In vitro reconstitution of the human pre-mRNA 3′ end processing machinery and mechanistic insights into endonuclease activation

Vytaute Boreikaite

Jesus College

September 2022

This thesis is submitted for the degree of Doctor of Philosophy

MRC Laboratory of Molecular Biology

UNIVERSITY OF CAMBRIDGE

Preface

This Thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the preface and specified in the text.

The work described in this Thesis is not substantially the same as any work that has already been submitted before for any degree or other qualification except as declared in the preface and specified in the text.

The majority of Chapter 1 has been submitted as a review article to *Annual Review of Biochemistry*:

> **Boreikaite V**, Passmore LA. 3′ end processing of eukaryotic mRNA: Machinery, regulation, and impact on gene expression.

Some of the work described in Chapter 2-4 has been published:

> **Boreikaite V**, Elliott TS, Chin JW, Passmore LA. 2022. RBBP6 activates the pre-mRNA 3' end processing machinery in humans. *Genes Dev.* 36:210–24.

This Thesis does not exceed the prescribed word limit for the Degree Committee of Biology.

Vytaute Boreikaite

Submitted for examination: September 2022

Vytaute Boreikaite

In vitro reconstitution of the human pre-mRNA 3′ end processing machinery and mechanistic insights into endonuclease activation

Summary

Maturation of protein-coding transcripts in eukaryotes involves several processing steps, including 5′ capping, splicing, and 3′ end processing. The latter entails endonucleolytic cleavage of the nascent pre-mRNA and addition of a poly(A) tail to the resultant free 3′ end. The poly(A) tail then facilitates nuclear export of mRNAs and controls their stability and translational efficiency in the cytoplasm. Pre-mRNA 3′ end processing is also tightly coupled to transcription termination. Defects in 3′ end processing cause a variety of human diseases, highlighting its critical role in gene expression.

In humans, 3′ end processing of most pre-mRNAs is carried out by a seven-subunit protein complex known as cleavage and polyadenylation specificity factor (CPSF; CPF in yeast). The polyadenylation activity of CPSF has been studied in detail biochemically, but our understanding of how CPSF cleaves the pre-mRNA remains limited. CPSF is an inherently inactive endonuclease on its own and was believed to require additional protein factors for its activation, enabling tight regulation of 3′ cleavage. To gain insight into the activation mechanism of the 3′ endonuclease, a minimal *in vitro* system with a well-defined protein composition is required, but this has eluded researchers for decades.

In this dissertation, I have reconstituted specific and efficient pre-mRNA cleavage activity by the human CPSF complex with purified recombinant proteins. I have determined that activation of the CPSF endonuclease requires three additional protein factors: cleavage stimulatory factor (CStF), cleavage factor IIm (CFIIm), and, importantly, a multidomain protein RBBP6. The role of RBBP6 in 3′ end processing in humans has been largely overlooked, and therefore, I studied this protein in more detail. The yeast orthologue of RBBP6, Mpe1, senses pre-mRNA binding and is a constitutive subunit of CPF. In contrast, by purifying endogenous CPSF from human cells, I show that RBBP6 is not a stable component of the human complex. Instead, my biochemical studies reveal that RBBP6 is likely recruited to CPSF in an RNA-dependent manner and that it also interacts with the CFIIm cleavage factor complex. My sequence and mutational analyses suggest that the role of RBBP6 in activating the CPSF endonuclease is conserved from yeast to human. I have also performed cryo-electron microscopy studies of some protein complexes involved in pre-mRNA 3′ end

processing in humans, aiming towards an atomic-level understanding of CPSF cleavage activity, which remains the major outstanding knowledge gap in the field of 3′ end processing.

Overall, the reconstitution of human pre-mRNA 3′ end processing with purified proteins described in this dissertation has enabled detailed mechanistic studies of CPSF structure and function, and may also facilitate the development of new therapeutics.

Acknowledgements

First and foremost, I would like to thank my doctoral supervisor Lori Passmore for her guidance throughout the past few years. As a summer student in her lab, I became inspired by Lori's enthusiasm for challenging scientific problems and soon knew that this environment would be an ideal one for my doctoral research. Indeed, her continuous support made my time at the LMB both a productive and enjoyable experience. I am incredibly grateful for the opportunity I had to drive such an interesting project, and for everything that I learnt along the way.

