified within the screen, the screen itself is capable of identifying position dependent effects. As a result, it appears more likely that there are few RBPs within position dependent effects on APA. Another RBP with position dependent effects on polyA site selection is Rnps1, which activates polyadenylation when tethered upstream of the polyadenylation site (fold change 1.7) and inhibits when tethered downstream of the polyadenylation site (fold change 0.6) (Figure 2.5). Rnps1 is a splicing activator with an extensive serine-rich region (Mayeda et al. 1999; Sakashita et al. 2004). Rnps1 has also previously been shown to activate 3' cleavage but the stimulatory affect was strongly dependent on the splicing of the last intron (McCracken et al. 2003). This dependency may explain why Rnps1 has a position dependent effect on APA: at the upstream position, Rnps1 is closer to the terminal exon of Renilla luciferase. If the cleavage activation is dependent on association of Rnps1 with members of the splicing machinery, it is possible that at the downstream position, Rnps1 is too far separated from the splicing machinery to activate cleavage. Importantly, Rnps1 is also part of the exon junction complex and its tethering to the 3' UTR of β-globin can trigger non-sense mediated decay so it will be essential to differentiate between the effects of Rnps1 on APA and mRNA degradation through direct analysis of mRNA by 3' RACE (Lykke-Andersen, Shu, and Steitz 2001).

2.4 DISCUSSION

In this chapter, a large-scale screen was performed to identify RNA binding proteins with novel roles in APA regulation. A pPASPORT reporter in which Renilla and Firefly luciferase are expressed on a bi-cistronic mRNA with the L3 viral polyadenylation site downstream of Renilla luciferase was modified to have two BoxB hairpins either upstream or downstream of the L3 cleavage site. λN-tagged RNA binding proteins were tethered at the polyA site and the change in the ratio of Renilla to Firefly luciferase is analyzed to identify RBPs that either activate or inhibit polyA site selection.

Our screen identified several RBPs with a potentially novel roles in APA. These proteins include, but are not limited to, several hnRNPs including hnRNP G, the hnRNP A family, as well as Musashi1, all of which were identified to be polyadenylation site repressors. Another family of proteins with a novel role in polyA site repression is the SR family of proteins. SR proteins were of particular interest in this study because they are structurally similar to cleavage factor I (CFIm), a polyadenylation site enhancer. SR proteins and CFIm both use RS domains to recruit members of the core splicing and polyadenylation machineries and thereby activate RNA processing (Zhu et al. 2017; Graveley, Hertel, and Maniatis 2001) (Figure 2.1). Because both CFIm and SR proteins use similar recruitment mechanisms, it was hypothesized that SR proteins may also be able to promote polyA sites. Surprisingly, however, all 9 of the 12 SR proteins utilized within this study inhibited polyA site selection. One potential explanation for this finding is that the RS domains of CFIm59 and CFIm68 are in fact different from that of SR proteins. The RS domain of SR proteins consists of RS dipeptide repeats. By contrast, the RS domain of CFIm59 and CFIm68 contains RE/D dipeptides in addition to RS dipeptides. Another difference is that the RS domain of SR proteins is phosphorylated, which is critical for its function as unphosphorylated SR proteins have limited interaction with U1-70K (Xiao and Manley 1997). While hyperphosphorylation inhibits the function of both SR proteins and CFIm, unphosphorylated CFIm59 and CFIm68 can still interact with Fip1 (Zhu et al. 2017). These differences in RS domain properties may explain why SR proteins repress polyadenylation sites instead of activating them, as was originally proposed. While all SR proteins tested inhibit polyadenylation, some have stronger effects than others, with the strongest effects for SRSF12, SRSF3, and SRSF5. Importantly, some studies have shown that different SR proteins have different phosphorylation levels, with SRSF3 having hypo-phosphorylation in vitro (Y. Long et al. 2019). If phosphorylation status is determined to be the mechanism behind the inhibition of polyadenylation by SR proteins, it will be interesting to evaluate whether SR proteins that are more extensively phosphorylated have differential effects than those that are hypo-phosphorylated.

Inhibition by SR proteins may also be explained by the fact that when polyadenylation sites lie within introns, there is a competition between the core polyadenylation machinery and the splicing machinery to determine whether intronic polyadenylation will occur or the intron will be removed by splicing (Tian et al. 2005). One line of evidence for this competition is mutations of intronic polyA sites enhance splicing and, conversely, mutations of splicing regulatory sequences enhance intronic polyadenylation. It is possible that repression by SR proteins exists as another mechanism to enhance splicing when intronic polyadenylation sites are present. In support of this, it has recently been shown that SRSF10 binds near intronic polyA sites and represses intronic polyA sites. SRSF10 is dysregulated in non-alcoholic fatty liver disease, leading to increased intronic polyadenylation and decreased expression of key metabolic genes. SRSF10 likely prevents the interaction of the 3’ processing machinery with intronic sites because there is increased interaction of the polyadenylation machinery with these sites when SRSF10 is depleted (Jobbins et al. 2022). Other SR proteins may also use similar mechanisms to repress polyadenylation sites as well.

While the findings of this study do identify potentially novel APA regulators, there are also several critical future experiments to both validate their regulatory activity as well as determine the precise mechanism for regulation. First, while reporter data suggests that several RBPs regulate APA when physically tethered to polyadenylation sites, it does not necessarily mean they regulate APA endogenously. In addition, even if the proposed regulation does occur endogenously, the identities of the mRNAs that are regulated are unknown. Crosslinking and immunoprecipitation sequencing (CLIP-seq) and polyadenylation site sequencing (PAS-seq) can be used to determine whether these RBPs bind to 3' UTRs and regulate APA of endogenous RNAs. CLIP-seq is utilized to identify the endogenous RNAs that a particular RBP interacts with (Figure 2.6A). If an RNA binding protein regulates 3' processing, we would predict to see it binding within 3' UTRs of endogenous RNAs near polyadenylation sites. It will also be important to test for direct regulation of endogenous APA by knocking down RBPs and performing PAS-seq. PAS-seq is a 3'-end sequencing method used to map polyadenylation sites globally and quantify the relative usage of alternative sites (Figure 2.7B) (Yoon, Soles, and Shi 2021). If an RBP regulates 3' processing, we would predict to see widespread APA changes. To verify that these APA changes are not due to off target effects, mRNAs that undergo PAS-seq would be overlapped with those that are bound by the RBP of interest; extensive overlap would suggest that the RBP directly regulates APA of these mRNAs.

After determining that the RBPs identified within this study directly regulate APA, the precise mechanism for this regulation can be determined. One aspect of this is to identify the domain(s) of each RNA binding protein that are critical for APA regulation. To do so, different domains within the RNA binding protein can be tested to see if they are either necessary or sufficient for APA regulation. To test if a domain is necessary, each domain can be individually deleted, and the domain deletion constructs can then be co-transfected with the pPASPORT reporter and tested for APA regulation. If a particular domain is necessary for APA regulation, when the domain is deleted, there will be reduced APA regulation in comparison to the wildtype protein. Similarly, to test if a domain is sufficient, each domain can be independently co-transfected with the dual luciferase reporter and again tested for APA regulation. If the domain is sufficient, there will be similar levels of APA regulation as with the wildtype protein. In addition to helping identify how individual RBPs regulate APA, this strategy can also be used to make predictions about RBPs that were not identified within this study that may also have the potential for APA regulation. Once the domains of individual RBPs that are necessary and/or sufficient for APA regulation have been identified, they can be compared across to identify commonalties. Identifying these commonalities will allow us to compare with the over 1500 human RBPs and predict those with higher likelihood for APA regulation.

Another critical aspect of investigating the mechanism for APA regulation will be to identify the core polyadenylation machinery components that these RBPs interact with. Previous mass spectrometry data for flag immunoprecipitated core polyadenylation machinery subunits provides some insight into these interactions (Table 2.2). For example, Fus, which was identified to be a position dependent APA activator when downstream of the cleavage site, interacts with CPSF160. Interestingly, Fus has also previously been shown to promote CPSF160 binding to polyadenylation sites at positions downstream of the cleavage site co-transcriptionally (Masuda, Takeda, and Ohno 2016; Masuda et al. 2015). Mass spectrometry data should be validated by flag immunoprecipitation of the identified RBPs to gain further insight into their interactions with the core polyadenylation machinery.

It is also important to note a potential caveat of this experiment. In the design of the dual luciferase reporter, the two boxB hairpins upstream of the polyadenylation site have been added to replace the two UGUA motifs that are normally present in the L3 viral polyadenylation site. UGUA is the binding site for CFIm so by removing the UGUA motif, any effects that are mediated by CFIm will not be observed. Future reporters may need to be designed with the boxB hairpins in a different location to investigate the effects of CFIm.

In summary, we have both identified putative regulators of APA and also established a system to screen for the roles of many other RNA binding proteins in APA. While further experiments are required to provide a mechanistic understanding of this regulation, our study provides a baseline for understanding how polyadenylation sites are selected, which cannot be determined by examining the core polyadenylation machinery alone. Importantly, we have also established the SR family of proteins as APA inhibitors, the mechanism for which will be important to elucidate. Finally, we have shown that the majority of RBPs tested have position independent effects on polyA site selection.

2.5 METHODS

Identification of RBPs with Enriched Binding Near PolyA sites

eCLIP sequencing for 90 unique RBPs was downloaded from the ENCODE Project and processed using deepTools to find average binding genome-wide (Van Nostrand et al. 2020; Ramírez et al. 2014). Within each 3' UTR, a region from 200 nucleotides upstream to 200 nucleotides downstream of the AAUAAA hexamer of a polyA site was selected and eCLIP signal at each position within the window was analyzed. RBPs were ranked by enrichment near polyadenylation sites.

Clones

Previously, a modified pcDNA3.1 construct was created with a λN tag added between restriction sites HindIII and Kpn1 and a FLAG tag added between restriction sites Kpn1 and BamHI. RNA binding proteins (Table 2.2) were cloned into λN-flag pcDNA3.1 using restriction sites BamHI and either XhoI or XbaI.

Reporters

A pPASPORT reporter was modified to have the L3 viral polyA site following Renilla luciferase (Lackford et al. 2014a). Two unique pPASPORT reporters were generated. In the first, two BoxB hairpins were added in the location of two UGUA motifs upstream of the AAUAAA hexamer, removing the UGUA motifs. In the second, 2 BoxB hairpins were added downstream of the GU rich region following the cleavage site.

Upstream BoxB Sequence

GGATCCTTCTTTTTGTCACTTGAAAAACAGGGCCCTGAAGAAGGGCCCAAAATAAG
GGCCCTGAAGAAGGGCCCTAGGAGACACTTTCAATAAAGGCAAATGTTTTATTTGT
ACACTCTCGGGTGATTATTTACCCCCCACCCTTGCCGTCTGCGAGGTACCGAGCTCG
AATTCT

Downstream BoxB Sequence

GGATCCTTCTTTTTGTCACTTGAAAAACATGTAAAAATAATGTACTAGGAGACACTT
TCAATAAAGGCAAATGTTTTATTTGTACACTCTCGGGTGATTATTTACCCCCCACCC
TTGCCGGGGCCCTGAAGAAGGGCCCTCTGCGGGGCCCTGAAGAAGGGCCCAGGTAC
CGAGCTCGAATTC

BoxB hairpin is highlighted in yellow. AAUAA hexamer is in red letters.

Luciferase Assay

110ng of λN-flag-tagged RBP was co-transfected with 110ng of pPASPORT dual luciferase reporter in triplicate. 110ng of each pPASPORT reporter was singly transfected as a control. 48 hours post-transfection, cells were washed with PBS and levels of Renilla and Firefly Luciferase were monitored with the Dual Luciferase Reporter Assay System according to standard protocol (Promega). Cells were resuspended in 50ul 1x Passive Lysis Buffer and shaken vigorously for 20 minutes. 10ul of lysate was mixed with 30ul of LAR II Reagent and Firefly luciferase luminescence was measured with luminometer. 30ul of Stop Solution was added and Renilla luciferase luminescence was measured with luminometer. The ratio of Renilla to Firefly luciferase was then calculated.

The remaining 40ng of cell lysate was mixed with 3x SDS loading dye and utilized for western blotting analysis to verify protein expression with an antibody specific for the FLAG tag.

Mass Spectrometry Analysis of PolyA Factors

Immunoprecipitations

Mass-spectrometry for many polyA factors previously performed within the lab was utilized for analysis of RBPs that interact (Serena Leong Chan 2014). Stable cell lines were generated in HEK 293T cells for subunits of CPSF (CPSF30, CPSF73, and CPSF160), CstF (CstF50, CstF64, CstF64tau, and CstF77) CFIm (CFIm25 and CFIm59) and CFIIm (Clp1 and Pcf11) using either flag-pcDNA3.1 or pCMV14-3x Flag mammalian expression vectors (Serena Leong Chan 2014). After selection with G418, individual clones were selected, expanded, and screened for expression of Flag-tagged proteins near endogenous protein levels.

Stable cell lines were used for FLAG immunoprecipitation. Five to ten 15cm plates of each cell line were spun down at 1.5krpm for 5 minutes at 4°C. Cell pellets were resuspended in 5x pellet volume Buffer A (10mm Hepes pH 7.9, 10mM KCl, 1.5mM $MgCl_2$, 10mM BME) and incubated on ice for 10 minutes before addition of NP-40 to 0.5%. Samples were spun at 4krpm for 10 minutes at 4°C and the nuclear pellet was resuspended in 1.5x the pellet volume with Buffer C (20mM Hepes pH 7.9, 420mM NaCl, 1.5mM MgCl2, 0.2mM EDTA, 25% glycerol, 10mM BME, protease inhibitor cocktail) and homogenized with an 18g needle. Nuclear extract was rotated at 4°C for 30 minutes and spun at 14krpm for 15 minutes. Supernatant was then mixed with Anti-FLAG M2 Affinity beads (Sigma) for 2 hours at 4°C. Beads were washed 3x with Buffer D300 (20mM Hepes pH 7.9, 300mM NaCl, 1mM $MgCl_2$, 0.2mM EDTA, 10mM BME) with 0.1% NP-40 and 1x with Buffer D100 (20mM Hepes pH 7.9, 300mM NaCl, 1mM $MgCl_2$, 0.2mM EDTA, 10mM BME) followed by elution in Buffer D100 + 3x FLAG peptide. Elutions were acetone precipitated and utilized for mass spectrometry.

For FLAG-CPSF30 stable cell line, proteins were immunoprecipitated from whole-cell lysate. Cells were centrifuged at 1.5krpm for 5 minutes at 4°C, resuspended in Buffer D300, and sonicated. Protease inhibitor cocktail and 0.1% NP-40 were added, and sample was rotated at 4°C for 30 minutes followed by centrifugation at 4°C at 14krpm for 10 minutes. Whole cell lysate was utilized for FLAG-immunoprecipitation as above.

Endogenous immunoprecipitation was used for Fip1 (Bethyl A301-462A), CPSF100 (Bethyl A301-581A), Symplekin (Bethyl A301-465A), Wdr33 (Bethyl BL4833), and CFIm68 (Bethyl A301-458A). 10ug of antibody was incubated with Protein A/G agarose beads (Pierce) for 1hr at room temperature. Beads were washed 2x with 0.2M sodium borate pH 9 followed by addition of 20mM DMP for 30 minutes to conjugate antibody to beads. Reaction was quenched with 0.2M ethanolamine pH 8 for 2hrs and beads were washed 2x with Buffer D300 + 0.1% NP-40. Beads were mixed with 1mL of HeLa cell nuclear extract for 2hrs at 4°C and subsequently washed 3x with Buffer D300 + 0.1% NP-40 and 1x with Buffer D100. Proteins were eluted 3x with 0.2M glycine pH 3.5 for 2 minutes. The entire protocol was repeated 2x. Elutions were combined, acetone precipitated, and analyzed by mass spectrometry.

Mass Spectrometry and Analysis

Immunoprecipitations were submitted to the Yates Lab at the Scripps Research Institute for high throughput liquid chromatography and tandem mass spectrometry (LC/MS/MS). Immunoprecipitations were digested with proteases to generate peptide fragments which were separated by LC followed by fragmentation in a tandem mass spectrometer. The mass:charge ratio of the peptides was measured followed by the mass:charge ratios of the daughter peptides. Fragmentation patterns were aligned to the genome to identify proteins immunoprecipitated. SAINT (significance analysis of interactome) was used to analyze mass spectrometry data by using probabilistic scoring to derive bona fide protein-protein interactions (Choi et al. 2010). The strengths between bait and prey were measured using sequence coverage. Next, Cytoscapes was used to visualize interactions with a probability of 0.95 and sequence coverage greater than 10%. Proteins identified were compared to the FLAG immunoprecipitation negative control and common proteins were filtered out to preserve only proteins that interacted with the 3' processing factor.

CHAPTER 3

CFIm REGULATES APA TO ATTENUATE GLOBAL GENE EXPRESSION AND MODULATE CELL FATE

3.1 SUMMARY

Of the core polyadenylation machinery, cleavage factor I (CFIm) is unique in that it is not essential for cleavage or polyadenylation (Boreikaite et al. 2022). Instead, CFIm enhances specific polyA sites by binding to UGUA enhancer motifs that are enriched upstream of distal polyA sites and recruiting the rest of the core polyadenylation machinery to nearby polyA sites (Zhu et al. 2017). As a result, CFIm is a critical regulator of alternative polyadenylation. In fact, it has been demonstrated that for an mRNA with multiple polyadenylation sites, binding of CFIm is the most indicative of the dominant cleavage and polyadenylation site of all members of the core polyadenylation machinery (G. Martin et al. 2012).

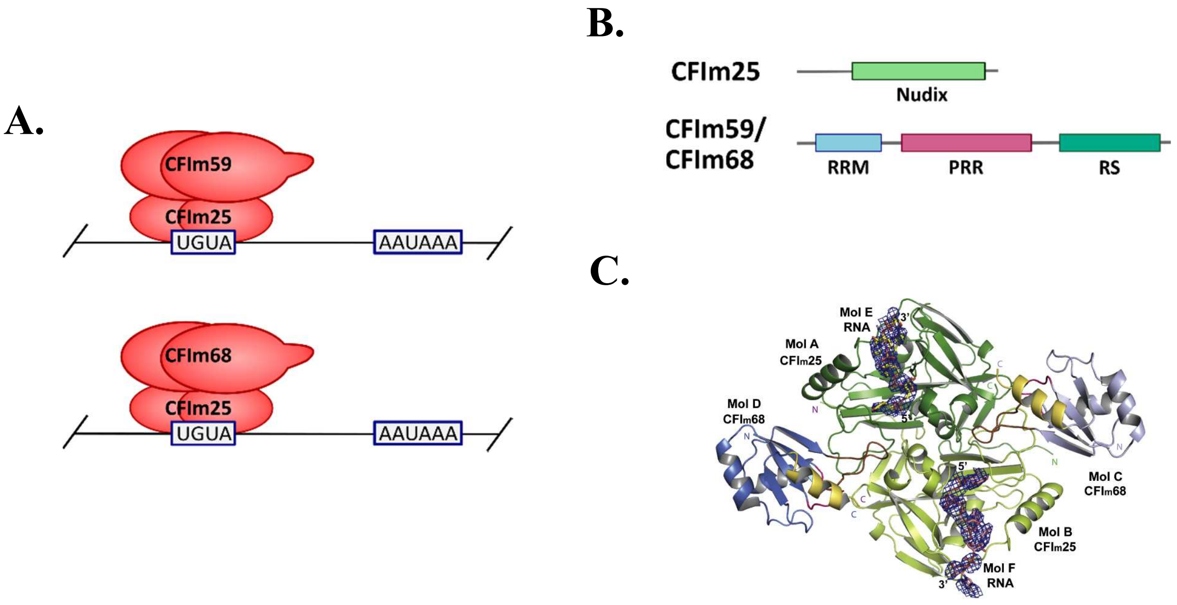
Because CFIm has a strong effect on the regulation of gene expression, we chose to characterize the effects of CFIm on APA genome-wide by knocking down CFIm25, the small subunit, and performing both RNA sequencing and polyadenylation site sequencing (PAS-Seq). In addition to confirming its known role in enhancing distal polyA sites, we also found a novel role for CFIm in activating intronic polyA sites, an alternative polyadenylation event in which one of the polyadenylation sites lies within the intronic region of a gene instead of within the 3' untranslated region. Intronic polyadenylation often results in mature mRNAs that are not translated; those that are translated create proteins that are severely truncated and lack functional domains. Strikingly, we found that CFIm promotes intronic polyadenylation within one subunit of each of the other core polyA machinery members, suggesting that CFIm controls the levels of the rest of the core polyA machinery.

In addition to its role in APA regulation, we and our collaborators recently showed that knockdown of CFIm25 leads to a 30-fold increase in the reprogramming of mouse embryonic fibroblasts into induced pluripotent stem cells. Importantly, CFIm25 knockdown still enhanced reprogramming when Myc was omitted, suggesting that CFIm depletion can substitute for myc overexpression (Justin Brumbaugh et al. 2018). Because knockdown of CFIm25 can eliminate the necessity of myc (a gene expression amplifier) (Bradner, Lee, and Young 2013; Nie et al. 2012) we propose that CFIm is a global gene attenuator that regulates cell fate through regulation of both 3' UTR APA and intronic polyadenylation.

3.2 INTRODUCTION

In addition to being an essential step in mRNA processing, 3'end processing can be alternatively regulated, leading to mature mRNAs with unique 3' ends and distinct regulatory properties. One member of the core polyadenylation machinery that is particularly relevant in the context of alternative polyadenylation is cleavage factor I (CFIm). CFIm is not necessary for reconstitution of 3'end processing in vitro and has no yeast homolog, indicating that it is not essential for cleavage or polyadenylation (Boreikaite et al. 2022; Schmidt et al. 2022). However, PAR-CLIP analysis of polyadenylation factors revealed that in cases of alternative polyadenylation, CFIm68 is the best predictor of the most frequently utilized polyA site, followed by CstF64 and CstF64tau, CFIm59, and CFIm25 (G. Martin et al. 2012). In addition, the polyA sites activated by CFIm were frequently distal polyA sites, indicating that while CFIm is not an essential member of the 3' processing machinery, it is a critical regulator of APA that likely enhances distal polyA sites (G. Martin et al. 2012).

CFIm consists of two subunits: a small subunit CFIm25 and one of two alternative large subunits, either CFIm59 or CFIm68. The CFIm complex is a hetero-tetramer formed by a homodimer of CFIm25 and the two large subunits (Kim et al. 2010) (Figure 3.1A). CFIm25 is a non-canonical member of the Nudix hydroxylase family of proteins. Unlike other members of the Nudix hydroxylase family, the Nudix domain of CFIm25 is catalytically inactive and instead is used to bind to the UGUA motif of polyA site RNAs through hydrogen bonding and stacking interactions, which provide sequence specificity (Brown and Gilmartin 2003; Q. Yang, Gilmartin, and Doublié 2010) (Figure 3.1B). The large subunits, CFIm59 and CFIm68, are SR-like proteins that contain an N-terminal RNA recognition motif (RRM), a proline-rich region (PRR), and a C-terminal RS domain (Figure 3.1B). Although both large subunits contain an RNA recognition motif, RNA binding in vitro is stronger for CFIm25 than either CFIm59 or CFIm68, indicating that CFIm25 is more important in RNA binding (Q. Yang, Gilmartin, and Doublié 2010). One explanation for this is that although both CFIm59 and CFIm68 have an RRM, it forms an interaction surface with CFIm25 instead of binding to RNA (Dettwiler et al. 2004). It is more likely that CFIm59 and CFIm68 modulate RNA binding by CFIm25 because crystal structures have revealed that CFIm25 forms a homodimer that binds to two UGUA motifs, which are clasped on opposite sites by two CFIm68 RRMs that enhance RNA binding and RNA looping (Q. Yang et al. 2011) (Figure 3.1C). The RS domain of CFIm59 and CFIm68 is involved in both protein-protein interactions with members of the core polyA machinery as well as other SR proteins in addition to nuclear localization in paraspeckles (Dettwiler et al. 2004; Zhu et al. 2017).



Yang et al (2011) Structure

Figure 3.1 Structure of the CFIm Complex

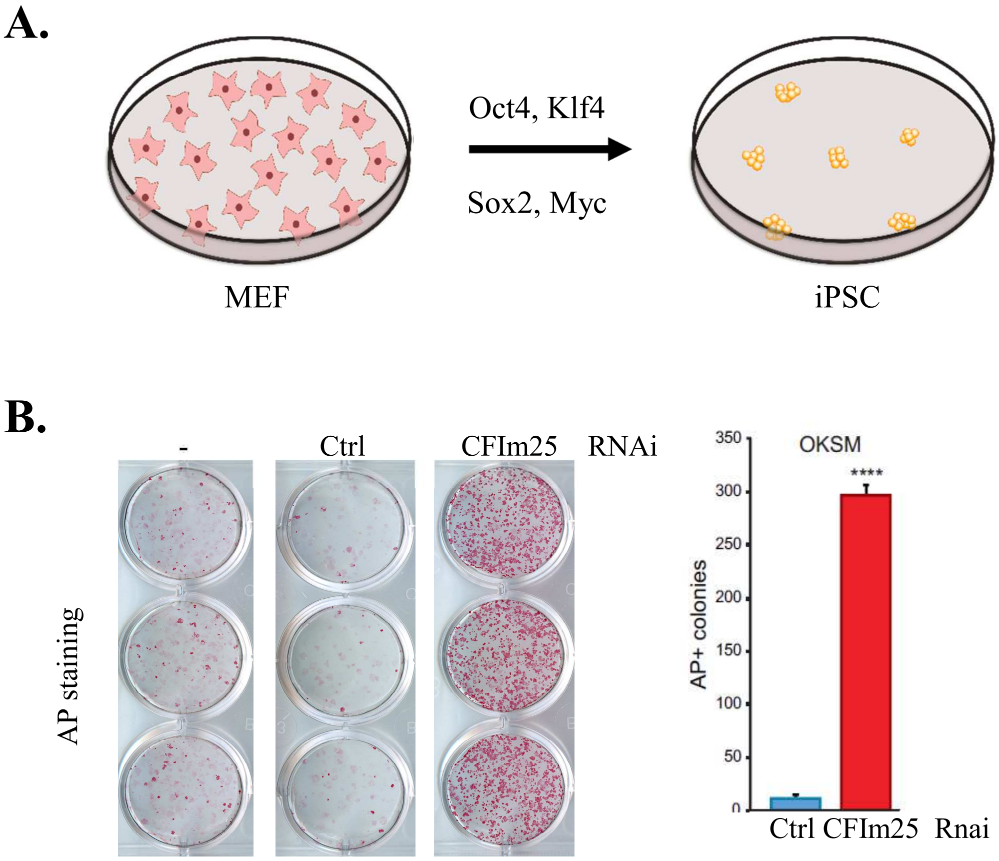
A. Schematic depicting the two variants of the CFIm complex. Each consists of two CFIm25 subunits with either CFIm59 or CFIm68. Both complexes bind to the UGUA motif upstream of polyA sites.

B. Schematic of domain structure of CFIm subunits.
RRM: RNA recognition motif PRR: Proline rich region
RS: Argine-Serine Rich Domain

C. Crystal structure demonstrating interaction of CFIm25-CFIm68 complex with RNA. RNA molecules are shown as a stick model.

Several lines of evidence indicate that CFIm is an enhancer of 3’ processing. First, preincubation of RNA substrates with CFIm enhances both the rate and efficiency of cleavage in vitro (Rüegsegger, Blank, and Keller 1998). In addition, several studies have indicated that depletion of CFIm25 leads to a widespread shift from distal to proximal polyA sites. Interestingly, knockdown of CFIm68 leads to a similar shift in APA profile, while depletion of CFIm59 does not, suggesting that although highly structurally similar, CFIm59 and CFIm68 may not be completely redundant for each other (G. Martin et al. 2012; Masamha et al. 2014; Zhu et al. 2017; W. Li et al. 2015; Kubo et al. 2006). Recently, our lab demonstrated that CFIm specifically enhances distal polyA sites. CFIm binds to the UGUA enhancer motif, which is enriched approximately 50 nucleotides upstream of distal polyA sites. It can then recruit the core polyA machinery to nearby polyA sites through interactions of the RS domain of CFIm68 with the RE/D domain of Fip1, a subunit of the cleavage complex CPSF (Figure 1.3). This model is not only consistent with but also provides the mechanistic basis for the previous observations. First, pre-incubation of RNA with CFIm enhances in vitro cleavage because it allows CFIm to recruit CPSF to the polyA site. Second, CFIm depletion causes a widespread shift from distal to proximal polyA sites because when CFIm levels are high, CFIm promotes distal polyA sites, leading to longer 3’ UTRs. However, when CFIm levels are depleted, there is no longer an enhancement of distal polyA sites, leading to preferential usage of proximal sites that are transcribed first and therefore have an inherent advantage.

In addition to its role in 3’ processing, CFIm also plays a critical role in cell fate determination. It has previously been shown that stem cells display a unique APA profile in comparison to differentiated cells. Specifically, stem cells have a high proportion of transcripts with shorter 3’ UTRs resulting from polyadenylation at proximal polyA sites. Conversely, during mouse



Adapted from Brumbaugh et al (2018) Cell

Figure 3.2 CFIm Regulates Cell Fate Decisions

A. Schematic depicting reprogramming of mouse embryonic fibroblasts (MEFs) into induced pluripotent stem cells (iPSC).
MEF: mouse embryonic fibroblast
iPSC: induced pluripotent stem cell

B. Reprogramming of MEFs into induced pluripotent stem cells in the presence or absence of CFIm. Reprogramming levels are measured by AP (alkaline phosphatase) staining.

embryonic development, there is progressive lengthening of 3' UTRs (Z. Ji et al. 2009). CFIm likely plays a role in this differential APA profile because it was recently shown to regulate reprogramming. During reprogramming, some differentiated cell types including fibroblasts can be reverted to a stem-cell like state known as induced pluripotent stem cells (iPSCS) by overexpression of transcription factors such as Oct4, Klf4, Sox2, and Myc. Like stem cells, iPSCS are pluripotent and can self-renew (Nakagawa et al. 2008; Takahashi and Yamanaka 2006) (Figure 3.2A). In addition, consistent with lengthening of 3' UTRs during cellular differentiation, there is also progressive shortening of 3' UTRs during reprogramming of mouse embryonic fibroblasts, with opposing APA changes during differentiation and reprogramming for many genes (Z. Ji and Tian 2009). Recent work with our collaborators revealed that knockdown of CFIm25 leads to a 30-fold increase in the ability to reprogram mouse embryonic fibroblasts into iPSCs (Figure 3.2B). Upon reprogramming, there were widespread APA changes, predominantly 3' UTR shortening, consistent with the role of CFIm in promoting distal polyA sites (Justin Brumbaugh et al. 2018). This data suggests that CFIm promotes a differentiated cell fate, which is relevant because current reprogramming protocols are highly inefficient, often resulting in only 0.1-0.3% reprogramming. The inefficiency of reprogramming has been hypothesized to be a result of “road-block genes” that protect cell fate and prevent aberrant cell identity changes. If this hypothesis is true, depletion of these road-block genes should enhance reprogramming, which is exactly the case with CFIm. Together, this indicates that CFIm is an important link between 3' processing and cell fate determination, but the precise mechanism is yet undetermined.

In this chapter, I further uncover the role of CFIm in both APA regulation and in cell fate determination. There are two competing hypotheses for the role of CFIm in linking APA and reprogramming. Under the first hypothesis, a small subset of APA events regulated by CFIm are relevant for reprogramming. Alternatively, CFIm may have a more global role in regulating gene expression and the sum of many APA changes together regulates reprogramming efficiency. To test this, I depleted CFIm and performed genome-wide sequencing analyses. Upon depletion of CFIm, I made the novel discovery that CFIm promotes intronic polyadenylation, in addition to corroborating previous studies that implicate CFIm in promoting distal polyA sites. Specifically, we found that CFIm regulates intronic polyadenylation within one factor of each of the other core polyA machinery members, suggesting that it is a master regulator of the 3' processing machinery. In addition, I found that CFIm depletion enhances gene expression on both the mRNA and protein level for a large number of genes, indicating that CFIm is a global attenuator of gene expression. Interestingly, the transcription factor c-myc, which is important but not essential for reprogramming to occur, amplifies gene expression (Bradner, Lee, and Young 2013; Nie et al. 2012). If CFIm has the opposing function as myc on gene expression, it may explain how CFIm depletion enhances reprogramming, providing novel insight into how CFIm links 3' processing to cell fate.

3.3 RESULTS

CFIm Regulates Both 3' UTR APA and Intronic Polyadenylation

To investigate the role of CFIm in regulating gene expression genome-wide, I knocked down the small subunit CFIm25 in human 293T cells and performed paired-end RNA sequencing. By knocking down CFIm25, CFIm68 was partially co-depleted and CFIm59 was co-depleted, indicating depletion of the CFIm complex as a whole (Figure 3.3A). Polyadenylation site usage was then compared between control and CFIm25 knockdown using DaPars, a computational

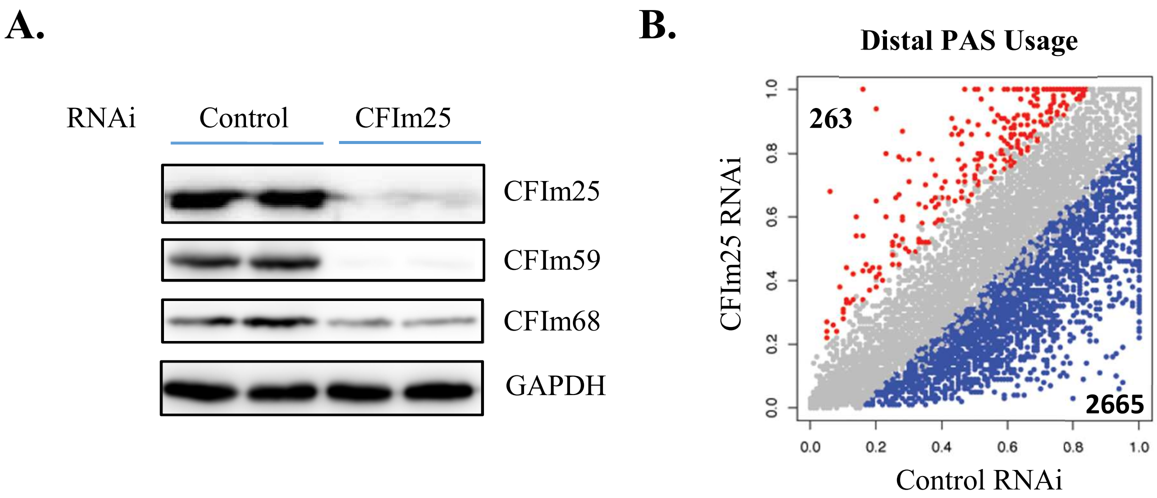


Figure 3.3 CFIm Mediates APA

A. Knockdown of CFIm25 in mammalian 293T cells using RNAi. Depletion of CFIm25 led to partial codepletion of CFIm68 and codepletion of CFIm59.
B. Scatterplot of CFIm-mediated APA events as determined by RNA sequencing and DaPARs analysis. Each point represents a specific the levels of distal polyA site usage for a specific polyA site. Control knockdown is on the x-axis and CFIm25 knockdown on the y-axis. Significant 3’UTR lengthening is indicated in red and shortening is indicated in blue.

program that performs de-novo polyA site identification and analyzes APA changes in RNA sequencing data (Xia et al. 2014). Consistent with data from both our group and others, there was a widespread shift from distal to proximal polyA site usage upon CFIm depletion, with 2665 genes showing a distal to proximal APA change and only 293 genes showing a proximal to distal APA change (G. Martin et al. 2012; Masamha et al. 2014; Zhu et al. 2017; W. Li et al. 2015) (Figure 3.3B). This bias towards distal to proximal APA shifts upon CFIm depletion is consistent with the known role of CFIm in polyadenylation regulation as CFIm is a polyadenylation enhancer and its binding sites are enriched upstream of distal polyA sites. Two examples of genes that undergo 3' UTR shortening regulated upon knockdown of CFIm25 include Vma21 and Spcs3 (Figure 3.4A). In both cases, within control cells there is strong usage of a downstream, distal polyA site. However, upon knockdown of CFIm25 there is shift to an upstream, proximal polyA site, indicating 3' UTR shortening. These APA changes were validated experimentally by 3' RACE, a cDNA amplification technique utilized for polyA site analysis, as shown with the example gene VMA21. In 3' RACE, cDNA is synthesized using a reverse oligo(dT) primer that recognizes the polyA tail and has a linker at the 3' end. cDNA is subsequently used for PCR using two forward primers in the same reaction that bind to either the shared region upstream of the proximal polyA site or a distal polyA site specific region and a reverse primer that binds to the linker added during reverse transcription. PCR products are then used for a second round of PCR to increase specificity. The ratio of common to extended isoforms is used to compare distal polyA site usage to proximal polyA site usage (Figure 3.4B). Consistent with the RNA-seq data, 3' RACE revealed Vma21 3' UTR shortening upon knockdown of CFIm25 (Figure 3.4C).

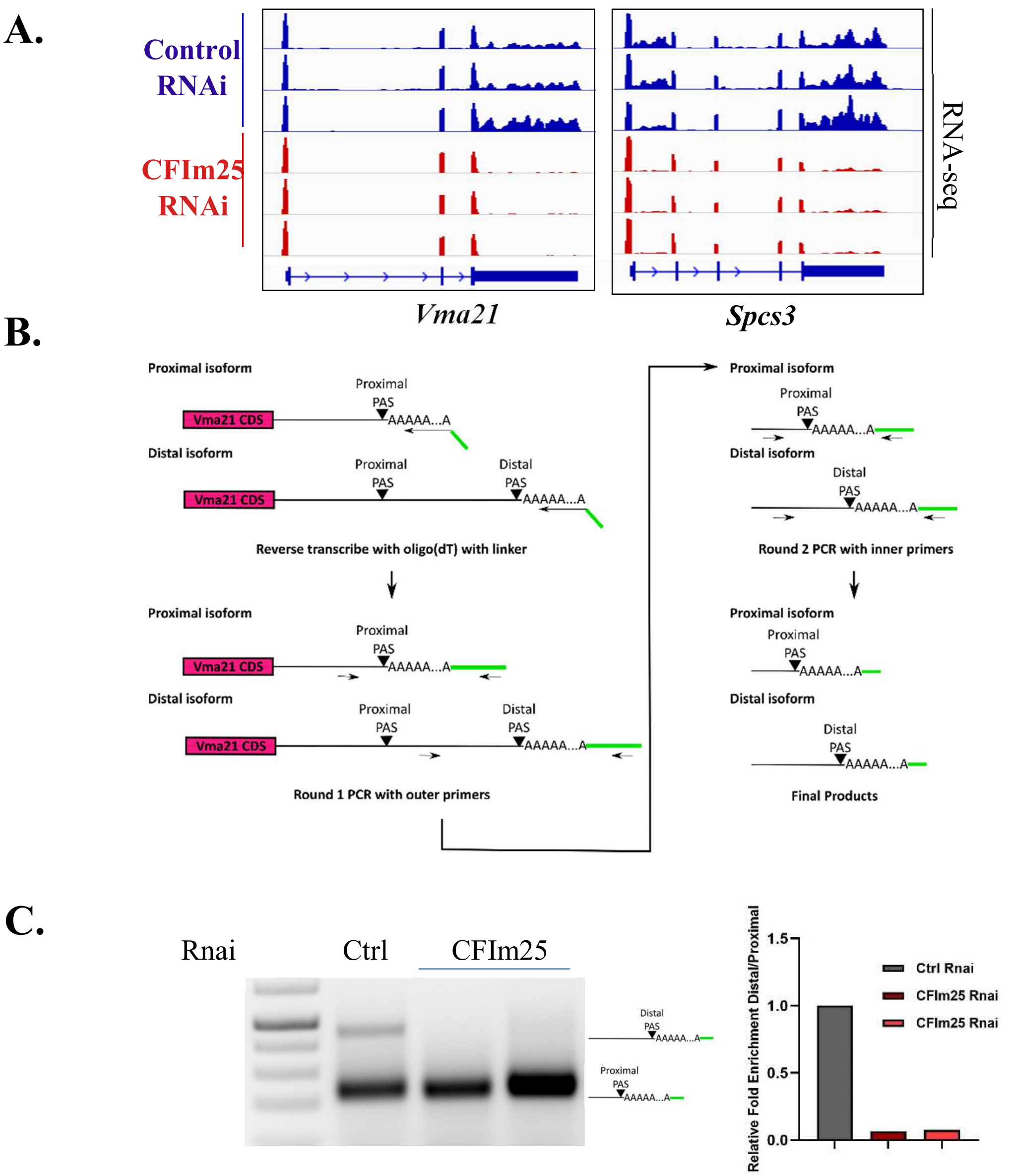


Figure 3.4 CFIm Enhances Distal PolyA Sites

A. RNA sequencing tracts demonstrating distal to proximal shift in polyA site usage upon knockdown of CFIm25 in Vma21 (left) and Spcs3 (right).
B. Schematic demonstrating procedure to verify APA changes in Vma21 by 3' RACE
C. 3' RACE for control and CFIm25 depletion cells to analyze polyA site usage of Vma21.

Despite this clear trend towards 3’ UTR shortening upon CFIm depletion, in 293 cases, there was a shift in the opposite direction from proximal to distal polyA site usage, the mechanism for which cannot be explained by the known role of CFIm in polyA site selection. Further investigation into these 293 cases revealed that for the majority of these non-canonical APA changes, CFIm promotes intronic polyadenylation. Intronic polyadenylation occurs when a polyA site is localized in an intronic region before the end of the protein coding region of the mRNA. When intronic polyadenylation occurs, it frequently causes nonsense-mediated decay although occasionally a truncated fragment may be produced if adenosines within the 3’ UTR create a stop codon (Vasudevan, Peltz, and Wilusz 2002; P. Yao et al. 2012). Two examples of genes that undergo CFIm mediated intronic polyadenylation include Wwp2 and CSTF3 (also known as CstF77) (Figure 3.5). In both cases, within control cells, there is strong usage of an intronic polyA site and limited expression of the full-length mRNA isoform. However, upon CFIm depletion, there is decreased usage of the intronic site and an increase in the full-length mRNA. Notably, these 293 cases are likely an underestimate because only Refseq annotated polyA sites were included within the analysis and many intronic polyA sites are missing from Refseq.