I would also like to express sincere gratitude to my second supervisor Joe Yeeles and my University supervisor Chris Smith. Their feedback, especially in the first year of my project, guided me through early setbacks and encouraged me to pursue difficult but exciting research questions.

My time as a doctoral student would not have been the same without the many amazingly talented and well-rounded people I met in Lori's lab. I am thankful to Chris Hill, who first initiated me into the field of 3′ end processing, and Ana Casañal and Ananth Kumar whose knowledge of protein biochemistry helped me to kick-start my project. I will dearly miss the lengthy discussions with Juan Rodriguez, Manuel Carminati and Holly Fagarasan – sharing ideas with you all was one of the best parts of my PhD days. I shall not forget the members of team "deadenylation" and team "Fanconi": Eva Absmeier, Terence Tang, James Stowell, Max Seidel, Cemre Manav, Tamara Sijacki, Pablo Alcon and Shabih Shakeel. I will always cherish the memories of our brunches, lunches, dinners, pandemic "Zoom" teas and all the fun we had inside and outside of the lab. You all have been *absolutely exquisite*!

My life was made so much easier by all the great scientific facilities at the LMB, including the media kitchen, the electron microscopy facility, the baculovirus facility led by the always kind and helpful Jianguo Shi and the mass spectrometry facility, especially Farida Begum and Sew-Yeu Peak-Chew. Thanks to their hard work, I could fully focus on my science. I would also like to thank Loic Carrique, Ervin Fodor, Steven West, Max Wilkinson and Thomas Elliott for sharing reagents with me, and John O'Donnell, Kelly Nguyen and Francis O'Reilly for helping me get started with mammalian cell culture work.

I am incredibly grateful to the Herchel Smith Fund that provided me with financial support for my doctoral studies. I greatly enjoyed the annual symposia where I met many brilliant

Herchel Smith scholars from both Cambridge and Harvard. Finally, I would like to thank the many staff members of Jesus College, Cambridge, for being extremely supportive in the challenging times of the pandemic.

I would not be where I am today without the continuous support of my parents, my gratitude to whom could not be expressed in words. I would also like to thank Chris for listening to of all my science stories and lab gossip, and patiently waiting for me to come home after a long day at the bench. I can honestly say that C helped make (hopefully soon-to-be) Dr Flute!

Table of Contents

Chapter 1: Introduction **21**

1.1 3′ end processing machinery **25**

1.1.2 mPSF or polymerase module 25

1.1.3 Recognition of the polyadenylation signal 26

1.1.4 Poly(A) polymerase 29

1.1.5 Nuclease module or mCF 31

1.1.6 RBBP6/Mpe1 33

1.1.7 Protein phosphatases of 3′ end processing complexes 34

1.2 Accessory factors regulate enzymatic activities of CPSF/CPF **36**

1.2.1 CStF, CFIIm/CF IA 36

1.2.2 CFIm/CF IB 36

1.2.3 Nuclear poly(A)-binding proteins control polyadenylation 37

1.3 3′ end processing of histone pre-mRNAs **38**

1.4 Regulation of 3′ end processing **39**

1.4.1 Mechanisms of alternative of polyadenylation 39

1.4.2 Regulation of 3′ end processing in disease 41

1.4.2.1 CPSF73 as a therapeutic target 41

1.4.2.2 3′ end processing machinery is targeted by viruses 42

1.5 Impact of 3′ end processing on gene expression **43**

1.5.1 Coordination with transcription termination 43

1.5.2 Coordination with splicing 45

1.6 Mechanism of CPSF endonuclease activation **47**

1.7 Aims of this Study **49**

Chapter 2: Purification of recombinant protein factors involved in human pre-mRNA 3′ end processing **50**

2.1 Purification and characterisation of recombinant human CPSF **52**

2.1.1 Purification of recombinant mPSF 52

2.1.2 Recombinant human mPSF is active in polyadenylation 55

2.1.3 Purification of recombinant mCF 57

2.1.4 Human mPSF and mCF form a stable CPSF complex in vitro 58

2.2 Purification and characterisation of recombinant human cleavage factors..... 60

2.2.1 Purification of recombinant CStF and CFIIm complexes 60

2.2.2 CStF and CFIIm form a stable complex..... 63

2.2.3 CStF enhances specificity of mPSF-dependent polyadenylation *in vitro* 65

2.2.4 Purification of recombinant CFIm 67

2.2.5 Purification of RBBP6 69

2.2.6 RBBP6 stimulates PAP activity independently of CPSF 71

2.3 Conclusions and perspectives 73

2.3.1 Production of human 3′ end processing factors requires deletions of several IDRs.....73

2.3.2 Covalent modifications may affect cleavage and polyadenylation activities of purified proteins 73

2.3.3 Several auxiliary factors regulate polyadenylation by mPSF..... 74

Chapter 3: *In vitro* reconstitution of CPSF endonuclease activity with purified proteins.....75