CFIm Complex Regulates APA

As CFIm is a multi-subunit protein complex, we were next interested in whether the large subunit, either CFIm59 or CFIm68, also regulates intronic polyadenylation. To directly compare the effects of CFIm25, CFIm59 and CFIm68 on intronic polyadenylation, all three subunits were knocked down in mammalian 293T cells and RNA was utilized for polyadenylation site sequencing (PAS-Seq) (Figure 3.6A). PAS-Seq is a 3’-end sequencing method used to map

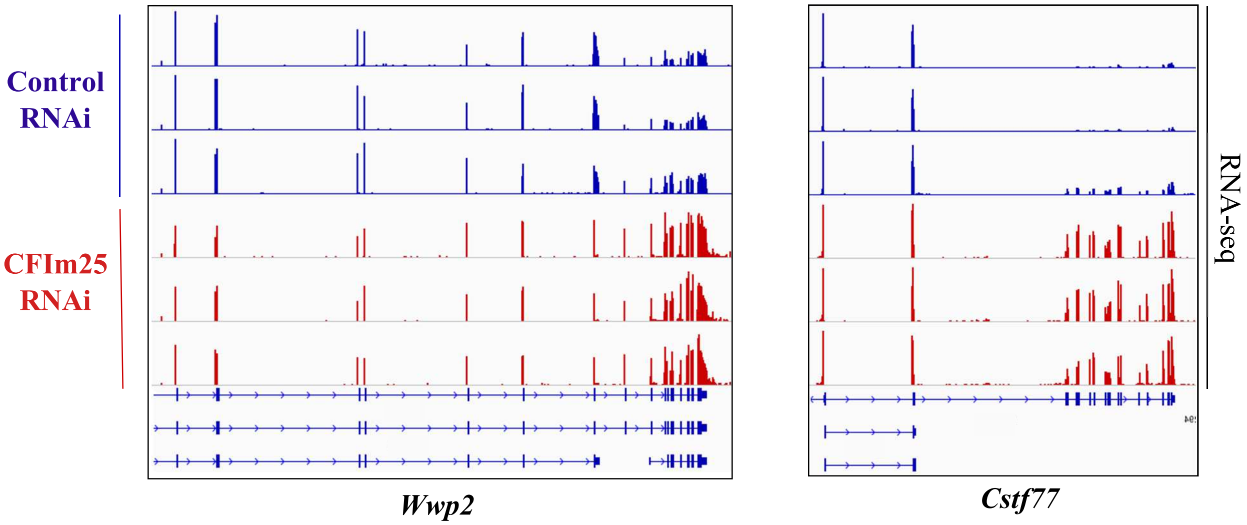
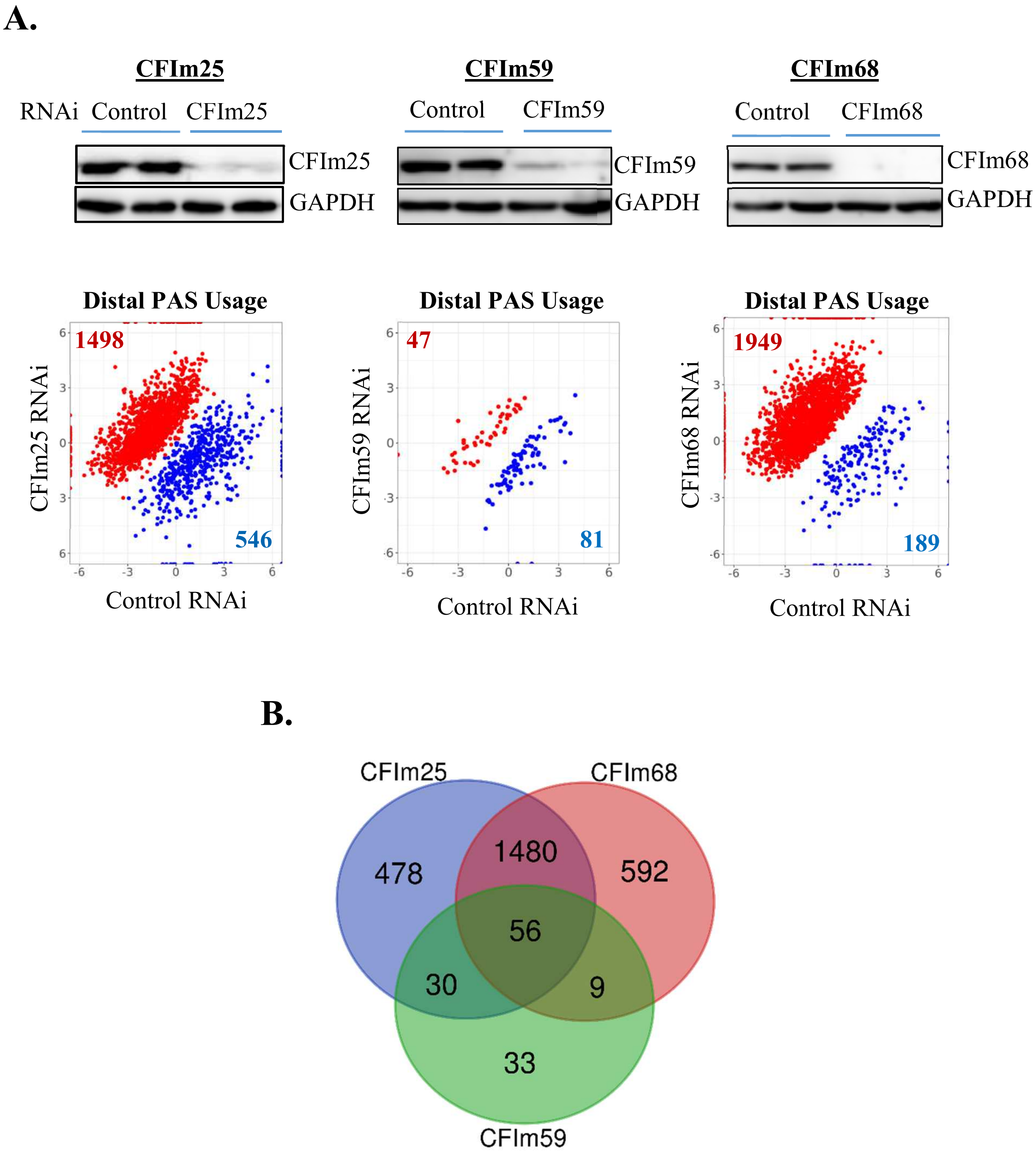


Figure 3.5 CFIm Promotes Intronic Polyadenylation

RNA sequencing tracts demonstrating decreased intronic polyA site usage upon knockdown of CFIm25 in Wwp2 (left) and CstF77 (right). In addition to decreased intronic polyA site usage, CFIm25 depletion RNAs exhibit increased full length RNA isoform levels.

polyadenylation sites globally and quantify the relative usage of alternative sites (Figure 2.8B) (Yoon, Soles, and Shi 2021). PAS-Seq analysis revealed that knockdown of all three subunits affected polyadenylation site selection, with 128, 2045, and 2139 polyadenylation sites showing differential regulation upon knockdown of CFIm59, CFIm25, and CFIm68 respectively (Figure 3.6A). 56 APA events were regulated by all three CFIm subunits (Figure 3.6B). For example, the genes Arl14ep and Lnpep both displayed distal to proximal shifts in polyA site usage upon knockdown of all CFIm subunits, as indicated by a shift an upstream polyA site (Figure 3.6A). Both genes also show similar 3' UTR shortening in RNA sequencing samples upon knockdown of CFIm25, confirming that RNA sequencing and PAS-seq both accurately identify APA (data not shown).

Although there were APA events regulated by all CFIm subunits, there were a greater number of APA events regulated by CFIm25 and CFIm68 than CFIm59, which suggests that CFIm25 and CFIm68 have a larger effect on alternative polyadenylation, a result that is consistent with previous data (Zhu et al. 2017) (Figure 3.6A). Even for Arl14ep and Lnpep, which exhibited significant APA changes, there was a larger change in APA upon knockdown of CFIm25 or CFIm68 than CFIm59. For example, in CFIm25 or CFIm68 knockdown cells there was either a complete shift from the distal to the proximal polyA site of Lnpep (CFIm68) or the proximal isoform became predominant (CFIm25). However, while CFIm59 knockdown increased proximal polyA site usage, there was an approximately equal ratio of the two isoforms (Figure 3.6C). Additionally, 1480 APA changes were regulated solely by CFIm25 and CFIm68, which is approximately 75% of all APA events regulated by CFIm25 (Figure 3.6B). Examples of genes with APA events regulated by both CFIm25 and CFIm68 include Dpy19l4 and Hipk1. Both genes displayed 3' UTR shortening upon knockdown of CFIm25 and CFIm68 but were not


A.
CFIm25
CFIm59
CFIm68
RNAi
Control
CFIm25
Control
CFIm59
Control
CFIm68
CFIm25
GAPDH
CFIm59
GAPDH
CFIm68
GAPDH
Distal PAS Usage
Distal PAS Usage
Distal PAS Usage
1498
546
CFIm25 RNAi
Control RNAi
47
81
CFIm59 RNAi
Control RNAi
1949
189
CFIm68 RNAi
Control RNAi
6
3
0
-3
-6
-6
-3
0
3
6
B.
CFIm25
CFIm68
CFIm59
478
1480
592
56
30
9
33

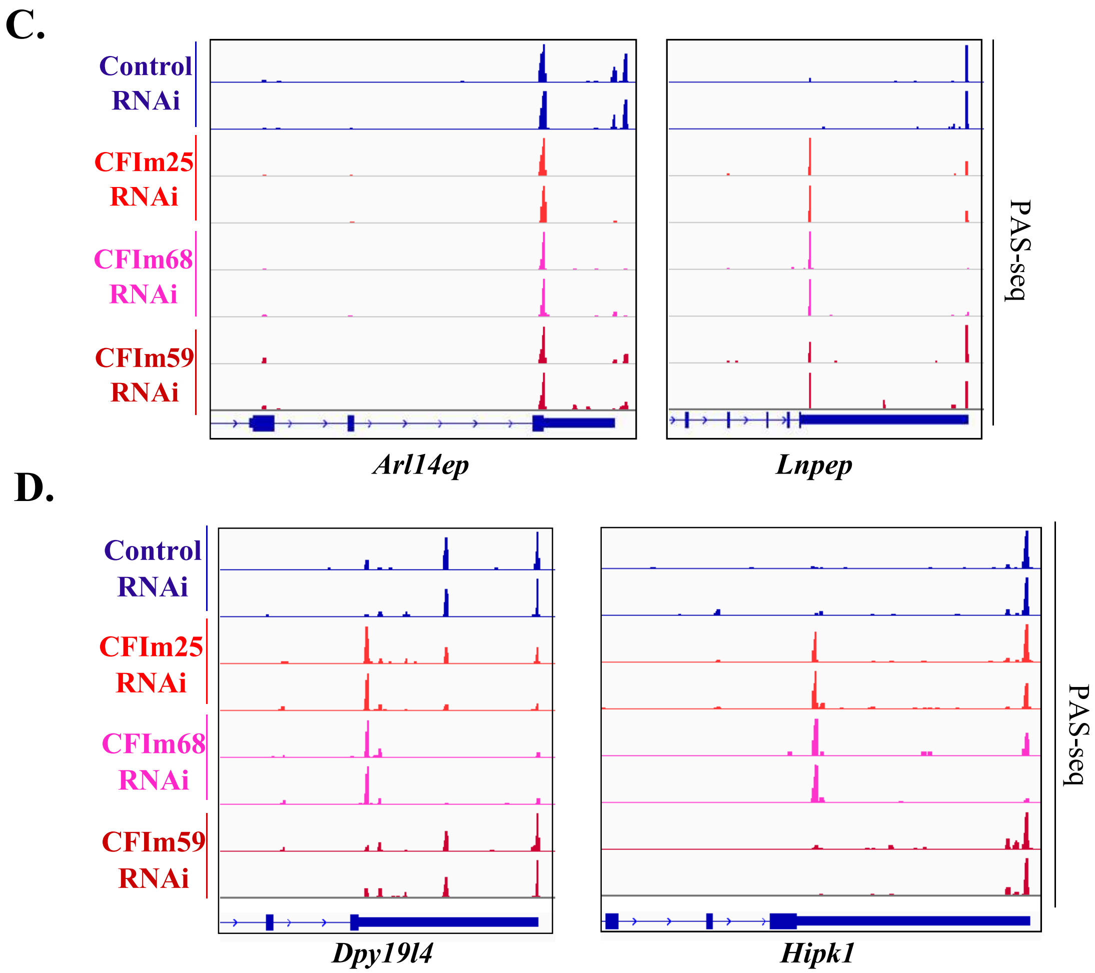


Figure 3.6 CFIm Subunits Have Unique and Overlapping Effects on 3' UTR APA

A. Knockdown of CFIm25, CFIm59 and CFIm68 in mammalian 293T cells (top) Scatterplot of CFIm-mediated APA events as determined by PAS-Seq. Each point represents the levels of distal polyA site usage for a specific polyA site. Control knockdown is on the x-axis and CFIm knockdown on the y-axis. Significant 3'UTR lengthening is indicated in red and shortening is indicated in blue.

B. Overlap of APA changes regulated by CFIm25, CFIm59, and CFIm68 as identified by PAS seq.

C. PAS sequencing tracts depicting APA changes caused by knockdown of CFIm25, CFIm59, and CFIm68 in example genes Alr14ep (left) and Lnpep (right). In both cases, all three subunits enhance distal polyA sites.

D. PAS sequencing tracts depicting APA changes caused by knockdown of CFIm25, CFIm59, and CFIm68 in example genes Dpy19l4 and Hipk1. For both genes, knockdown of CFIm25 and CFIm68 but not CFIm59 causes a distal to proximal polyA site usage shift.

significantly regulated by CFIm59, which looked more similar to control knockdown (Figure 3.6D). A final difference between CFIm59 and the other CFIm subunits was that knockdown of CFIm25 and CFIm68 induced a widespread shift from distal to proximal polyA site usage, with approximately 75% and over 90% of APA changes being 3’ UTR shortening for CFIm25 and CFIm68 respectively. Interestingly, however, the opposite was true upon knockdown of CFIm59: 47 APA changes showed a distal to proximal shift and 81 showed a proximal to distal shift (Figure 3.6A). Together, this suggests that two CFIm complexes have different effects on APA, which is consistent with previous studies that suggest that while CFIm59 and CFIm68 were once considered synonymous, they in fact play distinct roles and that the CFIm25-CFIm68 complex regulates APA more strongly than the CFIm25-CFIm59 complex (Zhu et al. 2017; Tseng et al. 2021).

RNA sequencing analysis also revealed a novel role for CFIm25 in promoting intronic polyadenylation, so we were next interested in which subunit or subunits are directly involved. PAS-Seq analysis of CFIm25 knockdown cells revealed that although the majority of APA changes were within the 3’ UTR, there were 362 cases of CFIm25-regulated intronic polyadenylation events. 155 or approximately 40% of CFIm-regulated IPA events showed a proximal to distal to shift, which is particularly striking as only approximately 25% of all CFIm-regulated APA events were distal to proximal shifts (Figure 3.7A). This finding is consistent with RNA sequencing data that suggests that CFIm25 promotes intronic, leading to increased full-length mRNA production upon knockdown of CFIm25. As an example, the gene Ankrd10 showed strong usage of an intronic site within control cells but showed decreased usage of this same intronic polyA site upon knockdown of CFIm25, CFIm59 or CFIm68 as well as increased usage of a polyA site within the terminal exon (Figure 3.7B).

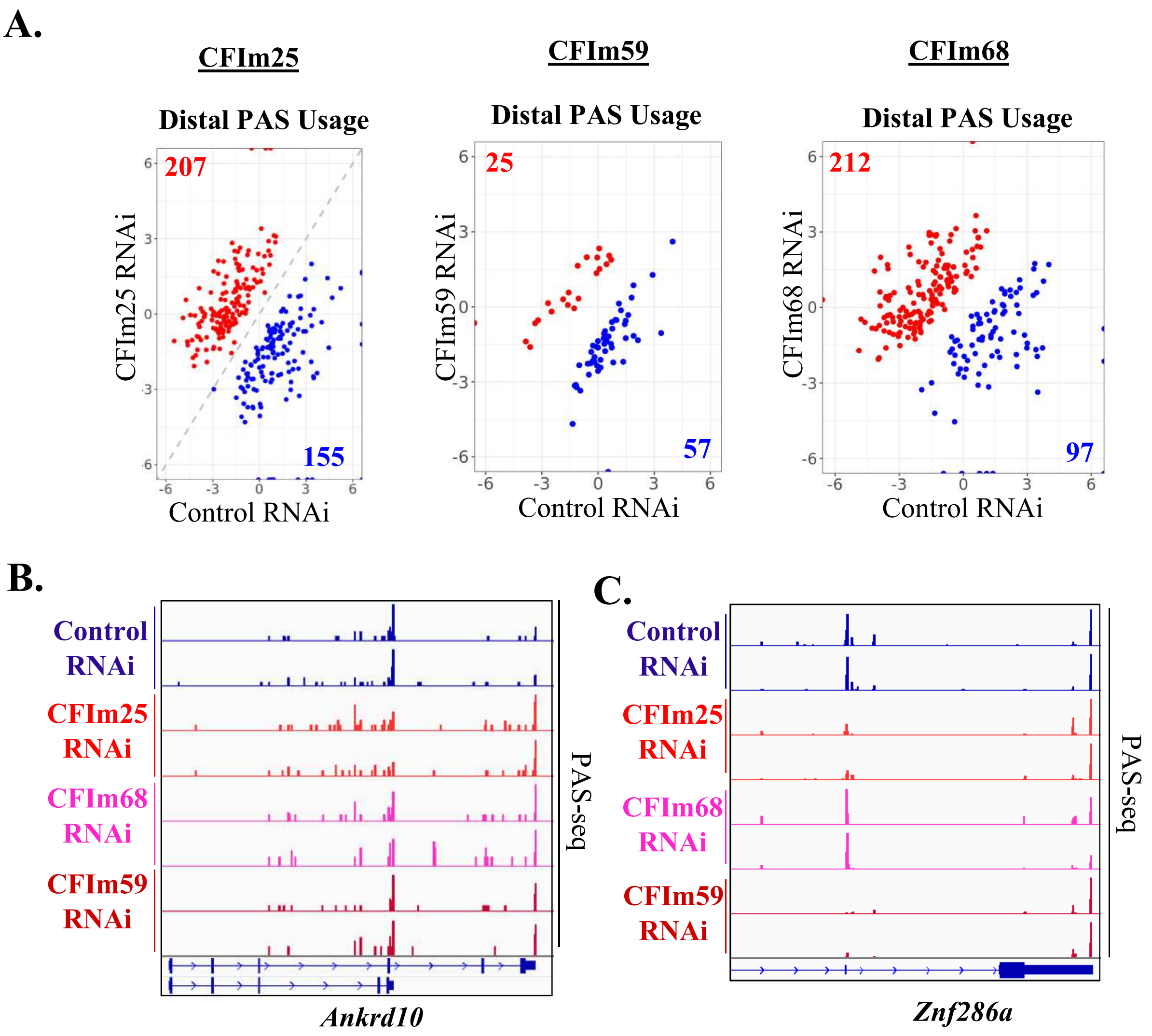


Figure 3.7 All CFIm Subunits Regulate Intronic Polyadenylation

A. Scatterplot of CFIm-mediated IPA events as determined by PAS-Seq. Each point represents the levels of distal polyA site usage for a specific polyA site. Control knockdown is on the x-axis and CFIm knockdown on the y-axis. Significant 3'UTR lengthening is indicated in red and shortening is indicated in blue.
B. PAS sequencing trats depicting APA changes caused by knockdown of CFIm25, CFIm59, and CFIm68. All three subunits promote intronic polyadenylation within Ankrd10.
C. PAS sequencing tracts depicting APA changes caused by knockdown of CFIm25, CFIm59, and CFIm68 in example gene Znf286a. Knockdown of CFIm25 and CFIm59 but not CFIm68 increase full length mRNA expression.

As with the case of Ankrd10, both CFIm59 and CFIm68 regulated intronic polyadenylation in addition to CFIm25. Of the 2115 APA events regulated by CFIm68, 311 are intronic events, a proportion that is similar to that of CFIm25 regulated intronic polyadenylation events. Approximately 30% of these intronic polyadenylation events exhibited proximal to distal shifts, which is proportionally lower than that of CFIm25-promoted IPA sites but still higher than the only 8% all CFIm68 regulated APA events that displayed proximal to distal shifts (Figure 3.7B). Interestingly, of the 128 APA events regulated upon knockdown of CFIm59, 82 included an intronic event. In addition, close to 70% of regulated intronic polyadenylation events displayed a distal to proximal shift and therefore increased full-length mRNA production. This indicates that the CFIm25-CFIm59 complex may have a larger effect on promoting intronic polyadenylation than the CFIm25-CFIm68 complex, further expounding on the differences between the two CFIm complexes. In fact, between 60%-70% of APA events that were regulated by both CFIm25 and CFIm59 were intronic events (19 out of 30 total events). For example, the Ips site of Znf286a is utilized in control and CFIm68 knockdown cells, but has decreased utilization upon knockdown of either CFIm59 or CFIm25, indicating that CFIm25-CFIm59 complex promotes intronic polyadenylation at this site (Figure 3.7C).

Together, our study indicates that the two CFIm complexes, CFIm25-CFIm59 and CFIm25-CFIm68, have partially overlapping but also unique functions in regulating APA. Overall, CFIm68 has a larger effect on APA than CFIm59, particularly within the 3' UTR. However, CFIm59 may play a more important role in promoting intronic polyadenylation, which is a novel function of CFIm in RNA processing regulation.

CFIm Binds Intronic polyA Sites to Regulate IPA

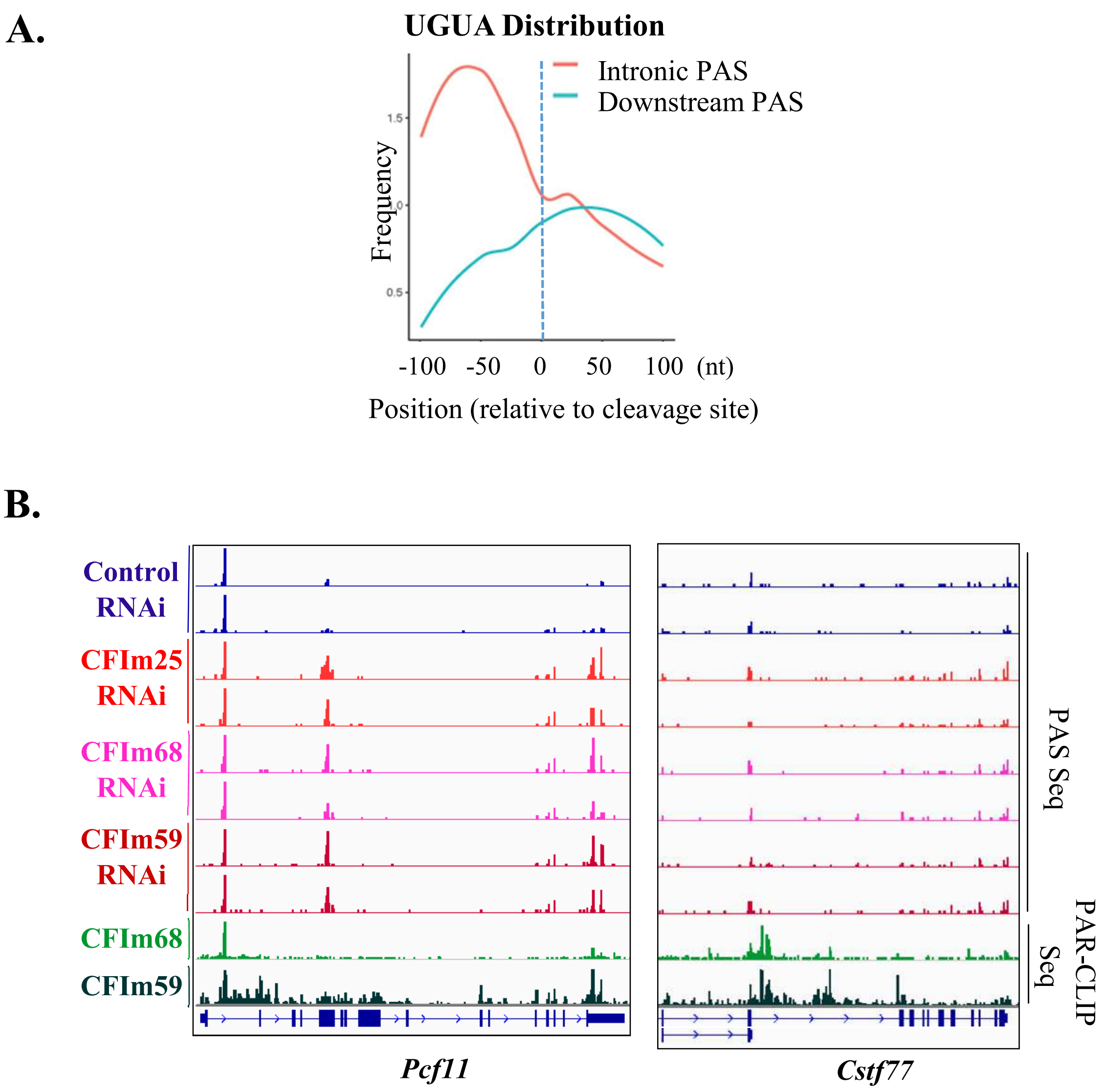


Figure 3.8 CFIm Binds to RNA to Regulate Intronic Polyadenylation

A. Genome-wide UGUA distribution at intronic and downstream polyA sites of genes containing CFIm-regulated IPA sites
B. PAS sequencing (top) and PAR-CLIP sequencing tracts (bottom) comparing CFIm-mediated intronic polyadenylation site regulation and CFIm binding.

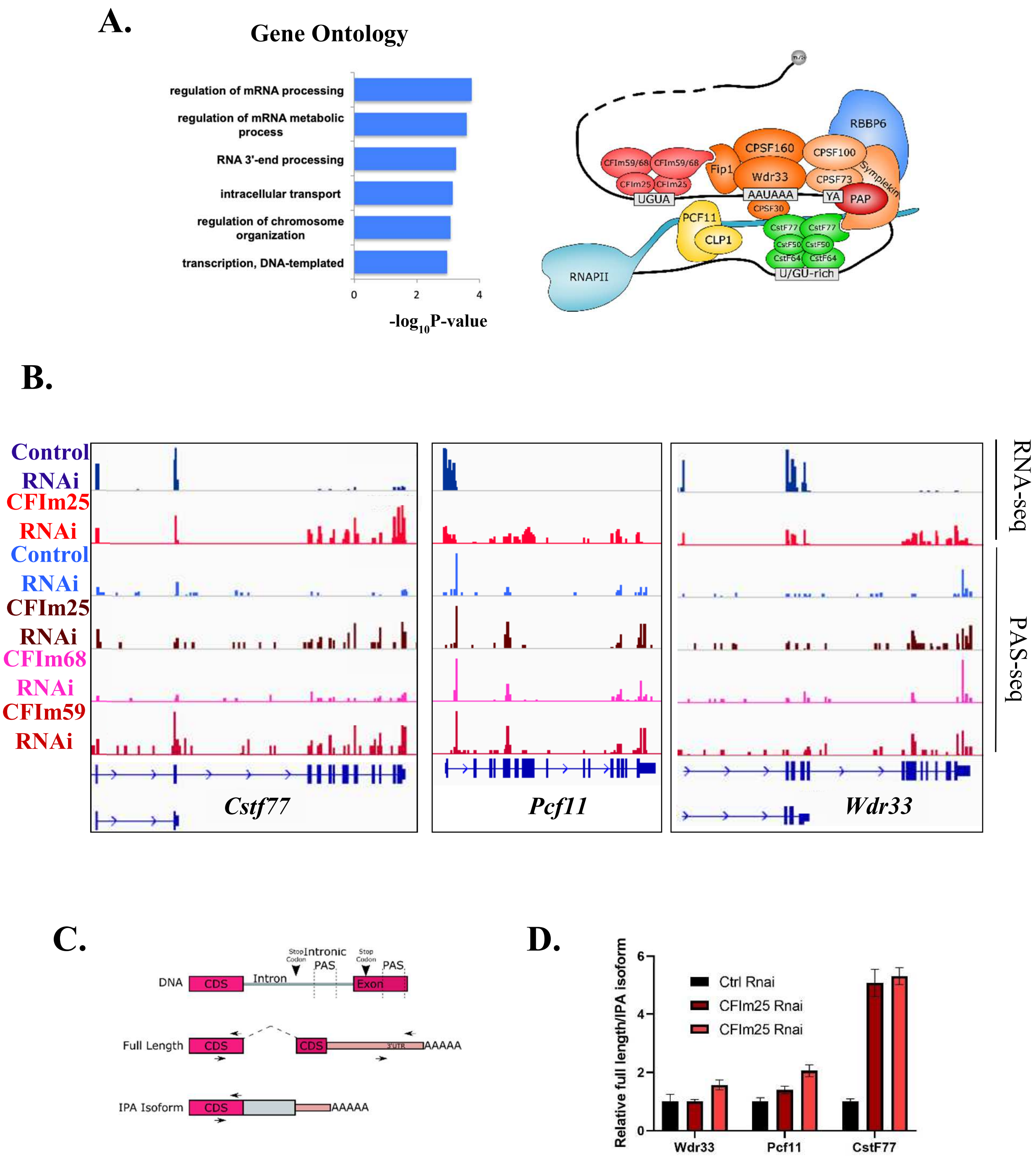
To investigate the mechanism for CFIm-mediated intronic polyadenylation regulation, we analyzed the enrichment of the CFIm-binding UGUA binding motif within these genes. For all genes containing a CFIm-regulated IPA site, we analyzed the frequency of UGUA from 100 nucleotides upstream of the cleavage site to 100 nucleotides downstream for both the intronic polyA site as well as for downstream, 3' UTR polyA sites in comparison to randomly selected sequences. Strikingly, there was an enrichment for UGUA approximately 50 nucleotides upstream of the cleavage site of these intronic polyA sites, consistent with previous studies that have shown that CFIm exerts maximal enhancer activity when bound 50 nucleotides upstream of cleavage sites (Figure 3.8A) (Zhu et al. 2017). This suggests that CFIm binds to and activates these intronic polyA sites. By contrast, there was no enrichment for UGUA upstream of the cleavage site within downstream polyA sites within the terminal exon. Instead, there was a modest enrichment of UGUA exon downstream of the polyA site (Figure 3.8A). This binding enrichment pattern indicates that CFIm may in fact compete with CstF for binding because CstF is known to bind to U or G/U rich regions downstream of the polyA site (Y Takagaki and Manley 1997). This provides further insight into how CFIm promotes intronic polyA sites as the intronic sites are stronger than downstream sites due to activation by CFIm. In addition, if CFIm competes with CstF for binding to the downstream element, CFIm may further inhibit downstream sites, providing a secondary mechanism by which intronic polyA sites are activated. Enrichment of UGUA motifs upstream of polyA sites is not direct evidence of binding so we analyzed publicly available PAR-CLIP data for both CFIm59 and CFIm68 to compare CFIm binding at IPA sites and downstream sites in the terminal exon (G. Martin et al. 2012). Like other iterations of CLIP, PAR-CLIP (photoactivatable ribonucleoside-enhanced crosslinking and immunoprecipitation) is a technique that utilizes UV-irradiation of cells to crosslink RNA binding proteins to RNA. The RBP of interest in is then immunoprecipitated and bound RNAs are sequenced (Figure 2.4A). What makes PAR-CLIP unique is that it allows for single nucleotide resolution (Danan, Manickavel, and Hafner 2016). Consistent with the enrichment data, PAR-CLIP analysis revealed that CFIm binds near the intronic polyA site of genes such as Pcf11 and CstF77, both of which show decreased intronic polyadenylation upon CFIm knockdown (Figure 3.8C) (G. Martin et al. 2012). In addition, for CstF77 there was stronger PAR-CLIP signal for both CFIm59 and CFIm68 at the intronic site than in the downstream 3' UTR; this was also observed for CFIm68 on Pcf11, although the CFIm59 PAR-CLIP signal was similar at the IPA and the downstream 3' UTR polyA site. Together, UGUA enrichment and PAR-CLIP analysis of CFIm59 and CFIm68 indicate that CFIm binds to intronic polyA.

CFIm is a Master Regulator of Other PolyA Factors

Upon determining that CFIm regulates intronic polyadenylation, we next investigated the types of genes that are affected. Gene ontology (GO Term) analysis of genes exhibiting CFIm-mediated intronic polyadenylation revealed moderate enrichment for "regulation of mRNA processing" and specifically "RNA 3'-end processing" (Figure 3.9). Analysis of the genes within these categories revealed that CFIm promotes intronic polyadenylation events within one component of each of the other core polyA complexes, specifically Wdr33 (CSPF), CstF77 (CstF), and Pcf11 (CFIIm)as well as both polyA polymerase (PAP) and Rbbp6 (Figure 3.9B). Consistently, PAS-seq analysis revealed that knockdown of all three CFIm subunits significantly decreased usage of an intronic polyA site within CstF77 and Wdr33.

To validate that CFIm regulates the levels of the polyA machinery on the mRNA level, RT-qPCR on CFIm25-depleted and control 293T cells was performed to compare the ratio of a region common to both the truncated and full-length isoforms to the 3'-UTR of the full-length isoform; this ratio indicates the ratio of the full-length isoform to the truncated isoform produced by intronic polyadenylation (Figure 3.9C). Consistent with the bioinformatic data, there was an increase in the ratio of full length Pcf11, Wdr33 and CstF77 in comparison to the truncated version produced by intronic polyadenylation upon CFIm depletion, validating that CFIm promotes intronic polyadenylation within one member of each other polyA complex (Figure 3.9D).

Upon determining that CFIm promotes intronic polyadenylation of polyA factors, we next evaluated polyA factors at the protein level. Numerous studies have indicated that changes at the mRNA and protein levels are not always directly correlated; in fact, across many studies, mRNA has approximately 40% explanatory power (De Sousa Abreu et al. 2009; Vogel and Marcotte 2012). Proteins are the functional unit for regulation of polyA site selection, so it is important to study how CFIm regulates expression levels at the protein level to gain a better understanding the true nature of CFIm's role in regulating 3' processing factors. Consistent with the mRNA results, western blotting analysis indicated that there was an increase in expression of full length Wdr33, CstF77, Pcf11 and PAP upon knockdown of CFIm25 (Figure 3.9E). Importantly, there was no increase in levels of other core polyA factors that do not contain a CFIm-regulated intronic polyA sites including CstF64 and CPSF100 as well as other RNA processing factors such as U170K (which is involved in splicing), indicating that the increase in expression is caused by regulation of intronic polyadenylation. This strongly suggests that CFIm specifically suppresses the expression of at least one subunit of each of the core polyA machinery complexes at both the mRNA and protein levels by promoting intronic polyadenylation.


A.
Gene Ontology
regulation of mRNA processing
regulation of mRNA metabolic process
RNA 3'-end processing
intracellular transport
regulation of chromosome organization
transcription, DNA-templated
0
2
4
-log10P-value
RBBP6
CPSF160
CPSF100
CFIm59/68
CFIm59/68
Fip1
Wdr33
CPSF73
Symplekin
CFIm25
CFIm25
UGUA
AAUAAA
YA
PAP
CPSF30
PCF11
CLP1
CstF77
CstF77
CstF50
CstF50
CstF64
CstF64
U/GU-rich
RNAPII
B.
Control RNAi
CFIm25 RNAi
Control RNAi
CFIm25 RNAi
CFIm68 RNAi
CFIm59 RNAi
Cstf77
Pcf11
Wdr33
RNA-seq
PAS-seq
C.
DNA
CDS
Intron
Stop Codon
Intronic PAS
Stop Codon
PAS
Exon
Full Length
CDS
CDS
3'UTR
AAAAA
IPA Isoform
CDS
AAAAA
D.
Relative full length/IPA isoform
0
2
4
6
Ctrl Rnai
CFIm25 Rnai
CFIm25 Rnai
Wdr33
Pcf11
CstF77

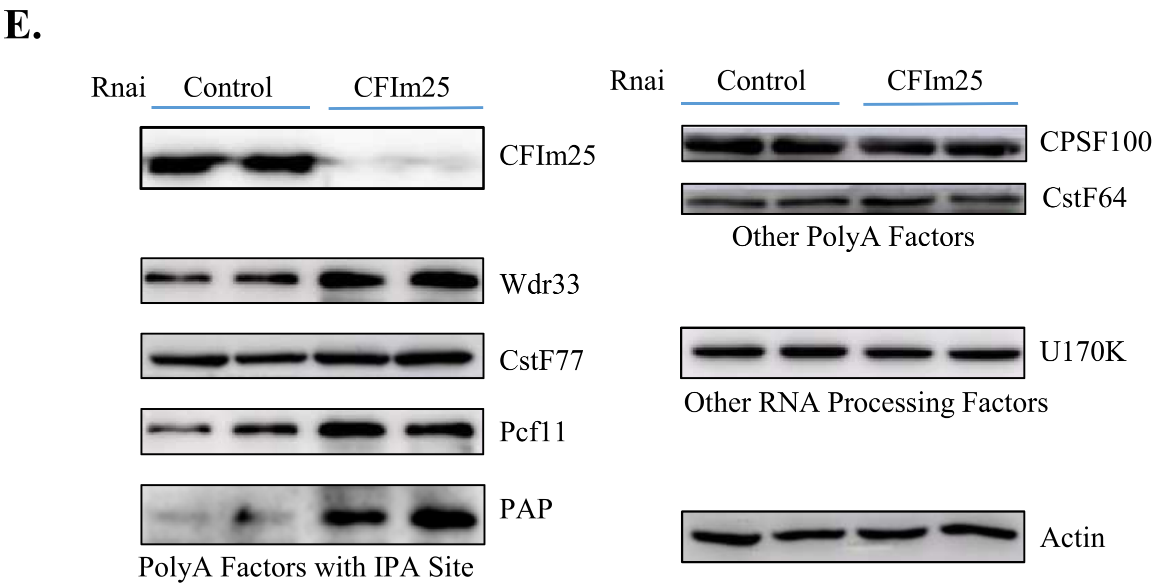


Figure 3.9 CFIm Regulates Expression of 3' Processing Factors

A. Gene Ontology (GO) analysis of genes exhibiting CFIm-promoted intronic polyA sites as determined by RNA-sequencing.
B. RNA sequencing (top) and PAS sequencing (bottom) tracts for 3' processing machinery components CstF77 (CstF), Wdr33 (CPSF), and Pcf11 (CFIIm). All three subunits contains CFIm-regulated intronic polyA sites.
C. Schematic depicting qPCR analysis of intronic polyA site usage. Primers bind to either a common region upstream of the IPA site or the 3' UTR of the full-length isoform. IPA is analyzed by the ratio of the common/extended isoform.
D. qPCR analysis of ratio between full length and truncated mRNA isoforms of Wdr33, Pcf11, and CstF77. All data is shown as relative to Ctrl Rnai for respective gene.
E. Western blotting analysis of expression levels of Wdr33, CstF77, Pcf11, and PAP following knockdown of CFIm25. Other polyA factors and RNA processing factors are included as controls.

CFIm-mediated regulation of core polyA machinery is a transient phenotype. A western blot time-course from 1 to 5 days post knockdown of CFIm25 indicated that at the protein level, there was a peak in levels of Wdr33, Pcf11, and CstF77 at early time points, followed by a return to levels closer to wildtype 293T cells at later time points (Figure 3.10A). The transient nature of this phenotype is likely due to known autoregulatory mechanisms within Pcf11 and CstF77 in which overexpression of the full length isoform induces increased production of the shorter isoform (Kamieniarz-Gdula et al. 2019; R. Wang et al. 2019; Z. Pan et al. 2006). This transient regulation is also recapitulated on the mRNA level because a similar time-course followed by RT-qPCR revealed that there is stronger activation of the full-length isoform of CstF77, Pcf11, and Wdr33 at early time points than later time points, particularly for Wdr33 and Pcf11 (Figure 3.10B).

If CFIm regulates levels of Wdr33, CstF77, PAP, and Pcf11 by promoting usage of intronic polyA sites, when these IPA sites are deleted, CFIm depletion should no longer regulate polyA factor levels. To test this hypothesis, the IPA site of Pcf11, Wdr33, PAP, and CstF77 were knocked out using CRISPR Cas9. Two guide RNAs were designed approximately 100 nucleotides upstream and downstream of the IPA site cleavage site to delete the IPA site as well as flanking regions, as performed previously; in total, a region of approximately 300 nucleotides was removed (Kamieniarz-Gdula et al. 2019) (Figure 3.11A). The only exception was PAP as there were no appropriate guides for removal of a 300 nucleotide region with adequate specificity and efficiency; as a result, a region of 600 nucleotides was removed. Colonies were screened by genomic DNA PCR for IPA site knockout using primers that bind to genomic regions upstream and downstream of the region removed by genome editing (Figure 3.11B).

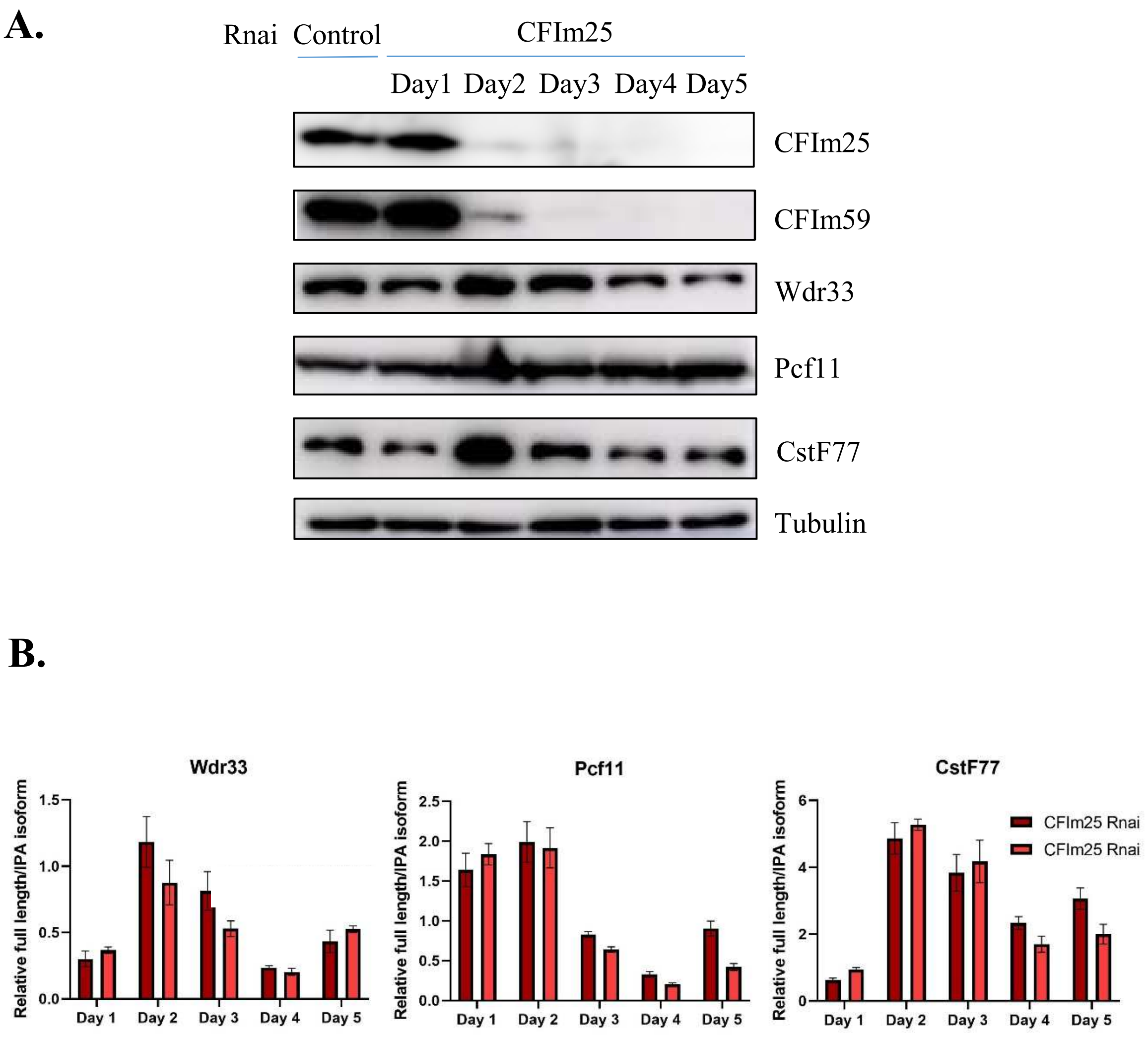


Figure 3.10 CFIm-Mediated Regulation of 3' Processing Machinery is Transient

A. Western blotting analysis of expression levels of Wdr33, Pcf11, and CstF77 from Days 1 to 5 post-CFIm25 knockdown.
B. qPCR analysis of the ratio between the full length and truncated mRNA isoform produced by IPA for Wdr33, Pcf11 and CstF77 from Days 1 to 5 post CFIm25 knockdown. All data is shown relative to control RNAi.

Following single colony selection, IPA site knockout cells were utilized for western blotting to analyze the effects of IPA site knockout on protein expression levels. It is predicted that knockout of an intronic polyadenylation site should increase the levels of both full-length mRNA and protein production as it prevents the utilization of the intronic site and therefore activates the downstream sites. Consistent with this hypothesis and previous reports, western blotting analysis revealed a 1.6-fold increase (standard deviation=0.25, n=3) in expression of Pcf11 upon knockout of the IPA site in 3 independent knockout cell lines (Kamieniarz-Gdula et al. 2019) (Figure 3.11C and 3.11D). However, there was no concomitant increase in other polyA factors including subunits of CPSF (CPSF100, CPSF73, and CPSF30), CstF (CstF50 and CstF77), and CFIm (CFIm25 and CFIm59), although two of the three Pcf11 IPA site knockout cell lines exhibited increased levels of Wdr33 and one of the three cell lines exhibited increased polyA polymerase levels (Figure 3.11C). Together, this evidence suggests that removal of the intronic polyA site of Pcf11 specifically promotes expression of the full-length isoform of Pcf11 and not a global increase in polyA factor levels, although some specific subunits may also display expression changes.

Similar to the increased expression of Pcf11 upon knockout of the IPA site, knockout of the IPA sites of Wdr33, CstF77 and PAP also increased expression of each factor, with average fold increases of 1.43 (standard deviation=0.65, n=10) for Wdr33, 1.48 (standard deviation=0.49, n=4) for CstF77, and 16.9 (standard deviation=10.3, n=5) for PAP (Figure 3.11D and 3.11E). Notably, knockout of the Wdr33 IPA site led to variable expression levels of Wdr33 with expression levels both higher and lower than wildtype 293T cells. This may be due to a compensatory mechanism or even an autoregulatory mechanism as other members of the core polyA machinery have previously been shown to autoregulate (R. Wang et al. 2019; Kamieniarz-

Gdula et al. 2019; Z. Pan et al. 2006), but this remains unknown and was not investigated further within this study. As a whole, however, this indicates that knockout of an intronic polyA site increases the levels of the full-length protein product.

Upon determining that deletion of an intronic polyA site controls expression levels of the protein, we next used IPA site knockout cells to test whether CFIm suppresses levels of other polyA factors by promoting intronic polyA sites. If this model is true, CFIm depletion should not increase protein levels of the specific polyA factor that has the deleted IPA site. However, deletion of this IPA site should not affect the regulation of other polyA factors; thus, CFIm depletion should increase the expression of other polyA factors with IPA sites to a similar extent as control 293T cells. Consistent with this hypothesis, western blotting analysis of a five-day time course post-CFIm25 knockdown revealed that control 293T cells exhibited increased expression of Pcf11 with highest levels on days 2, 3, and 4 post-knockdown. By contrast, knockout of the Pcf11 IPA site suppressed this phenotype, with Pcf11 levels remaining constant throughout the five-day time course. This indicates that CFIm25 regulates Pcf11 protein levels through its IPA site. Notably, however, both control 293T cells and Pcf11 IPA knockout cells experienced increased levels of PAP, CstF77 and Wdr33 at various points throughout the time course, indicating that the loss of regulation by CFIm is specific to the protein whose IPA site was knocked out (Figure 3.12). Together this data provides evidence that CFIm is a master regulator that controls expression of other polyA factors by regulating their intronic polyadenylation.

CFIm Regulates Global Gene Expression

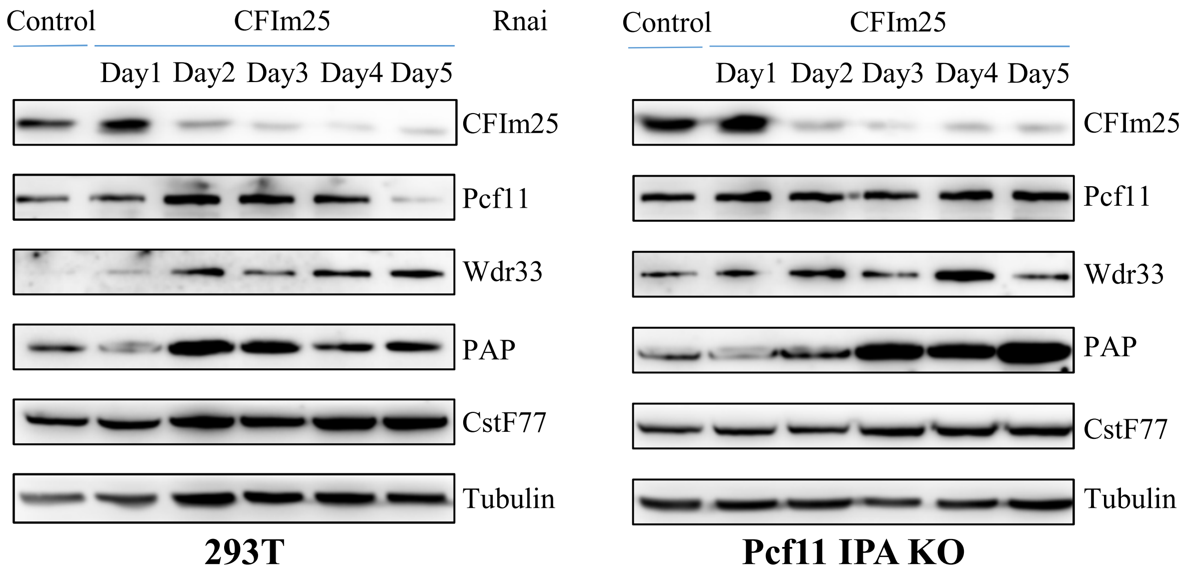


Figure 3.12 CFIm-Dependent Inhibition of Full Length Pcf11 Production is Mediated by IPA Site

Western blotting analysis of expression levels of Wdr33, Pcf11, and CstF77 from Days 1 to 5 post CFIm25 knockdown in control (left) and Pcf11 IPA Knockout 293T cells (right).

Our data indicates that CFIm regulates APA in at least two distinct ways. In over 2000 genes, CFIm promotes polyadenylation at distal polyA sites, leading to the production of mRNAs with long 3' untranslated regions (G. Martin et al. 2012; Masamha et al. 2014; Zhu et al. 2017; W. Li et al. 2015) (Figure 3.13A). This likely affects both mRNA stability and translation efficiency because 3' UTRs have a high frequency of binding sites for microRNA, which have previously been shown to destabilize mRNA and reduce translation efficiency (Sandberg et al. 2008; Hoffman et al. 2016; Fu et al. 2018). They also have a high frequency of destabilization elements that are bound by RNA binding proteins (Garneau, Wilusz, and Wilusz 2007). Because CFIm promotes production of mRNAs with longer 3'-UTRs, more microRNA and RBPs can bind, leading to an overall decrease in gene expression. In addition to promoting the production of mRNA isoforms with longer 3' UTRs, our data indicates that CFIm also promotes intronic polyadenylation, which results in mRNAs that are frequently degraded or produce truncated, non-functional protein (Figure 3.13A).

Although CFIm has two distinct roles in the regulation of alternative polyadenylation depending on the location of the polyA site, in both cases CFIm reduces gene expression. This global role in gene expression regulation is reminiscent of the role of myc in transcription regulation. Myc is a transcription factor that is highly overexpressed in many types of cancers as well as stem cells. One hypothesis for the role of myc in transcription regulation is that it is a gene expression amplifier that binds to the promoter of all active genes and upregulates transcription. As a result, when myc levels are high as in the case of stem cells or cancer, there is transcription amplification and RNA levels are globally higher (Nie et al. 2012; Bradner, Lee, and Young 2013). Because both of CFIm's roles in regulating alternative polyadenylation likely

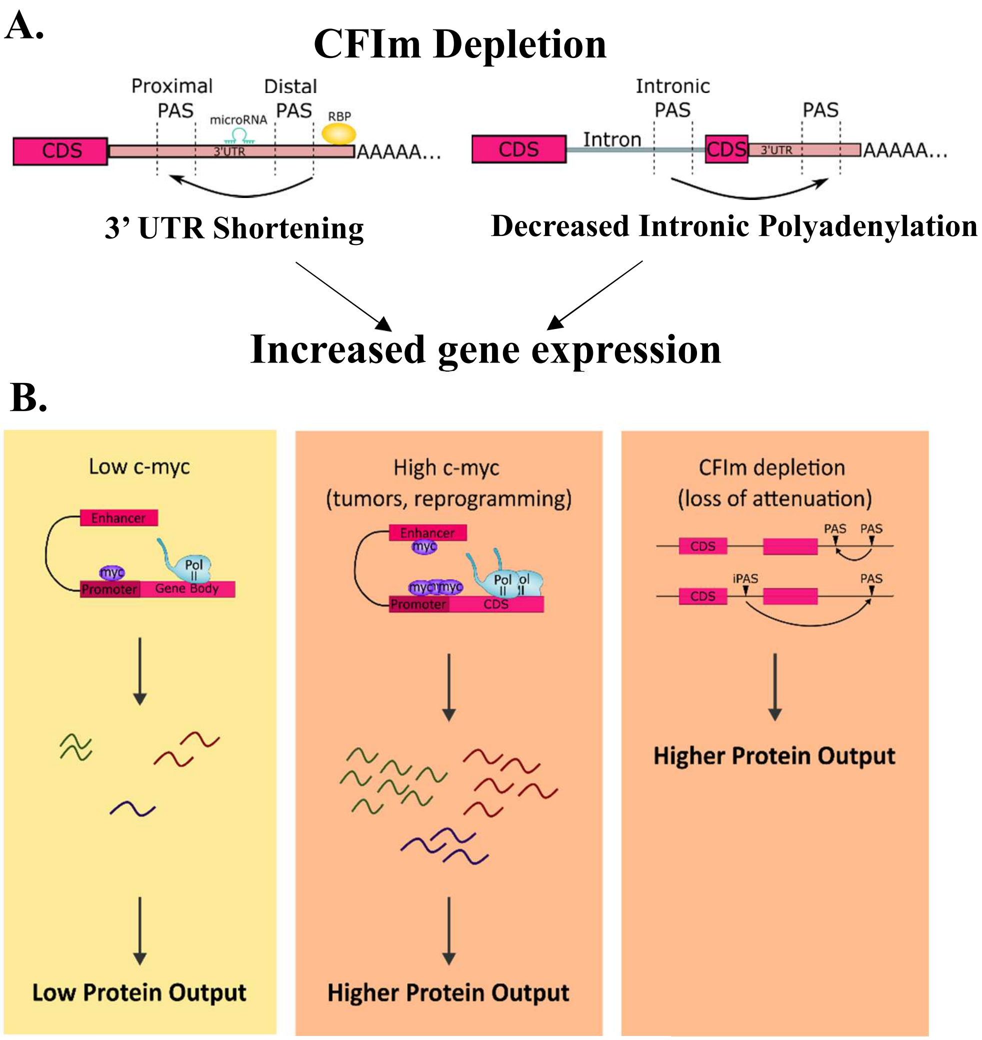


Image adapted from *Lin et al 2012*

Figure 3.13 CFIm Attenuates Gene Expression

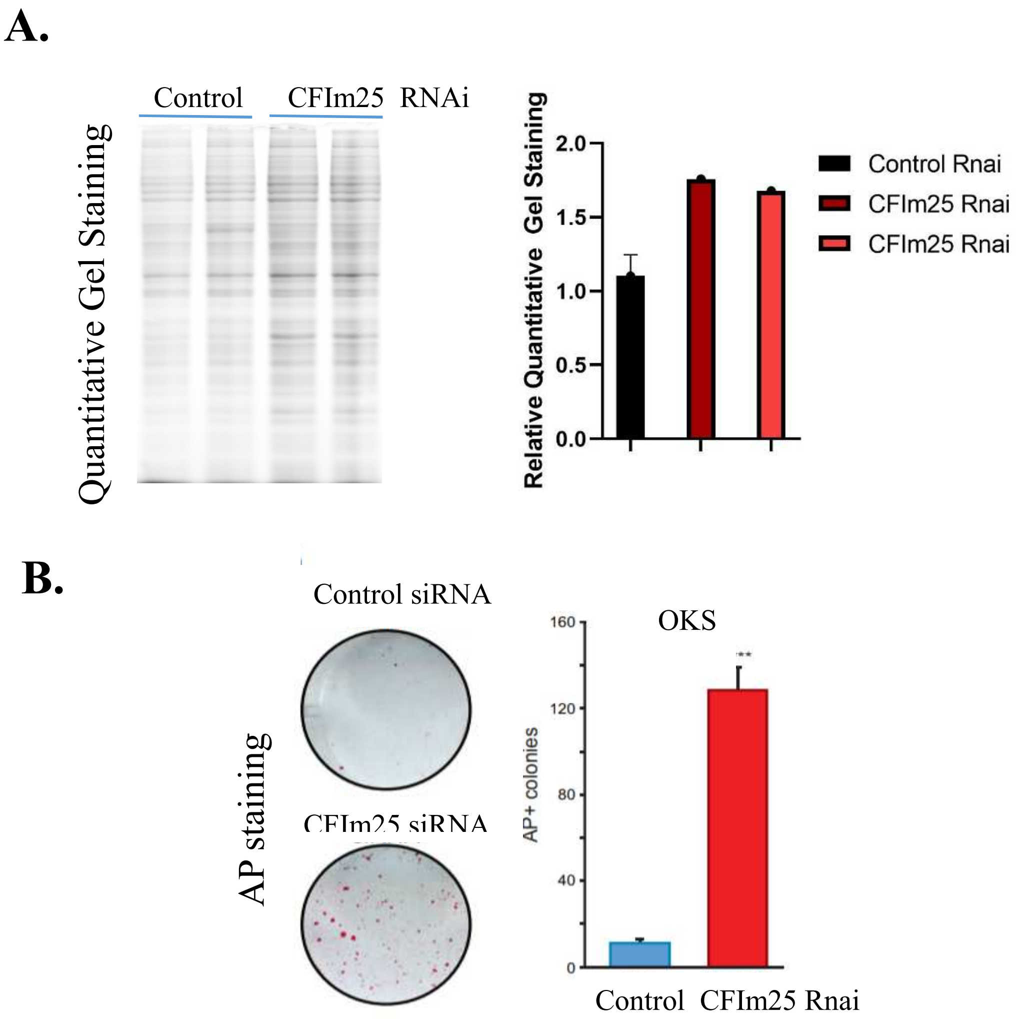
A. Model of effects of CFIm depletion on APA. CFIm promotes distal polyA sites. CFIm depletion causes a distal to proximal shift in polyA site usage. As 3' UTRs are enriched for binding sites of microRNA and RNA binding proteins, 3' UTR shortening may increase gene expression (left). CFIm also enhances intronic polyA sites so depletion of CFIm promotes full length gene expression (right). In both cases, CFIm depletion enhances gene expression, suggesting that CFIm is a gene expression attenuator.

B. Model comparing effects of overexpressing the transcription factor c-myc and depleting CFIm on overall gene expression levels. Myc is hypothesized to be a gene expression amplifier that binds to the promoter of all active genes and upregulates transcription (left). When myc is highly overexpressed as in stem cells or cancer, there is transcription amplification and RNA levels are globally higher (middle). Conversely, CFIm is an attenuator and CFIm depletion enhances gene expression (right).
PAS: polyA site iPAS: intronic polyA site CDS: coding

downregulate protein output, CFIm may have the opposite role as myc in regulating global gene expression (Figure 3.13B).

To test the role of CFIm in regulating global gene expression, CFIm25 was knocked down in 293T cells and global protein levels were analyzed by Cy5.5 quantitative gel staining. Interestingly, there was a 1.5-fold increase in global protein levels upon knockdown of CFIm25, indicating that CFIm attenuates the expression of a large number of genes (Figure 3.14A). It is also noteworthy that there is no single protein band that accounts for a substantial fraction of the increase in protein levels; instead, there are many bands that all show moderate increases, consistent with a global protein level increase. While this data is still preliminary, it indicates that CFIm is an anti-myc that attenuates global gene expression.

If CFIm has the opposite role as myc during reprogramming, we hypothesized that CFIm depletion may substitute for Myc overexpression during reprogramming. It has previously been shown that while Myc is not necessary for reprogramming, it strongly increases reprogramming efficiency because there are significantly more reprogrammed cells when Myc is included. This is unique because other reprogramming factors such as Oct4, Klf4 and Sox2 are essential and reprogramming cannot occur without their overexpression (Nakagawa et al. 2008). To test whether CFIm depletion may substitute for Myc overexpression, in collaboration with the Hochedlinger lab, MEFs were reprogrammed with Oct4, Klf4, Sox2 and either control or CFIm depletion. Strikingly, CFIm25 knockdown significantly enhanced reprogramming efficiency when myc is omitted, suggesting that CFIm depletion indeed is an alternative to transcriptional amplification by Myc during reprogramming (Figure 3.14C). Notably, CFIm depletion in combination with Myc overexpression had a greater effect on reprogramming than CFIm depletion alone (Figure 3.2 B and Figure 3.14C), indicating that CFIm and Myc may not act on



Adapted from Brumbaugh et al (2018) Cell

Figure 3.14 CFIm Depletion Substitutes for Myc Overexpression During Reprogramming

A. Quantitative Cy5.5 gel staining analyzing global protein amounts of control and CFIm25 knockdown protein samples for equivalent numbers of cells.
B. Reprogramming of mouse embryonic fibroblast into induced pluripotent stem cells in the presence of viral overexpression of Oct4, Klf4, Sox2 and c-Myc (OKSM) or lacking a single factor.
   OKS: Oct4, Klf4, Sox2
C. Reprogramming of mouse embryonic fibroblasts into induced pluripotent stem cells in the presence of OKS but absence of myc. Reprogramming is performed either in the presence or absence of CFIm.

the exact same subset of genes and that their combination activates a greater number of genes than CFIm or c-Myc alone (Justin Brumbaugh et al. 2018). However, this data provides strong evidence that CFIm attenuate gene expression and, as such, may substitute for c-Myc overexpression during reprogramming.

3.4 DISCUSSION

In this chapter, a genome-wide approach was utilized to investigate the roles of CFIm in alternative polyadenylation regulation and cell fate determination. Specifically, CFIm25 was knocked down in human 293T cells and used for by RNA and PAS-sequencing. Consistent with previous findings, CFIm depletion led to a distal to proximal polyA site shift in 3' UTRs. This study also made the novel finding that in addition to regulating 3' UTR APA, CFIm also promotes intronic polyadenylation in many genes including members of the core polyA machinery. We also observed a new role for CFIm in attenuating global gene expression, which may have important implications for reprogramming.

Many of the genes that intronic polyA sites regulated by CFIm were in fact other polyA factors including Wdr33 (CPSF), CstF77 (CstF), Pcf11 (CFIIm), and polyA polymerase (PAP), which is interesting on multiple levels. First, our study determined that promotion of full length polyA factor production induced by knockdown of CFIm25 is transient as production peaked at early time points then fell back to levels similar to control. Consistently, Pcf11 and CstF77 have known autoregulatory mechanisms through which overexpression of the full-length isoform increases production of the shorter isoform (Kamieniarz-Gdula et al. 2019; Z. Pan et al. 2006; R. Wang et al. 2019). This finding highlights the importance of regulating the levels of the core polyA machinery within the cell, which has interesting implications for cancer biology. In recent years, multiple studies on another RNA processing event, splicing, have indicated that Myc-driven cancers are particularly sensitive to inhibition of the spliceosome because oncogenic Myc-activation enhances production of both RNA and protein. Because mRNA levels are high, there is increased burden on the spliceosome to process RNA and genetic and pharmacological inhibition of the spliceosome impairs survival and metastasis of several Myc-driven cancer types (Hsu et al. 2015; Suda et al. 2017). As polyadenylation is also an essential RNA processing event, 3' processing, like splicing, may also be a rate-limiting step in cell growth and division, making it an important area of cancer research. Strikingly, our finding that CFIm promotes production of truncated polyA factors in combination with known autoregulatory mechanisms for core polyA factors indicates that intronic polyadenylation within 3' processing components may play a critical role in preventing aberrant cell growth. In addition, as CFIm regulates one subunit of each of the core polyadenylation factors, CFIm appears to be a master regulator of the core polyadenylation machinery and therefore may limit overall levels of 3' processing. As a result, CFIm may be an important therapeutic target for cancer treatment.

Secondly, the role of CFIm in regulating polyA factors indicates that the current model for CFIm-mediated regulation of APA within the terminal exon should be modified. The prevailing model is that CFIm binds to the UGUA motif, which is enriched upstream of distal polyA sites. When CFIm binds, it can recruit the rest of the core polyA machinery to nearby polyA sites through interactions of its RS domain with the RE/D domain of Fip1, a subunit of CPSF. As a result, distal polyA sites are preferentially used when CFIm levels are high but there is a shift to upstream polyA sites and therefore 3'UTR shortening when CFIm levels are low (Zhu et al. 2017). However, our study also found that CFIm regulates intronic polyadenylation within one member of each subunit of the core polyA machinery, indicating that CFIm is a master regulator of mRNA 3' end processing (Figure 3.15A). When CFIm25 are high, it causes intronic polyadenylation in other members of the polyA machinery and limited production of the full-length, functional protein. Because one member of each other component of the polyadenylation machinery is affected, there is limited levels of the core machinery available to perform 3' processing. As a result, there is increased dependency on CFIm to enhance polyadenylation site selection by recruiting CPSF and distal polyA sites are preferentially used because the UGUA recognition motif is enriched near distal polyA sites. However, when CFIm levels are low as in the case of CFIm25 knockdown, there is decreased intronic polyadenylation within Wdr33, CstF77, and Pcf11, so more polyadenylation machinery is available. When this occurs, the proximal site has an advantage not only because it is transcribed first but also because there is higher levels of the core polyA machinery (Figure 3.15B).

In addition to demonstrating the critical role that CFIm plays in intronic polyadenylation, we also demonstrate that the two CFIm complexes may in fact have distinct but overlapping functions. Consistent with previous studies, we determined that knockdown of CFIm68 had a greater effect on APA than knockdown of CFIm59, with over 2000 in comparison to 128 APA changes respectively (Zhu et al. 2017). Strikingly, however, while over 80% of APA changes regulated by CFIm68 were within the terminal exon, the opposite was true for knockdown of CFIm59: over 60% of APA changes occurred at intronic sites. This data indicates that although some APA events are shared, CFIm59 and CFIm68 may in fact also play more distinct roles. It will be interesting to further characterize the two CFIm complexes and critically analyze the roles they play in APA.

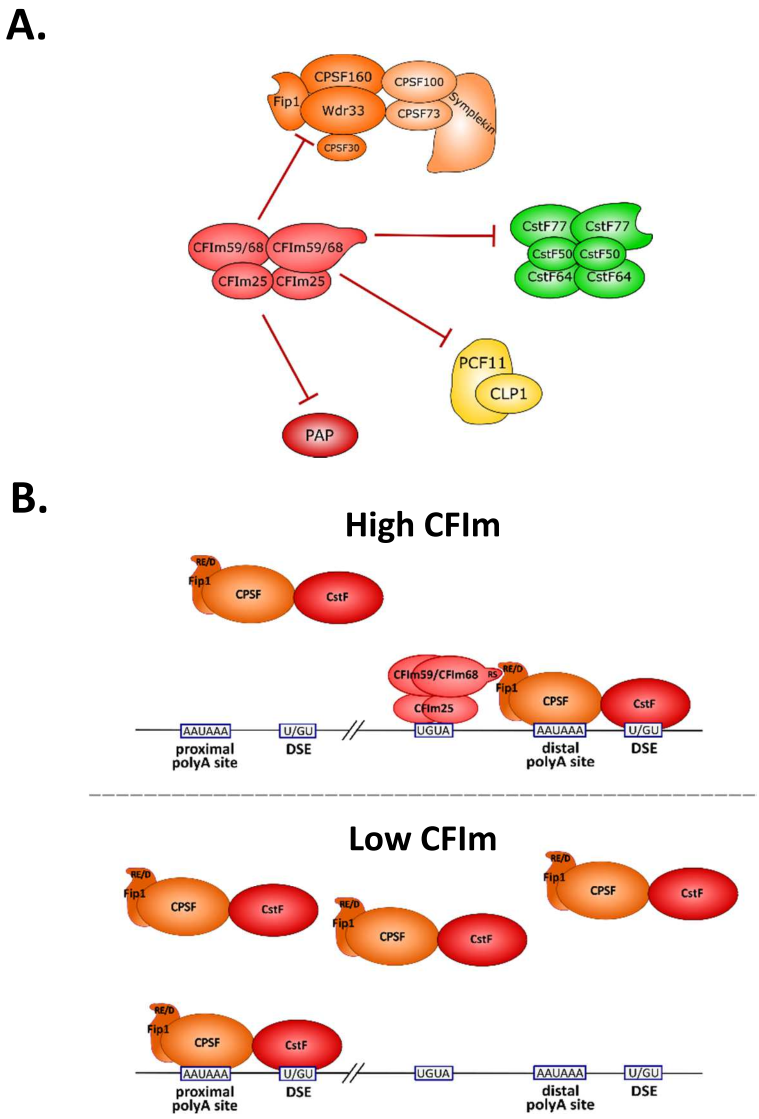


Figure 3.15 CFIm is a Master Regulator of APA

A. Schematic of CFIm-mediated regulation of core 3' processing machinery. CFIm limits levels of one of each member of the core 3' processing machinery: Wdr33 of CPSF, CstF77 of CstF, Pcf11 of CFIIm, and polyA polymerase.

B. Model comparing 3' UTR APA in high and low levels of CFIm. When CFIm levels are high, it promotes intronic polyadenylation within one subunit of each of the core polyA machinery complexes. This further promotes longer 3' UTRs as it makes the 3' end machinery more dependent on the enhancement of distal polyA sites by CFIm. However, when CFIm levels are low, there is high expression of the core polyA machinery, promoting binding at the proximal polyA sites which are transcribed first.

Finally, our study also suggests that CFIm may regulate cell fate by acting as a gene expression attenuator. Previously, we and our collaborators demonstrated that knockdown of CFIm25 led to a 30-fold increase in reprogramming of mouse embryonic fibroblasts into induced pluripotent stem cells. Although there were many APA changes, no single APA change could account for the observed change in reprogramming efficiency, indicating that it may be the additive result of many APA changes (Justin Brumbaugh et al. 2018). Our current study shows that CFIm depletion impacts APA in at least two distinct manners: enhancing distal polyA sites and promoting intronic polyA sites. Interestingly, both mechanisms have a similar end result: reducing gene expression. Genes with longer 3' UTRs have increased binding sites for microRNAs and also have decreased translation efficiency (Hoffman et al. 2016; Fu et al. 2018; Sandberg et al. 2008). Similarly, intronic polyadenylation leads to production of truncated mRNAs that are oftentimes not translated and if translated, produce truncated proteins. Consistent with this hypothesis, we observed that knockdown of CFIm25 both increases global gene expression and also can substitute for myc overexpression during reprogramming (Figure 3.15) (Justin Brumbaugh et al. 2018).

Future experiments will need to be performed to validate the role of CFIm as a gene expression attenuator. While our current data indicates that there are increased global protein levels upon knockdown of CFIm25, the exact nature of these proteins is unknown. One way to identify the proteins that have increased expression upon knockdown of CFIm25 would be through polysome profiling. Polysome profiling is a technique that infers translation efficiency of mRNAs. Several ribosomes can be associated with and translate a single mRNA because a particular ribosome only associates with about 20-30 nucleotides. As a result, the mass of the mRNP complex increases as a function of the number of bound ribosomes, making mRNAs that are translated more efficiently heavier than those that are translated less efficiently. During polysome profiling, ribosomes that are actively translating mRNA are stalled with cycloheximide, a drug that interferes with ribosome translocation. mRNP complexes are then run on a sucrose gradient, in which heavier polysomes sediment further in the gradient than less-translated, lighter monosomes. mRNP fractions are collected and RNA is sequenced, allowing for identification of mRNAs with high translation efficiency (Chassé et al. 2017).

If our hypothesis that CFIm is a global gene expression attenuator is correct, upon CFIm depletion, there should be increased translation of CFIm target genes. Because we hypothesize that CFIm depletion should increase translation, we predict to see a global increase in the number of mRNAs within the polysome fraction in comparison to wildtype cells. In addition, upon RNA sequencing, we predict that there will be overlap in genes showing increased translation with genes that showed either decreased intronic polyadenylation or distal to proximal shifts in the terminal exon upon CFIm25 knockdown as identified by PAS-Seq. This data will provide us with further insight the exact nature of CFIm regulation of gene expression as well as the mRNAs with the greatest impact on cell fate determination.

In conclusion, this chapter identified a novel role for CFIm in promoting intronic polyadenylation in many genes including members of the core polyA machinery. It provides evidence that CFIm links gene expression regulation with cell fate determination by attenuating gene expression. Our study highlights that although CFIm is not an essential member of the polyA machinery, it plays a critical role in promoting diverse 3’ processing events with important biological impacts.

3.5 METHODS

Knockdown of CFIm

CFIm25, CFIm59 and CFIm68 knockdown cells were generated from HEK 293T cells using lentiviral transduction of the pLKO.1 vector containing shRNA template followed by 1.25mg/mL puromycin. Cells were harvested for assays 5 days post-transduction unless otherwise noted within experimental design.

shRNA sequences: CFIm25: 5’ GAACCTCCTCAGTATCCATAT 3’
CFIm25: 5’ TGTACCCTCTTACCAATTATA 3’
CFIm59: 5’ AAGATATCATGAAGCGAAACA 3’
CFIm68: 5’ GGTGATTATGGGAGTGCTATT 3’
CFIm68: 5’ GTTGTAACTCCATGCAATAAA 3’

Sequencing and Analysis

RNA Sequencing

CFIm25 was knocked down using above protocol. RNA was extracted from cells using TRIzol (Invitrogen) and purified using standard protocol. Paired end libraries were prepared by UCI Genomic High-Throughput Facility (GHTF) using Illumina TruSeq Stranded Kit and sequenced using HiSeq4000. Samples were split between 2 lanes.

DaPars Analysis

RNA sequencing libraries were aligned to the human genome using STAR (Dobin et al. 2013) and analyzed for alternative polyadenylation using DaPars (Dynamic analyses of Alternative PolyAdenylation RNA-Seq) (Xia et al. 2014). DaPars is a bioinformatic algorithm that identifies APA events from RNA sequencing. RNA-seq data from all input samples were merged to show combined coverage. Next, DaPars identified the distal polyA site based upon continuous RNA-seq signal by identifying the region where coverage is less than 5% of coverage at the preceding exon. It then used linear regression to infer de novo proximal polyA sites that best explain localized read density change. Expression values of the two transcripts in different conditions were estimated and APA changes were quantified as the change in Percentage of Distal polyA Site Usage Index (ΔPDUI). APA events with FDR < 5%, PDUI with no greater mean difference than 0.2 within replicates, and mean fold change in PDUI greater than or equal to 1.5 were identified as significant. To avoid false positives, only genes with >30-fold mean coverage were included.

UGUA Enrichment Analysis

For each CFIm regulated IPA site, regions from 100 nucleotides upstream to 100 nucleotides downstream of the cleavage site were selected and the enrichment of the UGUA motif was calculated using the Bioconductor packages Biostrings, which performs pattern matching and pair-wise comparisons, and Seqpattern, which visualizes oligonucleotide patterns and motifs across large sets of sequences. The enrichment of UGUA within the same region at polyA sites within the 3' UTR was also calculated.

PAS-Seq Library Preparation and Analysis

CFIm25, CFIm59, and CFIm68 were knocked down using above protocol. PAS-Seq libraries were prepared according to PAS-Seq 2 protocol (Yoon, Soles, and Shi 2021). 1ug of total RNA was diluted in 50ul of H_2O and polyadenylated RNA was purified using NEBNext PolyA mRNA magnetic isolation module (NEB) following manufacturers protocol. Next, mRNA was fragmented at 70°C for 5 minutes with RNA Fragmentation Reagent (Invitrogen) and stopped with Stop solution. RNA was then ethanol precipitated precipitated with sodium acetate at -80°C. Pellets were resuspended in H_2O and reverse transcribed. For reverse transcription, oligo(dT) primer was bound at 72°C followed by reverse transcription with Superscript III and template switch oligo. cDNA was purified using AMPure XP beads (Beckman-Coulter) per manufacturer's protocol and amplified with 2x Phusion (NEB) with 10uM TruSeq universal adapter and 10uM TruSeq indexed adapter. Control knockdown was amplified using indexed adapter 22-24, CFIm25 knockdown was amplified using indexed adapter 13-15, CFIm59 knockdown was amplified using indexed adapter 19-21, and CFIm68 knockdown was amplified using indexed adapter 16-18. Following amplification, PCR products were run on a 2.5% low melt agarose gel (80V, 2.5hrs) and a 185-225 base pair band was gel extracted and submitted for sequencing at the 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.

Template Switching Oligo: CTACACGACGCTCTTCCGATCTCATrGrG+G
oligo(dT) primer:
TGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTTTTTTTTTTTTTTTTTV (V: A/C/G)
Indexed adapter 13:
CAAGCAGAAGACGGCATACGAGATTTGACTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC
Indexed adapter 14:
CAAGCAGAAGACGGCATACGAGATGGAACTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC
Indexed adapter 15:
CAAGCAGAAGACGGCATACGAGATTGACATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC
Indexed adapter 16:
CAAGCAGAAGACGGCATACGAGATGGACGGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC
Indexed adapter 17:
CAAGCAGAAGACGGCATACGAGATCTCTACGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC

Indexed adapter 18:
CAAGCAGAAGACGGCATACGAGATGCGGACGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC
Indexed adapter 19:
CAAGCAGAAGACGGCATACGAGATTTTCACGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC
Indexed adapter 20:
CAAGCAGAAGACGGCATACGAGATGGCCACGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC
Indexed adapter 21:
CAAGCAGAAGACGGCATACGAGATCGAAACGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC
Indexed adapter 22:
CAAGCAGAAGACGGCATACGAGATCGTACGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC
Indexed adapter 23:
CAAGCAGAAGACGGCATACGAGATCCACTCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC
Indexed adapter 24:
CAAGCAGAAGACGGCATACGAGATGCTACCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC

Following sequencing, data was analyzed as followed. Sequences were trimmed to remove the first six nucleotides of each read (which come from the template switch oligo) and consecutive As (which come from the oligo(dT)) with the program Cutadapt (M. Martin 2011). Trimmed sequences were then aligned to the reference genome using STAR (Dobin et al. 2013). Next, internal priming events were removed by identifying reads mapped to genomic sequences immediately following 6-consecutive As or 7 As out of 10 nucleotides which are a result of internal priming. 3' ends of each read were compared to a list of annotated polyA site based upon PolyA_DB (H. Zhang et al. 2005) and a count table was generated. Reads were assigned to a polyA site if they mapped within 40 nucleotides upstream or downstream. APA analysis was performed with edgeR using the "exon" mode and FDR scores were assigned to each polyA site.

To be considered APA, a gene had to have 1) a least 1 PAS reach FDR<0.05 and 2) greater than 15 percent change in usage between conditions.

3' RACE Analysis of APA

CFIm25 was knocked down in 293T cells using above protocol. The RNA was extracted from cells using TRIzol (Invitrogen) and purified using standard protocol. cDNA was then synthesized by reverse transcription using oligo(dT) and linker primer. Following cDNA synthesis, polyA site usage was analyzed in two rounds of PCR. In the first round of PCR, equivalent concentrations of primers Q0 (a reverse linker primer that recognizes all reverse transcribed cDNA) and two forward primers that bind and amplify either a common region of the VMA21 3' UTR or a region specific to the distal 3' UTR of mRNAs cleaved and polyadenylated at the distal site are added to the PCR reaction mi2. cDNA was amplified for 20 cycles to avoid saturation. One microliter of first round PCR was added to the second round of PCR with equivalent concentrations of Q1, (a reverse primer that binds within the PCR amplified in PCR1) and two competing forward primers that bind and amplify either a common region or the VMA21 3' UTR or a region specific to the distal 3' UTR of mRNAs cleaved and polyadenylated at the distal site. As before, cDNA was amplified for 20 cycles to avoid saturation. Following the second round of amplification, reactions were run on a 1% agarose gel. The proportion of distal to proximal site usage was calculated ($\frac{distal}{proximal}$) and compared to pLKO control to determine fold change upon CFIm25 knockdown.

3' RACE RT Primer:
5' CCAGTGAGCAGAGTGACGAGGACTCGAGCTCAAGCTTTTTTTTTTTTTTTTT 3'
3' RACE PCR1 R: CCAGTGAGCAGAGTGACG
3' RACE PCR1 R: GAGGACTCGAGCTCAAGC

VMA21proximal F 3’ RACE PCR1: GCTGCTATTGTTGCAGTGGTC
VMA21proximal F 3’ RACE PCR2: CGTCCATGTGGTGCTGG
VMA21distal F 3’ RACE PCR1: CTGGCCACACTTTCCTTGC
VMA21distal F 3’ RACE PCR2: CCCCACGTTCAGTGTAATCTC

qRT-PCR Analysis of IPA Site Usage of PolyA Factors

CFIm25 was knocked down in 293T cells using above protocol. RNA was extracted from cells using TRIzol (Invitrogen) and purified using standard protocol. Following treatment with RQ1 DNAse (Promega) following manufacturers protocol, cDNA was synthesized using All-In-One 5X RT MasterMix (Abmgood) following manufacturer’s protocol. Levels of intronic polyadenylation was compared between control and CFIm25 depletion cells using qRT-PCR using PowerUp SYBR Green qPCR Master Mix (Applied Biosystems). qPCR primers were designed to amplify either the 3’ UTR of the full-length isoform (extended) or a common region in the exon immediately upstream of the CFIm-regulated IPA site (common). Full length polyA factor production was calculated as the ratio of $\frac{common}{extended}$ and compared to a pLKO control to determine fold change.

Wdr33 common F: GTGAGGGCCATGACGTG
Wdr33 common R: CGCCTCCTTATGTGCCTG
Wdr33 extended F: GGAGGCTCAAGATTGCTTTAGG
Wdr33: extended R: GGTCTCTGGATCCGGTTTTAGC
Pcf11 common F: CAGTCATCGCTCGAAGACC
Pcf11 common R: CGGTTTGGGCCTCGATG
Pcf11 extended F: TTCCGAGAGAGCACCGTAGG
Pcf11 extended R: CTGGATGGTTCTCTATACCTGGAG
CstF77 common F: GCAGCTGAGTATGTCCCAGAG
CstF77 common R: TGCCTCTCGAATGAGAATGC
CstF77 extended F: GTGCCTCTATCACATGGTTCTT
CstF77 extended R: CTGCCACTTTGTACTGTTCTCA

Generation of Wdr33, Pcf11, CstF77, PAP, and Rbbp6 IPA Site Knockout Cell Lines

sgRNA were designed using an online tool (https://chopchop.cbu.uib.no/). Paired sgRNA were designed to flank the IPA site of Wdr33, Pcf11, CstF77, PAP, and Rbbp6 and remove the intervening region as described previously (R. Wang et al. 2019; Kamieniarz-Gdula et al. 2019). sgRNA for Pcf11 were designed from Kamieniarz-Gdula 2019; all other guides were designed specifically for this project using a similar method (Kamieniarz-Gdula et al. 2019). In all cases except PAP, each sgRNA was designed to target a region approximately 100-200nt upstream or downstream of the IPA site to remove a region of approximately 300 nucleotides total. For PAP, no sgRNA with high specificity and efficiency were present within the above parameters so they sgRNA were designed to remove a region approximately 600 nucleotides in total. sgRNA were annealed and cloned into pX330A using BbsI digestion. 1ug of each pX330A-sgRNA were transfected into a 6-well plate of HEK 293T cells using Lipofectamine 3000 (Invitrogen). 3 days post-transfection, cells were reseeded by limiting diluting into 96 well plates at a density of 0.5 cells/well. When colonies formed, they were screened by genomic DNA PCR with primers flanking region targeted for removal. Homozygous colonies were identified as those with a single band approximately 300 nucleotides in length shorter than the wildtype 293T cells (or 600 nucleotides in the case of PAP).

Pcf11 upstream sgRNA F: CACCACCGTCTCTAAACAATATAT
Pcf11 upstream sgRNA R: AAACATATATTGTTTAGAGACGGT
Pcf11 downstream sgRNA F: CACCACAAGATACACGGTTTCAGG
Pcf11 downstream sgRNA R: AAACCCTGAAACCGTGTATCTTGT
Pcf11 IPA KO gDNA colony screen F:
TCCCTGATAGCGAAGGAGTGTGTTGAGTATGACGAATGCTTCC
Pcf11 IPA KO gDNA colony screen R: TGTTGAGTATGACGAATGCTTCC

Wdr33 upstream sgRNA F: CACCTGAATTAACCGACAAGATAG
Wdr33 upstream sgRNA R: AAACCTATCTTGTCGGTTAATTCA
Wdr33 downstream sgRNA F: CCACGAGTTACGTGAATTCAGTGG

Wdr33 downstream sgRNA R: AAACCACTGAATTCACGTAACTC
Wdr33 IPA KO gDNA colony screen F: TCACAGTCATTTTGTAGGTTTTTAC
Wdr33 IPA KO gDNA colony screen R: AGGGTACGGAGTAGAAGGTAC

CstF77 upstream sgRNA F: CCACCATAGCCAATTAGGACAAGG
CstF77 upstream sgRNA R: AAACCCTTGTCCTAATTGGCTATG
CstF77 downstream sgRNA F: CCACTAAATCCTCCTTGTCCTAAT
CstF77 downstream sgRNA R: AAACATTAGGACAAGGAGGATTA
CstF77 IPA KO gDNA colony screen F: ACGGAAGACTTATGAACGCCT
CstF77 IPA KO gDNA colony screen R: AGTGCTTACATTGATAAACCATGC

PAP upstream sgRNA F: CCACTAAGTGCTTCGGGACAAAAA
PAP upstream sgRNA R: AAACTTTTTGTCCCGAAGCACTTA
PAP downstream sgRNA F: CCACTTGCCTTGGCCCAAAAGGTG
PAP downstream sgRNA R: AAACCACCTTTTGGGCCAAGGCAA
PAP IPA KO gDNA colony screen F: GATAGGTTGGGAAATCAGTTAC
PAP IPA KO gDNA colony screen R: TGGTGGCATAACCACATTTTC

Rbbp6 upstream sgRNA F: CCACTGTACAACACAGTGTCATAC
Rbbp6 upstream sgRNA R: AAACGTATGACACTGTGTTGTACA
Rbbp6 downstream sgRNA F: CCACTCATAGGGGGCCCCAGGTTC
Rbbp6 downstream sgRNA R: AAACGAACCTGGGGCCCCTATGA
Rbbp6 IPA KO gDNA colony screen F: CACATTGCTTTTACCTTTATAATGTAG
Rbbp6 IPA KO gDNA colony screen R: AGGTACAGCTTTCAATGGCC

Quantitative Gel Staining

Cells were resuspended to 1x10^6 cells per 100ul in RIPA buffer, vortexed and incubated on ice for 30 minutes. Following incubation, lysates were centrifuged at 14krpm for 15 minutes at 4°C and the supernatant was transferred to a new tube for further analysis. Lysates were labeled for quantitative staining with Amersham QuickStain (GE Healthcare). Samples were diluted 1:10 in labeling buffer (Tris-HCl labeling buffer with 0.1% SDS, pH 8.7 at 25°C) and 1ul Cy5 was added. Following a 30 minute incubation at room temperature, samples were mixed with 3x SDS loading dye containing lysine to quench the reaction as well as DTT and beta-mercaptoethanol for denaturation. Samples were heated at 95°C for 3 minutes and run on a 10%

polyacrylamide gel at 70V for 30 minutes followed by 120V. Gels were imaged on an Amersham Typhoon Biomolecular Imager (GE Healthcare).

CHAPTER 4

CFIm PLAYS A ROLE IN 3’ SPLICE SITE SELECTION AND REGULATES ALTERNATIVE SPLICING

4.1 SUMMARY

In addition to polyadenylation, splicing is another critical RNA processing event. One of the first steps in the assembly of the spliceosome is the recognition of the 3’ splice site by the U2 auxiliary factor (U2AF), which binds to the polypyrimidine tract sequence upstream of the 3’ splice site and, in some cases, the 3’ splice site itself. However, sequences preceding the 3’ splice sites are highly variable, and it remains unclear how U2AF can recognize such a wide variety of sequences.

In this chapter, we report the surprising finding that the polyadenylation factor cleavage factor I (CFIm) is responsible for recognizing a subset of 3’ splice sites and regulating thousands of alternative splicing events. Mechanistically, CFIm forms a complex with U2AF65 via an interaction between the RS domains of the CFIm large subunits (CFIm68 and CFIm59) and that of U2AF65. Such a CFIm-U2AF complex recognizes a subset of noncanonical 3’ splice sites, which lack polypyrimidine tracts, but have UC-rich sequences instead. Additionally, CFIm competes with other splicing factors for binding to U2AF and indirectly regulates additional alternative splicing events. Based on these results, we propose a model in which U2AF65 binds

to a number of RNA-binding proteins, including CFIm, to recognize the highly variable 3' splice site regions.

4.2 INTRODUCTION

Splicing, or the excision of noncoding introns and joining together of protein coding exons, is an essential step in the regulation of gene expression. It has been estimated that approximately 25% of the human genome consists of introns, which is 4-5 times that encoded exons, highlighting the necessity of the removal of these sequences (Jo and Choi 2015). Additionally, 92-94% of genes have alternatively spliced isoforms, with approximately 86% of genes having an alternative isoform with an expression of 15% or more. While the impact of alternatively spliced mRNA isoforms on proteomic diversity has historically been debated, recent studies indicate that splicing in fact makes significant contributions to proteomic diversity in humans (Liu et al. 2017; X. Wang et al. 2018). As a result, it is critical to study both RNA splicing in general and, more specifically, the mechanisms involved in alternative splicing.

Alternatively spliced exons frequently have weaker 5' and 3' splice sites than constitutive exons and are less efficiently recognized by the spliceosome (Baek and Green 2005; Garg and Green 2007; Zheng, Xiang-Dong, and Gribskov 2005). 3' splice sites consist of 3 elements: the YAG sequence directly preceding the intron-exon boundary, the branch point sequence with the consensus sequence YNYURAY (where Y represents a pyrimidine [C or U]), and the polypyrimidine tract characterized by a high frequency of pyrimidines (C or U) (Black 2003; Wilkinson, Charenton, and Nagai 2020; Moore, Query, and Sharp 1993). As the polypyrimidine tract is the most variable of these sequences, the primary determinant of 3' splice site strength is the ability of the U2 auxiliary factor (U2AF) to recognize the polypyrimidine tract. U2AF is a heterodimer of U2AF35, which recognizes the AG dinucleotide at the intron-exon boundary of the 3’ splice site, and U2AF65, which directly interacts with the polypyrimidine tract (Figure 2.3B). Factors that influence polypyrimidine tract recognition by U2AF65 include length and composition, with strong 3’ splice sites being defined by longer, contiguous polypyrimidine tracts with a high enrichment of uridine (Coolidge, Seely, and Patton 1997; Hertel 2008). Despite the importance of the polypyrimidine tract, its sequence is in fact highly variable. In vitro assays suggest that when polypyrimidine tracts are short and/or have a higher frequency of purine residues, the AG dinucleotide is essential for splicing but when the polypyrimidine tract is longer, the AG nucleotide is no longer required (Reed 1989; Coolidge, Seely, and Patton 1997). In addition, U2AF can only directly regulate approximately 88% of 3’ splice sites, indicating that there may be additional mechanisms involved in 3’ splice site selection (Shao et al. 2014). Remaining questions include how U2AF recognizes such a wide variety of polypyrimidine tract sequences, whether U2AF35 is still important for splice site recognition if the polypyrimidine tract is strong, and whether there are additional mechanisms involved in 3’ splice site selection of exons with weaker polypyrimidine tracts, particularly those that cannot be directly bound by U2AF65.