3.1 Determining the set of proteins required for CPSF endonuclease activation..... 76

3.2 Optimising conditions of recombinant CPSF endonuclease assay 79

3.2.1 Optimising buffer conditions 79

3.2.2 Chemical additives are not required for CPSF endonuclease activity..... 82

3.2.3 Optimising protein concentrations 83

3.3 Understanding the specificity of recombinant CPSF endonuclease activity..... 84

3.3.1 Recombinant CPSF endonuclease exhibits sequence specificity 84

3.3.2 Recombinant CPSF can cleave multiple substrates..... 87

3.3.3 CPSF73 is the only endonuclease within the CPSF complex 90

3.4 NS1 protein from Influenza A virus inhibits pre-mRNA 3′ end processing *in vitro* 92

3.4.1 NS1 inhibits cleavage activity of recombinant CPSF 92

3.4.2 Structural analysis of the mPSF-NS1 complex..... 94

3.5 A direct physical interaction between mPSF and mCF modules is not essential for endonuclease activation 98

3.6 Truncations of hFip1 and Pcf11 do not affect endonuclease activity of recombinant CPSF.....100

3.7 CFIm does not stimulate endonuclease activity of recombinant CPSF.....104

3.8 Recombinant CPSF catalyses coupled cleavage and polyadenylation.....107

3.9 CPSF and histone pre-mRNA 3′ end processing complexes are activated by different mechanisms 110

3.10 Conclusions and perspectives 112

3.10.1 CStF, CFIIm, RBBP6 and CPSF are required for pre-mRNA 3′ end cleavage 112

3.10.2 Endonuclease activation enforces the specificity of 3′ end processing 112

3.10.3 CPSF endonuclease could be activated under different conditions..... 113

3.10.4 Advantages and limitations of CPSF enzymatic activities reconstituted with purified proteins..... 115

Chapter 4: RBBP6 is a conserved activator of canonical pre-mRNA 3′ end processing 117

4.1 RBBP6 is not a constitutive subunit of human CPSF.....118

4.1.1 Over-expression of tagged subunits in mammalian cells leads to the purification of individual CPSF modules 118

4.1.2 Purifying endogenous CPSF from a stable cell line carrying a tagged subunit.... 122

4.1.3 RBBP6 is not a stable subunit of CPSF in human cells..... 124

4.2 Interactions between RBBP6 and CPSF.....127

4.2.1 RBBP6 is recruited to CPSF in an RNA-dependent manner..... 127

4.2.2 Structural analysis of the CPSF complex bound to RNA and RBBP6 130

4.2.3 RBBP6 may interact with CPSF73..... 133

4.2.4 RBBP6 may interact with mPSF..... 135

4.2.5 Interactions between RBBP6/Mpe1-PSR and mPSF/polymerase module may differ in yeast and humans 138

4.2.6 mPSF may hinder RBBP6 binding to CPSF73 in the context of the full CPSF complex.....142