One area of particular interest in understanding 3’ splice site selection is cross-regulation between splicing and polyadenylation. Components of the spliceosome, other splicing factors including as SR proteins and U2AF, and polyadenylation factors such as CPSF all associate with the RNA Polymerase II C-terminal domain (CTD), suggesting that both processes are co-transcriptional (Pandya-Jones and Black 2009; Listerman, Sapra, and Neugebauer 2006; Marvin and Inada 2018). As both processes are co-transcriptional, it seems likely that they may also be

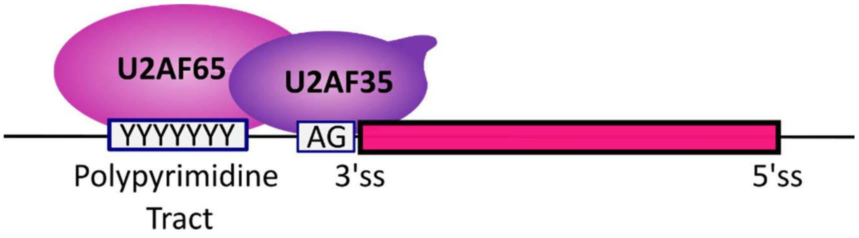


Figure 4.1 Recognition of the 3’ Splice Site by U2AF

Schematic depicting 3’ splice site recognition by the U2 auxiliary factor (U2AF). U2AF is a dimer of U2AF65, which recognizes the polypyrimidine tract, and U2AF35, which recognizes the AG dinucleotide at the intron/exon boundary.

able to cross-regulate as their machineries are likely closely associated. Previous studies also indicate that polyadenylation factors and splicing factors mutually regulate terminal exons. The exon definition model of splicing states that splicing occurs when U1 recognizes the downstream 5' splice site of an exon and U2AF recognizes the upstream 3' splice site. However, first and terminal exons each lack one of these features and terminal exons are regulated by both the splicing and polyadenylation machineries (Berget 1995). U2AF binds the 3' splice site through interactions with CFIm or PAP (Millevoi et al. 2006; Cooke, Hans, and Alwine 1999; Vagner, Vagner, and Mattaj 2000). In addition, terminal exon recognition is also regulated by interactions of CPSF with SF3B (part of the U2 snRNP) as the presence of CPSF was necessary for efficient splicing activity (Kyburz et al. 2006). While some studies suggest that the links between polyadenylation and splicing are exclusive to terminal exon recognition and processing, the extent of cross-regulation between splicing and polyadenylation at RNA processing events upstream of the terminal exon is still relatively unknown (Movassat et al. 2016).

One polyadenylation factor of particular interest as a putative global alternative splicing regulator is cleavage factor I (CFIm). Two independent affinity purifications of the spliceosome both identified members of the core polyadenylation machinery as co-purifying including subunits of CFIm, CPSF, and CstF (Zhou et al. 2002; Rappsilber et al. 2002) (Table 4.1). Interestingly, although these purifications used different purification methods as well as different pre-RNA splicing targets, the one complex that they had in common was CFIm, suggesting that CFIm may be an important link between polyadenylation and splicing. Additionally, the large subunits of CFIm (CFIm59 and CFIm68) are categorized as SR like proteins because they share several key functional domains with SR proteins, which regulate splicing (Figure 2.3A). SR proteins are characterized by the presence of C-terminal RS domain as well as one or more RNA

Table 4.1 Polyadenylation Factors Associated with Spliceosome

Rappsilber et al, (2002), Genome Res	Zhou et al, (2002), Nature
CFIm25	CFIm68
CPSF30	CFIm25
Similar to CFIm68	fSAP94 (related to Drosophila CstF64)
CPSF100	
CPSF73	
CPSF160	
CstF77	

recognition motif. In addition, they have the ability to complement splicing-deficient nuclear extract. SR-like proteins are similar in that they have the characteristic RS domain, but have a different class of RNA binding domain or cannot complement splicing-deficient nuclear extract (J. C. Long and Caceres 2009). SR-like proteins such as CFIm function in a variety of aspects of RNA metabolism including transport, stability, and translation (Wagner and Frye 2021).

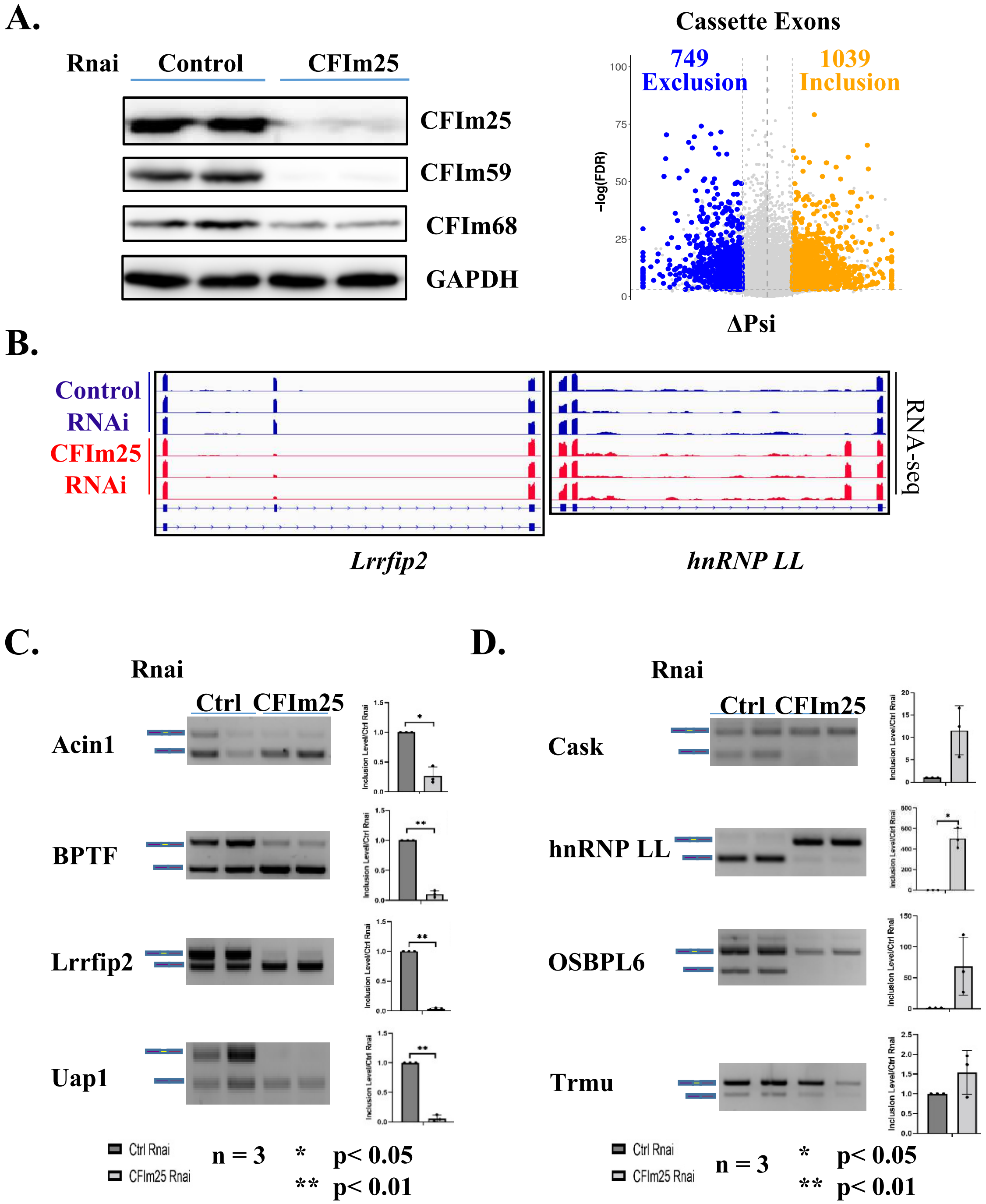
In addition to having structural similarities with SR proteins, CFIm and SR proteins also use similar mechanisms to regulate polyadenylation and splicing respectively. Recently, our lab demonstrated that CFIm is an activator of distal polyA sites. CFIm binds to the UGUA enhancer motif which is enriched approximately 50 nucleotides upstream of the AAUAAA hexamer through interactions of the Nudix domain of CFIm25 with RNA. CFIm then recruits the rest of the core polyadenylation machinery to the nearby polyA site through interactions of its RS domain with the RE/D domain of Fip1, a subunit of the cleavage and polyadenylation specificity factor (CPSF). Because the UGUA motif is specifically upstream of distal polyA sites, CFIm promotes mRNA isoforms with longer 3' UTRs. This is strikingly similar to the role that SR proteins play in splicing regulation. During splicing, SR proteins bind to exonic splicing enhancers (ESEs) and recruit U2AF to the 3' splice site and U170K to the 5' splice site of the exon through interactions of the RS domain of SR proteins with the RS domain of either U2AF35 or U170K (Cho et al. 2011; Saha and Ghosh 2022; J. Y. Wu and Maniatis 1993; Graveley, Hertel, and Maniatis 2001). As both CFIm and SR proteins use their RS domains to regulate polyadenylation and splicing respectively, we hypothesized that CFIm plays a global role in alternative splicing regulation that is not limited to terminal exons (Figure 2.3B).

In this study, I used genome-wide approaches to investigate the role of CFIm in alternative splicing and found that it regulates thousands of alternative splicing events. Mechanistically, CFIm interacts with U2AF65 and recruits it to weak polypyrimidine tracts, leading to increased exon inclusion. In addition, we propose a model by which a variety of RNA binding proteins compete for interaction with U2AF65 to recognize non-canonical 3' splice sites.

4.3 RESULTS

CFIm Regulates Alternative Splicing

To investigate the role of CFIm in alternative splicing, the small subunit, CFIm25, was knocked down in human HEK 293T cells and paired-end RNA sequencing was performed. By knocking down CFIm25, CFIm68 was partially co-depleted and CFIm59 was co-depleted, indicating depletion of the CFIm complex as a whole (Figure 4.2A). Alternative splicing was then analyzed using rMATS (S. Shen et al. 2014). rMATS is a computational tool that uses replicate RNA sequencing data to analyze the percentage of inclusion or exclusion of different features including exons, introns, and 5' or 3' splice sites and compare between different conditions to identify alternative splicing changes. Knockdown of CFIm25 induced over 2000 alternative cassette exons. 749 exons showed decreased exon inclusion upon CFIm depletion, indicating that CFIm activates exon inclusion for this subset. There were also 1039 exons that showed increased inclusion, indicating that CFIm represses exon inclusion for this subset of exons (Figure 4.2A). Two examples of genes with CFIm mediated alternative cassette exons include Lrrfip2 and hnRNP LL. Lrrfip2 contains a CFIm activated exon as there was high exon inclusion within control cells but decreased exon inclusion upon knockdown of CFIm25. The exact opposite is true for hnRNP LL which contains a CFIm repressed cassette exon. In this case, the cassette exon showed increased inclusion upon CFIm depletion (Figure 4.2B). Both CFIm activated and CFIm repressed alternative cassette exons could be validated by RT PCR


A.
Rnai
Control
CFIm25
CFIm25
CFIm59
CFIm68
GAPDH
Cassette Exons
749
Exclusion
1039
Inclusion
-log(FDR)
100
75
50
25
0
ΔPsi
B.
Control
RNAi
CFIm25
RNAi
RNA-seq
Lrrfip2
hnRNP LL
C.
Rnai
Ctrl
CFIm25
Acin1
BPTF
Lrrfip2
Uap1
Inclusion Level/Ctrl Rnai
1.5
1.0
0.5
0.0
*
**
Ctrl Rnai
CFIm25 Rnai
n = 3
* p< 0.05
** p< 0.01
D.
Rnai
Ctrl
CFIm25
Cask
hnRNP LL
OSBPL6
Trmu
Inclusion Level/Ctrl Rnai
20
15
10
5
0
800
600
400
200
150
100
50
2.5
2.0
Ctrl Rnai
CFIm25 Rnai
n = 3
* p< 0.05
** p< 0.01

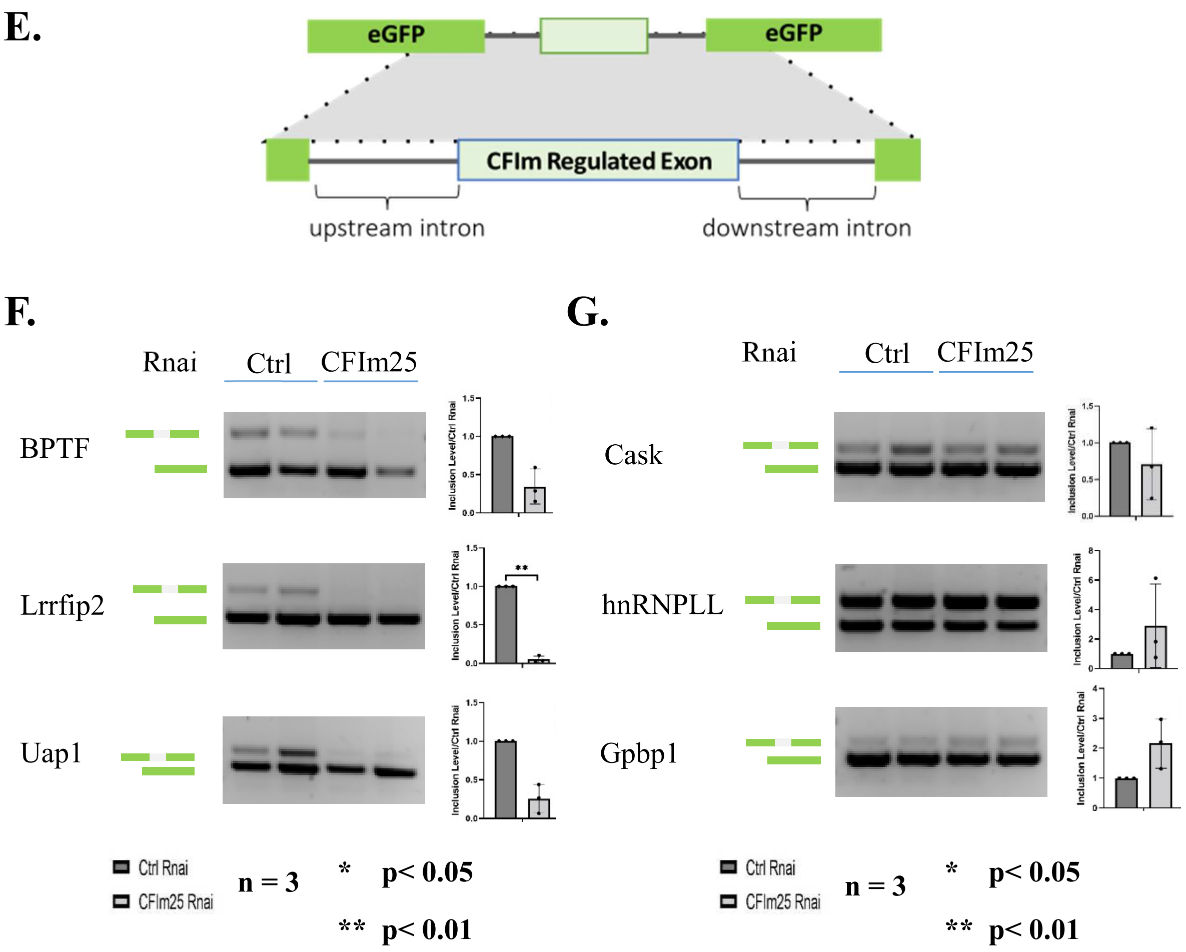


Figure 4.2 CFIm Regulates Alternative Cassette Exon Splicing

A. Knockdown of CFIm25 in human 293T cells (left). RNA sequencing analysis of alternative splicing changes induced by knockdown of CFIm25 (right). Shown is alternative cassette exon splicing.
B. RNA sequencing tracts depicting examples of CFIm activated Lrrfip2 (left) and CFIm repressed hnRNP LL (right) alternative cassette exons.
C. RT-PCR analysis of CFIm activated cassette exons. Quantification of 3 replicates.
D. RT-PCR analysis of CFIm repressed cassette exons. Quantification of 3 replicates.
E. Schematic representation of split eGFP reporter to analyze usage of CFIm regulated alternative cassette exons.
F. RT-PCR analysis of CFIm cassette exon eGFP reporter splicing for CFIm activated exons. Quantification of 3 replicates.
G. RT-PCR analysis of CFIm cassette exon eGFP reporter splicing for CFIm repressed exons. Quantification of 3 replicates.

using primers that bind to the exons immediately upstream and downstream of the alternative cassette exon (Figures 4.2C and 4.2D).

To further validate that the effects seen on alternative cassette exon splicing are a direct effect of CFIm25 knockdown, a series of split eGFP reporters were created. CFIm-regulated cassette exons as well as upstream and downstream intronic regions were cloned between 2 regions of an eGFP gene sequence to interrupt the eGFP reading frame. This reporter was then transfected into control or CFIm25 knockdown cells and splicing was analyzed through either eGFP expression or RT-PCR with primers that bind to the eGFP regions upstream and downstream of the inserted CFIm-regulated exon (Figure 4.2E). For CFIm activated exons, there was a high level of consistency between the eGFP reporters and both RNA sequencing and endogenous RT-PCR as represented by RT-PCR results for BPTF, Lrrfip2, and Uap1, which all showed decreased exon inclusion upon knockdown of CFIm25. This indicates that these alternative cassette exons are likely direct targets of CFIm (Figure 4.2F). However, for CFIm repressed exons, many of the eGFP reporter results were inconsistent with the RNA sequencing and endogenous RT-PCR data as represented by eGFP reporter RT-PCR results for Cask, hnRNP LL, and Gpbp1. For both hnRNP LL and Gpbp1, there was a modest increase in exon inclusion, consistent with the RNA sequencing data. However, the increase was variable across replicates and not significant. In addition, Cask exhibited decreased exon inclusion, contradictory to the RNA-seq data (Figure 4.2G). This suggests that CFIm-repressed exons may not in fact be direct targets of CFIm, mechanism for which will be proposed later.

Importantly, knockdown of CFIm25 did not only affect alternative cassette exons. Instead, other categories of alternative splicing events were also affected, although they were not the focus within this study. Approximately 100 genes exhibited changes in intron retention, with

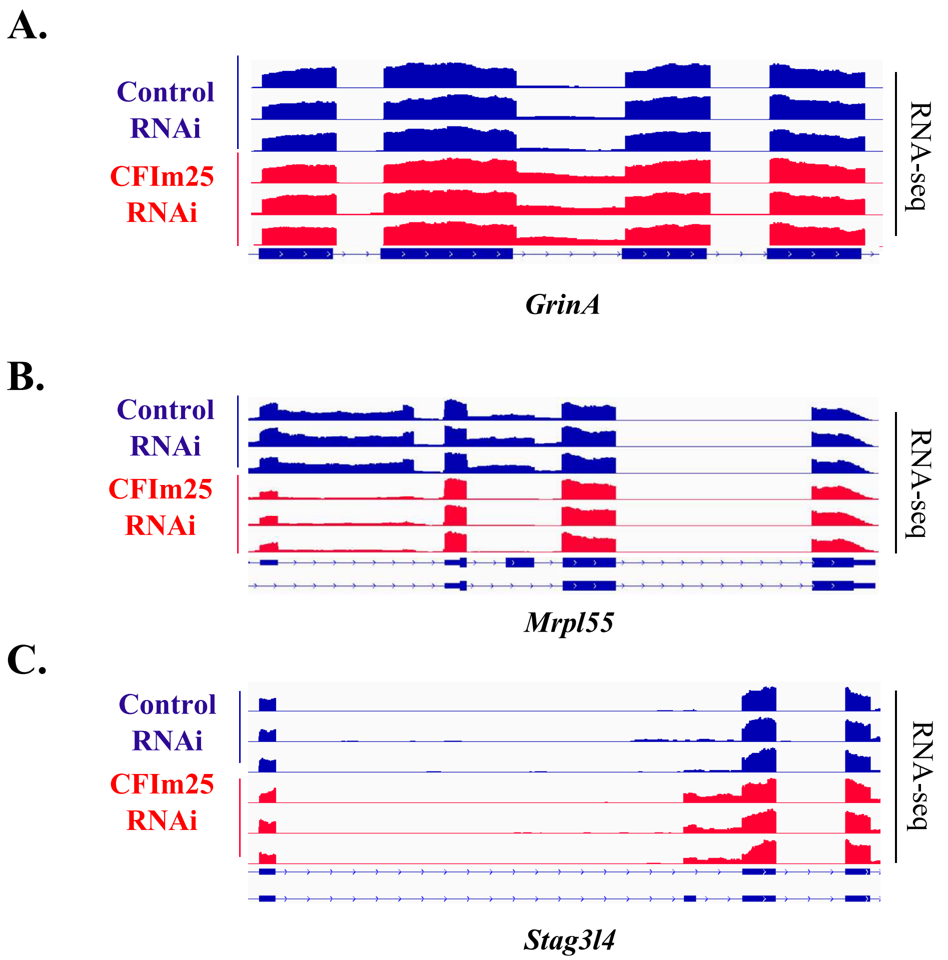


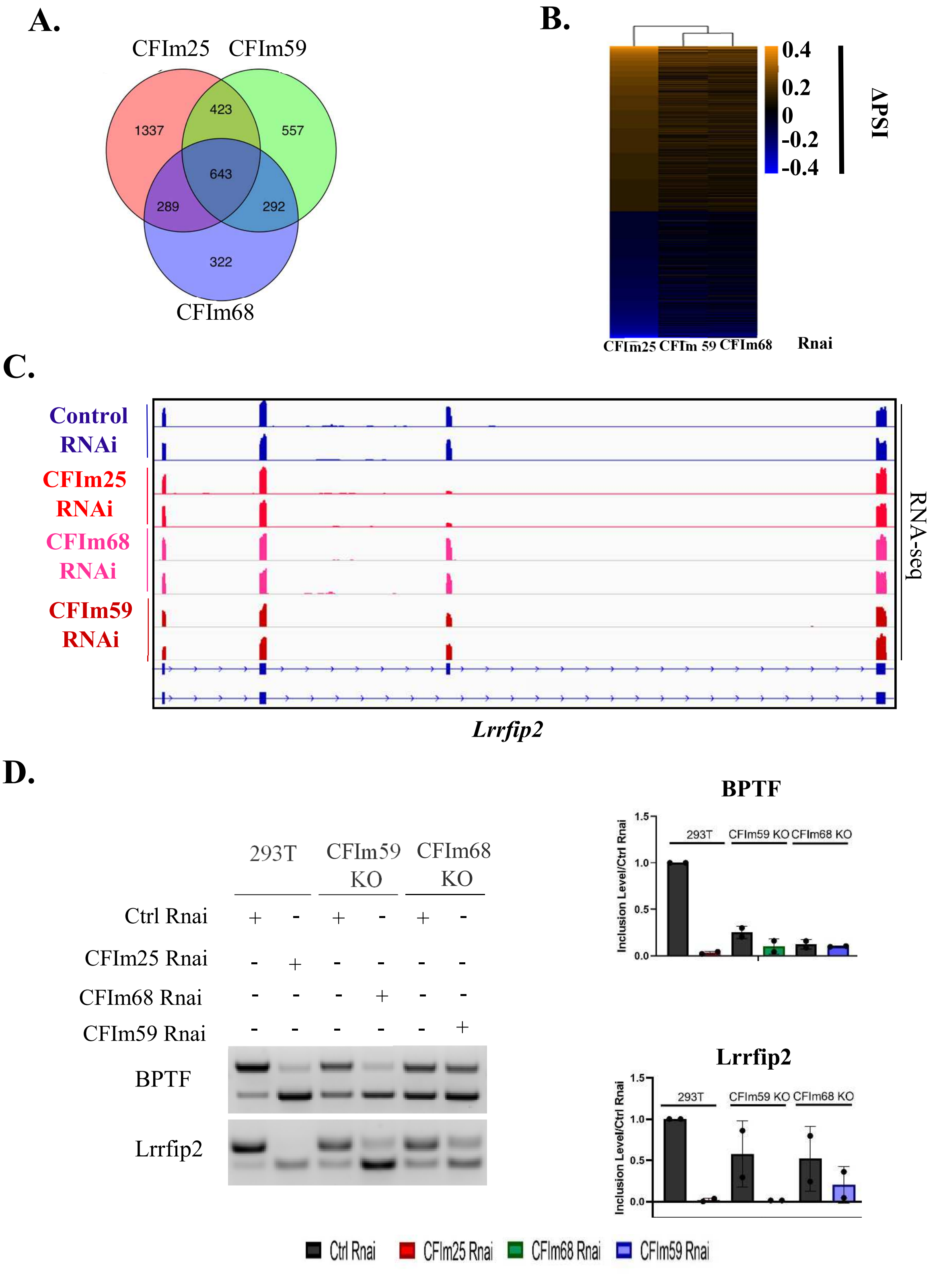
Figure 4.3 CFIm Regulates Other Alternative Splicing Events

A. RNA sequencing tract depicting example of CFIm-dependent intron retention event with GrinA.
B. RNA sequencing tract depicting CFIm-dependent alternative 5' splice site selection within gene Mrpl55.
C. RNA sequencing tract depicting CFIm-dependent alternative 3' splice site selection within gene Stag3l4.

approximately equal numbers showing increased or decreased intron inclusion. For example, the gene GrinA contains an intron that showed increase retention upon knockdown of CFIm25 (Figure 4.3A). In addition, CFIm depletion altered both alternative 5' and 3' splice site usage. CFIm regulated approximately 200 alternative 5' splice sites. As an example, the gene Mrlp55 contains a 5' splice regulated by CFIm because upon knockdown of CFIm25, there was a shift to an upstream 5' splice site (Figure 4.3B). CFIm also regulated approximately 200 alternative 3' splice sites with the example gene of Stag314 showing a shift to an upstream 3' splice site upon knockdown of CFIm25 (Figure 4.3C). Together, this data suggests that CFIm regulates the alternative splicing of thousands of alternative splicing events. While it primarily regulates alternative splicing of cassette exons, it also can regulate other splicing events including intron retention and alternative 5' and 3' splice sites.

All Subunits of CFIm Participate in Splicing Regulation

As CFIm is a multiple subunit protein complex, the role of the large subunit, either CFIm59 or CFIm68, was then evaluated. To test this, previously generated CFIm59 and CFIm68 knockout 293T cells were utilized for paired-end RNA sequencing (Zhu et al. 2017). As before, splicing analysis was performed using rMATs with FDR <0.05 and the absolute value of ΔPSI (percent spliced in) greater than 0.15. Consistent with the hypothesis that the large subunit directly regulates alternative splicing through CFIm59 and CFIm68, there were alternative splicing changes upon knockout of CFIm59 and CFIm68, with 653 alternative splicing changes being regulated by all three conditions: knockdown of CFIm25 and knockout of both CFIm59 and CFIm68. Importantly, however, the greatest number of alternative splicing changes induced by knockdown of CFIm25, with other over 1300 changes solely regulated by CFIm25 in


A.
CFIm25
CFIm59
CFIm68
1337
423
557
643
289
292
322
B.
0.4
0.2
0
-0.2
-0.4
ΔPSI
CFIm25 CFIm 59 CFIm68
Rnai
C.
Control RNAi
CFIm25 RNAi
CFIm68 RNAi
CFIm59 RNAi
RNA-seq
Lrrfip2
D.
293T
CFIm59 KO
CFIm68 KO
Ctrl Rnai + - + - + -
CFIm25 Rnai - + - - - -
CFIm68 Rnai - - - + - -
CFIm59 Rnai - - - - - +
BPTF
Lrrfip2
BPTF
Inclusion Level/Ctrl Rnai
1.5
1.0
0.5
0.0
293T
CFIm59 KO
CFIm68 KO
Lrrfip2
Inclusion Level/Ctrl Rnai
1.5
1.0
0.5
0.0
293T
CFIm59 KO
CFIm68 KO
Ctrl Rnai
CFIm25 Rnai
CFIm68 Rnai
CFIm59 Rnai

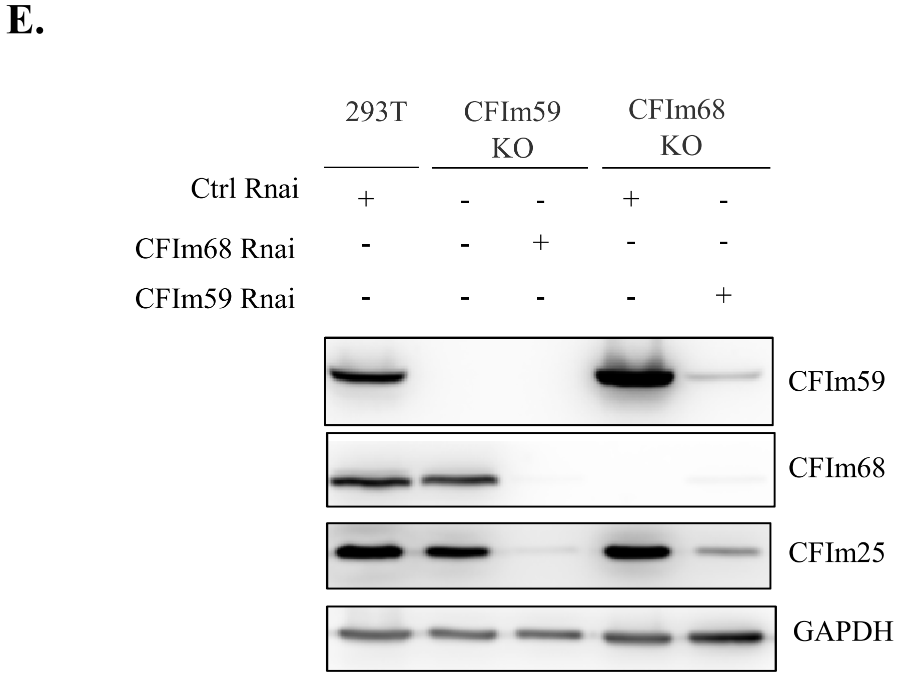


Figure 4.4 All CFIm Subunits Regulate Alternative Splicing

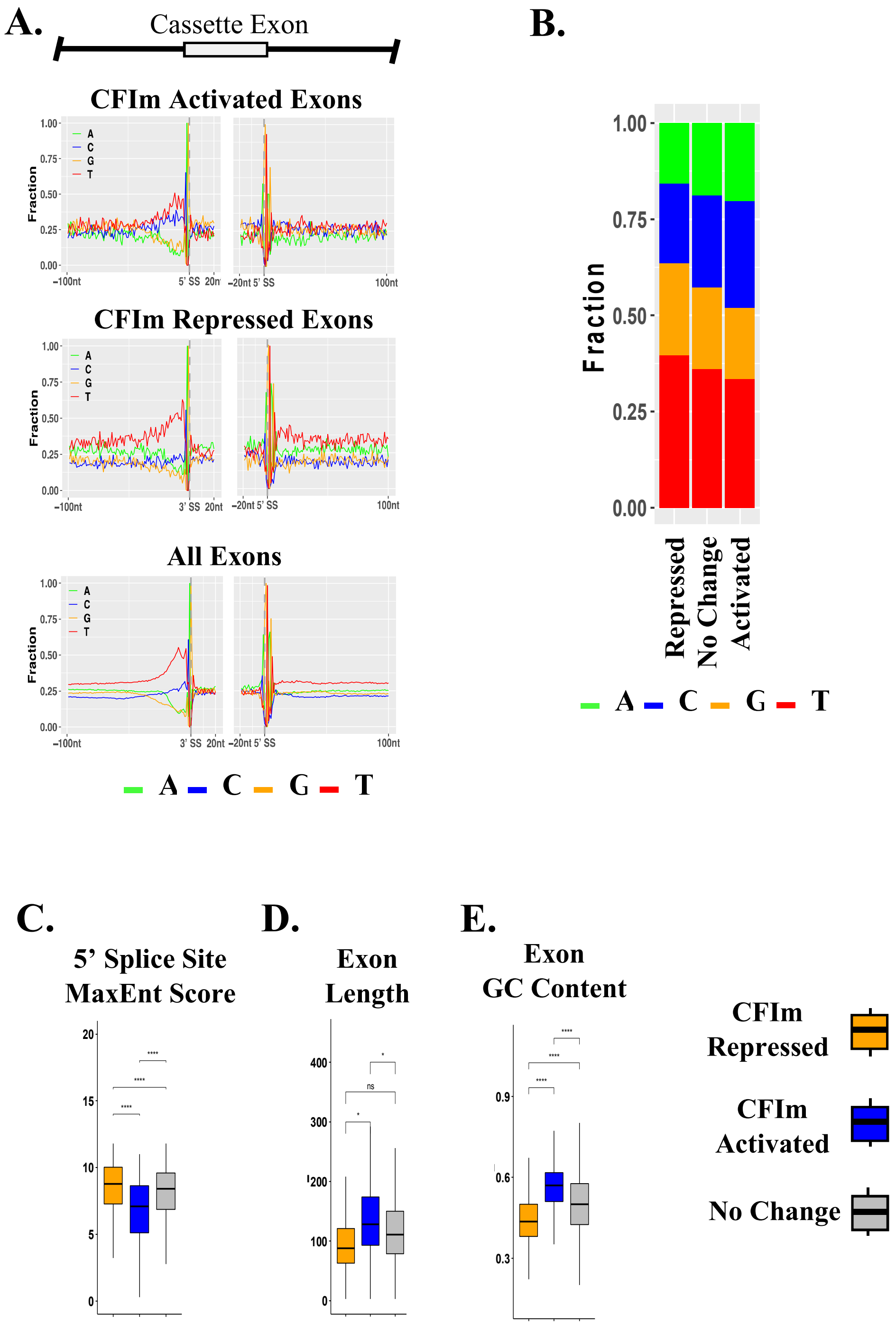
A. Overlap of alternative cassete exon splicing events regulated by CFIm25 (pink), CFIm59 (green), and CFIm68 (blue).
B. Heat map depicting changes in percent spliced in (ΔPSI) upon CFIm25 knockdown, CFIm59 knockout, or CFIm68 knockout for alternative cassette exons.
C. RNA sequencing tract depicting alternative cassette usage upon knockdown of CFIm25 or knockout of CFIm59 or CFIm68 for Lrrfip2.
D. RT-PCR analysis of alternative cassette exon usage of CFIm activated exon BPTF and Lrrfip2 upon co-depletion of CFIm59 and CFIm68.
E. Western blotting analysis of CFIm59 and CFIm68 co-depletion. Both co-depletion methods also deplete CFIm25.

comparison to only 557 for CFIm59 knockout and 322 for CFIm68 knockout (Figure 4.4A). In addition, when alternative splicing changes were compared genome-wide, for the majority of alternative splicing changes, there was a greater magnitude of change upon knockdown of CFIm25 than knockout of either CFIm59 or CFIm68 as measured by ΔPSI (Figure 4.4B). For example, Lrrfip2 contained a CFIm activated exon that showed decreased exon inclusion upon knockdown of CFIm25. Although Lrrfip2 also experienced decreased exon inclusion upon knockout of CFIm59 or CFIm68, the percent of change is smaller (Figure 4.4C). The stronger regulation by CFIm25 knockdown than either CFIm59 or CFIm68 knockout likely indicates that the two alternative larger subunits are partially redundant for each other. When CFIm25 was knocked down, both CFIm59 and CFIm68 were co-depleted. However, when CFIm59 and CFIm68 were knocked out, the levels of the alternative larger subunit as well as CFIm25 remained unchanged. As a result, if CFIm59 and CFIm68 are partially redundant, it would explain why there is a greater magnitude of change for CFIm25 knockdown. However, the possibility that the differences in affect size are caused by differences in the depletion system cannot be ruled out. With the development of CRISPR-Cas9 genome-editing technologies in molecular biology, scientists have observed discrepancies between knockdown and knockout experiments. While this was originally believed to be due to off-target effects, it has been shown that cells can also compensate for knockout of some proteins, often by upregulating other proteins with similar functions (Salanga and Salanga 2021). This is likely not the case with CFIm59 and CFIm68 knockout as levels of the alternative large subunit as well as CFIm25 remain equal to wildtype 293T cells, but this possibility has not been followed up further and therefore cannot be ruled out entirely.

Next, co-depletion studies were performed to further investigate whether CFIm59 and CFIm68 are partially redundant for each other. For co-depletion experiments, the alternative large subunit was knocked down in the CFIm59 and CFIm68 knockout cell lines. Co-depletion of the alternative large subunit further altered splicing to an extent more similar to knockdown of CFIm25. For example, for the CFIm activated exons Lrrfip2 and BPTF, while knockout of CFIm59 or CFIm68 individually decreased exon inclusion, the magnitude of change was less than that of knockdown of CFIm25. However, when CFIm59 was co-depleted in CFIm68 knockout cells, there was decreased exon inclusion, with similar results upon co-depletion of CFIm68 in CFIm59 knockout cells (Figure 4.4D). While this data suggests that CFIm59 and CFIm68 mediate changes in alternative splicing and that they are partially redundant for each other, western blotting analysis revealed that co-depletion of CFIm59 and CFIm68 also depleted CFIm25, likely through a change in stability (Figure 4.4E). As a result, the possibility that the results seen are due to the decrease in CFIm25 levels cannot be eliminated.

CFIm Regulated Exons Have Distinct Sequence Features

Upon determining that CFIm regulates alternative splicing, the precise mechanism was determined. To do so, cassette exons were sorted as either CFIm activated exons, CFIm repressed exons, or exons with no change and the percent enrichment of each nucleotide was calculated from 100 nucleotides upstream and 20 nucleotides downstream of the 3’ splice as well as 20 nucleotides upstream to 100 nucleotides downstream of the 5’ splice site. For unchanged exons as well as CFIm repressed exons, the most highly enriched nucleotide was U, consistent with strong polypyrimidine tracts (Coolidge, Seely, and Patton 1997; Reed 1989; Hertel 2008). However, CFIm activated exons had an enrichment of both C and U, indicating weaker



A.
Cassette Exon
CFIm Activated Exons
Fraction
1.00
0.75
0.50
0.25
0.00
A
C
G
T
−100nt
5' SS
20nt
−20nt 5' SS
100nt
CFIm Repressed Exons
3' SS
All Exons
A C G T
B.
Fraction
1.00
0.75
0.50
0.25
0.00
Repressed
No Change
Activated
A C G T
C.
5' Splice Site
MaxEnt Score
20
15
10
5
0
****
D.
Exon
Length
400
300
200
100
0
*
ns
E.
Exon
GC Content
0.9
0.6
0.3
CFIm
Repressed
CFIm
Activated
No Change

Figure 4.5 CFIm Regulated Exons Have Distinct Sequence Features

A. Genome-wide enrichment of nucleotides from 100 nucleotides upstream of 3’ splice site to 20 nucleotides downstream of 3’ splice site as well as from 20 nucleotides upstream to 100 nucleotides downstream of 5’ splice site for CFIm activated exons (top), repressed (middle), or exons with no change (bottom).
B. Genome-wide nucleotide enrichment as fraction of all nucleotides from 100 nucleotides upstream of 3’ splice site to 20 nucleotides downstream of 3’ splice site for CFIm repressed exons (left), CFIm activated exons (right), or exons with no change (middle).
C. Genome-wide analysis of 5’ splice site MaxEnt Scores for CFIm repressed (left), no change (middle) or CFIm activated (right) exons.
D. Genome-wide analysis of exons length for CFIm repressed (left), no change (middle) or CFIm activated (right) exons.
E. Genome-wide analysis of exon GC content for CFIm repressed (left), no change (middle) or CFIm activated (right) exons.

polypyrimidine tracts (Figure 4.5A). In addition, when exons were categorized as either activated, repressed, or unchanged, CFIm activated exons have the lowest enrichment of U and the highest enrichment of C. This suggests that CFIm may be important for activating exons with weaker polypyrimidine tracts that are more C/T rich (Figure 4.5B)

Additional features of CFIm regulated exons including 5' splice site strength, exon GC content, and exon length were also analyzed. CFIm activated exons had the lowest 5' splice site strengths as measured by MaxEnt Scores (Yeo and Burge 2004), suggesting that in additional to non-canonical 3' splice sites with weak polypyrimidine tracts, CFIm activated exons also have weak 5' splice sites. This further supports a model in which CFIm promotes exon inclusion of exons with poor splice sites (Figure 4.5C). CFIm activated exons were also longer and had a higher GC content that either CFIm repressed or non-regulated exons. CFIm repressed exons also had significantly lower GC content than non-regulated exons. (Figure 4.5D and 4.5E). Interestingly, it has previously been shown that exon GC content is related to splicing efficiency, with more GC rich exons being more efficiently spliced. While it is originally proposed this differential splicing is mediated by nucleosome deposition, these exons also seem to rely on CFIm to promote their exon inclusion, so it is possible that CFIm is also important for the recognition of these GC-rich exons (Amit et al. 2012).

CFIm Binds Near 3' Splice Sites and Interacts with U2AF

To further explore the mechanism for CFIm-mediated splicing regulation, publicly available PAR-CLIP (Figure 2.4A) data for both CFIm59 and CFIm68 was utilized to investigate where in the genome CFIm binds (G. Martin et al. 2012). Excluding 3' UTRs, PAR-CLIP analysis revealed that over 50% of CFIm binds within introns (Figure 4.6A). Aligning PAR-CLIP signal to cassette exons genome-wide revealed an enrichment of CFIm binding at both 3' and 5' splice sites, which is very reminiscent of U2AF and U1 in recognizing 3' and 5' respectively. Next, U2AF65 iCLIP was performed to compare its binding distribution with that of CFIm. Strikingly, there was a similar enrichment of U2AF65 at 3' splice sites (Figure 4.6B). When average iCLIP signal is compared along cassette exons genome-wide, the peak binding for CFIm and U2AF65 overlap specifically at 3' splice sites, with a stronger iCLIP signal for U2AF65 than for CFIm59 or CFIm68 (Figure 4.6C). When cassette exons were additionally sorted as either CFIm activated or CFIm repressed exons, it became apparent that CFIm activated and repressed exons were regulated differently. For CFIm activated exons, there was a stronger enrichment of CFIm68 than CFIm59 with a peak in binding near the 3' splice site. The exact opposite is true for CFIm repressed exons: enrichment for CFIm59 binding is stronger than CFIm68 with stronger enrichment near the 5' splice site (Figure 4.6D). This suggests that both the nature of the CFIm complex as well as the location of binding can impact how an exon is regulated. It is not entirely unexpected that CFIm59 and CFIm68 may have unique roles in the regulation of alternative splicing. While the two CFIm heterodimers were originally considered to be equivalent, recent reports have revealed that they have overlapping and unique roles in the regulation of alternative polyadenylation (Zhu et al. 2017; Tseng et al. 2021).

As CFIm and U2AF both bind near 3' splice sites, they are likely to interact. FLAG co-immunoprecipitations performed with flag-U2AF65-overexpressing 293T cells in the presence of RNase revealed that in addition to interacting with the small subunit U2AF35, U2AF65 interacts with CFIm25 as well as both CFIm59 and CFIm68, with stronger interaction with CFIm68 than CFIm59 (Figure 4.7A).

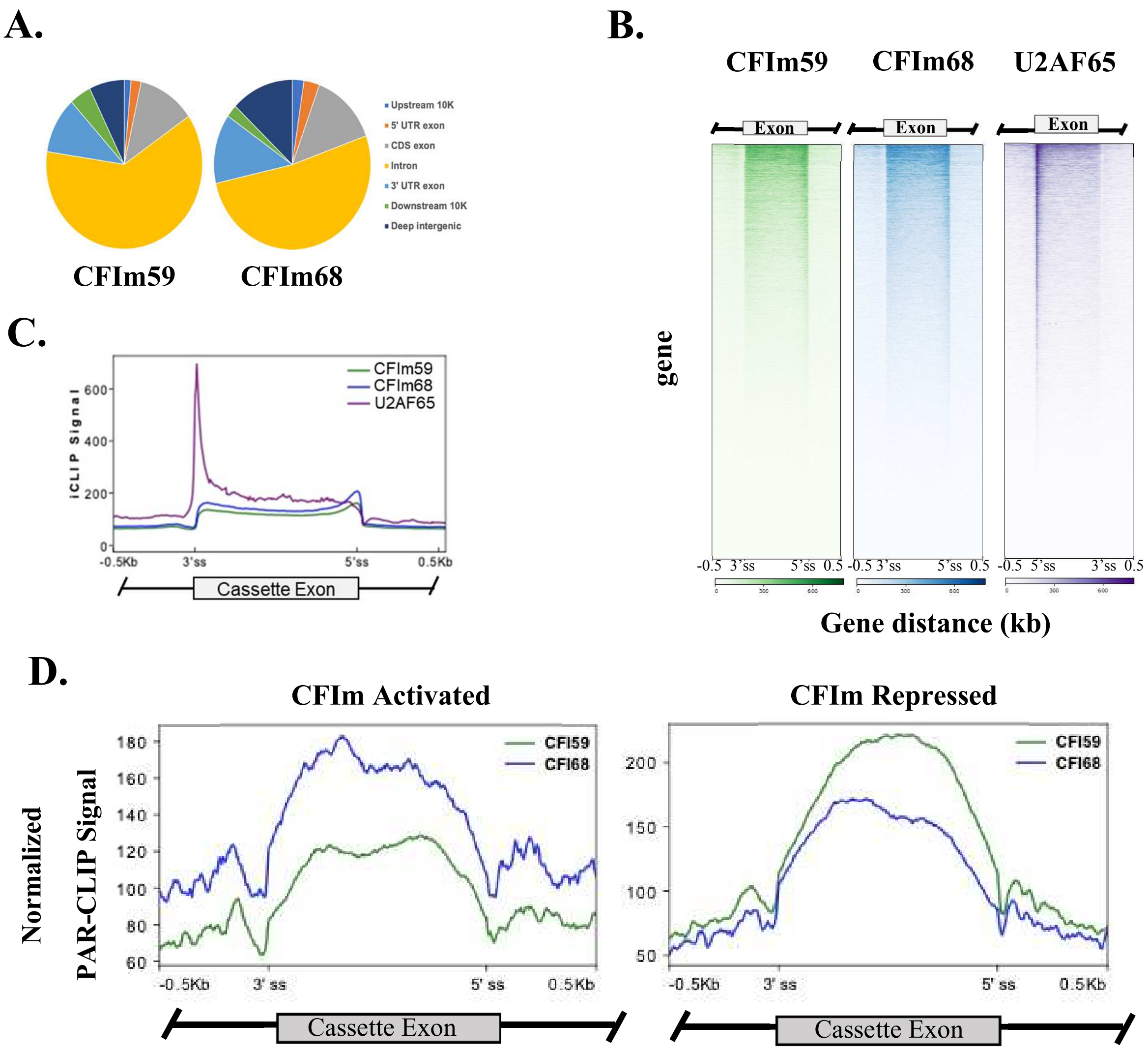
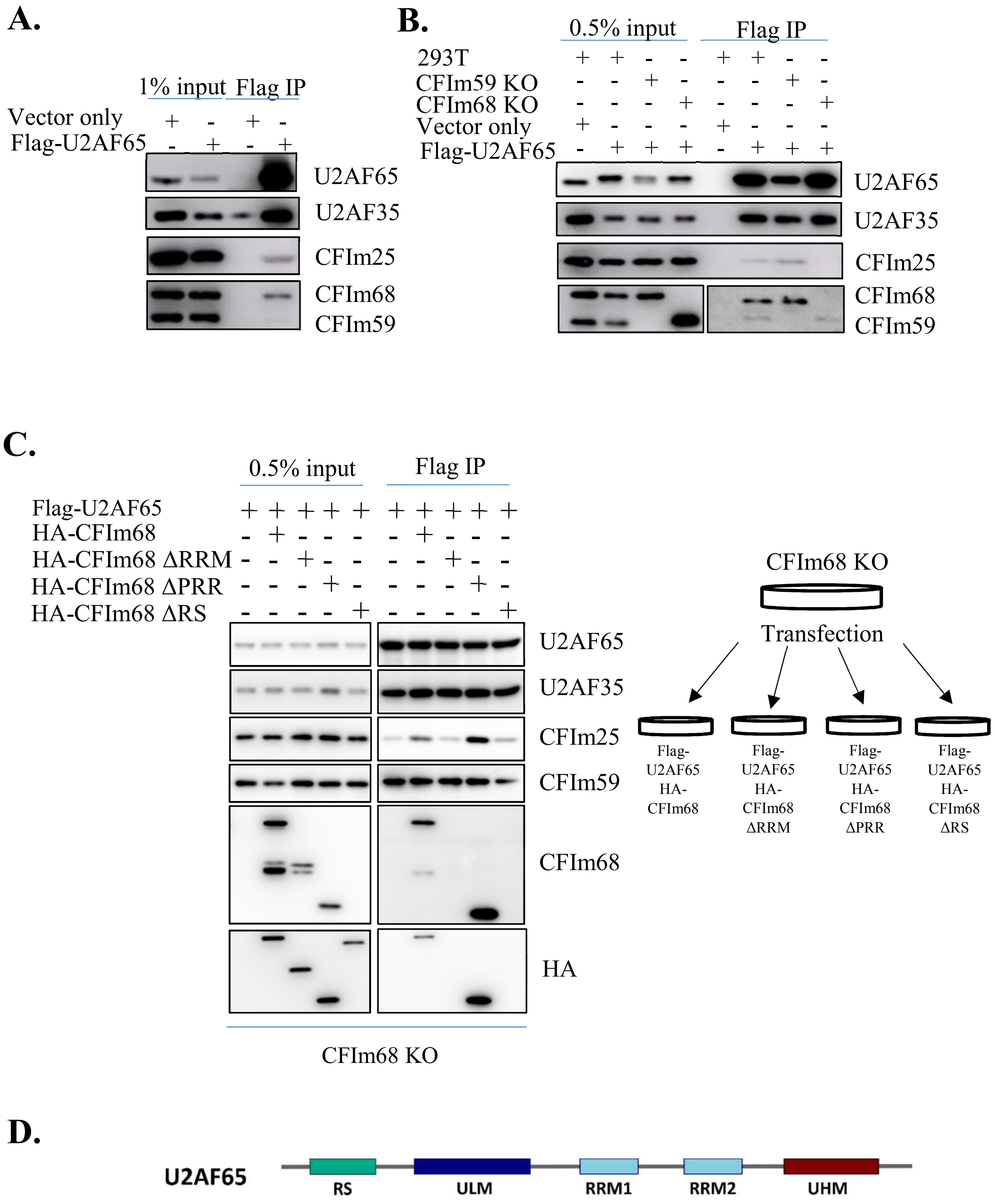


Figure 4.6 CFIm Binds to 3' and 5' Splice Sites

A. Genome-wide PAR-CLIP enrichment for CFIm59 and CFIm68 in different genomic regions excluding 3' UTRs.
B. PAR-CLIP signal for CFIm59 (green) and CFIm68 (blue) of iCLIP signal for U2AF65 (purple) across cassette exons genome-wide from 500nt upstream to 500nt downstream.
C. Average genome-wide iCLIP signal of CFIm59 (green), CFIm68 (blue), and U2AF65 (purple) across CFIm-regulated cassette exons genome-wide from 500nt upstream of 3' splice site to 500nt downstream of 5' splice site.
D. Average genome-wide PAR-CLIP signal of CFIm59 (green) and CFIm68 (blue) CFIm-regulated cassette exons genome-wide from 500nt upstream of 3' splice site to 500nt downstream of 5' splice site. Cassette exons are sorted as either CFIm activated (left) or CFIm repressed (right).


A.
1% input
Flag IP
Vector only + - + -
Flag-U2AF65 - + - +
U2AF65
U2AF35
CFIm25
CFIm68
CFIm59
B.
0.5% input
Flag IP
293T + + - - + + - -
CFIm59 KO - - + - - - + -
CFIm68 KO - - - + - - - +
Vector only + - - - + - - -
Flag-U2AF65 - + + + - + + +
U2AF65
U2AF35
CFIm25
CFIm68
CFIm59
C.
0.5% input
Flag IP
Flag-U2AF65 + + + + + + + + + +
HA-CFIm68 - + - - - - + - - -
HA-CFIm68 ΔRRM - - + - - - - + - -
HA-CFIm68 ΔPRR - - - + - - - - + -
HA-CFIm68 ΔRS - - - - + - - - - +
U2AF65
U2AF35
CFIm25
CFIm59
CFIm68
HA
CFIm68 KO
CFIm68 KO
Transfection
Flag-U2AF65 HA-CFIm68
Flag-U2AF65 HA-CFIm68 ΔRRM
Flag-U2AF65 HA-CFIm68 ΔPRR
Flag-U2AF65 HA-CFIm68 ΔRS
D.
U2AF65
RS
ULM
RRM1
RRM2
UHM

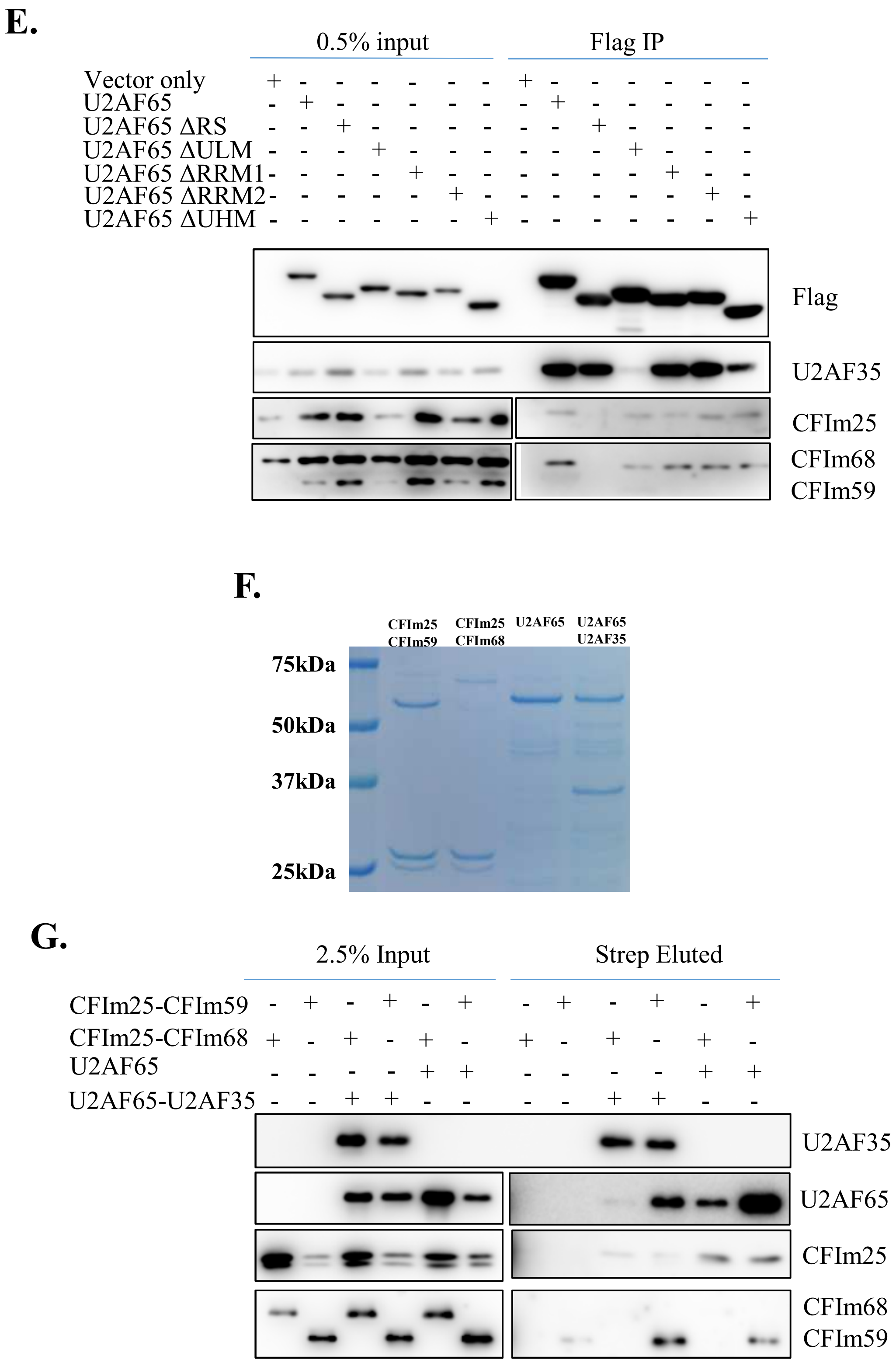

E.
0.5% input
Flag IP
Vector only + - - - - - - + - - - - - -
U2AF65 - + - - - - - - + - - - - -
U2AF65 ΔRS - - + - - - - - - + - - - -
U2AF65 ΔULM - - - + - - - - - - + - - -
U2AF65 ΔRRM1 - - - - + - - - - - - + - -
U2AF65 ΔRRM2 - - - - - + - - - - - - + -
U2AF65 ΔUHM - - - - - - + - - - - - - +
Flag
U2AF35
CFIm25
CFIm68
CFIm59
F.
CFIm25
CFIm59
CFIm25
CFIm68
U2AF65
U2AF65
U2AF35
75kDa
50kDa
37kDa
25kDa
G.
2.5% Input
Strep Eluted
CFIm25-CFIm59 - + - + - + - + - + - +
CFIm25-CFIm68 + - + - + - + - + - + -
U2AF65 - - - - + + - - - - + +
U2AF65-U2AF35 - - + + - - - - + + - -
U2AF35
U2AF65
CFIm25
CFIm68
CFIm59

Figure 4.7 CFIm Interacts with U2AF Through the RS Domain

A. Western blotting analysis of U2AF65 flag-immunoprecipitation.

B. Western blotting analysis of U2AF65 flag-immunoprecipitation for either control 293T, CFIm59 knockout 293T, or CFIm68 knockout 293T cells. Loading is normalized to by flag-U2AF65 intensity.

C. Western blotting analysis of U2AF65 flag-immunoprecipitation from either control or CFIm68 knockout 293T cells. CFIm68 knockout is rescued with either HA-CFIm68 or a HA-CFIm68 construct lacking a single domain of CFIm68. Loading is normalized to by flag-U2AF65 intensity.
RRM: RNA recognition motif PRR: proline rich region
RS: arginine-serine rich domain

D. Schematic representation of domains of U2AF65.
RS: arginine-serine rich domain ULM: U2AF ligand motif
RRM: RNA recognition motif UHM: U2AF homology motif

E. Western blotting analysis of U2AF65 flag-immunoprecipitation with either full-length U2AF65 or U2AF65 lacking a single domain. Loading is normalized to by flag-U2AF65 intensity.

F. Coomassie staining of recombinant His-tagged CFIm25-CFIm59 complex, His-tagged CFIm25-CFIm68 complex, Strep-tagged U2AF65 and Strep-tagged U2AF65-U2AF35 complex.

G. Western blotting analysis of recombinant strep-U2AF65 affinity purification. U2AF65 or the U2AF65-U2AF35 complex is incubated with either CFIm25-CFIm59 or CFIm25-CFIm68 to analyze effects on interaction in vitro.

As CFIm is a heterodimeric protein complex, the hypothesis that it is the large subunit of CFIm that interacts with U2AF65 was tested by repeating the U2AF65 FLAG immunoprecipitation in CFIm59 and CFIm68 knockout 293T cells. In CFIm59 knockout cells, there was no change in the interaction of U2AF65 with CFIm25. However, upon knockout of CFIm68, there was decreased interaction of CFIm25 with U2AF65, indicating that CFIm68 is necessary for the interaction with U2AF65 (Figure 4.7B). CFIm68 contains an N-terminal RNA recognition motif, a proline-rich region, and a C-terminal RS domain (Figure 2.2A). To interrogate the domain(s) of CFIm68 that are responsible for interacting with U2AF65, the FLAG immunoprecipitation was repeated in CFIm68 knockout cells that were rescued with either full length HA-tagged CFIm68 or a HA-tagged CFIm68 mutant that lacks a single domain. When the CFIm68 knockout cells were rescued with either wildtype CFIm68 or a mutant CFIm68 lacking the proline rich region, the interaction of CFIm25 with U2AF65 was restored, indicating that the proline rich region is not necessary for interacting with U2AF65 (Figure 4.7C). In contrast, when CFIm68 knockout cells were rescued with mutants lacking either the RNA recognition motif or RS domain of CFIm68, the interaction between U2AF65 and CFIm25 was not rescued. The deletion of the RNA recognition motif likely does not rescue the interaction between CFIm25 and U2A65 because the RNA recognition motif of CFIm68 interacts with CFIm25 (Q. Yang et al. 2011). As a result, it seems likely that the RS domain of CFIm68 is necessary for interacting with U2AF65.

U2AF65 also contains multiple domains: an N-terminal RS domain, a U2AF ligand motif (ULM), two RNA recognition motifs (RRM), and a C-terminal U2AF homology motif (UHM) (Figure 4.7D). UHMs are protein-protein interaction domains that are structurally related to RNA recognition motifs and interact with ULMs. Like RRMs, UHMs have the characteristic βαββαβ topology, but the first helix is extended. During splicing, the two canonical RRMs of U2AF65 bind to the polypyrimidine tract. The interaction of U2AF65 is strengthened by the interaction of the UHM of U2AF65 with the ULM in Sf1, which recognizes the branch point. In addition, the ULM of U2AF65 interacts with the UHM of U2AF35 (Kielkopf, Lücke, and Green 2004).

To investigate the domain(s) of U2AF65 that interact with CFIm, the U2AF65 FLAG immunoprecipitation was performed with constructs lacking a single domain of U2AF65. Consistent with an interaction between U2AF65 and U2AF35 through the ULM of U2AF65, deletion of the ULM of U2AF65 disrupted the interaction with U2AF35. In addition, deletion of the RS domain of U2AF65 impaired the interaction with CFIm25, CFIm59, and CFIm68, indicating that the RS domain of U2AF65 is critical for interacting with CFIm. Importantly, deletion of the RS domain did not impair the interaction of U2AF65 with U2AF35, suggesting that the change in interaction with CFIm is not due to a change in the structural integrity of U2AF65 upon deletion of the RS domain (Figure 4.7E).

While the flag immunoprecipitation results indicate that CFIm and U2AF interact, it is unknown whether the interaction is direct or indirect. To test for direct interaction between CFIm and U2AF, recombinant CFIm and U2AF were purified from Sf9 insect cells. For U2AF, either strep-tagged U2AF65 alone (denoted U2AF65) or in complex with his-tagged U2AF35 (denoted U2AF65-U2AF35) was purified. For CFIm, his-tagged CFIm25 was purified in complex with either CFIm59 or in complex with CFIm68 (denoted CFIm25-CFIm59 or CFIm25-CFIm68) (Figure 4.7F). An excess of CFIm was incubated with U2AF followed by strep purification. Western blotting analysis revealed that there was a direct interaction between U2AF and CFIm as CFIm can be co-purified with U2AF (Figure 4.7G). There is no apparent difference between the two CFIm complexes in interaction with U2AF. Both the U2AF65-U2AF35 and U2AF65 alone interacted with both CFIm complexes, with slightly stronger interaction with U2AF65 than U2AF65-U2AF35.

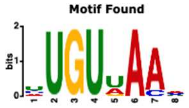
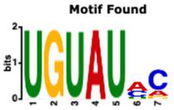
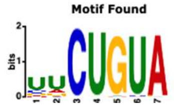
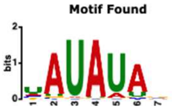
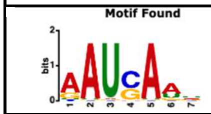
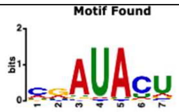
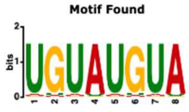
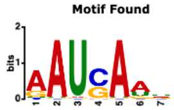
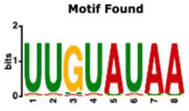
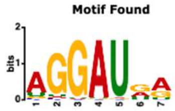
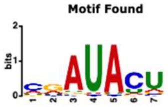
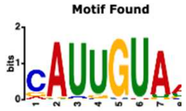
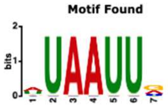
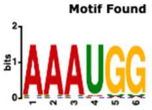
Together, the flag immunoprecipitation data and in vitro pulldown data suggest that there is a direct interaction between CFIm and U2AF65 through the RS domain of CFIm68 and the RS domain of U2AF65.

CFIm and U2AF Collectively Regulate Alternative Splicing

After determining that CFIm and U2AF both bind 3’ splice sites and interact with each other, we investigated how CFIm binding at regulated exons may regulate alternative splicing. To do so, the BPTF cassette exon was selected as a representative CFIm-activated exon and a series of CFIm binding mutations were created. To determine the sequence of CFIm binding sites near regulated exons, PAR-CLIP data for both CFIm59 and CFIm68 was analyzed to identify the most highly enriched sequences for CFIm binding near regulated exons (Table 4.2). For CFIm59, the most highly enriched sequence was UGU followed by various iterations of the UGUA motif that CFIm is known to bind to in 3’ UTRs. Similarly, the most highly enriched sequence for CFIm68 binding was UGUAU. Both CFIm59 and CFIm68 data validate that CFIm binds to sequences related to the canonical UGUA sequence when bound near cassette exons, confirming that CFIm binds similar sequences to regulate splicing and polyadenylation.

Several single or double mutations were created within an eGFP splicing reporter that either altered a UGUA motif, a sequence related to the UGUA motif such as UGUC that is likely still bound by CFIm, or the polypyrimidine tract (Figure 4.8A). The polypyrimidine tract of BPTF is noncanonical and contains a G-rich region. By mutating the C-rich region to a G-rich region, it

Table 4.2 CFIm Binding Motifs at Cassette Exons

CFIm68		CFIm59	
Motif	Enrichment	Motif	Enrichment
	1.3e-2420		2.06e-2644
	6.5e-1164		4.5e-536
	2.7e-935		1.9e-311
	2.4e-884		7.4e-291
	4.2e-776		1.1e-290
	2.1e-506		4.8e-248
	2.2e-505		2.9e-235

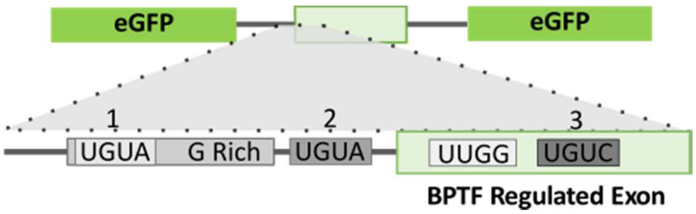
was predicted that the percentage of exon inclusion may increase. At the same time, G to C mutation also changes a UGUA motif to UCUA, making this splicing event independent of CFIm. By contrast, mutations that are necessary for CFIm binding will likely decrease exon inclusion as BPTF is a CFIm activated exon that requires CFIm for recruitment of BPTF to the 3' splice site.

BPTF mutant reporters were transfected into 293T cells and the effects on alternative splicing were evaluated by RT PCR. The polypyrimidine tract mutation that mutated the G rich sequence to a C rich sequence increased cassette exon inclusion, consistent with the BPTF polypyrimidine tract being weak and non-canonical. Additional single mutations had limited effect on alternative splicing, while double mutations had larger effect sizes (Figure 4.8B). Interestingly, mutating exonic UUGG and UGUC sequences increased exon inclusion. It is possible that these sequences are not critical for CFIm binding, particularly as they are not canonical CFIm binding sites; any observed result may be caused by an alternative mechanism. In addition, as these sequences are exonic, it may indicate that it is the region upstream of the 3' splice site not the exonic region itself that CFIm binds to and regulates recruitment of U2AF.

CFIm Regulates U2AF-RNA Interactions

As CFIm and U2AF interact and both bind near 3' splice sites, it seemed likely that CFIm may regulate U2AF-RNA interactions. To test this, U2AF65 iCLIP was performed in CFIm25 knockdown and control knockdown cells. Upon segregation for CFIm activated and repressed exons, it can be observed that specifically for CFIm activated exons, there is decreased

A.



B.

Lane	Mutant
1	Wildtype
2	G -> C
3	UGUA2 -> UCUA
4	UUGG -> UUCC
5	UGUC3 -> CAUC
6	UGUC3 -> UCUC
7	G -> C & UGUA2 -> UCUA
8	UUGG -> CCGG & UGUC3 –> CAUC
9	UUGG -> CCGG & UGUA3 -> UCUA

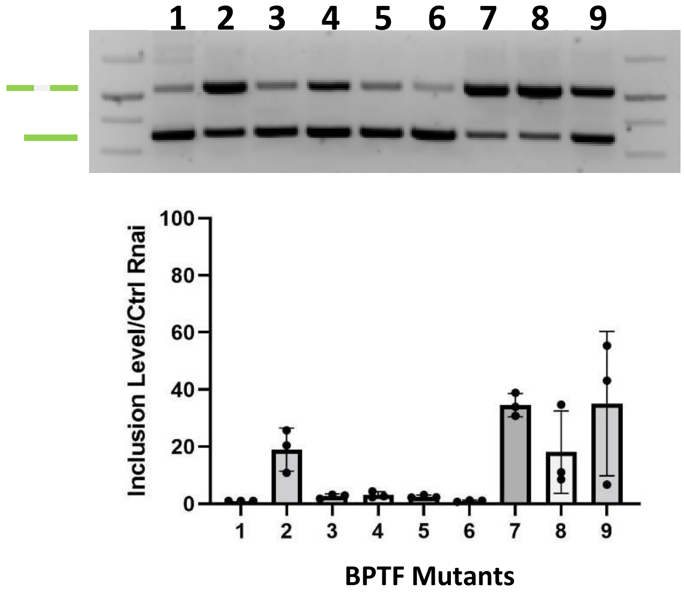


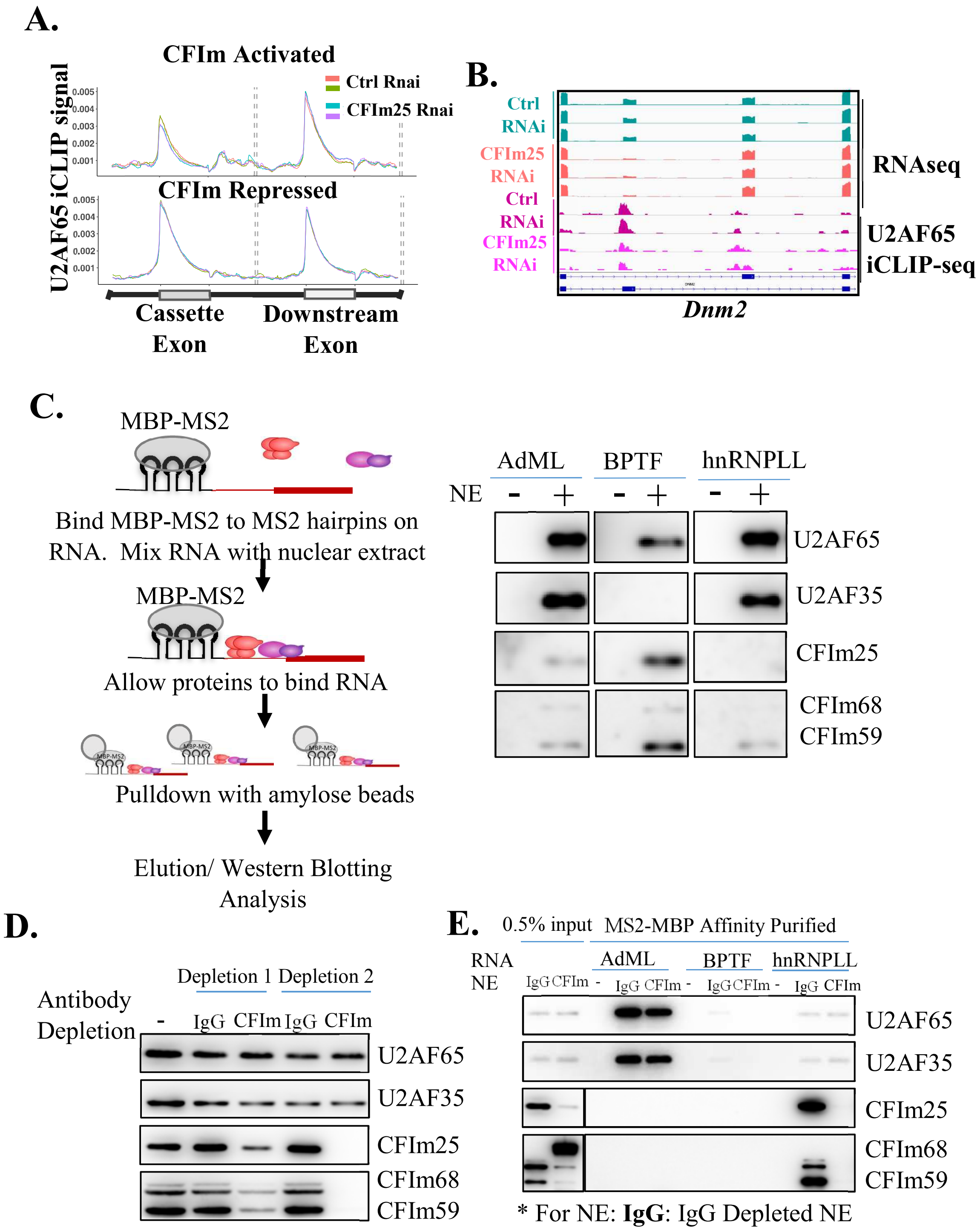
Figure 4.8 CFIm Binding Sites at BPTF Regulated Exon Affect Alternative Splicing

A. eGFP reporter splicing reporter with CFIm-regulated cassette exon of BPTF. Region has 3 putative CFIm binding sites (1, 2, and 3), a UUGG sequence that may also be bound by CFIm, and a G-rich polypyrimidine overlapping the first UGUA element.

B. Left: Table of single or double mutants created for eGFP BPTF reporter.
Right: RT-PCR analysis for wildtype and mutant eGFP reporters.

interaction of U2AF65 with RNA at 3’ splice sites. The decrease in U2AF65 binding caused by CFIm depletion is likely a result of loss of CFIm binding because U2AF65-RNA interactions of downstream exons is not affected (Figure 4.9A). This indicates that CFIm activated exons may require CFIm to recruit U2AF to nearby 3’ splice sites, which is interesting as CFIm activated exons had weaker polypyrimidine tracts than either control or CFIm repressed exons (Figure 4.4A). By contrast, there was no observed effect of U2AF65 binding on CFIm repressed exons, indicating that these may be indirect effects whereas the CFIm may directly recruit U2AF65 to CFIm activated exons (Figure 4.9A). An example of a gene that exhibits decreased U2AF65 binding upon knockdown of CFIm25 is Dmn2. Dmn2 contains a CFIm activated exon and U2AF65 iCLIP reveals that upon knockdown of CFIm25, there is decreased binding of U2AF65 to the 3’ splice site of the regulated exon (Figure 4.9B).

The role of CFIm in regulating U2AF65-RNA interactions was subsequently evaluated in vitro. Candidate CFIm activated and repressed exons were selected for MS2-MBP affinity purification. Publicly available PAR-CLIP data for CFIm59 and CFIm68 binding was analyzed to determine the region of BPTF (a CFIm activated exon) and hnRNP LL (a CFIm repressed exon) that were bound by CFIm. Based upon the PAR-CLIP analysis, for BPTF, a region from approximately 50 nucleotides upstream of the 3’ splice site to the first 40 nucleotides of the regulated exon were cloned downstream of 3 MS2 hairpins to allow for in vitro transcription of an the BPTF regulated exon tagged with MS2 hairpins. Similarly, for hnRNP LL, a CFIm repressed exon, a region from approximately 70 nucleotides upstream of the 3’ splice site to the first 30 nucleotides of the regulated exon was cloned. First, we tested whether these candidate RNAs are bound by CFIm and U2AF. RNA was in-vitro transcribed and used for an MS2-MBP affinity



* For NE: **IgG**: IgG Depleted NE
CFIm: CFIm Depleted NE

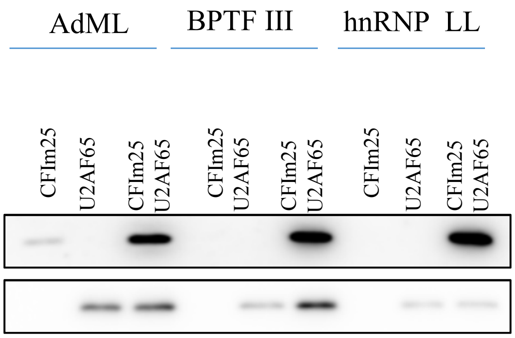
Figure 4.9 CFIm Regulates U2AF-RNA Interactions

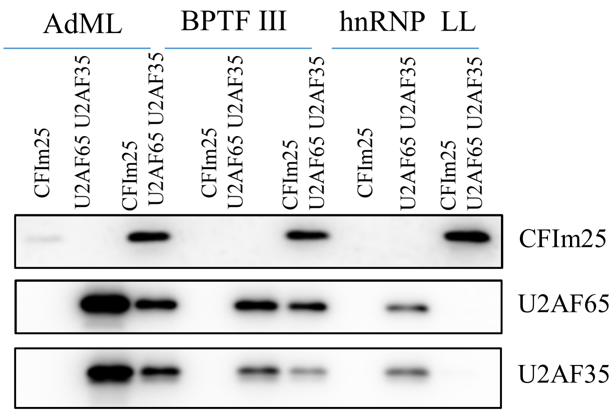
A. Genome-wide average of U2AF65 iCLIP-seq on either control or CFIm25 knockdown cells normalized to downstream exon. Exons are segregated as either CFIm activated or CFIm repressed.
B. Sequencing tracts comparing of RNA sequencing (top) and U2AF65 iCLIP-seq (bottom) upon knockdown of CFIm25 on example CFIm activated exon Dnm2.
C. Left: Schematic depicting MS2-MBP affinity purification strategy for CFIm activated and CFIm repressed exons.
Right: Western blotting analysis for MS2-MBP affinity purification of control exon AdML, CFIm activated exon BPTF, and CFIm repressed exon hnRNP LL 3' splice sites
D. Western blotting analysis depicting depletion of CFIm from HeLa cell nuclear extract. Nuclear extract was incubated with protein A/G agarose beads bound to either IgG or CFIm25 antibody for two rounds of depletion.
E. Western blotting analysis for MS2-MBP affinity purification of control exon AdML, CFIm activated exon BPTF, and CFIm repressed exon hnRNP LL 3' splice sites. Affinity purification was performed with either IgG or CFIm25 depleted nuclear extract.

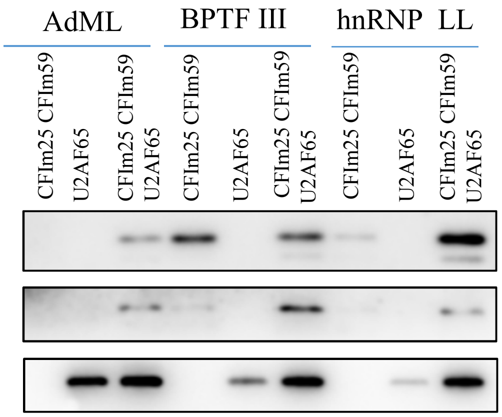
purification assay in which it was bound to MS2-MBP protein, incubated with HeLa cell nuclear extract and purified using amylose beads which bind to MS2-MBP. RNA and the associated proteins were then eluted with maltose. MS2-MBP affinity purification from nuclear extract indicated that both exons can be bound by all subunits of both CFIm and U2AF (Figure 4.9C). AdML, a commonly used in vitro splicing substrate, is used as a control as it contains a strong polypyrimidine tract that should not be dependent on CFIm. AdML was strongly bound by both U2AF65 and U2AF35 but more weakly bound by CFIm subunits, as would be predicted (Figure 4.9C).

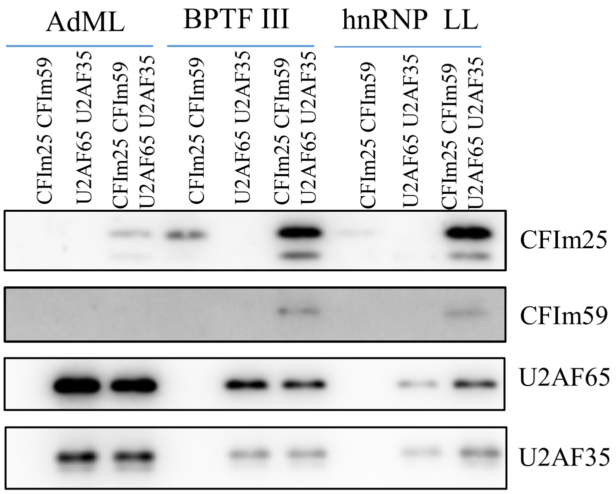
After determining that CFIm and U2AF can both bind to CFIm regulated exons, the binding of endogenous U2AF in the presence and absence of CFIm was compared. CFIm was depleted from HeLa cell nuclear extract by incubating with either a CFIm25 antibody or an IgG antibody covalently bound to protein A/G agarose beads. Two rounds of depletion were used to further reduce CFIm levels. Importantly, while CFIm was depleted from nuclear extract, other proteins including U2AF65 and U2AF35 were not co-depleted, indicating that depletion of CFIm was specific (Figure 4.9D). Following CFIm depletion, MS2-MBP affinity purification was performed for the AdML, BPTF, and hnRNP LL RNAs as above with either HeLA cell nuclear extract or an equal volume of CFIm-depleted nuclear extract. Western blotting analysis revealed that while both U2AF65 and U2AF35 binding to AdML and hnRNP LL RNAs are not affected by depletion of CFIm, depletion of CFIm reduces binding of both U2AF65 and U2AF35 to the BPTF RNA, indicating that CFIm may be important for recruiting U2AF to this region (Figure 4.9E)

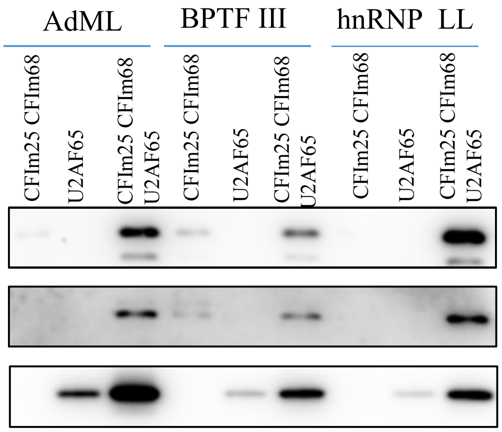
Constructs were then used for in-vitro pulldown assays using recombinant U2AF and CFIm to identify which subunits are responsible for interactions with RNA. Recombinant CFIm25 was


AdML
BPTF III
hnRNP LL
CFIm25
U2AF65
CFIm25 U2AF65
CFIm25
U2AF65
CFIm25 U2AF65
CFIm25
U2AF65
CFIm25 U2AF65


AdML
BPTF III
hnRNP LL
CFIm25
U2AF65 U2AF35
CFIm25 U2AF65 U2AF35
CFIm25
U2AF65 U2AF35
CFIm25 U2AF65 U2AF35
CFIm25
U2AF65 U2AF35
CFIm25 U2AF65 U2AF35
CFIm25
U2AF65
U2AF35


AdML
BPTF III
hnRNP LL
CFIm25 CFIm59
U2AF65
CFIm25 CFIm59 U2AF65
CFIm25 CFIm59
U2AF65
CFIm25 CFIm59 U2AF65
CFIm25 CFIm59
U2AF65
CFIm25 CFIm59 U2AF65


AdML
BPTF III
hnRNP LL
CFIm25 CFIm59
U2AF65 U2AF35
CFIm25 CFIm59 U2AF65 U2AF35
CFIm25 CFIm59
U2AF65 U2AF35
CFIm25 CFIm59 U2AF65 U2AF35
CFIm25 CFIm59
U2AF65 U2AF35
CFIm25 CFIm59 U2AF65 U2AF35
CFIm25
CFIm59
U2AF65
U2AF35


AdML
BPTF III
hnRNP LL
CFIm25 CFIm68
U2AF65
CFIm25 CFIm68 U2AF65
CFIm25 CFIm68
U2AF65
CFIm25 CFIm68 U2AF65
CFIm25 CFIm68
U2AF65
CFIm25 CFIm68 U2AF65

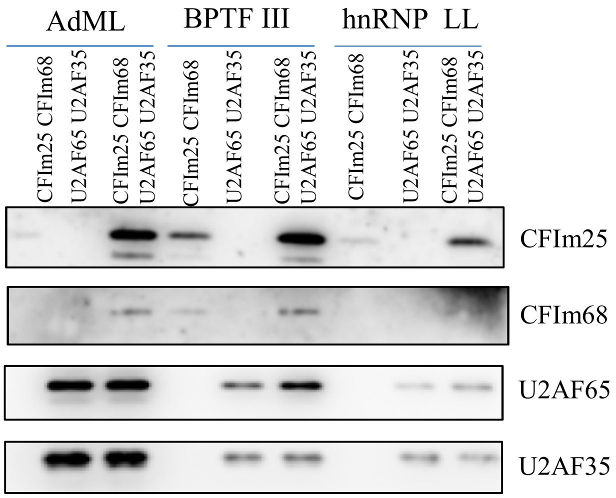

AdML
BPTF III
hnRNP LL
CFIm25 CFIm68
U2AF65 U2AF35
CFIm25 CFIm68 U2AF65 U2AF35
CFIm25 CFIm68
U2AF65 U2AF35
CFIm25 CFIm68 U2AF65 U2AF35
CFIm25 CFIm68
U2AF65 U2AF35
CFIm25 CFIm68 U2AF65 U2AF35
CFIm25
CFIm68
U2AF65
U2AF35

Figure 4.10 CFIm and U2AF Collectively Regulate RNA Binding

Western blotting analysis of MS2-MBP affinity purification of control AdmL, CFIm activated exon BPTF, and CFIm repressed exon hnRNP LL 3’ splice sites. RNA was incubated with recombinant CFIm25 alone (top) or either the CFIm25-CFIm59 (middle) or CFIm25-CFIm68 (bottom) complexes as well as either U2AF65 alone (left) or the U2AF65-U2AF35 complex (right).

purified from E. coli whereas recombinant CFIm complexes and U2AF were purified from Sf9 insect cells. For U2AF, strep-tagged U2AF65 (denoted U2AF65) was purified alone or in a complex with his-tagged U2AF35 (denoted U2AF65-U2AF35) was purified. For CFIm, his-tagged CFIm25 was purified as a single subunit (denoted CFIm25) or in a complex with CFIm59 or CFIm68 (denoted CFIm25-CFIm59 or CFIm25-CFIm68) (Figure 4.10). As with the experiments with endogenous CFIm and U2AF, RNA was bound by MS2-MBP, incubated with recombinant CFIm and U2AF, and purified using amylose beads. For both substrates, the presence of U2AF increases binding of CFIm, indicating that binding of CFIm is partially dependent on U2AF. We also see that U2AF increases binding of CFIm on the control substrate AdML. As AdML is not likely to be dependent on CFIm for splicing activation and binding of CFIm to the AdML RNA is quite weak in the absence of U2AF, this provides further evidence that U2AF may recruit CFIm through direct interactions (Figure 4.10).

There are also distinct differences between U2AF65 alone and the U2AF65-U2AF35 complex. For the CFIm activated substrate BPTF, while CFIm25 promotes U2AF65 binding, it decreases binding of the U2AF65-U2AF35 complex, suggesting that there may be some competition between CFIm25 and U2AF35. Similarly, for the CFIm repressed exon hnRNP LL, while U2AF35 increases U2AF65 binding to the substrate in the absence of CFIm25, when CFIm25 is present, binding of U2AF65 alone is unchanged while binding of both U2AF subunits is abolished in the presence of U2AF35 (Figure 4.10). Similar results are also seen for both CFIm heterodimers. For both BPTF and hnRNP LL, when either CFIm heterodimer is incubated with U2AF65, there is increased interaction of U2AF65 with the RNA, indicating that CFIm recruits U2AF65 to these 3' splice sites. However, when the CFIm heterodimers are incubated with the U2AF65-U2AF35 complex, there is a less substantial change in recruitment of U2AF to these

RNAs or no change (Figure 4.10). This indicates that CFIm likely promotes U2AF65 recruitment, but competes with U2AF35, potentially for interacting with U2AF65. Alternatively, U2AF65 binding may be stronger in the presence of U2AF35 than in the absence, making CFIm less important for U2AF65 binding when abundant U2AF35 is present. As U2AF65 interacts with CFIm through its RS domain and U2AF35 through its ULM and these domains are consecutive, it is possible U2AF65 cannot interact with both proteins at the same time, causing competition between the different proteins.