4.3 Interactions between RBBP6 and cleavage factors 145

4.3.1 RBBP6 does not interact with CStF but binds weakly to CFIIm..... 145

4.3.2 mCF and CFIIm form a stable complex..... 147

4.3.3 RBBP6 interacts with the mCF-CFIIm complex in the absence of RNA 149

4.3.4 Structural studies of the mCF-CFIIm-RBBP6 complex..... 149

4.3.5 Identification of the interaction sites in the mCF-CFIIm-RBBP6 complex 152

4.4 Assembly of the complete active human pre-mRNA 3' end processing machinery *in vitro* 156

4.5 Conclusions and perspectives 158

4.5.1 RBBP6 is not a constitutive subunit of CPSF in humans 158

4.5.2 More CPSF interactors may await identification 159

4.5.3 RBBP6 interacts with CPSF and CFIIm 159

4.5.4 RBBP6 may mediate cross-talk between various components of the pre-mRNA 3' end processing machinery 160

4.5.5 Approaches to improve the preparation of CPSF complexes for cryoEM 161

4.5.4 AlphaFold is a powerful tool in studying protein complexes 162

Chapter 5: Conclusions and perspectives 164

5.1 Summary of this Thesis 165

5.2 Coordination of 3' end processing with other steps of mRNA biogenesis 166

5.3 Molecular mechanism of cleavage and polyadenylation 168

5.4 Final conclusions 171

Chapter 6: Materials and Methods 172

6.1 Cloning 173

6.1.1 General cloning methods 173

6.1.1.1 Polymerase chain reaction (PCR) and gel electrophoresis of DNA 173

6.1.1.2 Gibson assembly 173

6.1.1.3 Transformation of competent cells 174

6.1.1.4 Plasmid amplification and purification 174

6.1.1.5 Gene synthesis and Sanger sequencing 174

6.1.2 Cloning individual proteins and protein complexes for expression in insect cells 175

6.1.2.1 Tagging genes in pACEBAC vectors 175

6.1.2.2 hFip1$_{iso4}$ 175

6.1.2.3 Site-directed mutagenesis of CPSF73, CPSF100, CPSF30 and RBBP6 175

6.1.2.4 biGBac cloning 176

6.1.2.5 Cloning NS1 protein and its effector domain for co-expression with mPSF insect cells 176

6.1.3 Cloning for expression in *E. coli* 179

6.1.3.1 Cloning SSU72 179
6.1.3.2 Cloning NS1 and its effector domain for expression in *E. coli* 179
6.1.4 Cloning for expression in mammalian cells 179
6.1.4.1 CRISPR-Cas9 gene targeting in mammalian cells 179
6.1.4.2 Mammalian vectors for transient overexpression 179

6.2 Protein expression 180

6.2.1 Baculovirus-mediated protein expression in insect cells 180
6.2.1.1 Bacmid preparation 180
6.2.1.2 Propagation of baculovirus 181
6.2.2 Protein expression in *E. coli* 181

6.3 Protein purification 182

6.3.1 Protein analysis and quantification 182
6.3.2 mPSF-hFip1$_{iso4}$-SII: on its own or co-expressed with either NS1$_{R38A/K41A}$ or NS1-ED 182
6.3.3 mPSF-hFip1$_{FL}$-SII and mPSF-ΔhFip1-SII 183
6.3.4 mCF-SII, mCF-CPSF73$_{D75N/H76A}$-SII, mCF-symplekin$_{ΔNTD}$-SII, mCF-CPSF100$_{PIM\ MUT}$, mCF-CPSF100$_{ΔPIM}$, mCF-SII bound to CStF64 183
6.3.5 CStF-SII 183
6.3.6 CFIIm-SII 184
6.3.7 RBBP6-SII, RBBP6$_{Y228G}$-SII, RBBP6$_{P195G}$-SII, RBBP6$_{D43K\ R74E}$-SII 184
6.3.8 RBBP6-UBL-SII 184
6.3.9 CFIm-SII 184
6.3.10 PAP-SII 185
6.3.11 His$_6$-SSU72 185
6.3.12 NS1$_{R38A/K41A}$-MBP 186
6.3.13 NS1-ED-MBP 186

6.4 Preparation of RNA substrates 187

6.5 Assays with recombinant CPSF 190

6.5.1 Polyadenylation assays 190
6.5.2 Cleavage assays 190
6.5.4 Sequencing of 5' cleavage products 191
6.5.5 Assays with JTE-607 acid compound and assay quantification 192