CFIm Competes for Interaction with U2AF

Unlike CFIm activated exons which are likely direct targets for CFIm binding and recruitment of U2AF, iCLIP revealed that CFIm repressed exons did not exhibit changes in U2AF65 binding upon knockdown of CFIm25, indicating potential indirect effects on alternative splicing (Figure 4.9A). One potential source of indirect effects could be explained by competition with other RNA binding proteins for interaction with U2AF65, similar to what was seen in in-vitro assays with U2AF35 (Figure 4.10). As U2AF65 contains a U2AF ligand motif (ULM), it interacts with U2AF homology motifs (UHMs) of many RNA binding proteins. This group consists of the small subunit U2AF35 as well as many others including Rbm39, Puf60 and Spf45, many of which are implicated in 3' splice site selection. For example, as many as 20% of 3' splice sites can be regulated by both U2AF and Rbm39 (Mai et al. 2016). While it was originally believed that these proteins only interact with U2AF65 through the U2AF65 ULM domain, recent studies indicate that the RS domain of U2AF65 also plays a role in interaction beyond its known role in nuclear localization (Stepanyuk et al. 2016; Kralovicova et al. 2015). As CFIm also interacts with U2AF65 using the RS domain, these proteins may be in direct competition with each other

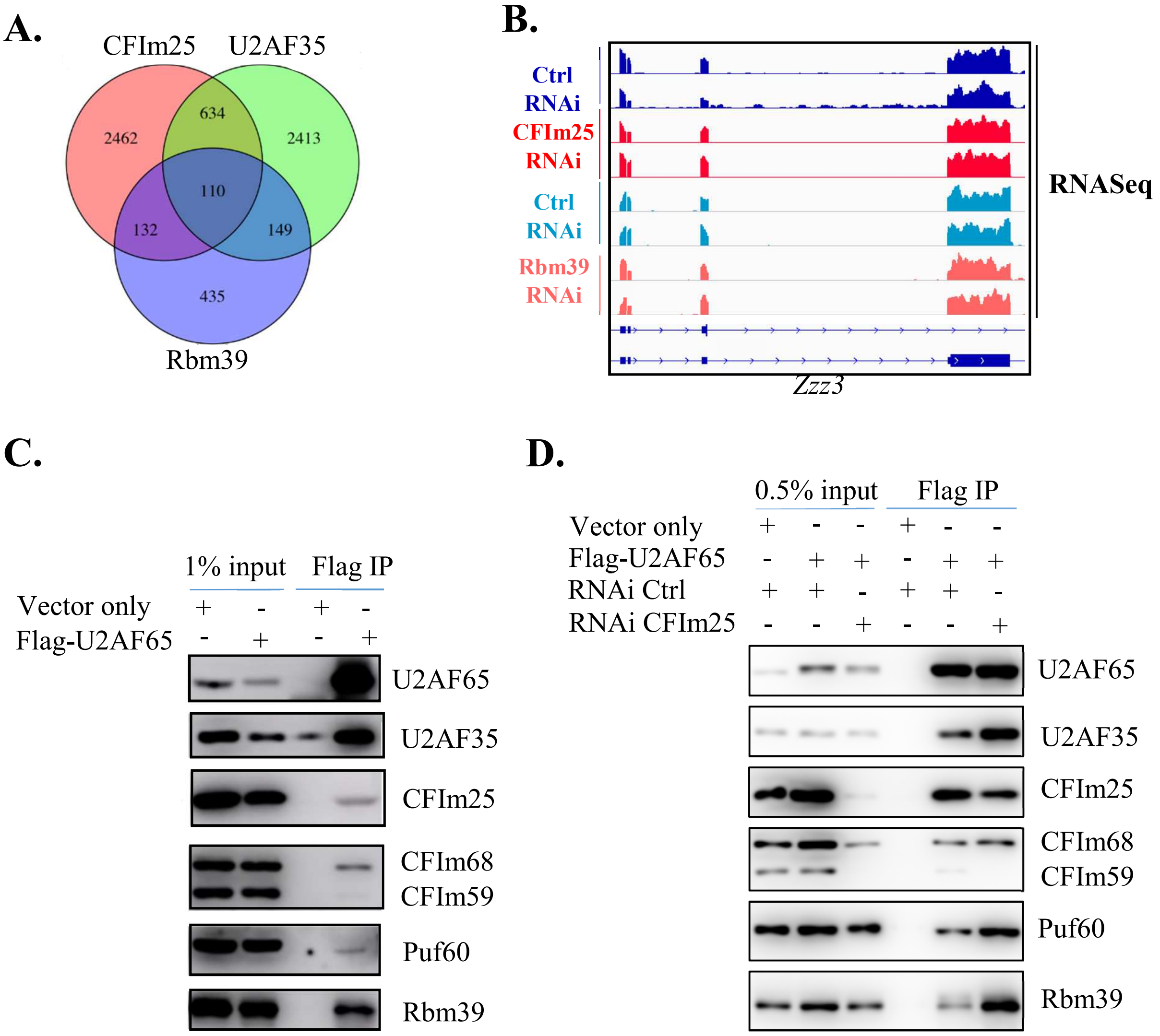


Figure 4.11 CFIm Competes with Other RNA Binding Proteins for Interaction with U2AF65

A. Overlap of RNA sequencing alternative splicing results upon knockdown of CFIm25, U2AF35, and Rbm39.
B. RNA sequencing tract for CFIm regulated exon in Zzz3 in CFIm25 and Rbm39 knockdown. Upon knockdown of CFIm25, there is increased exon inclusion but there is decreased exon inclusion upon knockdown of Rbm39.
C. Western blotting analysis of flag immunoprecipitation of U2AF65.
D. Western blotting analysis for flag immunoprecipitation of U2AF65 in either control or CFIm25 knockdown cells. Loading is normalized to FLAG-U2AF65 signal.

for interaction with U2AF65 and regulation of alternative splicing. A competition model could also explain why hnRNP LL (a CFIm repressed exon) and BPTF (a CFIm activated exon) showed similar results from in vitro assays (Figure 4.10); because only CFIm and U2AF were present and there were no other RBPs to compete with CFIm for binding 3' splice sites, CFIm was still able to bind to the 3'ss of hnRNP LL.

Alternative splicing changes upon knockdown of CFIm25 were compared with published datasets for Rbm39 and U2AF35 knockdown. While there are over 100 alternative splicing changes shared among the three conditions, the majority of alternative splicing changes upon knockdown of CFIm25 were unique (Figure 4.11A). Shared changes have a higher likelihood of being indirect whereas splicing changes unique to CFIm25 are more likely to be direct. The overlap in alternative splicing changes upon knockdown of CFIm25 with knockdown of Rbm39 also indicates that they may compete for 3' splice site selection. An example of a gene that demonstrated potential competition between CFIm25 and Rbm39 is Zzz2. Upon knockdown of CFIm25, there was increased exon inclusion but upon knockdown of Rbm39 there was decreased exon inclusion (Figure 4.11B). This indicates that CFIm may inhibit inclusion of the exon by interfering with 3' splice site selection by Rbm39.

To verify that these additional RNA binding proteins interact with U2AF65, U2AF65 FLAG immunoprecipitation was performed. In addition to interacting with U2AF35 and CFIm, U2AF65 also interacted with Rbm39 and Puf60 (Figure 4.11C). Next, to investigate whether CFIm and other RNA binding proteins are in competition for binding to U2AF65, U2AF65 FLAG immunoprecipitation was performed in either control or CFIm25 knockdown cells. Upon knockdown of CFIm25, there was increased interaction of U2AF65 with U2AF35, Puf60, and

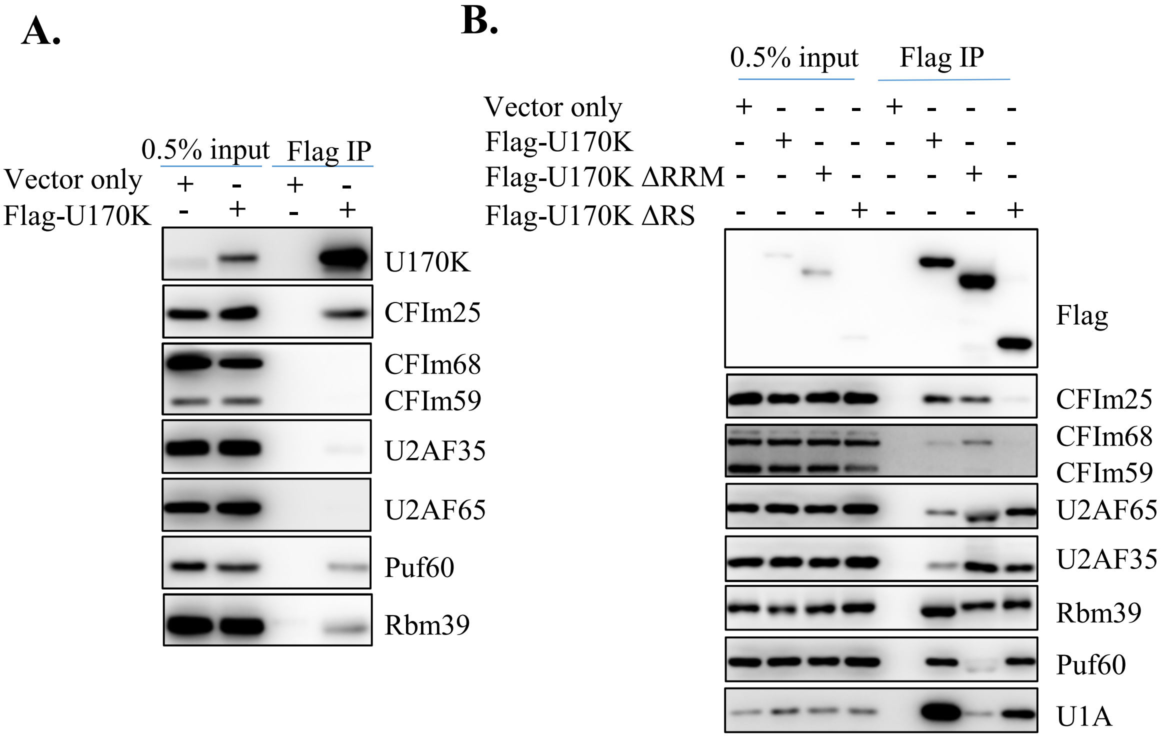


Figure 4.12 CFIm Interacts with U170K Through the RS Domain

A. Western blotting analysis of U170K flag immunoprecipitation.
B. Western blotting analysis of U170K flag-immunoprecipitation with either full-length U170K or U170K lacking a single domain. Loading is normalized to Flag U170K signal.

Rbm39 (Figure 4.11D). This supports a competition model because depletion of one competitor (CFIm) increased the binding of other competitors for U2AF65.

CFIm Binds Near 5' Splice Sites and Interacts with U170K

Analysis of PAR-CLIP data revealed that in addition to binding near 3' splice sites genome-wide, both CFIm59 and CFIm68 also bind near 5' splice sites (Figure 4.6B). 5' splice sites are recognized by the U1 snRNP through base pairing of the U1 snRNA with the 5' splice site. One of the proteins that composes the U1 snRNP is U170K, which, similar to U2AF at 3' splice sites, is recruited to 5' splice site by SR proteins bound to exonic splicing enhancers (Cho et al. 2011). As SR proteins can interact with both U2AF and U170K and thereby regulate both 3' and 5' splice sites, CFIm may also regulate 5' splice site selection by interacting with U170K in addition to regulating 3' splice site selection by interacting with U2AF.

To investigate whether CFIm and U170K interact, a U170K FLAG immunoprecipitation was performed. Western blotting analysis revealed that in addition to interacting with known interaction partners including Rbm39 and U2AF35, U170K interacted with CFIm25, CFIm59, and CFIm68, with stronger interaction with CFIm25 and CFIm68 than with CFIm59 (Královičová et al. 2018; Campagne et al. 2022; Day et al. 2012) (Figure 4.12A).

U170K is composed of an N-terminal RRM and two C-terminal RS domains. To determine the mechanism for the CFIm-U170K interaction, either both RS domains or the RRM domain of U170K was deleted and the FLAG U170K flag immunoprecipitation was repeated. Western blotting analysis revealed that the RS domain of U170K was necessary for interacting with CFIm because there was decreased interaction of all CFIm subunits with U170K when the RS domain was deleted. However, deletion of the RNA recognition motif had no effect on coimmunoprecipitation of CFIm, indicating that it is not necessary for interaction (Figure

4.12B). Deletion of neither domain of U170K completely eliminated the interaction of U170K with Rbm39. However, both domains may contribute to the interaction because both deletions RRM and RS domain deletions decreased the level of interaction with Rbm39, consistent with previous findings that the RRM domain of U170K is important for the interaction but it is supported by additional RS/RS interactions (Královičová et al. 2018; Campagne et al. 2022). Additionally, FLAG immunoprecipitation reveals that U170K also interacts with Puf60, potentially through the RNA recognition motif, a finding that has not been previously reported.

4.4 DISCUSSION

In this chapter, CFIm, a polyadenylation factor, is identified as an alternative splicing regulator involved in 3' splice site selection. When CFIm25 was knocked down in 293T cells, there were widespread changes in alternative splicing, with cassette exons that showed increased and decreased exon inclusion. Mechanistically, the RS domain of CFIm interacts with the RS domain of U2AF65. By doing so, CFIm activates a subset of cassette exons by recruiting U2AF to non-canonical 3' splice sites that are C-rich and have weaker polypyrimidine tracts. By contrast, CFIm repressed exons may be indirectly regulated as CFIm competes with other RNA binding proteins including Rbm39 and Puf60 for interaction with U2AF65 and therefore 3' splice site selection. As a result, when CFIm is depleted, there is increased exon inclusion for these subsets of cassette exons.

Our findings are novel for several reasons. The accepted model for 3' splice site selection is that it is regulated by the U2AF complex consisting of U2AF65 and U2AF35, with U2AF65 binding to polypyrimidine tracts and U2AF35 binding to the AG dinucleotide at the intron-exon boundary. However, this does not explain how U2AF65 can recognize a wide variety of

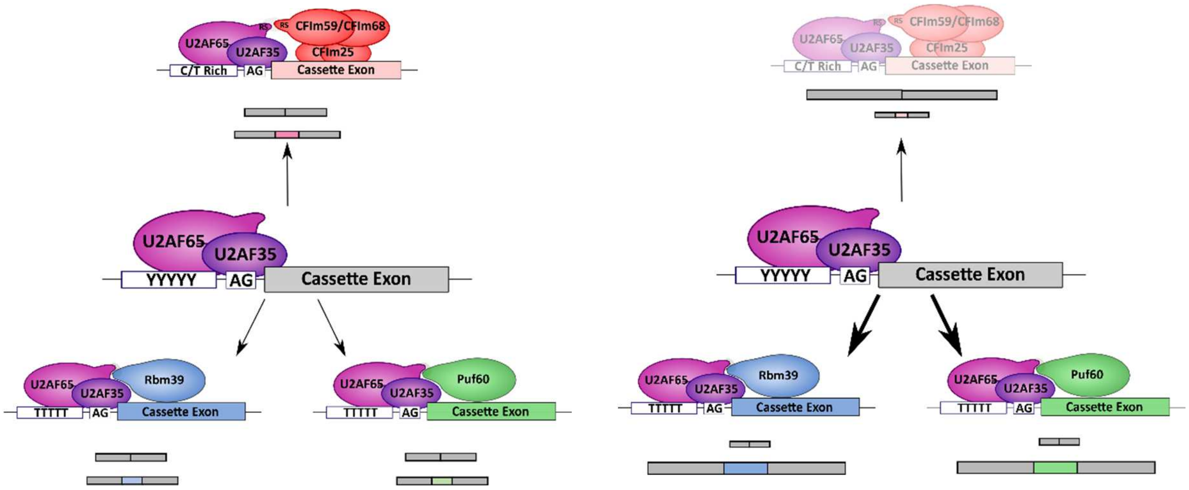


Figure 4.13 CFIm and Other RNA Binding Proteins Compete for Interaction with U2AF65 and 3' Splice Site Selection

3' splice site selection is mediated by the U2AF complex in combination with a variety of RNA binding proteins including CFIm, Rbm39, and Puf60. Each subcomplex recognizes a subset of 3' splice sites and therefore regulates the inclusion of this subset of cassette exons. However, when one factor is depleted as in the case of CFIm25 knockdown, there is decreased exon inclusion for this subset of cassette exons. At the same time, there is now more U2AF available to bind other competitors including Rbm39 and Puf60, leading to increased 3' splice site selection and therefore increased cassette exon inclusion for these subsets.

polypyridine tract sequences, especially non-canonical sequences. While there have previously Puf60 (T. Wu and Fu 2015), our competition model of 3' splice site selection is unique in that it been reported individual cases of proteins regulating alternative 3' splice site selection including suggests that there are many RNA binding proteins that can all regulate a subset 3' splice sites and promote the inclusion of these cassette exons. When one of these proteins is depleted, as in the case of CFIm25 knockdown, there are at least two types of changes. First, for 3' splice sites that are recognized by CFIm in combination with U2AF, there is decreased 3' splice site recognition and therefore decreased exon inclusion for this subset of cassette exons. At the same time, there is now more U2AF available to bind to other RNA binding proteins, leading to increased 3' splice site selection and these subsets of cassette exons experience increased exon inclusion (Figure 4.13). Our data expands on the classical model by not only explaining how a single protein such as CFIm can simultaneously activate a subset of cassette exons and repress others, but also by demonstrating the important role of U2AF binding partners in 3' splice site selection particularly in cases with non-canonical splice sites.

This model is particularly relevant in the context of cancer development. One of the hallmarks of cancer is mis-regulated splicing, with analysis of over 8,000 tumors from 32 cancer types revealing thousands of alternative splicing events not found within non-tumorigenic tissues (Bonnal, López-Oreja, and Valcárcel 2020). RNA binding proteins are known to be a key family of proteins that are dysregulated during tumor progression, with RBPs showing significantly increased and decreased expression. In fact, several studies in different cancer types indicate that differential expression of RBPs can potentially be used for a diagnostic for cancer as well as cancer prognosis (Qin et al. 2020; J. Li et al. 2021; Kang, Lee, and Lee 2020). For example, CFIm levels have been shown to be downregulated in glioblastoma leading to enhanced tumorigenic properties, Puf60 is overexpressed in breast cancer with Puf60 high-expressing patients having lower survival than Puf60 low-expression patients, and Rbm39 regulates splicing of transcription regulators in acute myeloid leukemia (AML) with its knockout increasing overall survival of AML patients (Xu et al. 2021; D. Sun et al. 2019; Masamha et al. 2014). The competition model for 3' splice site selection emphasizes that it is not just the expression of different RNA binding proteins that is critical during cancer progression but also their relative expression to one another to prevent aberrant alternative splicing changes.

There are also still many unanswered questions. First, it will be essential to determine the precise relationship between U2AF35 and CFIm. In-vitro assays suggest that the U2AF65-U2AF35 complex behaves differently from U2AF65 by itself because there is increased binding of U2AF65 to reporters RNAs in vitro when CFIm is present, but this same increase is not apparent when the U2AF65-U2AF35 complex is incubated with the same reporter RNAs. Additionally, in vivo there is increased interaction of U2AF35 to U2AF65 upon depletion of CFIm (Figure 4.10). Together this indicates that CFIm competes with U2AF35 for binding to U2AF65 and that U2AF65 has increased dependence on CFIm for binding to 3' splice site when U2AF35 is absent. Notably, U2AF35 is one of the most frequently mutated splicing protein in cancer, with mutational hotspots at S34 and Q157 (Y. Zhang et al. 2021). As a result, CFIm may be particularly essential for 3' splice site selection during conditions such as cancer when U2AF35 is mutated and splicing patterns have been altered.

Our in vitro assays also found few differences between the interactions of U2AF and CFIm on hnRNP LL and BPTF, which was a surprising finding as hnRNP LL is a CFIm repressed exon whereas BPTF is a CFIm activated exon (Figure 4.10). However, this may simply be a caveat of performing this experiment in vitro with recombinant protein. By the competition model, CFIm represses exons by limiting the levels of U2AF available to interact with other proteins such as Puf60 and Rbm39. However, the in vitro system only contains CFIm and U2AF, removing any potential competition with other RNA binding proteins. As a result, if CFIm can bind to the hnRNP LL exon or 3' splice site region, it may still recruit U2AF as they interact. One way to further test the competition model would be to test for the levels of Puf60 and Rbm39 that bind to the BPTF and hnRNP LL RNAs in control and CFIm depleted nuclear extract. If the competition model is true, we would predict to see increased levels of Puf60 and Rbm39 binding to hnRNP LL when CFIm is depleted but not necessarily to BPTF.

Next, our preliminary data suggests that CFIm also binds 5' splice sites and interacts with U170K, suggesting that in addition to regulating 3' splice site selection, CFIm regulates 5' splice sites. It remains to be determined whether CFIm regulates both 5' and 3' splice sites equally, or whether CFIm binding at either the 5' or 3' splice site has a stronger influence on alternative splicing. It has been hypothesized that both SR proteins and Rbm39 bridge 5' and 3' splice sites, which in the case of Rbm39 is transcription dependent (Královičová et al. 2018; Day et al. 2012; Campagne et al. 2022). As CFIm also interacts with both U2AF65 and U170K, it raises the question of whether CFIm may also serve as a bridge between splice sites. It also remains to be determined whether multiple CFIm complexes are required or if a single CFIm binding event may be sufficient. CFIm is a heterodimer of two large subunits (CFIm59 or CFIm68) and two small subunits (CFIm25), which raises the question of whether because there are two large subunits in each complex, the two large subunits can simultaneously activate both 5' and 3' splices sites by recruiting U1 and U2AF (Figure 4.14). In addition, if CFIm can bridge 5' and 3' splice sites, it may suggest that there are many other SR-like proteins that function similar to SR proteins in alternative splicing regulation even if they have additional, previously described

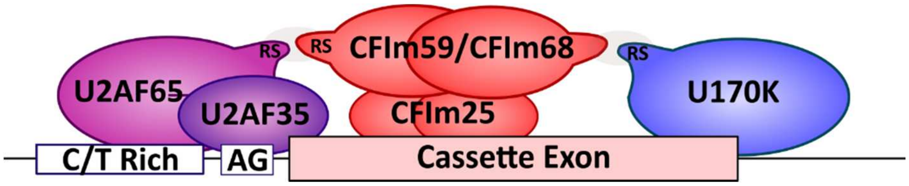


Figure 4.14 CFIm May Bridge 3’ and 5’ Splice Sites by Interacting with U2AF and U170K

CFIm interacts with both U170K and U2AF through the RS domain of the large subunit. CFIm is a heterotetramer and therefore contains 2 copies of either CFIm59 or CFIm68. One explanation for the ability of CFIm to interact with proteins at both the 5’ and 3’ splice sites is that one copy of the large subunit interacts with each protein.

functions. In 2001, a genome-wide survey in metazoans estimated that there are over 240 proteins containing an RS domain, including proteins involved in transcription, chromatin remodeling, phosphorylation, and cell structure (Boucher et al. 2001), highlighting the vast number of proteins that may have an uncharacterized role in splicing regulation.

Finally, as with the case of polyadenylation, CFIm59 and CFIm68 may have unique roles in splice site selection. Our data suggests that CFIm59 and CFIm68 affect alternative cassette exons different, as CFIm68 interacts with CFIm activated exons with a PAR-CLIP peak near the 3' splice site whereas CFIm59 interacts with CFIm repressed exons with highest binding near the 5' splice site (Figure 4.7D). Additionally, CFIm68 knockout inhibited interaction of U2AF with CFIm25 whereas CFIm59 had no effect (Figure 4.8B) and CFIm68 may also have stronger interaction with U170K than CFIm59 (Figure 4.12). Together, this data indicates that the CFIm25-CFIm68 complex may be responsible for alternative splicing regulation, which is consistent with reports that the CFIm25-CFIm68 has a stronger effect on alternative polyadenylation (Zhu et al. 2017). This also suggests that CFIm59 may repress exons by reducing the amount of CFIm25 available to interact with CFIm68.

In conclusion, this chapter establishes that CFIm is not merely a regulator of alternative last exons but is in fact a general alternative splicing regulator that can both activate and repress exons. We also provide evidence that the model for 3' splice site selection should be changed to recognize the importance of RNA binding proteins in recognizing non-canonical 3' splice sites as CFIm is just one example of an RNA binding protein that can recruit U2AF to weak polypyridine tracts. These U2AF complexes each recognize a specific subset of polypyrimidine tracts and compete for interacting with U2AF, leading to widespread changes in alternative splicing when a competitor is depleted. Finally, these findings indicate the importance of CFIm, a polyadenylation regulator, in alternative splicing, further highlighting the interconnected relationship between splicing and polyadenylation regulation.

4.4 METHODS

Knockdown of CFIm

CFIm25, CFIm59 and CFIm68 knockdown cells were generated from HEK 293T cells using lentiviral transduction of the pLKO.1 vector containing shRNA template followed by 1.25mg/mL puromycin. Cells were harvested for assays 5 days post-transduction unless otherwise noted within experimental design.

shRNA sequences: CFIm25: 5’ GAACCTCCTCAGTATCCATAT 3’
CFIm25: 5’ TGTACCCTCTTACCAATTATA 3’
CFIm59: 5’ AAGATATCATGAAGCGAAACA 3’
CFIm68: 5’ GGTGATTATGGGAGTGCTATT 3’
CFIm68: 5’ GTTGTAACTCCATGCAATAAA 3’

Sequencing and Bioinformatic Analysis

RNA Sequencing

CFIm25 was knocked down using above protocol. For CFIm25 knockdown, CFIm59 knockout, and CFIm68 knockout, RNA was extracted from cells using TRIzol (Invitrogen) and purified using standard protocol. Paired end libraries were prepared by UCI Genomic High-Throughput Facility (GHTF) using Illumina TruSeq Stranded Kit and sequenced using HiSeq4000. Samples were split between 2 lanes.

rMATS Analysis of Alternative Splicing

Alternative splicing was analyzed from RNA sequencing for CFIm25 knockdown, CFIm59 knockout, CFIm68 knockout, U2AF35 knockdown, and Rbm39 knockdown. CFIm datasets were generated from this study whereas Rbm39 and U2AF35 were previously published. RNA sequencing libraries were aligned to the human genome using STAR (Dobin et al. 2013). Next, alternative splicing was analyzed using rMATS, a bioinformatic analysis that uses RNA-Seq reads mapped to different mRNA isoforms to estimate isoform proportion. rMATS pipeline can identify alternative cassette exons, alternative 5' or 3' splice sites, mutually exclusive exons, and intron retention. rMATS calculates percent spliced in (PSI) values as the percentage of exon inclusion transcripts divided by the total of exon inclusion and exon exclusion transcripts. With the example of alternative cassette exons, exon inclusion reads include reads from the upstream splice junction, the alternative exon, and the downstream splice junction. Exon skipping reads are reads that directly connect the upstream exon to the downstream exon. rMATS then uses a binomial distribution model to estimate the uncertainty of the PSI value.

To identify alternative splicing between conditions, rMATS compares the binomial distribution for the uncertainty in replicates and the normal distribution for variability among replicates. It then uses a likelihood-ratio test to calculate the likelihood that for each alternative splicing event between conditions, there is a ΔPSI value greater than 0.15 (S. Shen et al. 2014). Alternative cassette exons with $\Delta PSI > 0.15$ or < -0.15 and $FDR < 0.05$ were considered significant.

Analysis of Alternative Cassette Exon Features

To identify sequence features of alternative cassette exons, rMATS was first used to identify exons with increased inclusion (CFIm repressed exons), exons with decreased inclusion (CFIm activated exons) or exons with no change in splicing upon knockdown of CFIm25. Matt was then used to analyze sequence features of regulated exons. Matt is a UNIX toolkit for alternative splicing analysis that utilizes alternative splicing data to analyze features such as motif enrichment, 5' and 3' splice site strength, GC content and exon length. Features are extracted from each subset of exons and Mann-Whitney U tests are used to compare feature distributions between groups (Gohr and Irimia 2019). The command get_vast was used to extract alternative splicing events, ΔPSI values, and gene IDs. Next, def_cats was used to identifying CFIm activated, CFIm repressed, and unregulated exons.

Sequence enrichment analysis from 100 nucleotides upstream to 20 nucleotides downstream of the 3' splice site as well as from 20 nucleotides upstream to 100 nucleotides downstream of the 5' splice site were calculated using Matt command get_seqs.

Exon length, 5' and 3' splice site strength were all calculated using Matt command get_efeatures and box plots comparing the distributions of each feature were generated using cmpr_exons.

CFIm25 Knockdown and U2AF65 iCLIP Library Preparation and Analysis

iCLIP was performed according to previously published protocol (Konig et al. 2011). CFIm25 was knocked down using above protocol. 5 days post-transduction, cells were washed with PBS and irradiated with 150mJ/cm^2 at 254nm in a Stratalinker. -UV crosslinking and IgG antibody were used as controls.

Cells were resuspended in lysis buffer (50mM Tris-HCl pH 7.4, 100mM NaCl, 1% NP-40, 0.1% SDS, 0.5% sodium deoxycholate) and sonicated at 1 Amp for 1 second. Lysate was digested with 10ul RQ1 DNAse and 10ul of 1:25 diluted RNase I at 37°C 1200rpm for 5 minutes. Digested lysate was centrifuged at 15,000xg for 20 minutes.

For immunoprecipitation, 80ul of Protein A Dynabeads was added to lysis buffer with 5ug of U2AF65 antibody (Sigma U4758) and rotated at room temperature for 45 minutes followed by washing. Cell lysate was mixed with antibody-bound beads and rotated at 4°C for 4 hours followed by 2x wash with high salt wash buffer (50mM Tris HCl pH 7.4, 1M NaCl, 1mM EDTA, 1% NP-40, 0.1% SDS), 1x wash with wash buffer (50mM Tris-HCl pH 7.4, 10mM $MgCl_2$, 0.2% Tween-20), and rinse with PNK buffer (70mM Tris-HCl pH 7.4, 10mM $MgCl_2$). 3' cyclic phosphates on RNA were removed by treating with T4 PNK (NEB) at 37°C for 20 minutes followed by 2x wash with high salt wash buffer, 1x wash with wash buffer and rinse with RNA ligase buffer (50mM Tris-HCl pH 7.5, 10mM $MgCl_2$). To ligate 3' linker, beads were incubated at 16°C overnight with T4 RNA ligase and iCLIP 3' RNA linker.

Beads were washed with wash buffer and rinsed with PNK Buffer followed by addition of 10 units of T4 PNK and 20uCi of [γ-^{32}P]ATP and incubation at 37°C for 10 minutes. 2ul of 10mM ATP were added for 10 minutes followed by wash with wash buffer. SDS loading dye was added to beads and heated at 70°C for 5 minutes to elute sample, which was run at 140V for approximately 2 hours on Nu-Page Bis-Tris 4-12% precast gel (Invitrogen) in NuPage MOPS SDS Running Buffer. Sample was transferred to nitrocellulose at 390A for 1hr at 4°C and membrane was exposed to phosphor screen. Following imaging, the region 20-70kDa above the protein was removed and digested with 200ug of protease K at 37°C 1200rpm for 20 minutes. PK/Urea buffer (PK buffer + 7M Urea) was added at 37°C 1200rpm 20 minutes to extract RNA before phenol chloroform extraction and ethanol precipitation.

RNA was pelleted and reverse transcribed by SuperScript III (Invitrogen) following manufacturer's protocol with RT Primer that contains 2 cleavage adapter regions and a barcode. RNase A/T1 was used to digest at 37°C for 20 minutes followed by ethanol precipitation.

Remove free RT primer, cDNA pellet was resuspended in H_2O and loading dye, heated at 75°C, and run on 8% urea page gel at 800V for 40 minutes. The region from 80-300 nucleotides was extracted and eluted in TE (10mM Tris-HCl pH 7.4, 1mM EDTA) at 37°C 1100 rpm for 2 hours followed by ethanol precipitation.

To ligate primer to 5' end of cDNA, cDNA pellet is circularized using CircLigase II (Lucigen). After 2hrs at 60°C, BamHI annealing oligo was annealed by heating at 95°C for 2 minutes followed by cooling to room temperature. cDNA was then digested with BamHI (NEB) at 37°C for 30 minutes to linearize and ethanol precipitated.

cDNA was amplified using barcode adapter primers and Phusion Mastermix (NEB) with annealing at 65°C for 30 cycles. To prepare libraries, PCR products were purified using Ampure XP beads (Beckman Coulter) following manufacturer's protocol. Libraries were sequenced at the UCI Genomics High Throughput Facility.

The CLIP Tool kit (CTK) was used to map protein-RNA interactions genome-wide. CTK is a software package that analyzes CLIP data from raw reads by filtering and mapping reads, collapsing PCR duplicates, and calling peaks (Shah et al. 2017). Reads were aligned with the CTK tool Burrows Wheeler Aligner. After PCR duplicates were collapsed, CTK performed peak calling using a "valley seeking" algorithm that calculates the number of overlapping CLIP tags at each genomic position, finds a local maximum, and calls two local maxima as different peaks only if they are separated by a valley with sufficient depth based upon the height of the two peaks and the read coverage of the valley. This improves peak calling as CLIP peaks do not always have clear separation, especially if a transcript is abundant.

Next, CFIm activated and repressed exons as determined by rMATS were compared with CLIP data using deepTools to identify binding of U2AF65 along CFIm activated or CFIm repressed exons genome-wide (Ramírez et al. 2014). For each exon, a region from 500 nucleotides upstream of the 3’ splice site to 500 nucleotides downstream of the 5’ splice site was selected and iCLIP signal at each position within the window was analyzed. A similar analysis was performed for exons downstream of each regulated exon. Finally, U2AF65 iCLIP signal was compared between CFIm25 and control knockdown and a p value was calculated for each position within the window.

PAR-CLIP Analysis

PAR-CLIP data for CFIm59 and CFIm68 was downloaded from Martin et al and processed using deepTools to find average binding genome-wide (G. Martin et al. 2012; Ramírez et al. 2014). 3’ UTRs were excluded and the proportion of CFIm binding in the region 10k upstream of a gene, 5’ UTR exons, CDS exons, introns, 3’ UTR exons, 10k downstream of a gene, and intragenic regions was calculated.

Next, all CFIm regulated exons as identified by rMATS were compared with PAR-CLIP data using deepTools to identify binding of CFIm along regulated exons genome-wide (Ramírez et al. 2014). For each exon, a region from 500 nucleotides upstream of the 3’ splice site to 500 nucleotides downstream was selected and PAR-CLIP signal at each position within the window was analyzed.

RT PCR Analysis

Endogenous Splicing

CFIm25, CFIm59, and CFIm68 were knocked down using above protocol.

For CFIm25 knockdown and endogenous splicing, 293T cells were transduced with lentivirus and RNA was extracted using standard TRIzol (Invitrogen) protocol 5 days post transfection and cDNA was generated using All-In-One 5X RT MasterMix (Abmgood). RT-PCR was performed using primers that bind to the exons upstream and downstream of the cassette exons and run on a 2% agarose gel to resolve.

For CFIm59 and CFIm68 co-depletion RT PCR, CFIm59 knockout 293T cells were transduced with CFIm68 lentivirus or CFIm68 knockout 293T cells were transduced with CFIm59 lentivirus. cDNA synthesis and RT-PCR were performed as above.

OSBPL6 regulated exon F:	ATCTTGCACATTGCCAGTC
OSBPL6 regulated exon R:	CTTTCTGTGCGATTAGACAAAA
Gpbp1 regulation exon F:	CCGTCTTTAAATCCTGAGTATGAG
Gpbp1 regulated exon R:	TTTTGTAGGTGGAGCAGCAG
Hmgxb4 regulated exon F:	GATAGTGAACTTTACTTCTTGGGGA
Hmgxb4 regulated exon R:	TTTCATTTTTAAACCATCAGGCT
Acin1 regulated exon F:	ATTGGTGAGGAAATGAGCCA
Acin1 regulated exon R:	CTGTCTGACCCTAGATGATCG
BPTF regulated exon F:	AAGAAATTTTGGAATCCATAAGAGC
BPTF regulated exon R:	CTATCTCTTCCTGAATAGAGACAGG
Lrrfip2 regulated exon F:	CCTTCATCTCGAAATTCTGCC
Lrrfip2 regulated exon R:	CTTTGATTTTTCTTCATTTTCTCTATAAAATT
Uap1 regulated exon F:	GAAGTTTGTGGTATATGAAGTATTGC
Uap1 regulated exon R:	TTCTCCAGCATAGGAGATAAGAG
Cask regulated exon F:	AGGGAAATGCGGGGGA
Cask regulated exon R:	CTGTCGTCCTTTTGGTTGG
hnRNP LL regulated exon F:	CCAACTCGTCTAAATGTTATTAG
hnRNP LL regulated exon R:	CCATAGCCATCATGTCTAAA
Trmu regulated exon F:	GGTTACCAGGTGACAGGG
Trmu regulated exon R:	CATTTCTAACTTCAAACCGATTTCT

Wildtype eGFP Reporter Splicing

Split eGFP reporter was generated from pcDNA3.1 plasmid using EcoRI and XhoI restriction sites. CFIm-regulated cassette exons as well as upstream and downstream intronic regions were cloned between 2 regions of an eGFP gene sequence to interrupt the eGFP reading frame. For amplification of CFIm regulated exons, genomic DNA was purified from 293T cells using DNA QuickExtract (Lucigen) and amplified using the primers below.

293T cells were transduced with lentivirus for CFIm25 knockdown following protocol above. 3 days post-transfection, cells were transfected with 250ng eGFP reporter using PEI per 24 well. 5

days post-transduction, RNA was extracted using standard TRIzol (Invitrogen) protocol and cDNA was generated using All-In-One 5X RT MasterMix (Abmgood). RT-PCR was performed with eGFP primers and resolved on a 2% agarose gel.

hnRNP LL eGFP F:	CACGAATTCATTTGGGAAAATTGGCATGC
hnRNP LL eGFP R:	GCTCTCGAGCCAATACAACAATTCATAAAAGAAA
Trmu eGFP F:	CACGAATTCAGAGGCAGGGTTTTAGTG
Trmu eGFP R:	GCTCTCGAGTCACGCTTGTAATCTCAGC
Cask eGFP F:	CACGAATTCAGGGTTTTAATATCAAGTCACTC
Cask eGFP R:	GCTCTCGAGTACATCCTTTGCATAGGTTTGA
Osbpl6 eGFP F:	CACGAATTCAGCTTTGAAGATATGCTTTAGATAC
Osbpl6 eGFP R:	GCTCTCGAGTTTTAAGGTAACTTTTTGTTAAAATATATTTTAATG
Gpbp1 eGFP F:	CACGAATTCAAACACATGTTCATTTGTATTTTTCT
Gpbp1 eGFP R:	GCTCTCGAGCTGAAACAGTAGAACCAAAATCT
Acin1 eGFP F:	CACGAATTCTGCTTTCATAGCCCATGAAG
Acin1 eGFP R:	GCTCTCGAGCTGGTGAAGAGATGAAAAAGA
BPTF eGFP F:	CACGAATTCTGCCATTTCCCTATGAAAAAGA
BPTF eGFP R:	GCTCTCGAGAATTAAAACCAAAATATCTGCCACTATG
Lrrfip2 eGFP F:	CACGAATTCATATAGCTTTTTCCTCCATATATAGC
Lrrfip2 eGFP R:	GCTCTCGAGATATAGCTTTTTCCTCCATATATAGC
Uap1 eGFP F:	CACGAATTCTACTGGTACATAACTGGTTTTACAA
Uap1 eGFP R:	GCTCTCGAGCCCTGACCTGTAACCTG

eGFP RT PCR F:	GGCAAGCTGACCCTGA
eGFP RT PCR R:	CTTGTCGGCCATGATATAG

BPTF Mutant eGFP Reporter Splicing

CFIm binding sequences surrounding regulated exons were determined by comparing CFIm25 knockdown rMATS analysis with CFIm59 and CFIm68 PAR-CLIP. CFIm regulated exons were classified as either CFIm activated or CFIm repressed exons and compared with PAR-CLIP data using DeepTools to identify CFIm59 and CFIm68 binding motifs along activated and repressed exons genome-wide.

Upon determining CFIm binding sites, a series of eGFP BPTF mutant reporters were generated from eGFP BPTF splicing reporter above. Mutations included mutating the C rich region in the polypyrimidine tract to Gs (G -> C), mutating the second UGUA to UCUA (UGUA2 -> UCUA), mutating the third UGUC to both CAUC or UCUA (UGUA3 -> CAUC and UGUA3 -> UCUA), and mutating the exonic TTGG sequence to CCGG (TTGG -> CCGG). In addition to single mutations, double mutations of G->C and UGUA2 -> UCUA as well as UGUA3 -> CAUC and TTGG -> CCGG were created to compare the effects of mutating more than one CFIm binding site on splicing.

Single mutations were created by PCR linearization using eGFP BPTF reporter as a template. Primers were 5' phosphorylated by T4 PNK (Fisher) according to manufacturer's protocol. Following PCR amplification with Phusion HF Mastermix (NEB) according to manufacturer's protocol, PCR products were digested with DpnI (NEB) to remove template DNA and ligated with T4 DNA ligase (NEB). Double mutatants were created as above except substituting either 1) eGFP BPTF UGUA2 -> UCUA or 2) eGFP BPTF UUGG -> CCGG single mutant plasmids to create 1) eGFP BPTF G->C and UGUA2 -> UCUA or 2) eGFP BPTF UGUA3 -> CAUC and UUGG -> CCGG double mutant reporters respectively.

250ng Reporters were transfected into a 24-well of 293T cells. 2 days post transfection, RNA was purified and utilized for RT-PCR following protocol for wildtype reporters above.

GAATTCGCCATTTCCCTATGAAAAGAACTATTCTTTAATATTATACCAGAGTATAC
ACTTTTGTATGTGATAAAATGTTCATTTTTATATTAATTACTTAAGCATTCTTAATAT
ATTATTAAATATTCTTAAATACTCTCTGAATTACCTATTGCAGCAATATCCGAGAATC
TACTTGCTTTAAAATGGCCGTTCAATGGAAGCAGGTTTTTTGTGATTATTTAGTAGTT
GACTGAATGTGGCCATTTGCCCTGAAGCAATTTTAAAGAATATCTTTGAAGTTTT***G***TT
GT***G***CATTTTG***C****TGTA*GAGCCAACAGAAG***TTGG***GGATAAAGGTAACTCTG*TGTC*AGCA
AATCTTGGCGACAACACAACAAATGCAACTTCAGAAGAGACTAGTCCCTCTGAAGG
GAGGAGCCCTGTGGGGTGTCTCTCAGAAACCCCCGATAGCAGCAACATGGCAGAGA
AGAAGGTGGCATCTGAGCTCCCCCAGGATGTGCCAGGTACAGAGGGCAGCGTATCA
ATGCCTCTGTAATGGGGGAATCCTTCCCTTTTGTAGTAAAAGCCGAATGTCACCTA
AAACCTTAAACTATGTGTTCATTCATGCTGCTTGCTTACAGTGCCATCCCCATTTGCT
AAATTGTCACCTAACATTTGGAGTTTGATAAATGTCCTCGCAGTTAGTGCTTGAAAC
TCATCATAATTTTCATGCTTCTTAAACTTCATTTAGCTGTTTTTGTTTAATAGATTTAT
TTTTATATTTTATCACCAGACCATAGTGGCAGATATTTTGGTTTTAATTACTCGAG

Sequence of BPTF region cloned into eGFP reporter with intron unhighlighted and exon highlighted in yellow. Locations of mutations are underlined, bolded, and italicized. Red is location of G->C mutation, green is UGUA2->UCUA mutation, turquoise is TTGG ->CCGG mutation, and purple is TGUC3 -> CAUC or UCUC mutation.

Immunoprecipitations

Flag Immunoprecipitations

U2AF65, U170K, and U2AF35 were cloned into flag pcDNA3.1 using restriction sites BamHI and XbaI (U2AF65, U170K) or BamHI and XhoI (U2AF35). 20ug of flag pcDNA3.1 plasmid was transfected into a 15cm plate of 293T cells using PEI. 2 days post transfection, cells were resuspended in PBS and spun down at 1.5krpm at 4°C for 5 minutes. Cell pellets were resuspended in 1mL of cold Buffer A (10mM Hepes pH 7.9, 10mM KCl, 1.5mM $MgCl_2$, 10% glycerol) and allowed to swell on ice for 10 minutes before addition of NP-40 to 0.5% to lyse cells. Samples were spun down at 4krpm for 5 minutes and cytoplasmic supernatant was removed. Nuclear pellet was resuspended in 300ul of Buffer D 300 (20mM Hepes pH 7.9, 300mM NaCl, 1mM $MgCl_2$, 0.1mM EDTA) and homogenized through 18g needle 5 times. Phosphatase inhibitor, proteinase inhibitor, and 0.2ug/ul RNAse were added and nuclear extract was rotated at 4°C for 30 minutes. Samples were spun down at 10krpm for 5 minutes. Pellet was removed and supernatant was used as nuclear extract. Nuclear extract was mixed with 300ul Buffer D 100 (20mM Hepes pH 7.9, 100mM NaCl, 1mM $MgCl_2$, 0.1mM EDTA) and rotated with 20ul Anti-FLAG M2 Affinity Beads (Sigma) overnight at 4°C. Following overnight incubation, beads were washed with Buffer D150 (20mM Hepes pH 7.9, 150mM NaCl, 1mM $MgCl_2$, 0.1mM EDTA) with 0.05% NP-40 4x, 10 minute each. Proteins were eluted in Buffer

D100 (20mM Hepes pH 7.9, 100mM NaCl, 1mM $MgCl_2$, 0.1mM EDTA, 10% glycerol) with 3x FLAG Peptide. Elutions were combined and precipitated overnight at -20°C in acetone followed by pelleting at 14krpm for 20 minutes, resuspension in 1x SDS loading dye, denaturation at 95°C for 10 minutes, and western blotting analysis.

CFIm59 and CFIm68 Knockout U2AF65 Flag Immunoprecipitations

Flag-pcDNA3.1 U2A65 was transfected into CFIm59 and CFIm68 knockout 293T cells and harvested for flag immunoprecipitation according to above protocol.

For rescue experiments, 15ug of flag-pcDNA3.1 U2AF65 was cotransfected with 15ug of previously generated pcDNA wildtype HA-CFIm68 or HA-CFIm68 lacking either the RNA recognition motif (RRM), proline rich region (PRR) or RS domain into a 15cm plate of CFIm68 knockout 293T cells. 2 days post transfection, cells were harvested for flag immunoprecipitation according to above protocol. Loading of samples was normalized to levels of FLAG signal for each domain deletion construct.

Domain Deletion Flag Immunoprecipitations

Domain deletion U2AF65 and U170K constructs were cloned into flag pcDNA3.1 using BamHI and XbaI restriction sites with a nuclear localization signal added to the C-terminus. For U2AF65, the RS (amino acids 17-47), U2AF ligand motif (ULM) (amino acids 78-110), RNA recognition motif 1 (RRM1) (amino acids 249-231), RNA recognition motif 2 (RRM2) (amino acids 259-337), and U2AF homology motif (UHM) (amino acids 385-466) were each deleted in separate constructs using overlap extension PCR. Similarly, for U170K, the RRM (amino acids 102-187) or the RS domain (amino acids 231-377) were each deleted in separate constructs using overlap extension PCR. Domain deletion constructs were transfected into 293T cells and used for flag immunoprecipitation according to above protocol. Loading of samples was normalized to levels of FLAG signal for each domain deletion construct.

U2AF65 Flag Immunoprecipitations upon Knockdown of CFIm25

CFIm25 was knocked down according to above protocol. 3 days post transduction, 20ug of flag U2AF65 was transfected into a 15cm plate of CFIm25 knockdown or control cells. 5 days post transfection, cells were harvested for flag immunoprecipitation according to above protocol. Loading of samples was normalized to flag signal for each sample.

Recombinant Protein Expression and Purification

CFIm25

CFIm25 was cloned into pRSET and transformed into BL21 competent cells. 100ul of overnight culture was added to 500mL of LB media and grown at 37°C 225rpm until OD 0.5. IPTG was added at 0.2mM to induce protein expression at 16°C overnight. Bacteria were spun down at 4krpm 10 minutes, resuspended in 30mL of lysis buffer (20mM Hepes pH 7.9, 300mM NaCl, 10mM imidazole, 10% glycerol), and sonicated at for 10 cycles of 10 amps for 10 seconds. Following sonication, lysate was spun down at 10krpm for 20 minutes at 4°C. Supernatant was incubated with 1ml HisPur^{TM} Ni-NTA Superflow Agarose beads (Fisher) overnight at 4°C. Beads were washed 3x with 1M NaCl wash buffer (20mM Hepes pH 7.9, 1M NaCl, 10mM imidazole, 10% glycerol) and 1x with 300mM NaCl wash buffer (20mM Hepes pH 7.9, 300mM NaCl, 10mM imidazole, 10% glycerol) for 10 minutes with rotation. Protein was eluted 5x in elution buffer (20mM Hepes pH 7.9, 300mM NaCl, 200mM imidazole, 10% glycerol) and elutions were frozen at -80°C. Elutions were combined and dialyzed to Buffer D 100 (20mM

Hepes pH 7.9, 100mM NaCl, 1mM $MgCl_2$, 0.1mM EDTA, 10% glycerol). Protein concentration was determined by Coomassie stain and comparison with a BSA standard.

CFIm Complexes

Recombinant CFIm complexes (CFIm25-CFIm59 or CFIm25-CFIm68) had previously been expressed using the MultiBac expression system (Pelosse et al. 2017). MultiBac is a protein expression system in insect cells that is utilized for expression and purification of multi-protein complexes by enabling infinite gene insertions with an assembly of restriction sites. Plasmids are transformed into DH10-Bac competent cells for preparation of bacmid, a large bacterial plasmid that contains the baculovirus genome, and bacmid is transfected into Sf9 insect cells to create baculovirus following standard protocol(Pelosse et al. 2017). After expression of complex in Sf9 insect cells, cells were resuspended in lysis buffer (20mM Hepes pH 7.9, 300mM NaCl, 10mM imidazole, 10% glycerol), sonicated at 2 Amp for 5 cycles of 5 seconds and centrifuged at 10,000xg for 20 minutes. Lysate was added to HisPur^{TM} Ni-NTA Superflow beads and incubated at 4°C for 2 hours. Beads were washed twice with high salt wash buffer (20mM Hepes pH 7.9, 1M NaCl, 10mM imidazole, 10% glycerol) and once with low salt wash buffer (20mM Hepes pH 7.9, 500mM NaCl, 10mM imidazole, 10% glycerol). Protein was eluted five times in elution buffer (10mM Hepes pH 7.9, 300mM NaCl, 20mM imidazole, 10% glycerol) and stored at -80°C. Protein concentration was determined by Coomassie stain and comparison with BSA standard.

U2AF65 and U2AF65-U2AF35 Complex

Strep-U2AF65 was cloned into pFastBac using BamHI and XbaI restriction sites. U2AF35 was cloned into pFastBac HTB using BamHI and NotI restriction sites. Plasmids were transformed into DH10-Bac competent cells for preparation of bacmid and bacmid was transfected into Sf9 insect cells for production of baculovirus. For U2AF65-U2AF35 complex, insect cells were co-infected with baculovirus for U2AF65 and U2AF35. Following baculovirus infection, cells were resuspended in Buffer W (100mM Tris pH 8, 150mM NaCl, 1mM EDTA), sonicated for 6 cycles of 10 Amps for 30 seconds and centrifuged at 15,000xg for 30 minutes. After 30 minutes, the supernatant was removed and centrifuged at 15,000xg for 30 minutes. Lysate was applied to a 20mL column (BioRad) pre-packed with 2mL of Streptactin XT-4flow suspension (IBA Lifesciences) (1mL column volume) and allowed to flow-through column. Column was washed with 5x column volume Buffer W. Protein was eluted six times in 0.5x column volume Buffer BXT (100mM Tris pH 8, 150mM NaCl, 1mM EDTA, 50mM Biotin) and stored at -80°C. Protein concentration was determined by Coomassie stain with BSA standard.

In Vitro Interaction of CFIm and U2AF

3ug of CFIm25 in complex with either CFIm59 or CFIm68 (CFIm25-CFIm59 and CFIm25-CFIm68) were pre-incubated with either Buffer D100, 1ug U2AF65, or 1ug of U2AF65 in complex with U2AF35 (U2AF65 and U2AF65-U2AF35) as well as 100ug BSA, and 0.1% NP-40 for 2 hours at 4°C. Following pre-incubation, protein was incubated with 100ul Streptactin Sepharose (IBA Lifesciences) for 2 hours at 4°C. Beads were washed 3x with short wash in Strep Wash Buffer (100mM Tris HCl pH 7.9, 150mM NaCl, 0.1% NP-40). Proteins were eluted 3x in Buffer E (IBA) containing desthiobiotin. 3x volumes of acetone and 1ul of Glycoblue were added and proteins were precipitated overnight at -20°C. Following acetone precipitation, proteins were centrifuged at max speed for 20 minutes. Pellets were resuspended in 20ul 1x SDS loading dye and used for western blotting analysis.

MBP-MS2 Affinity Purification

MBP-MS2 Affinity Purification from Nuclear Extract

AdML, BPTF, and hnRNP LL were cloned into a pBlueScript plasmid downstream of 3MS2 binding sites. Plasmid was linearized by digestion with EcoRV (NEB) and DNA was purified by phenol chloroform extraction. AdML, BPTF, and hnRNP LL substrates were in vitro transcribed using T7 RNA polymerase following standard protocol.

Following transcription, 3.75pmol of RNA was mixed with 28.125pmol MBP-MS2 for 7.5-fold excess MS2-MBP. Following incubation, ATP was added to a final concentration of 1mM, creatine phosphate to 20mM, and tRNA to 100ng/ul. 100ul of HeLa cell nuclear extract to reach 40% of the total reaction volume of 250ul. Samples were incubated at 30°C for 20 minutes. Samples were incubated with 25ul of amylose resin (NEB) for 1hr at 4°C followed by washing 4x with wash buffer + 0.5% NP-40 (10mM Hepes pH 7.9, 100mM KCl, 1mM $MgCl_2$, 0.1mM EDTA) and 1x with wash buffer -NP-40. Samples were eluted in wash buffer + 20mM maltose and acetone precipitated at -20°C. Following acetone precipitation, samples were centrifuged at max speed to pellet and resuspended in 1X SDS loading dye for western blotting analysis for U2AF65, U2AF35, CFIm25, CFIm59, and CFIm68.

AdML:
TTTCCTTGAAGCTTTCGTGCTGACCCTGTCCCTTTTTTTTCCACAGCTGCAGGTCGAC
GTTGAGGACAAACTCTTCGCGGTCTTTCCAGTACTCTTGGATCC

BPTF:
GCCCTGAAGCAATTTTAAAGAATATCTTTGAAGTTTTGTTGTGCATTTTGCTGTAGAG
CCAACAGAAGTTGGGGATAAAGGTAACTCTGTGTCAGCAAATCT~~TGGCGACAACAC~~
~~AACAAATGCAACTTCAGAAGAGACTAGTCCCTCTGAAGGGAGGAGCCCTGTGGGGT~~
~~GTCTCTCAGAAACCCCCGATAGCAGCAACATGGCAGAGAAGAAGGTGGCATCTGAG~~
~~CTCCCCCAGGATGTGCCAG~~

hnRNP LL:
CAGGTGTTTGTAAATATGTGCATATACTTATAAAATACTCTGTGTTTACTGATTGTGA

AACCATGTAATACAGGTTGTTGTTGGCTTATGCTGAAACTTCTTGAAA~~GAAGCCCACTTTGAAAATGTGTTGCTGTTGAAGCTAATAGTATGATTCAG~~

Yellow highlighted region represents the regulated exon of AdML, BPTF, and hnRNP LL. Gray strikethrough region represents region of the CFIm regulated exon of BPTF and hnRNP LL that is not included within the in vitro substrate.

MBP-MS2 Affinity Purification from CFIm Depleted Nuclear Extract

CFIm was depleted from nuclear extract by immunoprecipitation of CFIm25. CFIm25 antibody (Proteintech 10322-1-AP) of an IgG control was coupled to 20ul of protein A/G agarose beads. 5ug of CFIm25 was incubated with 20ul of protein A/G agarose beads for 1 hour at room temperature. Beads were washed with 10 volumes of 0.2M sodium borate pH9 and 20mM DMP (which reacts with primary amines) was added to covalently crosslink antibody to beads for 30 minutes at room temperature. The reaction was quenched with 0.2M ethanolamine pH 8 for 2 hours at room temperature.

Following coupling, antibody-bound beads were incubated with nuclear extract at 4°C for 2 hours followed by removal of nuclear extract. Beads were recycled by incubation with 0.1M glycine pH 3.5 for 5 minutes. CFIm levels in nuclear extract were analyzed by western blotting. Depletion was repeated using protocol above until CFIm levels were less than 10% of HeLa nuclear extract.

Following generation of CFIm depleted nuclear extract, MBP-MS2 affinity purification occurred as above with either CFIm or IgG depleted nuclear extract.

MS2-MBP Affinity Purification with Recombinant CFIm and U2AF

15pmol of AdML, BPTF, and hnRNP LL substrates in vitro transcribed as above were incubated with 112.5pmol of MBP-MS2 on ice for 30 minutes. BSA was added to 500ug/mL and tRNA was added to 100ng/mL. 8pmol of CFIm and/or U2AF were added to each reaction according to table below.

1	2	3	4	5	6	7	8	9	10	11
CFIm25	CFIm25 CFIm59	CFIm25 CFIm68			CFIm25	CFIm25	CFIm25 CFIm59	CFIm25 CFIm59	CFIm25 CFIm68	CFIm25 CFIm68
			U2AF65	U2AF65 U2AF35	U2AF65	U2AF65 U2AF35	U2AF65	U2AF65 U2AF35	U2AF65	U2AF65 U2AF35

Reactions were incubated at 30 degrees for 20 minutes and added to 70ul of amylose resin. Proteins were incubated with amylose resin, washed, and eluted as above. Following acetone precipitation, samples were resuspended in 1x SDS loading dye and utilized for western blotting for U2AF65, U2AF35, CFIm25, CFIm59, and CFIm68.

CHAPTER 5

OVERALL CONCLUSIONS AND PERSPECTIVES

RNA processing events are dynamic, interconnected, and highly regulated. In this manuscript, I have shown that RNA binding proteins play critical roles in regulating RNA processing events, many of which have not yet been discovered. In chapter 2, RNA binding proteins with no known role in human APA were identified as putative polyadenylation regulators by were tethered near polyadenylation sites to test for a role in polyA site selection. The identified regulators included Musashi1, hnRNP G, and hnRNP A0, all of which inhibit 3' processing. One of the most surprising results was that all 9 of the 12 SR proteins tested inhibit polyA site selection from both upstream and downstream of polyA sites, indicating that the SR family as a whole represses polyadenylation. SR proteins are traditionally studied as splicing activators that use a similar activation mechanism as CFIm (a member of the core polyA machinery) uses to promote polyA sites (Graveley, Hertel, and Maniatis 2001; Zhu et al. 2017). In addition, SR proteins have position dependent effects on splicing as they activate splicing when bound to exonic splicing enhancers but inhibit splicing when bound to intronic splicing silencers. As this study found that SR proteins have a position independent inhibitory effect on polyA site selection, it suggests that the role that SR proteins play in 3' processing is unique from that in splicing. One possible explanation for this phenomenon is that the ability of SR proteins to repress polyA sites evolved to compete with the polyadenylation machinery at intronic sites.

When polyadenylation sites lie within introns, there is a competition between the core polyadenylation machinery and the splicing machinery to determine whether intronic polyadenylation will occur or the intron will be removed by splicing (Tian et al. 2005). Several models have been proposed to explain how this competition occurs including U1 telescripting, which occurs when U1 snRNA suppresses premature polyadenylation (Kaida et al. 2010; Berg et al. 2012). It is possible that repression by SR proteins exists as another mechanism to enhance splicing when intronic polyadenylation sites are present and, consistently, this has been reported for SRSF10 (Jobbins et al. 2022)

Ultimately, to understand exactly how SR proteins and other RNA binding proteins regulate 3' processing, further mechanistic studies will have to be performed. Currently, RBPs were physically tethered to a reporter polyadenylation site, indicating that they *can* regulate APA but not whether they *do endogenously*. Endogenous APA regulation can be analyzed by depletion of these proteins and performing PolyAdenylation Site Sequencing (PAS-seq), which will not only confirm that these RBPs have a role in APA but also provide mechanistic information about both the types of RNAs and APA changes that are regulated. Regardless of being in its initial stages, this study supports the hypothesis that many RBPs outside of the core polyadenylation machinery play roles in 3' processing. There are over 1500 RNA binding proteins in humans, many of which have tissue or other context specific expression (Kang, Lee, and Lee 2020). Identifying those proteins with the capacity to directly influence polyA site selection will greatly increase our ability to predict APA patterns.

In the next two chapters, I show that the 3' processing factor cleavage factor I (CFIm) has uncharacterized functions in both 3' processing as well as another RNA processing event: splicing. In chapter 3, CFIm is shown to promote intronic polyA sites of many genes in addition to its role in enhancing distal polyA sites in the 3' UTR. Interestingly, CFIm was shown to regulate intronic polyadenylation within one subunit of each other complex in the core 3' processing machinery: Wdr33 (CPSF), CstF77 (CstF), and Pcf11 (CFIIm) as well as polyA polymerase and Rbbp6 on both the mRNA and protein level. This finding supports a modified model for 3' UTR APA. By this model, when CFIm levels are high, there is IPA within members of the core polyadenylation machinery and therefore limited availability of each complex. This increases the dependency on CFIm to enhance polyadenylation sites; distal polyA sites are preferentially used because of the enrichment of the UGUA recognition motif. However, when CFIm levels are low, there is decreased intronic polyadenylation within Wdr33, CstF77, and Pcf11, increasing levels of the polyadenylation machinery available to perform 3'end processing. When this occurs, the proximal site has an advantage because it is transcribed first (Figure 3.17B).

The findings from chapter 3 also suggest a mechanism for CFIm to link 3' processing with cell fate decisions. Previously, it was demonstrated that depletion of CFIm enhances reprogramming of mouse embryonic fibroblasts into induced pluripotent stem cells 30-fold, indicating that CFIm is an important cell fate regulator (Justin Brumbaugh et al. 2018). Strikingly, the findings of our current study suggest that both CFIm-mediated 3' UTR APA and IPA regulation reduce protein production. In 3' UTR APA, CFIm promotes production of mRNA transcripts with longer 3' UTRs and therefore increased binding sites for microRNAs and RNA binding proteins that can decrease mRNA stability and translation efficiency (Mayr and Bartel 2009; Tushev et al. 2018; Hoffman et al. 2016; Garneau, Wilusz, and Wilusz 2007). We also find that CFIm promotes intronic polyadenylation within many genes, leading to production of truncated mRNAs that are oftentimes degraded or produce non-functional protein products. As a result, upon CFIm depletion, we would predict to see a global increase in protein production by both mechanisms, which was seen with quantitative protein staining (Figure 3.15). This finding provides insight into the role of CFIm in cell fate determination because CFIm has the opposite role of myc, a gene expression amplifier that is also one of the Yamanaka factors for reprogramming (Takahashi and Yamanaka 2006; Nie et al. 2012; Bradner, Lee, and Young 2013).

Importantly, there is still no direct link between CFIm mediated APA regulation and the increase in protein production seen upon knockdown of CFIm25, which is necessary to support our hypothesis that CFIm is a global gene expression attenuator. To test this, polysome profiling should be performed to identify the exact identities of the RNAs that undergo increased protein production and compared to the mRNAs that undergo CFIm-mediated APA. However, our study still furthers our understanding of how 3' processing and cell fate determination are linked.

In chapter 4, I investigate another novel role for CFIm, this time in splicing regulation. Genome-wide sequencing after CFIm25 knockdown revealed that CFIm regulates alternative splicing genome wide, with exons that are both activated and repressed by CFIm. Mechanistically, we establish that CFIm interacts with both U2AF and U170K. In addition, CFIm regulates U2AF65-RNA interacts at 3' splice sites for CFIm activated exons, which typically having weaker polypyrimidine tracts with a higher enrichment of C. While it has been previously established that CFIm interacts with U2AF65 and can terminal exon splicing, this is the first study to show that CFIm is a general alternative splicing regulator (Movassat et al. 2016; Millevoi et al. 2006).

Our findings also support a new model for recognition of 3' splice site by U2AF. Under our proposed model, CFIm and other RNA binding proteins all compete to interact with U2AF65. Each U2AF65 complex recognizes a subset of cassette exons and their exon inclusion. When one factor is depleted such as in the case of CFIm25 knockdown, there is decreased exon inclusion for the subset recognized by CFIm but there is simultaneously more U2AF65 available to interact with other binding partners, leading to increased exon inclusion (Figure 4.13). This model not only explains how CFIm can both activate and repress cassette exons, but also suggests that there are many other unknown RNA binding proteins that can interact with U2AF and also play a role in 3' splice site selection and alternative splicing.

Future experiments will be needed to directly test whether CFIm regulates splicing in vitro. One way to do so would be to perform in vitro splicing assays with CFIm depleted nuclear extract to test whether there is a decrease exon inclusion for CFIm activated exons. In addition, as U2AF levels do not change upon CFIm depletion from nuclear extract (Figure 4.8D), we could also test whether there is increased exon inclusion for CFIm repressed exons which are likely regulated by other RBPs.

Strikingly, we also establish that CFIm uses a common mechanism to regulate both polyadenylation and splicing. In both cases, the RS domain of CFIm is necessary for interactions; in the case of polyadenylation, it interacts with the RE/D domain of Fip1 and in splicing, it interacts with the RS domain of both U2AF65 and U170K. CFIm also has previously been shown to enhance distal polyA sites and we additionally show that it increases U2AF binding to weak polypyrimidine tracts (Zhu et al. 2017). The findings of chapters 3 and 4 indicate that CFIm uses a unified model for regulate of polyadenylation and splicing.

Similar to our findings on CFIm, chapters 2 and 4 have also established that splicing and polyadenylation are highly interconnected, with SR proteins and hnRNPs regulating APA and the converse: CFIm regulating splicing. It will therefore be critical to redefine how we study these RNA processing events to account for their cross-regulation instead of treating them as two independent events, as they have traditionally been studied. Remaining questions include the precise timing of splicing and polyadenylation in relation with each other and whether it is distinct regulatory proteins that cross-regulate or whether it is the core machineries themselves.

Overall, the work in this manuscript has identified novel roles for many RNA binding proteins. Sometimes, it is merely a redefined role, as with CFIm regulating intronic polyadenylation in addition to 3' UTR APA. At other times, it can be a role in a unique RNA processing event such as SR proteins and other RBPs regulating APA or CFIm regulating alternative splicing. With the over 1500 RBPs encoded within the human genome, it is highly likely that we will continue to identify new regulators of RNA processing. Interestingly, another source of novel RNA processing regulators may in fact be DNA binding proteins were once considered to be a functionally distinct class of proteins and therefore studied independently but in more recent years have been shown to be involved in RNA translation, miRNA biogenesis, and telomere maintenance (Hudson and Ortlund 2014). As a result, it is critical to not limit our understanding of these nucleic acid binding proteins to one specific type of processing event or even one specific species of nucleic acid.

References

Alkan, Serkan A., Kathleen Martincic, and Christine Milcarek. 2006. “The HnRNPs F and H2 Bind to Similar Sequences to Influence Gene Expression.” *Biochemical Journal* 393 (1): 361–71. https://doi.org/10.1042/BJ20050538.

Amit, Maayan, Maya Donyo, Dror Hollander, Amir Goren, Eddo Kim, Sahar Gelfman, Galit Lev-Maor, et al. 2012. “Differential GC Content between Exons and Introns Establishes Distinct Strategies of Splice-Site Recognition.” *Cell Reports* 1 (5): 543–56. https://doi.org/10.1016/J.CELREP.2012.03.013.

Arora, Ankita, Raeann Goering, Hei Yong G. Lo, Joelle Lo, Charlie Moffatt, and J. Matthew Taliaferro. 2022. “The Role of Alternative Polyadenylation in the Regulation of Subcellular RNA Localization.” *Frontiers in Genetics* 12 (January): 2791. https://doi.org/10.3389/FGENE.2021.818668/BIBTEX.

Baek, Daehyun, and Phil Green. 2005. “Sequence Conservation, Relative Isoform Frequencies, and Nonsense-Mediated Decay in Evolutionarily Conserved Alternative Splicing.” *Proceedings of the National Academy of Sciences of the United States of America* 102 (36): 12813–18. https://doi.org/10.1073/PNAS.0506139102.

Batra, Ranjan, Konstantinos Charizanis, Mini Manchanda, Apoorva Mohan, Moyi Li, Dustin J. Finn, Marianne Goodwin, et al. 2014. “Loss of MBNL Leads to Disruption of Developmentally Regulated Alternative Polyadenylation in RNA-Mediated Disease.” *Molecular Cell* 56 (2): 311–22. https://doi.org/10.1016/j.molcel.2014.08.027.

Beaudoing, Emmanuel, Susan Freier, Jacqueline R. Wyatt, Jean Michel Claverie, and Daniel Gautheret. 2000. “Patterns of Variant Polyadenylation Signal Usage in Human Genes.” *Genome Research* 10 (7): 1001. https://doi.org/10.1101/GR.10.7.1001.

Bejar, Rafael. 2016. “Splicing Factor Mutations in Cancer.” *Advances in Experimental Medicine and Biology* 907: 215–28. https://doi.org/10.1007/978-3-319-29073-7_9.

Berg, Michael G., Larry N. Singh, Ihab Younis, Qiang Liu, Anna Maria Pinto, Daisuke Kaida, Zhenxi Zhang, et al. 2012. “U1 SnRNP Determines MRNA Length and Regulates Isoform Expression.” *Cell* 150 (1): 53–64. https://doi.org/10.1016/J.CELL.2012.05.029.

Berget, Susan M. 1995. “Exon Recognition in Vertebrate Splicing.” *Journal of Biological Chemistry* 270 (6): 2411–14. https://doi.org/10.1074/JBC.270.6.2411.

Bittencourt, Danielle, and Didier Auboeuf. 2012. “Analysis of Co-Transcriptional RNA Processing by RNA-ChIP Assay.” *Methods in Molecular Biology (Clifton, N.J.)* 809: 563–77. https://doi.org/10.1007/978-1-61779-376-9_36.

Black, Douglas L. 2003. “Mechanisms of Alternative Pre-Messenger RNA Splicing.” *Annual Review of Biochemistry* 72: 291–336. https://doi.org/10.1146/ANNUREV.BIOCHEM.72.121801.161720.

Bonnal, Sophie C., Irene López-Oreja, and Juan Valcárcel. 2020. “Roles and Mechanisms of Alternative Splicing in Cancer — Implications for Care.” *Nature Reviews Clinical Oncology 2020 17:8* 17 (8): 457–74. https://doi.org/10.1038/s41571-020-0350-x.

Boreikaite, Vytaute, Thomas S. Elliott, Jason W. Chin, and Lori A. Passmore. 2022. “RBBP6 Activates the Pre-MRNA 3′ End Processing Machinery in Humans.” *Genes & Development* 36 (3–4): 210–24. https://doi.org/10.1101/GAD.349223.121.

Boucher, L., C. A. Ouzounis, A. J. Enright, and B. J. Blencowe. 2001. “A Genome-Wide Survey of RS Domain Proteins.” *RNA* 7 (12): 1693. /pmc/articles/PMC1370209/?report=abstract.

Bradner, James E, Tong Ihn Lee, and Richard A Young. 2013. “Transcriptional Amplification in Tumor Cells with Elevated C-Myc - 1-S2.0-S0092867412010574-Main.Pdf” 151 (1): 56–67. https://doi.org/10.1016/j.cell.2012.08.026.Transcriptional.

Brown, Kirk M., and Gregory M. Gilmartin. 2003. “A Mechanism for the Regulation of Pre-MRNA 3′ Processing by Human Cleavage Factor Im.” *Molecular Cell* 12 (6): 1467–76. https://doi.org/10.1016/S1097-2765(03)00453-2.

Busch, Anke, and Klemens J. Hertel. 2012. “Evolution of SR Protein and HnRNP Splicing Regulatory Factors.” *Wiley Interdisciplinary Reviews: RNA* 3 (1): 1–12. https://doi.org/10.1002/WRNA.100.

Callebaut, Isabelle, Despina Moshous, Jean Paul Mornon, and Jean Pierre De Villartay. 2002. “Metallo-β-Lactamase Fold within Nucleic Acids Processing Enzymes: The β-CASP Family.” *Nucleic Acids Research* 30 (16): 3592. https://doi.org/10.1093/NAR/GKF470.

Campagne, Sébastien, Daniel Jutzi, Florian Malard, Maja Matoga, Ksenija Romane, Miki Feldmuller, Martino Colombo, Marc-David Ruepp, and Frédéric H-T. Allain. 2022. “The Cancer-Associated RBM39 Bridges the Pre-MRNA, U1 and U2 SnRNPs to Regulate Alternative Splicing.” *BioRxiv*, August, 2022.08.30.505862. https://doi.org/10.1101/2022.08.30.505862.

Chan, Serena, Eun A. Choi, and Yongsheng Shi. 2011. “Pre-MRNA 3’-End Processing Complex Assembly and Function.” *Wiley Interdisciplinary Reviews. RNA* 2 (3): 321–35. https://doi.org/10.1002/WRNA.54.

Chan, Serena L., Ina Huppertz, Chengguo Yao, Lingjie Weng, James J. Moresco, John R. Yates, Jernej Ule, James L. Manley, and Yongsheng Shi. 2014. “CPSF30 and Wdr33 Directly Bind to AAUAAA in Mammalian MRNA 3′ Processing.” *Genes & Development* 28 (21): 2370–80. https://doi.org/10.1101/GAD.250993.114.

Chan, Serena Leong. 2014. “UC Irvine UC Irvine Electronic Theses and Dissertations Title.” https://escholarship.org/uc/item/0m16c8r4.

Chassé, Héloïse, Sandrine Boulben, Vlad Costache, Patrick Cormier, and Julia Morales. 2017. “Analysis of Translation Using Polysome Profiling.” *Nucleic Acids Research* 45 (3): e15–e15. https://doi.org/10.1093/NAR/GKW907.

Chen, Fan, Clinton C. Macdonald, and Jeffrey Wilusf. 1995. “Cleavage Site Determinants in the Mammalian Polyadenylation Signal.” *Nucleic Acids Research* 23 (14): 2614. https://doi.org/10.1093/NAR/23.14.2614.

Cho, Suhyung, Amy Hoang, Rahul Sinha, Xiang Yang Zhong, Xiang Dong Fu, Adrian R. Krainer, and Gourisankar Ghosh. 2011. “Interaction between the RNA Binding Domains of Ser-Arg Splicing Factor 1 and U1-70K SnRNP Protein Determines Early Spliceosome

Assembly." *Proceedings of the National Academy of Sciences of the United States of America* 108 (20): 8233–38. https://doi.org/10.1073/PNAS.1017700108/-/DCSUPPLEMENTAL/PNAS.1017700108_SI.PDF.

Choi, Hyungwon, Brett Larsen, Zhen Yuan Lin, Ashton Breitkreutz, Dattatreya Mellacheruvu, Damian Fermin, Zhaohui S. Qin, Mike Tyers, Anne Claude Gingras, and Alexey I. Nesvizhskii. 2010. "SAINT: Probabilistic Scoring of Affinity Purification–Mass Spectrometry Data." *Nature Methods 2010 8:1* 8 (1): 70–73. https://doi.org/10.1038/nmeth.1541.

Colgan, D. F., K. G.K. Murthy, C. Prives, and J. L. Manley. 1996. "Cell-Cycle Related Regulation of Poly(A) Polymerase by Phosphorylation." *Nature* 384 (6606): 282–85. https://doi.org/10.1038/384282A0.

Cooke, Charles, Holly Hans, and James C. Alwine. 1999. "Utilization of Splicing Elements and Polyadenylation Signal Elements in the Coupling of Polyadenylation and Last-Intron Removal." *Molecular and Cellular Biology* 19 (7): 4971–79. https://doi.org/10.1128/MCB.19.7.4971.

Coolidge, Candace J., Raymond J. Seely, and James G. Patton. 1997. "Functional Analysis of the Polypyrimidine Tract in Pre-MRNA Splicing." *Nucleic Acids Research* 25 (4): 888. https://doi.org/10.1093/NAR/25.4.888.

Coseno, Molly, Georges Martin, Christopher Berger, Gregory Gilmartin, Walter Keller, and Sylvie Doublié. 2008. "Crystal Structure of the 25 KDa Subunit of Human Cleavage Factor Im." *Nucleic Acids Research* 36 (10): 3474–83. https://doi.org/10.1093/NAR/GKN079.

Dai, Weijun, Gen Zhang, and Eugene V. Makeyev. 2012. "RNA-Binding Protein HuR Autoregulates Its Expression by Promoting Alternative Polyadenylation Site Usage." *Nucleic Acids Research* 40 (2): 787–800. https://doi.org/10.1093/NAR/GKR783.

Dalziel, Martin, Nuno Miguel Nunes, and Andre Furger. 2007. "Two G-Rich Regulatory Elements Located Adjacent to and 440 Nucleotides Downstream of the Core Poly(A) Site of the Intronless Melanocortin Receptor 1 Gene Are Critical for Efficient 3′ End Processing." *Molecular and Cellular Biology* 27 (5): 1568–80. https://doi.org/10.1128/MCB.01821-06/ASSET/7654AE36-7E74-414F-8F04-B14B55BABF60/ASSETS/GRAPHIC/ZMB0050765960008.JPEG.

Danan, Charles, Sudhir Manickavel, and Markus Hafner. 2016. "PAR-CLIP: A Method for Transcriptome-Wide Identification of RNA Binding Protein Interaction Sites." *Methods in Molecular Biology (Clifton, N.J.)* 1358: 153. https://doi.org/10.1007/978-1-4939-3067-8_10.

Dass, Brinda, Steve Tardif, Yeon Park Ji, Bin Tian, Harry M. Weitlauf, Rex A. Hess, Kay Carnes, Michael D. Griswold, Christopher L. Small, and Clinton C. MacDonald. 2007. "Loss of Polyadenylation Protein TCstF-64 Causes Spermatogenic Defects and Male Infertility." *Proceedings of the National Academy of Sciences of the United States of America* 104 (51): 20374–79. https://doi.org/10.1073/PNAS.0707589104/SUPPL_FILE/07589FIG6.JPG.

Davis, Ryan, and Yongsheng Shi. 2014. "The Polyadenylation Code: A Unified Model for the

Regulation of MRNA Alternative Polyadenylation." *Journal of Zhejiang University. Science. B* 15 (5): 429–37. https://doi.org/10.1631/JZUS.B1400076.

Day, Irene S., Maxim Golovkin, Saiprasad G. Palusa, Alicia Link, Gul S. Ali, Julie Thomas, Dale N. Richardson, and Anireddy S.N. Reddy. 2012. "Interactions of SR45, an SR-like Protein, with Spliceosomal Proteins and an Intronic Sequence: Insights into Regulated Splicing." *The Plant Journal : For Cell and Molecular Biology* 71 (6): 936–47. https://doi.org/10.1111/J.1365-313X.2012.05042.X.

Dettwiler, Sabine, Chiara Aringhieri, Stefano Cardinale, Walter Keller, and Silvia M.L. Barabino. 2004. "Distinct Sequence Motifs within the 68-KDa Subunit of Cleavage Factor Im Mediate RNA Binding, Protein-Protein Interactions, and Subcellular Localization." *The Journal of Biological Chemistry* 279 (34): 35788–97. https://doi.org/10.1074/JBC.M403927200.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. "STAR: Ultrafast Universal RNA-Seq Aligner." *Bioinformatics* 29 (1): 15. https://doi.org/10.1093/BIOINFORMATICS/BTS635.

Dominguez, Daniel, Peter Freese, Maria S Alexis, Gene W Yeo, Brenton R Graveley, and Christopher B Burge. 2018. "Sequence, Structure, and Context Preferences of Human RNA Binding Proteins." *Molecular Cell* 70: 854–67. https://doi.org/10.1016/j.molcel.2018.05.001.

Dubbury, Sara J., Paul L. Boutz, and Phillip A. Sharp. 2018. "CDK12 Regulates DNA Repair Genes by Suppressing Intronic Polyadenylation." *Nature 2018 564:7734* 564 (7734): 141–45. https://doi.org/10.1038/s41586-018-0758-y.

Dvinge, Heidi. 2018. "Regulation of Alternative MRNA Splicing: Old Players and New Perspectives." *FEBS Letters* 592 (17): 2987–3006. https://doi.org/10.1002/1873-3468.13119.

Erkelenz, Steffen, William F. Mueller, Melanie S. Evans, Anke Busch, Katrin Schöneweis, Klemens J. Hertel, and Heiner Schaal. 2013. "Position-Dependent Splicing Activation and Repression by SR and HnRNP Proteins Rely on Common Mechanisms." *Rna* 19 (1): 96–102. https://doi.org/10.1261/rna.037044.112.

Erson-Bensan, Ayse Elif, and Rbps Apa. 2016. "Alternative Polyadenylation and RNA-Binding Proteins." *Journal of Molecular Endocrinology* 57 (2): F29–34. https://doi.org/10.1530/JME-16-0070.

Fischl, Harry, Jonathan Neve, Zhiqiao Wang, Radhika Patel, Alastair Louey, Bin Tian, and Andre Furger. 2019. "HnRNPC Regulates Cancer-Specific Alternative Cleavage and Polyadenylation Profiles." *Nucleic Acids Research* 47 (14): 7580. https://doi.org/10.1093/NAR/GKZ461.

Forouzanfar, Mahboobeh, Liana Lachinani, Kianoush Dormiani, Mohammad Hossein Nasr-Esfahani, Ali Osmay Gure, and Kamran Ghaedi. 2020. "Intracellular Functions of RNA-Binding Protein, Musashi1, in Stem and Cancer Cells." *Stem Cell Research and Therapy* 11 (1): 1–10. https://doi.org/10.1186/S13287-020-01703-W/FIGURES/3.

Fu, Yonggui, Liutao Chen, Chengyong Chen, Yutong Ge, Mingjing Kang, Zili Song, Jingwen Li, et al. 2018. “Crosstalk between Alternative Polyadenylation and MiRNAs in the Regulation of Protein Translational Efficiency.” *Genome Research* 28 (11): 1656–63. https://doi.org/10.1101/gr.231506.117.

Gandhi, T. K.B., Jun Zhong, Suresh Mathivanan, L. Karthick, K. N. Chandrika, S. Sujatha Mohan, Salil Sharma, et al. 2006. “Analysis of the Human Protein Interactome and Comparison with Yeast, Worm and Fly Interaction Datasets.” *Nature Genetics 2006 38:3* 38 (3): 285–93. https://doi.org/10.1038/ng1747.

Garg, Kavita, and Phil Green. 2007. “Differing Patterns of Selection in Alternative and Constitutive Splice Sites.” *Genome Research* 17 (7): 1015. https://doi.org/10.1101/GR.6347907.

Garneau, Nicole L., Jeffrey Wilusz, and Carol J. Wilusz. 2007. “The Highways and Byways of MRNA Decay.” *Nature Reviews Molecular Cell Biology* 8 (2): 113–26. https://doi.org/10.1038/nrm2104.

Geuens, Thomas, Delphine Bouhy, and Vincent Timmerman. 2016. “The HnRNP Family: Insights into Their Role in Health and Disease.” *Human Genetics* 135 (8): 851–67. https://doi.org/10.1007/s00439-016-1683-5.

Ghazy, Mohamed A., James M.B. Gordon, Susan D. Lee, Badri Nath Singh, Andrew Bohm, Michael Hampsey, and Claire Moore. 2012. “The Interaction of Pcf11 and Clp1 Is Needed for MRNA 3’-End Formation and Is Modulated by Amino Acids in the ATP-Binding Site.” *Nucleic Acids Research* 40 (3): 1214–25. https://doi.org/10.1093/NAR/GKR801.

Giammartino, Dafne Campigli Di, Wencheng Li, Koichi Ogami, Jossie J. Yashinskie, Mainul Hoque, Bin Tian, and James L. Manley. 2014. “RBBP6 Isoforms Regulate the Human Polyadenylation Machinery and Modulate Expression of MRNAs with AU-Rich 3′ UTRs.” *Genes & Development* 28 (20): 2248. https://doi.org/10.1101/GAD.245787.114.

Gohr, André, and Manuel Irimia. 2019. “Matt: Unix Tools for Alternative Splicing Analysis.” *Bioinformatics* 35 (1): 130–32. https://doi.org/10.1093/BIOINFORMATICS/BTY606.

Graveley, Brenton R., Klemens J. Hertel, and T. O.M. Maniatis. 2001. “The Role of U2AF35 and U2AF65 in Enhancer-Dependent Splicing.” *RNA* 7 (6): 806. https://doi.org/10.1017/S1355838201010317.

Hanada, Toshikatsu, Stefan Weitzer, Barbara Mair, Christian Bernreuther, Brian J. Wainger, Justin Ichida, Reiko Hanada, et al. 2013. “CLP1 Links TRNA Metabolism to Progressive Motor-Neuron Loss.” *Nature 2013 495:7442* 495 (7442): 474–80. https://doi.org/10.1038/nature11923.

Heinrich, Bettina, Zhaiyi Zhang, Oleg Raitskin, Michael Hiller, Natalya Benderska, Annette M. Hartmann, Laurent Bracco, et al. 2009. “Heterogeneous Nuclear Ribonucleoprotein G Regulates Splice Site Selection by Binding to CC(A/C)-Rich Regions in Pre-MRNA.” *Journal of Biological Chemistry* 284 (21): 14303–15. https://doi.org/10.1074/JBC.M901026200.

Hertel, Klemens J. 2008. “Combinatorial Control of Exon Recognition.” *Journal of Biological*

Chemistry 283 (3): 1211–15. https://doi.org/10.1074/JBC.R700035200.

Hocine, Sami, Robert H. Singer, and David Grünwald. 2010. "RNA Processing and Export." *Cold Spring Harbor Perspectives in Biology* 2 (12). https://doi.org/10.1101/CSHPERSPECT.A000752.

Hoffman, Yonit, Debora Rosa Bublik, Alejandro P Ugalde, Ran Elkon, Tammy Biniashvili, Reuven Agami, Moshe Oren, and Yitzhak Pilpel. 2016. "3'UTR Shortening Potentiates MicroRNA-Based Repression of Pro-Differentiation Genes in Proliferating Human Cells." *PLoS Genetics* 12 (2): e1005879. https://doi.org/10.1371/journal.pgen.1005879.

Hsu, Tiffany Y.T., Lukas M. Simon, Nicholas J. Neill, Richard Marcotte, Azin Sayad, Christopher S. Bland, Gloria V. Echeverria, et al. 2015. "The Spliceosome Is a Therapeutic Vulnerability in MYC-Driven Cancer." *Nature* 525 (7569): 384. https://doi.org/10.1038/NATURE14985.

Huang, Jingjing, Tingting Weng, Junsuk Ko, Ning-yuan Chen, Yu Xiang, Kelly Volcik, Leng Han, Michael R. Blackburn, and Xiang Lu. 2018. "Suppression of Cleavage Factor Im 25 Promotes the Proliferation of Lung Cancer Cells through Alternative Polyadenylation." *Biochemical and Biophysical Research Communications* 503 (2): 856–62. https://doi.org/10.1016/j.bbrc.2018.06.087.

Hudson, William H., and Eric A. Ortlund. 2014. "The Structure, Function and Evolution of Proteins That Bind DNA and RNA." *Nature Reviews Molecular Cell Biology 2014 15:11* 15 (11): 749–60. https://doi.org/10.1038/nrm3884.

Huelga, Stephanie C., Anthony Q. Vu, Justin D. Arnold, Tiffany D. Liang, Patrick P. Liu, Bernice Y. Yan, John Paul Donohue, et al. 2012. "Integrative Genome-Wide Analysis Reveals Cooperative Regulation of Alternative Splicing by HnRNP Proteins." *Cell Reports* 1 (2): 167–78. https://doi.org/10.1016/J.CELREP.2012.02.001.

Hwang, Hun Way, Christopher Y. Park, Hani Goodarzi, John J. Fak, Aldo Mele, Michael J. Moore, Yuhki Saito, and Robert B. Darnell. 2016. "PAPERCLIP Identifies MicroRNA Targets and a Role of CstF64/64tau in Promoting Non-Canonical Poly(A) Site Usage." *Cell Reports* 15 (2): 423–35. https://doi.org/10.1016/J.CELREP.2016.03.023.

Ji, Xinjun, Ji Wan, Melanie Vishnu, Yi Xing, and Stephen A. Liebhaber. 2013. "ACP Poly(C) Binding Proteins Act as Global Regulators of Alternative Polyadenylation." *Molecular and Cellular Biology* 33 (13): 2560–73. https://doi.org/10.1128/MCB.01380-12/SUPPL_FILE/ZMB999100020SO1.PDF.

Ji, Zhe, Ju Youn Lee, Zhenhua Pan, Bingjun Jiang, and Bin Tian. 2009. "Progressive Lengthening of 3' Untranslated Regions of MRNAs by Alternative Polyadenylation during Mouse Embryonic Development." *Proceedings of the National Academy of Sciences of the United States of America* 106 (17): 7028–33. https://doi.org/10.1073/PNAS.0900028106.

Ji, Zhe, and Bin Tian. 2009. "Reprogramming of 3′ Untranslated Regions of MRNAs by Alternative Polyadenylation in Generation of Pluripotent Stem Cells from Different Cell Types." *PLOS ONE* 4 (12): e8419. https://doi.org/10.1371/JOURNAL.PONE.0008419.

Jo, Bong-Seok, and Sun Shim Choi. 2015. "Introns: The Functional Benefits of Introns in

Genomes." *Genomics & Informatics* 13 (4): 112. https://doi.org/10.5808/GI.2015.13.4.112.

Jobbins, Andrew M, Nejc Haberman, Natalia Artigas, Christopher Amourda, Helen A B Paterson, Sijia Yu, Samuel J I Blackford, et al. 2022. "Dysregulated RNA Polyadenylation Contributes to Metabolic Impairment in Non-Alcoholic Fatty Liver Disease." *Nucleic Acids Research* 50 (6): 3379–93. https://doi.org/10.1093/NAR/GKAC165.

Justin Brumbaugh, Authors, Bruno Di Stefano, Xiuye Wang, Guang Hu, Yongsheng Shi, Konrad Hochedlinger Correspondence, Justin Brumbaugh, et al. 2018. "Nudt21 Controls Cell Fate by Connecting Alternative Polyadenylation to Chromatin Signaling Article Nudt21 Controls Cell Fate by Connecting Alternative Polyadenylation to Chromatin Signaling." *Cell* 172: 106–20. https://doi.org/10.1016/j.cell.2017.11.023.

Kaida, Daisuke, Michael G. Berg, Ihab Younis, Mumtaz Kasim, Larry N. Singh, Lili Wan, and Gideon Dreyfuss. 2010. "U1 SnRNP Protects Pre-MRNAs from Premature Cleavage and Polyadenylation." *Nature* 468 (7324): 664–68. https://doi.org/10.1038/nature09479.

Kamieniarz-Gdula, Kinga, Michal R. Gdula, Karin Panser, Takayuki Nojima, Joan Monks, Jacek R. Wiśniewski, Joey Riepsaame, Neil Brockdorff, Andrea Pauli, and Nick J. Proudfoot. 2019. "Selective Roles of Vertebrate PCF11 in Premature and Full-Length Transcript Termination." *Molecular Cell* 74 (1): 158-172.e9. https://doi.org/10.1016/j.molcel.2019.01.027.

Kang, Donghee, Yerim Lee, and Jae Seon Lee. 2020. "RNA-Binding Proteins in Cancer: Functional and Therapeutic Perspectives." *Cancers* 12 (9): 1–33. https://doi.org/10.3390/CANCERS12092699.