6.6 Pull-downs of endogenous CPSF from mammalian cells 193

6.6.1 Pull-downs using transient transfection of a tagged subunit 193

6.6.2 Pull-down using tagged endogenous WDR33 subunit 193

6.7 Pull-downs from insect cells 195

6.8 Analytical gel filtration chromatography 195

6.9 Electromobility shift assay (EMSA) 196

6.10 *In vitro* pull-downs on M2-L3 pre-mRNA 196

6.11 Cryo-electron microscopy (cryoEM) 197

6.11.1 mPSF-NS1$_{R38A/K41A}$ 197

6.11.2 CPSF-RBBP6-RNA 197

6.11.3 mCF-CFIIm-RBBP 198

Chapter 7: References 199

Chapter 8: Appendices 214

List of Figures

Figure 1.1 3′ end processing depends on multiple protein complexes and cis-regulatory elements..... 24
Figure 1.2 mPSF/polymerase module specifically recognises the PAS RNA. 28
Figure 1.3 hFip1/Fip1 flexibly tethers poly(A) polymerase to mPSF/polymerase module. 30
Figure 1.4 mCF/nuclease is flexibly tethered to mPSF/polymerase module. 32
Figure 1.5 3′ ends of histone pre-mRNAs are processed by a specialised ribonucleoprotein machinery. 39
Figure 1.6 Selection of 3′ cleavage sites of eukaryotic pre-mRNAs is highly regulated..... 41
Figure 1.7 Eukaryotic pre-mRNA 3′ end processing is tightly coordinated transcription termination..... 45
Figure 1.8 Eukaryotic pre-mRNA 3′ end processing is likely coordinated with splicing. . 46
Figure 1.9 Activation of 3′ processing endonucleases requires accessory factors. 48
Figure 2.1 Purification of recombinant human mPSF..... 55
Figure 2.2 Recombinant human mPSF is active in polyadenylation. 56
Figure 2.3 Purification of recombinant human mCF. 57
Figure 2.4 mPSF and mCF assemble into the CPSF complex. 59
Figure 2.5 Purification of recombinant human CStF. 61
Figure 2.6 Purification of recombinant human CFIIm. 63
Figure 2.7 Human CStF and CFIIm complexes interact directly..... 64
Figure 2.8 CStF inhibits polyadenylation of suboptimal substrate RNAs. 66
Figure 2.9 Purification of recombinant human CFIm. 68
Figure 2.10 Purification of recombinant human RBBP6. 70
Figure 2.11 RBBP6 stimulates polyadenylation by PAP independent of mPSF or CPSF..... 72
Figure 3.1 CStF, CFIIm and RBBP6 are required to activate CPSF endonuclease. 78
Figure 3.2 Optimising buffer conditions of the recombinant CPSF cleavage assay. 81
Figure 3.3 Chemical additives are not required for endonuclease activity of recombinant CPSF. 82
Figure 3.4 Optimising protein concentrations in the recombinant CPSF cleavage assay. 83
Figure 3.5 Recombinant CPSF endonuclease is sequence specific. 86
Figure 3.6 Recombinant CPSF can cleave multiple pre-mRNA substrates. 88
Figure 3.7 Cleavage efficiency of recombinant CPSF is dependent on RNA length..... 89
Figure 3.8 CPSF73 is the only active endonuclease within CPSF. 91
Figure 3.9 NS1 inhibits endonuclease activity of recombinant CPSF..... 93

Figure 3.10 Preparation of mPSF-NS1 complex for cryoEM analysis. 95
Figure 3.11 Schematic representation of the processing pipeline of the mPSF-NS1 complex in Relion 3.1. 97
Figure 3.12 Physical tethering of mCF to mPSF is not essential for CPSF endonuclease activation. 99
Figure 3.13 FEGP repeats of Pcf11 are not required for endonuclease activity of recombinant CPSF. 101
Figure 3.14 IDR of hFip1 is not required for endonuclease activity of recombinant CPSF. 103
Figure 3.15 CFIm does not stimulate endonuclease activity of recombinant CPSF. 106
Figure 3.16 Recombinant CPSF can catalyse coupled cleavage and polyadenylation reactions. 109
Figure 3.17 CPSF and histone pre-mRNA 3′ end processing complex are activated by different mechanisms. 111
Figure 4.1 Transient over-expression of a tagged subunit does not allow purification of intact CPSF. 121
Figure 4.2 Purification of native CPSF from a stable cell line. 123
Figure 4.3 Subunit composition of endogenous CPSF. 126
Figure 4.4 CPSF interacts with RBBP6 in an RNA-dependent manner. 129
Figure 4.5 CPSF bound to RNA recruits RBBP6. 130
Figure 4.6 Structural characterisation of the CPSF-RBBP6-RNA complex. 132
Figure 4.7 RBBP6 interacts with CPSF73. 135
Figure 4.8 RBBP6 may interact with the mPSF module. 138
Figure 4.9 Interactions between RBBP6/Mpe1 and mPSF/polymerase module may differ between yeast and humans. 142
Figure 4.10 mPSF may block RBBP6 binding to mCF in the context of full CPSF. 144
Figure 4.11 RBBP6 does not interact with CStF but may bind CFIIm. 146
Figure 4.12 CFIIm and mCF form a stable complex. 148
Figure 4.13 RBBP6 stably interacts with the CFIIm-mCF complex. 151
Figure 4.14 Predicted architecture of the mCF-CFIIm-RBBP6 complex. 154
Figure 4.15 Experimental evidence for direct interactions between CFIIm subunits and either mCF or RBBP6. 156
Figure 4.16 Assembly of the complete active pre-mRNA 3′ end processing machinery with human proteins. 157
Figure 4.17 RBBP6 interacts with multiple components of the 3′ end processing machinery. 161