Katz, Yarden, Eric T. Wang, Edoardo M. Airoldi, and Christopher B. Burge. 2010. "Analysis and Design of RNA Sequencing Experiments for Identifying Isoform Regulation." *Nature Methods 2010 7:12* 7 (12): 1009–15. https://doi.org/10.1038/nmeth.1528.

Kaufmann, Isabelle, Georges Martin, Arno Friedlein, Hanno Langen, and Walter Keller. 2004. "Human Fip1 Is a Subunit of CPSF That Binds to U-Rich RNA Elements and Stimulates Poly(A) Polymerase." *The EMBO Journal* 23 (3): 616–26. https://doi.org/10.1038/SJ.EMBOJ.7600070.

Kawahara, Hironori, Takao Imai, Hiroaki Imataka, Masafumi Tsujimoto, Ken Matsumoto, and Hideyuki Okano. 2008. "Neural RNA-Binding Protein Musashi1 Inhibits Translation Initiation by Competing with EIF4G for PABP." *Journal of Cell Biology* 181 (4): 639–53. https://doi.org/10.1083/JCB.200708004.

Keller, W., S. Bienroth, K. M. Lang, and G. Christofori. 1991. "Cleavage and Polyadenylation Factor CPF Specifically Interacts with the Pre-MRNA 3′ Processing Signal AAUAAA." *The EMBO Journal* 10 (13): 4241–49. https://doi.org/10.1002/J.1460-2075.1991.TB05002.X.

Kielkopf, Clara L, Stephan Lücke, and Michael R Green. 2004. "U2AF Homology Motifs: Protein Recognition in the RRM World NIH Public Access." *Genes Dev* 18 (13): 1513–26.

Kim, Sol, Junichi Yamamoto, Yexi Chen, Masatoshi Aida, Tadashi Wada, Hiroshi Handa, and Yuki Yamaguchi. 2010. "Evidence That Cleavage Factor Im Is a Heterotetrameric Protein

Complex Controlling Alternative Polyadenylation." *Genes to Cells : Devoted to Molecular & Cellular Mechanisms* 15 (9): 1003–13. https://doi.org/10.1111/J.1365-2443.2010.01436.X.

Konig, Julian, Kathi Zarnack, Gregor Rot, Tomaz Curk, Melis Kayikci, Blaz Zupan, Daniel J. Turner, Nicholas M. Luscombe, and Jernej Ule. 2011. "ICLIP - Transcriptome-Wide Mapping of Protein-RNA Interactions with Individual Nucleotide Resolution." *Journal of Visualized Experiments*, no. 50. https://doi.org/10.3791/2638.

Kralovicova, Jana, Marcin Knut, Nicholas C.P. Cross, and Igor Vorechovsky. 2015. "Identification of U2AF(35)-Dependent Exons by RNA-Seq Reveals a Link between 3′ Splice-Site Organization and Activity of U2AF-Related Proteins." *Nucleic Acids Research* 43 (7): 3747–63. https://doi.org/10.1093/NAR/GKV194.

Královičová, Jana, Ivana Ševčíková, Eva Stejskalová, Mina Obuča, Michael Hiller, David Staněk, and Igor Vořechovský. 2018. "PUF60-Activated Exons Uncover Altered 3' Splice-Site Selection by Germline Missense Mutations in a Single RRM." *Nucleic Acids Research* 46 (12): 6166–87. https://doi.org/10.1093/NAR/GKY389.

Krecic, Annette M., and Maurice S. Swanson. 1999. "HnRNP Complexes: Composition, Structure, and Function." *Current Opinion in Cell Biology* 11 (3): 363–71. https://doi.org/10.1016/S0955-0674(99)80051-9.

Kubo, Tomohiro, Tadashi Wada, Yuki Yamaguchi, Akira Shimizu, and Hiroshi Handa. 2006. "Knock-down of 25 KDa Subunit of Cleavage Factor Im in Hela Cells Alters Alternative Polyadenylation within 3'-UTRs." *Nucleic Acids Research* 34 (21): 6264–71. https://doi.org/10.1093/NAR/GKL794.

Kyburz, Andrea, Arno Friedlein, Hanno Langen, and Walter Keller. 2006. "Direct Interactions between Subunits of CPSF and the U2 SnRNP Contribute to the Coupling of Pre-MRNA 3' End Processing and Splicing." *Molecular Cell* 23 (2): 195–205. https://doi.org/10.1016/J.MOLCEL.2006.05.037.

Lackford, Brad, Chengguo Yao, Georgette M. Charles, Lingjie Weng, Xiaofeng Zheng, Eun A. Choi, Xiaohui Xie, et al. 2014a. "Fip1 Regulates MRNA Alternative Polyadenylation to Promote Stem Cell Self-Renewal." *The EMBO Journal* 33 (8): 878–89. https://doi.org/10.1002/EMBJ.201386537.

Lackford, Brad, Chengguo Yao, Georgette M Charles, Lingjie Weng, Xiaofeng Zheng, Eun-A Choi, Xiaohui Xie, et al. 2014b. "Fip1 Regulates MRNA Alternative Polyadenylation to Promote Stem Cell Self-Renewal." *EMBO J* 33: 878–89. https://doi.org/10.1002/embj.201386537.

Lee, Shih-Han, Irtisha Singh, Sarah Tisdale, Omar Abdel-Wahab, Christina S. Leslie, and Christine Mayr. 2018. "Widespread Intronic Polyadenylation Inactivates Tumour Suppressor Genes in Leukaemia." *Nature* 561 (7721): 127–31. https://doi.org/10.1038/s41586-018-0465-8.

Legnini, Ivano, Jonathan Alles, Nikos Karaiskos, Salah Ayoub, and Nikolaus Rajewsky. 2019. "FLAM-Seq: Full-Length MRNA Sequencing Reveals Principles of Poly(A) Tail Length Control." *Nature Methods 2019 16:9* 16 (9): 879–86. https://doi.org/10.1038/s41592-019-

0503-y.

Li, Junyi, Tao Pan, Liuxin Chen, Qi Wang, Zhenghong Chang, Weiwei Zhou, Xinhui Li, et al. 2021. "Alternative Splicing Perturbation Landscape Identifies RNA Binding Proteins as Potential Therapeutic Targets in Cancer." *Molecular Therapy - Nucleic Acids* 24 (June): 792–806. https://doi.org/10.1016/j.omtn.2021.04.005.

Li, Wencheng, Bei You, Mainul Hoque, Dinghai Zheng, Wenting Luo, Zhe Ji, Ji Yeon Park, et al. 2015. "Systematic Profiling of Poly(A)+ Transcripts Modulated by Core 3' End Processing and Splicing Factors Reveals Regulatory Rules of Alternative Cleavage and Polyadenylation." Edited by Eric Wagner. *PLOS Genetics* 11 (4): e1005166. https://doi.org/10.1371/journal.pgen.1005166.

Lin, Yuefeng, Zhihua Li, Fatih Ozsolak, Sang Woo Kim, Gustavo Arango-Argoty, Teresa T. Liu, Scott A. Tenenbaum, et al. 2012. "An In-Depth Map of Polyadenylation Sites in Cancer." *Nucleic Acids Research* 40 (17): 8460–71. https://doi.org/10.1093/NAR/GKS637.

Listerman, Imke, Aparna K. Sapra, and Karla M. Neugebauer. 2006. "Cotranscriptional Coupling of Splicing Factor Recruitment and Precursor Messenger RNA Splicing in Mammalian Cells." *Nature Structural & Molecular Biology 2006 13:9* 13 (9): 815–22. https://doi.org/10.1038/nsmb1135.

Liu, Yansheng, Mar Gonzà Lez-Porta, Sergio Santos, Ruedi Aebersold, Ashok R Venkitaraman, Vihandha O Wickramasinghe, Alvis Brazma, and John C Marioni. 2017. "Impact of Alternative Splicing on the Human Proteome In Brief Resource Impact of Alternative Splicing on the Human Proteome." *Cell Reports* 20. https://doi.org/10.1016/j.celrep.2017.07.025.

Long, Jennifer C., and Javier F. Caceres. 2009. "The SR Protein Family of Splicing Factors: Master Regulators of Gene Expression." *Biochemical Journal* 417 (1): 15–27. https://doi.org/10.1042/BJ20081501.

Long, Yunxin, Weng Hong Sou, Kristen Wing Yu Yung, Haizhen Liu, Stephanie Winn Chee Wan, Qingyun Li, Chuyue Zeng, et al. 2019. "Distinct Mechanisms Govern the Phosphorylation of Different SR Protein Splicing Factors." *Journal of Biological Chemistry* 294 (4): 1312–27. https://doi.org/10.1074/JBC.RA118.003392/ATTACHMENT/A0FCE18E-BD45-455F-A0A2-9FA00E3428F1/MMC1.ZIP.

Lukong, Kiven E., Kai wei Chang, Edouard W. Khandjian, and Stéphane Richard. 2008. "RNA-Binding Proteins in Human Genetic Disease." *Trends in Genetics* 24 (8): 416–25. https://doi.org/10.1016/j.tig.2008.05.004.

Luo, Wenting, Zhe Ji, Zhenhua Pan, Bei You, Mainul Hoque, Wencheng Li, Samuel I. Gunderson, and Bin Tian. 2013. "The Conserved Intronic Cleavage and Polyadenylation Site of CstF-77 Gene Imparts Control of 3′ End Processing Activity through Feedback Autoregulation and by U1 SnRNP." *PLoS Genetics* 9 (7). https://doi.org/10.1371/journal.pgen.1003613.

Lykke-Andersen, J., M. D. Shu, and J. A. Steitz. 2001. "Communication of the Position of Exon-Exon Junctions to the MRNA Surveillance Machinery by the Protein RNPS1." *Science* 293

(5536): 1836–39. https://doi.org/10.1126/SCIENCE.1062786/ASSET/559641CE-BD49-4DB3-A27F-3AB921248959/ASSETS/GRAPHIC/SE3519734004.JPEG.

MacDonald, Clinton C., and José Luis Redondo. 2002. “Reexamining the Polyadenylation Signal: Were We Wrong about AAUAAA?” *Molecular and Cellular Endocrinology* 190 (1–2): 1–8. https://doi.org/10.1016/S0303-7207(02)00044-8.

Madison, Bethany J., Kathleen A. Clark, Niraja Bhachech, Peter C. Hollenhorst, Barbara J. Graves, and Simon L. Currie. 2018. “Electrostatic Repulsion Causes Anticooperative DNA Binding between Tumor Suppressor ETS Transcription Factors and JUN–FOS at Composite DNA Sites.” *Journal of Biological Chemistry* 293 (48): 18624–35. https://doi.org/10.1074/JBC.RA118.003352/ATTACHMENT/C0EFF5F5-D2A4-443D-89E7-0E03F8591F4B/MMC1.ZIP.

Mai, Sanyue, Xiuhua Qu, Ping Li, Qingjun Ma, Cheng Cao, and Xuan Liu. 2016. “Global Regulation of Alternative RNA Splicing by the SR-Rich Protein RBM39.” *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* 1859 (8): 1014–24. https://doi.org/10.1016/J.BBAGRM.2016.06.007.

Mandel, C. R., Y. Bai, and L. Tong. 2007. “Protein Factors in Pre-MRNA 3′-End Processing.” *Cellular and Molecular Life Sciences 2007 65:7* 65 (7): 1099–1122. https://doi.org/10.1007/S00018-007-7474-3.

Mandel, Corey R., Syuzo Kaneko, Hailong Zhang, Damara Gebauer, Vasupradha Vethantham, James L. Manley, and Liang Tong. 2006. “Polyadenylation Factor CPSF-73 Is the Pre-MRNA 3’-End-Processing Endonuclease.” *Nature* 444 (7121): 953–56. https://doi.org/10.1038/NATURE05363.

Mariella, Elisa, Federico Marotta, Elena Grassi, Stefano Gilotto, and Paolo Provero. 2019. “The Length of the Expressed 3′ UTR Is an Intermediate Molecular Phenotype Linking Genetic Variants to Complex Diseases.” *Frontiers in Genetics* 10 (JUL): 714. https://doi.org/10.3389/FGENE.2019.00714/BIBTEX.

Martin, Georges, Andreas R Gruber, Walter Keller, and Mihaela Zavolan. 2012. “Genome-Wide Analysis of Pre-MRNA 3&prime; End Processing Reveals a Decisive Role of Human Cleavage Factor I in the Regulation of 3&prime; UTR Length.” https://doi.org/10.1016/j.celrep.2012.05.003.

Martin, Marcel. 2011. “Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads.” *EMBnet.Journal* 17 (1): 10–12. https://doi.org/10.14806/EJ.17.1.200.

Marvin, Bonnie, and Maki Inada. 2018. “Co-Transcriptional MRNA Processing in Eukaryotes.” *Molecular Life Sciences*, 137–46. https://doi.org/10.1007/978-1-4614-1531-2_41.

Masamha, Chioniso P., Zheng Xia, Jingxuan Yang, Todd R. Albrecht, Min Li, Ann-Bin Shyu, Wei Li, and Eric J. Wagner. 2014. “CFIm25 Links Alternative Polyadenylation to Glioblastoma Tumour Suppression.” *Nature* 510 (7505): 412–16. https://doi.org/10.1038/nature13261.

Masuda, Akio, Jun ichi Takeda, and Kinji Ohno. 2016. “FUS-Mediated Regulation of Alternative RNA Processing in Neurons: Insights from Global Transcriptome Analysis.”

Wiley Interdisciplinary Reviews: RNA 7 (3): 330–40. https://doi.org/10.1002/WRNA.1338.

Masuda, Akio, Jun Ichi Takeda, Tatsuya Okuno, Takaaki Okamoto, Bisei Ohkawara, Mikako Ito, Shinsuke Ishigaki, Gen Sobue, and Kinji Ohno. 2015. "Position-Specific Binding of FUS to Nascent RNA Regulates MRNA Length." *Genes & Development* 29 (10): 1045–57. https://doi.org/10.1101/GAD.255737.114.

Mayeda, Akila, Joseph Badolato, Ryuji Kobayashi, Michael Q. Zhang, Edith M. Gardiner, and Adrian R. Krainer. 1999. "Purification and Characterization of Human RNPS1: A General Activator of Pre-MRNA Splicing." *The EMBO Journal* 18 (16): 4560–70. https://doi.org/10.1093/EMBOJ/18.16.4560.

Mayr, Christine. 2019. "What Are 3′ UTRs Doing?" *Cold Spring Harbor Perspectives in Biology* 11 (10): a034728. https://doi.org/10.1101/CSHPERSPECT.A034728.

Mayr, Christine, and David P. Bartel. 2009. "Widespread Shortening of 3′UTRs by Alternative Cleavage and Polyadenylation Activates Oncogenes in Cancer Cells." *Cell*. https://doi.org/10.1016/j.cell.2009.06.016.

McCracken, Susan, Nova Fong, Krassimir Yankulov, Scott Ballantyne, Guohua Pan, Jack Greenblatt, Scott D. Patterson, Marvin Wickens, and David L. Bentley. 1997. "The C-Terminal Domain of RNA Polymerase II Couples MRNA Processing to Transcription." *Nature 1997 385:6614* 385 (6614): 357–61. https://doi.org/10.1038/385357a0.

McCracken, Susan, Dasa Longman, Iain L. Johnstone, Javier F. Cáceres, and Benjamin J. Blencowe. 2003. "An Evolutionarily Conserved Role for SRm160 in 3′-End Processing That Functions Independently of Exon Junction Complex Formation." *Journal of Biological Chemistry* 278 (45): 44153–60. https://doi.org/10.1074/jbc.M306856200.

Millevoi, Stefania, Clarisse Loulergue, Sabine Dettwiler, Sarah Zeïneb Karaa, Walter Keller, Michael Antoniou, and Stéphan Vagner. 2006. "An Interaction between U2AF 65 and CF I m Links the Splicing and 3′ End Processing Machineries." *EMBO Journal* 25 (20): 4854–64. https://doi.org/10.1038/sj.emboj.7601331.

Minvielle-Sebastia, Lionel, Katrin Beyer, Annette M. Krecic, Ron E. Hector, Maurice S. Swanson, and Walter Keller. 1998. "Control of Cleavage Site Selection during MRNA 3' End Formation by a Yeast HnRNP." *The EMBO Journal* 17 (24): 7454–68. https://doi.org/10.1093/EMBOJ/17.24.7454.

Moore, Melissa J, Charles C Query, and Phillip A Sharp. 1993. "Splicing of Precursors to MRNA by the Spliceosome." www.cshlpress.com/copyright.

Movassat, Maliheh, Tara L. Crabb, Anke Busch, Chengguo Yao, Derrick J. Reynolds, Yongsheng Shi, and Klemens J. Hertel. 2016. "Coupling between Alternative Polyadenylation and Alternative Splicing Is Limited to Terminal Introns." *RNA Biology* 13 (7): 646–55. https://doi.org/10.1080/15476286.2016.1191727/SUPPL_FILE/KRNB_A_1191727_SM6624.ZIP.

Murthy, K. G.K., and J. L. Manley. 1992. "Characterization of the Multisubunit Cleavage-Polyadenylation Specificity Factor from Calf Thymus." *Journal of Biological Chemistry*

267 (21): 14804–11. https://doi.org/10.1016/S0021-9258(18)42111-4.

———. 1995. “The 160-KD Subunit of Human Cleavage-Polyadenylation Specificity Factor Coordinates Pre-MRNA 3’-End Formation.” *Genes & Development* 9 (21): 2672–83. https://doi.org/10.1101/GAD.9.21.2672.

Nakagawa, Masato, Michiyo Koyanagi, Koji Tanabe, Kazutoshi Takahashi, Tomoko Ichisaka, Takashi Aoi, Keisuke Okita, Yuji Mochiduki, Nanako Takizawa, and Shinya Yamanaka. 2008. “Generation of Induced Pluripotent Stem Cells without Myc from Mouse and Human Fibroblasts.” *Nature Biotechnology* 26 (1): 101–6. https://doi.org/10.1038/nbt1374.

Neve, Jonathan, Kaspar Burger, Wencheng Li, Mainul Hoque, Radhika Patel, Bin Tian, Monika Gullerova, and Andre Furger. 2016. “Subcellular RNA Profiling Links Splicing and Nuclear DICER1 to Alternative Cleavage and Polyadenylation.” *Genome Research* 26 (1): 24–35. https://doi.org/10.1101/GR.193995.115.

Neve, Jonathan, Radhika Patel, Zhiqiao Wang, Alastair Louey, and Andr Martin Furger. 2017. “Cleavage and Polyadenylation: Ending the Message Expands Gene Regulation.” https://doi.org/10.1080/15476286.2017.1306171.

Newnham, Catherine M, Tyra Hall-Pogar, Songchun Liang, Jing Wu, Bin Tian, Jim Hu & Carol, and S S Lutz. 2010. “Alternative Polyadenylation of MeCP2: Influence of Cis-Acting Elements and Trans-Acting Factors.” *Www.Landesbioscience.Com RNA Biology* 7 (3): 361–72. https://doi.org/10.4161/rna.7.3.11564.

Nie, Zuqin, Gangqing Hu, Gang Wei, Kairong Cui, Arito Yamane, Wolfgang Resch, Ruoning Wang, et al. 2012. “C-Myc Is a Universal Amplifier of Expressed Genes in Lymphocytes and Embryonic Stem Cells.” *Cell* 151 (1): 68–79. https://doi.org/10.1016/j.cell.2012.08.033.

Nostrand, Eric L. Van, Gabriel A. Pratt, Alexander A. Shishkin, Chelsea Gelboin-Burkhart, Mark Y. Fang, Balaji Sundararaman, Steven M. Blue, et al. 2016. “Robust Transcriptome-Wide Discovery of RNA-Binding Protein Binding Sites with Enhanced CLIP (ECLIP).” *Nature Methods 2016 13:6* 13 (6): 508–14. https://doi.org/10.1038/nmeth.3810.

Nostrand, Eric L. Van, Gabriel A. Pratt, Brian A. Yee, Emily C. Wheeler, Steven M. Blue, Jasmine Mueller, Samuel S. Park, et al. 2020. “Principles of RNA Processing from Analysis of Enhanced CLIP Maps for 150 RNA Binding Proteins.” *Genome Biology* 21 (1): 1–26. https://doi.org/10.1186/S13059-020-01982-9/FIGURES/8.

Nunes, Nuno Miguel, Wencheng Li, Bin Tian, and André Furger. 2010. “A Functional Human Poly(A) Site Requires Only a Potent DSE and an A-Rich Upstream Sequence.” *The EMBO Journal* 29 (9): 1523–36. https://doi.org/10.1038/EMBOJ.2010.42.

Ogorodnikov, Anton, Michal Levin, Surendra Tattikota, Sergey Tokalov, Mainul Hoque, Denise Scherzinger, Federico Marini, et al. 2018. “Transcriptome 3′end Organization by PCF11 Links Alternative Polyadenylation to Formation and Neuronal Differentiation of Neuroblastoma.” *Nature Communications 2018 9:1* 9 (1): 1–16. https://doi.org/10.1038/s41467-018-07580-5.

Oliveira, Camila, Helisson Faoro, Lysangela Ronalte Alves, and Samuel Goldenberg. 2017. “RNA-Binding Proteins and Their Role in the Regulation of Gene expressionin

Trypanosoma Cruzi and Saccharomyces Cerevisiae.” *Genetics and Molecular Biology* 40 (1): 22. https://doi.org/10.1590/1678-4685-GMB-2016-0258.

Pan, Qun, Ofer Shai, Leo J. Lee, Brendan J. Frey, and Benjamin J. Blencowe. 2008. “Deep Surveying of Alternative Splicing Complexity in the Human Transcriptome by High-Throughput Sequencing.” *Nature Genetics* 40 (12): 1413–15. https://doi.org/10.1038/NG.259.

Pan, Zhenhua, Haibo Zhang, Lisa K. Hague, Ju Youn Lee, Carol S. Lutz, and Bin Tian. 2006. “An Intronic Polyadenylation Site in Human and Mouse CstF-77 Genes Suggests an Evolutionarily Conserved Regulatory Mechanism.” *Gene* 366 (2): 325–34. https://doi.org/10.1016/J.GENE.2005.09.024.

Pandya-Jones, Amy, and Douglas L. Black. 2009. “Co-Transcriptional Splicing of Constitutive and Alternative Exons.” *RNA* 15 (10): 1896–1908. https://doi.org/10.1261/RNA.1714509.

Pelosse, Martin, Hannah Crocker, Barbara Gorda, Paul Lemaire, Jens Rauch, and Imre Berger. 2017. “MultiBac: From Protein Complex Structures to Synthetic Viral Nanosystems.” *BMC Biology* 15 (1): 1–10. https://doi.org/10.1186/S12915-017-0447-6/FIGURES/6.

Qin, Hai, Haiwei Ni, Yichen Liu, Yaqin Yuan, Tao Xi, Xiaoman Li, and Lufeng Zheng. 2020. “RNA-Binding Proteins in Tumor Progression.” *Journal of Hematology and Oncology* 13 (1): 1–23. https://doi.org/10.1186/S13045-020-00927-W/TABLES/2.

Ramírez, Fidel, Friederike Dündar, Sarah Diehl, Björn A. Grüning, and Thomas Manke. 2014. “DeepTools: A Flexible Platform for Exploring Deep-Sequencing Data.” *Nucleic Acids Research* 42 (Web Server issue): W187. https://doi.org/10.1093/NAR/GKU365.

Rappsilber, Juri, Ursula Ryder, Angus I Lamond, and Matthias Mann. 2002. “Large-Scale Proteomic Analysis of the Human Spliceosome.” *Genome Research* 12 (8): 1231–45. https://doi.org/10.1101/gr.473902.

Reed, R. 1989. “The Organization of 3’ Splice-Site Sequences in Mammalian Introns.” *Genes & Development* 3 (12b): 2113–23. https://doi.org/10.1101/GAD.3.12B.2113.

Roca, Xavier, Ravi Sachidanandam, and Adrian R. Krainer. 2005. “Determinants of the Inherent Strength of Human 5′ Splice Sites.” *RNA* 11 (5): 683. https://doi.org/10.1261/RNA.2040605.

Ross, Nathan T., Felix Lohmann, Seth Carbonneau, Aleem Fazal, Wilhelm A. Weihofen, Scott Gleim, Michael Salcius, et al. 2020. “CPSF3-Dependent Pre-MRNA Processing as a Druggable Node in AML and Ewing’s Sarcoma.” *Nature Chemical Biology* 16 (1): 50–59. https://doi.org/10.1038/S41589-019-0424-1.

Rüegsegger, Ursula, Katrin Beyer, and Walter Keller. 1996. “Purification and Characterization of Human Cleavage Factor Im Involved in the 3’ End Processing of Messenger RNA Precursors.” *The Journal of Biological Chemistry* 271 (11): 6107–13. https://doi.org/10.1074/JBC.271.11.6107.

Rüegsegger, Ursula, Diana Blank, and Walter Keller. 1998. “Human Pre-MRNA Cleavage Factor Im Is Related to Spliceosomal SR Proteins and Can Be Reconstituted In Vitro from Recombinant Subunits.” *Molecular Cell* 1 (2): 243–53. https://doi.org/10.1016/S1097-

2765(00)80025-8.

Ryan, Kevin, and David L.V. Bauer. 2008. “Finishing Touches: Post-Translational Modification of Protein Factors Involved in Mammalian Pre-MRNA 3′ End Formation.” *The International Journal of Biochemistry & Cell Biology* 40 (11): 2384. https://doi.org/10.1016/J.BIOCEL.2008.03.016.

Saha, Kaushik, and Gourisankar Ghosh. 2022. “Cooperative Engagement and Subsequent Selective Displacement of SR Proteins Define the Pre-MRNA 3D Structural Scaffold for Early Spliceosome Assembly.” *Nucleic Acids Research* 50 (14): 8262–78. https://doi.org/10.1093/NAR/GKAC636.

Sakashita, Eiji, Sawako Tatsumi, Dieter Werner, Hitoshi Endo, and Akila Mayeda. 2004. “Human RNPS1 and Its Associated Factors: A Versatile Alternative Pre-MRNA Splicing Regulator In Vivo.” *Molecular and Cellular Biology* 24 (3): 1174–87. https://doi.org/10.1128/MCB.24.3.1174-1187.2004/ASSET/FAE1F7B7-0E87-43C3-ADC8-DC240B4BFB1E/ASSETS/GRAPHIC/ZMB00304094200T2.JPEG.

Salanga, Cristy M., and Matthew C. Salanga. 2021. “Genotype to Phenotype: CRISPR Gene Editing Reveals Genetic Compensation as a Mechanism for Phenotypic Disjunction of Morphants and Mutants.” *International Journal of Molecular Sciences* 22 (7). https://doi.org/10.3390/IJMS22073472.

Sandberg, Rickard, Joel R Neilson, Arup Sarma, Phillip A Sharp, and Christopher B Burge. 2008. “Proliferating Cells Express MRNAs with Shortened 3’ Untranslated Regions and Fewer MicroRNA Target Sites.” *Science (New York, N.Y.)* 320 (5883): 1643–47. https://doi.org/10.1126/science.1155390.

Schmidt, Moritz, Florian Kluge, Felix Sandmeir, Uwe Kühn, Peter Schäfer, Christian Tüting, Christian Ihling, Elena Conti, and Elmar Wahle. 2022. “Reconstitution of 3’ End Processing of Mammalian Pre-MRNA Reveals a Central Role of RBBP6.” *Genes and Development* 36 (3–4): 195–209. https://doi.org/10.1101/GAD.349217.121/-/DC1.

Schwich, Oliver Daniel, Nicole Blümel, Mario Keller, Marius Wegener, Samarth Thonta Setty, Melinda Elaine Brunstein, Ina Poser, et al. 2021. “SRSF3 and SRSF7 Modulate 3′UTR Length through Suppression or Activation of Proximal Polyadenylation Sites and Regulation of CFIm Levels.” *Genome Biology* 22 (1): 1–34. https://doi.org/10.1186/S13059-021-02298-Y/FIGURES/8.

Scotti, Marina M., and Maurice S. Swanson. 2015. “RNA Mis-Splicing in Disease.” *Nature Reviews Genetics 2015 17:1* 17 (1): 19–32. https://doi.org/10.1038/nrg.2015.3.

Shah, Ankeeta, Yingzhi Qian, Sebastien M. Weyn-Vanhentenryck, and Chaolin Zhang. 2017. “CLIP Tool Kit (CTK): A Flexible and Robust Pipeline to Analyze CLIP Sequencing Data.” *Bioinformatics* 33 (4): 566–67. https://doi.org/10.1093/BIOINFORMATICS/BTW653.

Shao, Changwei, Bo Yang, Tongbin Wu, Jie Huang, Peng Tang, Yu Zhou, Jie Zhou, et al. 2014. “Mechanisms for U2AF to Define 3′ Splice Sites and Regulate Alternative Splicing in the Human Genome.” *Nature Structural & Molecular Biology* 21 (11): 997. https://doi.org/10.1038/NSMB.2906.

Shen, Shihao, Juw Won Park, Zhi-xiang Lu, Lan Lin, Michael D Henry, Ying Nian Wu, Qing Zhou, and Yi Xing. 2014. "RMATS: Robust and Flexible Detection of Differential Alternative Splicing from Replicate RNA-Seq Data." *Proceedings of the National Academy of Sciences of the United States of America* 111 (51): E5593-601. https://doi.org/10.1073/pnas.1419161111.

Shen, Ting, Huan Li, Yifang Song, Li Li, Jinzhong Lin, Gang Wei, and Ting Ni. 2019. "Alternative Polyadenylation Dependent Function of Splicing Factor SRSF3 Contributes to Cellular Senescence." *Aging (Albany NY)* 11 (5): 1356. https://doi.org/10.18632/AGING.101836.

Shepard, Peter J., and Klemens J. Hertel. 2009. "The SR Protein Family." *Genome Biology* 10 (10): 242. https://doi.org/10.1186/GB-2009-10-10-242/FIGURES/6.

Shi, Yongsheng, Dafne Campigli Di Giammartino, Derek Taylor, Ali Sarkeshik, William J. Rice, John R. Yates, Joachim Frank, and James L. Manley. 2009. "Molecular Architecture of the Human Pre-MRNA 3' Processing Complex." *Molecular Cell* 33 (3): 365–76. https://doi.org/10.1016/J.MOLCEL.2008.12.028.

Shi, Yongsheng, and James L. Manley. 2015. "The End of the Message: Multiple Protein???RNA Interactions Define the MRNA Polyadenylation Site." *Genes and Development* 29 (9): 889–97. https://doi.org/10.1101/gad.261974.115.

Sousa Abreu, Raquel De, Luiz O. Penalva, Edward M. Marcotte, and Christine Vogel. 2009. "Global Signatures of Protein and MRNA Expression Levels." *Molecular BioSystems* 5 (12): 1512. https://doi.org/10.1039/B908315D.

Stepanyuk, Galina A., Pedro Serrano, Eigen Peralta, Carol L. Farr, Herbert L. Axelrod, Michael Geralt, Debanu Das, et al. 2016. "UHM–ULM Interactions in the RBM39–U2AF65 Splicing-Factor Complex." *Acta Crystallographica. Section D, Structural Biology* 72 (Pt 4): 497. https://doi.org/10.1107/S2059798316001248.

Suda, Kenichi, Leslie Rozeboom, Hui Yu, Kim Ellison, Christopher J. Rivard, Tetsuya Mitsudomi, and Fred R. Hirsch. 2017. "Potential Effect of Spliceosome Inhibition in Small Cell Lung Cancer Irrespective of the MYC Status." *PLoS ONE* 12 (2). https://doi.org/10.1371/JOURNAL.PONE.0172209.

Sun, Dongying, Wei Lei, Xiaodong Hou, Hui Li, and Wenlu Ni. 2019. "PUF60 Accelerates the Progression of Breast Cancer through Downregulation of PTEN Expression." *Cancer Management and Research* 11: 821. https://doi.org/10.2147/CMAR.S180242.

Sun, Yadong, Yixiao Zhang, Keith Hamilton, James L. Manley, Yongsheng Shi, Thomas Walz, and Liang Tong. 2018. "Molecular Basis for the Recognition of the Human AAUAAA Polyadenylation Signal." *Proceedings of the National Academy of Sciences of the United States of America* 115 (7): E1419–28. https://doi.org/10.1073/PNAS.1718723115/SUPPL_FILE/PNAS.1718723115.SAPP.PDF.

Takagaki, Y., J. L. Manley, C. C. MacDonald, J. Wilusz, and T. Shenk. 1990. "A Multisubunit Factor, CstF, Is Required for Polyadenylation of Mammalian Pre-MRNAs." *Genes & Development* 4 (12A): 2112–20. https://doi.org/10.1101/GAD.4.12A.2112.

Takagaki, Y, and J L Manley. 1997. "RNA Recognition by the Human Polyadenylation Factor CstF." *Molecular and Cellular Biology* 17 (7): 3907–14. https://doi.org/10.1128/mcb.17.7.3907.

Takagaki, Yoshio, and James L. Manley. 2000. "Complex Protein Interactions within the Human Polyadenylation Machinery Identify a Novel Component." *Molecular and Cellular Biology* 20 (5): 1515–25. https://doi.org/10.1128/MCB.20.5.1515-1525.2000.

Takahashi, Kazutoshi, and Shinya Yamanaka. 2006. "Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors." *Cell* 126 (4): 663–76. https://doi.org/10.1016/j.cell.2006.07.024.

Tian, Bin, Jun Hu, Haibo Zhang, and Carol S. Lutz. 2005. "A Large-Scale Analysis of MRNA Polyadenylation of Human and Mouse Genes." *Nucleic Acids Research* 33 (1): 201–12. https://doi.org/10.1093/NAR/GKI158.

Tian, Bin, and James L Manley. 2017. "Alternative Polyadenylation of MRNA Precursors HHS Public Access." *Nat Rev Mol Cell Biol* 18 (1): 18–30. https://doi.org/10.1038/nrm.2016.116.

Tseng, Hsin-Wei, Anthony Mota-Sydor, Rania Leventis, Ivan Topisirovic, Thomas F Duchaine, Morris Goodman, and Gerald Bronfman. 2021. "Distinct, Opposing Functions for CFIm59 and CFIm68 in MRNA Alternative Polyadenylation of Pten and in the PI3K/Akt Signalling Cascade." *BioRxiv*, September, 2021.09.09.459613. https://doi.org/10.1101/2021.09.09.459613.

Tushev, Georgi, Caspar Glock, Maximilian Heumüller, Anne Biever, Marko Jovanovic, and Erin M. Schuman. 2018. "Alternative 3′ UTRs Modify the Localization, Regulatory Potential, Stability, and Plasticity of MRNAs in Neuronal Compartments." *Neuron* 98 (3): 495-511.e6. https://doi.org/10.1016/J.NEURON.2018.03.030.

Vagner, Stéphan, Christine Vagner, and Iain W. Mattaj. 2000. "The Carboxyl Terminus of Vertebrate Poly(A) Polymerase Interacts with U2AF 65 to Couple 3′-End Processing and Splicing." *Genes & Development* 14 (4): 403. https://doi.org/10.1101/gad.14.4.403.

Vasudevan, Shobha, Stuart W. Peltz, and Carol J. Wilusz. 2002. "Non-Stop Decay--a New MRNA Surveillance Pathway." *BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology* 24 (9): 785–88. https://doi.org/10.1002/BIES.10153.

Venkataraman, Krishnan, Kirk M. Brown, and Gregory M. Gilmartin. 2005. "Analysis of a Noncanonical Poly(A) Site Reveals a Tripartite Mechanism for Vertebrate Poly(A) Site Recognition." *Genes & Development* 19 (11): 1315. https://doi.org/10.1101/GAD.1298605.

Veraldi, Kristen L., George K. Arhin, Kathleen Martincic, Ling-Hsiu Chung-Ganster, Jeffrey Wilusz, and Christine Milcarek. 2001. "HnRNP F Influences Binding of a 64-Kilodalton Subunit of Cleavage Stimulation Factor to MRNA Precursors in Mouse B Cells." *Molecular and Cellular Biology* 21 (4): 1228. https://doi.org/10.1128/MCB.21.4.1228-1238.2001.

Verfaillie, Catherine M. 2004. "'Adult' Stem Cells: Tissue Specific or Not?" *Handbook of Stem Cells* 2 (January): 13–20. https://doi.org/10.1016/B978-012436643-5/50092-4.

Vogel, Christine, and Edward M. Marcotte. 2012. "Insights into the Regulation of Protein

Abundance from Proteomic and Transcriptomic Analyses." *Nature Reviews Genetics 2012 13:4* 13 (4): 227–32. https://doi.org/10.1038/nrg3185.

Vries, Henk De, Ursula Rüegsegger, Wolfgang Hübner, Arno Friedlein, Hanno Langen, and Walter Keller. 2000. "Human Pre-MRNA Cleavage Factor II(m) Contains Homologs of Yeast Proteins and Bridges Two Other Cleavage Factors." *The EMBO Journal* 19 (21): 5895–5904. https://doi.org/10.1093/EMBOJ/19.21.5895.

Wagner, Rebecca E., and Michaela Frye. 2021. "Noncanonical Functions of the Serine-Arginine-Rich Splicing Factor (SR) Family of Proteins in Development and Disease." *BioEssays* 43 (4): 2000242. https://doi.org/10.1002/BIES.202000242.

Wang, Ruijia, Dinghai Zheng, Lu Wei, Qingbao Ding, and Bin Tian. 2019. "Regulation of Intronic Polyadenylation by PCF11 Impacts MRNA Expression of Long Genes." *Cell Reports* 26 (10): 2766-2778.e6. https://doi.org/10.1016/j.celrep.2019.02.049.

Wang, Xiaojing, Simona G Codreanu, Bo Wen, Kai Li, Matthew C Chambers, Daniel C Liebler, and Bing Zhang. 2018. "Detection of Proteome Diversity Resulted from Alternative Splicing Is Limited by Trypsin Cleavage Specificity* □ S." *Molecular & Cellular Proteomics* 17: 422–30. https://doi.org/10.1074/mcp.RA117.000155.

Weiße, Jonas, Julia Rosemann, Vanessa Krauspe, Matthias Kappler, Alexander W. Eckert, Monika Haemmerle, and Tony Gutschner. 2020. "RNA-Binding Proteins as Regulators of Migration, Invasion and Metastasis in Oral Squamous Cell Carcinoma." *International Journal of Molecular Sciences* 21 (18): 1–28. https://doi.org/10.3390/IJMS21186835.

Wilkinson, Max E, Clément Charenton, and Kiyoshi Nagai. 2020. "Annual Review of Biochemistry RNA Splicing by the Spliceosome." https://doi.org/10.1146/annurev-biochem-091719.

Wilusz1, Jeffrey, and Thomas Shenk2. 1990. "A Uridylate Tract Mediates Efficient Heterogeneous Nuclear Ribonucleoprotein C Protein-RNA Cross-Linking and Functionally Substitutes for the Downstream Element of the Polyadenylation Signal." *Molecular and Cellular Biology* 10 (12): 6397–6407. https://doi.org/10.1128/MCB.10.12.6397-6407.1990.

Wu, Jane Y., and Tom Maniatis. 1993. "Specific Interactions between Proteins Implicated in Splice Site Selection and Regulated Alternative Splicing." *Cell* 75 (6): 1061–70. https://doi.org/10.1016/0092-8674(93)90316-I.

Wu, Tongbin, and Xiang Dong Fu. 2015. "Genomic Functions of U2AF in Constitutive and Regulated Splicing." *RNA Biology* 12 (5): 479. https://doi.org/10.1080/15476286.2015.1020272.

Xia, Zheng, Lawrence A. Donehower, Thomas A. Cooper, Joel R. Neilson, David A. Wheeler, Eric J. Wagner, and Wei Li. 2014. "Dynamic Analyses of Alternative Polyadenylation from RNA-Seq Reveal a 3′-UTR Landscape across Seven Tumour Types." *Nature Communications* 5 (1): 5274. https://doi.org/10.1038/ncomms6274.

Xiao, Shou Hua, and James L. Manley. 1997. "Phosphorylation of the ASF/SF2 RS Domain Affects Both Protein-Protein and Protein-RNA Interactions and Is Necessary for Splicing." *Genes & Development* 11 (3): 334–44. https://doi.org/10.1101/GAD.11.3.334.

Xu, Caipeng, Xiaohua Chen, Xuetian Zhang, Dapeng Zhao, Zhihui Dou, Xiaodong Xie, Hongyan Li, et al. 2021. “RNA-Binding Protein 39: A Promising Therapeutic Target for Cancer.” *Cell Death Discovery 2021 7:1* 7 (1): 1–9. https://doi.org/10.1038/s41420-021-00598-7.

Yang, Qin, Molly Coseno, Gregory M. Gilmartin, and Sylvie Doublié. 2011. “Crystal Structure of a Human Cleavage Factor CFI(m)25/CFI(m)68/RNA Complex Provides an Insight into Poly(A) Site Recognition and RNA Looping.” *Structure (London, England : 1993)* 19 (3): 368–77. https://doi.org/10.1016/J.STR.2010.12.021.

Yang, Qin, Gregory M Gilmartin, and Sylvie Doublié. 2010. “Structural Basis of UGUA Recognition by the Nudix Protein CFI(m)25 and Implications for a Regulatory Role in MRNA 3’ Processing.” *Proceedings of the National Academy of Sciences of the United States of America* 107 (22): 10062–67. https://doi.org/10.1073/pnas.1000848107.

Yang, Wen, Peter L. Hsu, Fan Yang, Jae Eun Song, and Gabriele Varani. 2018. “Reconstitution of the CstF Complex Unveils a Regulatory Role for CstF-50 in Recognition of 3′-End Processing Signals.” *Nucleic Acids Research* 46 (2): 493–503. https://doi.org/10.1093/NAR/GKX1177.

Yao, Chengguo, Jacob Biesinger, Ji Wan, Lingjie Weng, Yi Xing, Xiaohui Xie, and Yongsheng Shi. 2012. “Transcriptome-Wide Analyses of CstF64-RNA Interactions in Global Regulation of MRNA Alternative Polyadenylation.” *Proceedings of the National Academy of Sciences of the United States of America* 109 (46): 18773–78. https://doi.org/10.1073/PNAS.1211101109/SUPPL_FILE/SD01.XLSX.

Yao, Peng, Alka A. Potdar, Abul Arif, Partho Sarothi Ray, Rupak Mukhopadhyay, Belinda Willard, Yichi Xu, Jun Yan, Gerald M. Saidel, and Paul L. Fox. 2012. “Coding Region Polyadenylation Generates a Truncated TRNA Synthetase That Counters Translation Repression.” *Cell* 149 (1): 88–100. https://doi.org/10.1016/J.CELL.2012.02.018.

Yeo, Gene, and Christopher B. Burge. 2004. “Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals.” *Journal of Computational Biology : A Journal of Computational Molecular Cell Biology* 11 (2–3): 377–94. https://doi.org/10.1089/1066527041410418.

Yoon, Yoseop, Lindsey V. Soles, and Yongsheng Shi. 2021. “PAS-Seq 2: A Fast and Sensitive Method for Global Profiling of Polyadenylated RNAs.” *Methods in Enzymology* 655 (January): 25–35. https://doi.org/10.1016/BS.MIE.2021.03.013.

Zeng, Changqing, and Susan M. Berget. 2000. “Participation of the C-Terminal Domain of RNA Polymerase II in Exon Definition during Pre-MRNA Splicing.” *Molecular and Cellular Biology* 20 (21): 8290. https://doi.org/10.1128/MCB.20.21.8290-8301.2000.

Zhang, Haibo, Jun Hu, Michael Recce, and Bin Tian. 2005. “PolyA_DB: A Database for Mammalian MRNA Polyadenylation.” *Nucleic Acids Research* 33 (Database issue). https://doi.org/10.1093/NAR/GKI055.

Zhang, Yuanjiao, Jinjun Qian, Chunyan Gu, and Ye Yang. 2021. “Alternative Splicing and Cancer: A Systematic Review.” *Signal Transduction and Targeted Therapy 2021 6:1* 6 (1): 1–14. https://doi.org/10.1038/s41392-021-00486-7.

Zheng, Christina L., F. U. Xiang-Dong, and Michael Gribskov. 2005. “Characteristics and Regulatory Elements Defining Constitutive Splicing and Different Modes of Alternative Splicing in Human and Mouse.” *RNA* 11 (12): 1777. https://doi.org/10.1261/RNA.2660805.

Zhou, Zhaolan, Lawrence J. Licklider, Steven P. Gygi, and Robin Reed. 2002. “Comprehensive Proteomic Analysis of the Human Spliceosome.” *Nature* 419 (6903): 182–85. https://doi.org/10.1038/nature01031.

Zhu, Yong, Xiuye Wang, Elmira Forouzmand, Joshua Jeong, Feng Qiao, Gregory A. Sowd, Alan N. Engelman, Xiaohui Xie, Klemens J. Hertel, and Yongsheng Shi. 2017. “Molecular Mechanisms for CFIm-Mediated Regulation of MRNA Alternative Polyadenylation.” *Molecular Cell*, 1–13. https://doi.org/10.1016/j.molcel.2017.11.031.