Figure 5.1 Model of molecular mechanisms of cleavage and polyadenylation by human CPSF. 170
Figure 6.1 Schematic representation of the biGBac cloning protocol. 178

List of Tables

Table 3.1 Comparison between this Study and Schmidt *et al* of the conditions of the CPSF endonuclease assays reconstituted from purified recombinant human proteins. 115
Table 6.1 List of pBig1 constructs used in this study. 177
Table 6.2 List RNA substrates used in this study. 189

List of Appendices

Appendix Table 8.1 Canonical pre-mRNA 3′ end processing factors in humans and budding yeast, their functions and the multi-subunit protein complexes they belong to. 215
Appendix Figure 8.1 WDR33 contains a long C-terminal IDR. 217
Appendix Figure 8.2 RBBP6 contains a long C-terminal IDR. 218
Appendix Figure 8.3 Statistics of the AlphaFold structure predictions discussed in this Thesis. 220
Appendix Figure 8.4 mCF does not catalyse endonucleolytic cleavage in the absence of mPSF. 221
Appendix Figure 8.5 ATP contamination in CPSF cleavage assays does not explain the differences between this Study and Schmidt *et al*. 221
Appendix Figure 8.6 mPSF interacts with RBBP6 in an RNA-dependent manner. 222
Appendix Figure 8.7 Disrupting the UBL-CPSF73 interface abolished RBBP6 binding to the mCF-CFIIm complex. 223
Appendix Figure 8.8 Parts of sequence alignments of RBBP6 and Clp1 orthologues. 223
Appendix Figure 8.9 CPSF can be cross-linked *in situ*. 224

List of Abbreviations

2X TY	yeast extract tryptone media
6-FAM	6-carboxyfluorescein
APA	alternative polyadenylation
APT	associated with Pta1 complex
ATP	adenosine triphosphate
biGBac	Gibson assembly-based baculovirus cloning protocol
BP	β-propeller
BS3	bis-sulfosuccinimidyl-suberate
CAN	canonical
CF	cleavage factor
CID	RNA polymerase II CTD-interacting domain
CP	creatine phosphate
CPF	cleavage and polyadenylation factor
CPSF	cleavage and polyadenylation specificity factor
CryoEM	cryo-electron microscopy
CStF	cleavage stimulatory factor
CTD	C-terminal domain
Dim	dimerisation domain
DNA	deoxyribonucleic acid
dNTPs	deoxynucleotide triphosphates
DPAS	distal PAS
DTT	dithiothreitol
ED	effector domain
FEGP	phenylalanine, glutamate, glycine, proline-rich repeats
FL	full-length
FT	flow through
HAT	half a tetratricopeptide repeat domain
HCC	histone cleavage complex
HDE	histone downstream element
HEK	human embryonic kidney cells
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
HRP	horse radish peroxidase
HTBH	His_6/TEV protease cleavage site/biotin acceptor peptide/His_6 tag

IDR	intrinsically-disordered region
INTAC	Integrator-PP2A complex
IR	Pcf11-interacting region
Iso	isoform
K_d	dissociation constant
LCR	low complexity region
mCF	mammalian cleavage factors
mPSF	mammalian polyadenylation specificity factor
MS	mass spectrometry
MUT	mutant
Mw	molecular weight
MβL	metallo-β-lactamase domain
NE	nuclear extract
NMR	nuclear magnetic resonance
NNS	Nrd1-Nab3-Sen1 complex
NS1	non-structural protein 1
NTD	N-terminal domain
NUDIX	nudix hydroxylase-like domain
OAc	acetate
P53	p53-binding domain
PABP/Pab	poly(A)-binding protein
PAE	predicted alignment error
PAGE	polyacrylamide gel electrophoresis
PAP/Pap	poly(A) polymerase
PAS	polyadenylation signal
PCR	polymerase chain reaction
PEG	polyethylene glycol
PIM	mPSF-interacting motif
pLDDT	per-residue confidence score of AlphaFold structure predictions
Pol II	RNA polymerase II
Poly(A)	polyadenylate
PP	protein phosphatase
PPAS	proximal PAS
PR-5′-HK	polyribonucleotide 5′-hydroxyl-kinase domain
Pre-mRNA	precursor messenger RNA
Pro	proline-rich domain

PSR	pre-mRNA-sensing region
PVA	polyvinyl alcohol
Rb	retinoblastoma protein-binding domain
RBBP6	retinoblastoma protein-binding protein 6
RBD	RNA-binding domain
RE/D	arginine, glutamate/aspartate-rich domain
RING	RING finger domain
RNA	ribonucleic acid
RRM	RNA recognition motif
RS	arginine, serine-rich domain
SDS	sodium dodecyl sulphate
SII	Strep-II tag
SLBP	stem loop-binding protein
snoRNA	small nucleolar RNA
snRNA	small nuclear RNA
SOB	super optimal broth media
Sulfo-SDA	sulfo-succinimidyl-diazirine
SV40	simian virus 40
TAE	tris-acetate-EDTA buffer
TAPS	tandem affinity purification tag
TBE	tris-borate-EDTA buffer
TCEP	tris(2-carboxyethyl)phosphine
TEV	tobacco etch virus
TRiC	T-complex protein ring complex
Tris	tris-hydromethyl-aminomethane
TSS	transcription start site
UBL	ubiquitin-like domain
UTR	untranslated region
UV	ultraviolet
WD40	β-propeller with 40 amino acids of tryptophan, aspartate repeats
WT	wild-type
ZnF	zinc finger domain
ZnK	zinc knuckle domain
β-CASP	metallo-β-lactamase, Artemis, CPSF, Pso2 domain

Chapter 1:

Introduction

The complexity of the eukaryotic cell is enabled by multiple intricate gene regulatory mechanisms. In the nucleus, these include regulation of transcription as well as extensive co-transcriptional processing of precursor messenger RNAs (pre-mRNAs). Pre-mRNA processing in eukaryotes includes 5′ capping with a 7-methylguanylate cap, removal of intronic sequences by the splicing machinery, and 3′ end processing (1). All three steps must be completed for the mRNA to be efficiently exported out of the nucleus and translated in the cytoplasm, and thus, pre-mRNA processing is critical to the production of functional protein-coding transcripts. Regulation of pre-mRNA processing increases the diversity of protein products generated from a single gene and modifies how their expression is regulated post-transcriptionally (2). In particular, recent years have witnessed major progress in our understanding of how pre-mRNAs are processed at their 3′ ends, and how this is regulated (3).

3′ end processing of eukaryotic pre-mRNAs includes a co-transcriptional endonucleolytic cleavage event at a specific site in the pre-mRNA (Figure 1.1A). This releases the nascent transcript from RNA polymerase II (Pol II) and generates a free 3′ end, which can act as a substrate for addition of a polyadenine (poly(A)) tail that is required for nuclear export and efficient translation of the mRNA (Figure 1.1B) (4). The site of cleavage defines the 3′ end of the mature transcript and will therefore determine the C-terminal sequence of the protein product and/or the length and sequence of the 3′ UTR of the mature mRNA. 3′ UTR sequences regulate translational efficiency, localisation and stability of the mRNA, as well as localisation and activity of its protein product (2). 3′ end processing is also intimately linked to other co-transcriptional processes such as splicing and transcription termination (5).

Most of our understanding of pre-mRNA 3′ end processing stems from studies of human and budding yeast systems, and the general mechanisms are likely to be highly conserved across eukaryotes. Over 80 proteins have been identified as part of the eukaryotic 3′ end processing machinery, either directly involved in cleavage and polyadenylation reactions or coordinating them with other nuclear processes (6).

3′ end processing in eukaryotes requires many protein factors that recognise various cis-regulatory elements surrounding the cleavage site (Figure 1.1C&D). In particular, cleavage and polyadenylation are carried out by a large multi-subunit protein complex, termed cleavage and polyadenylation specificity factor (CPSF) in humans and cleavage and polyadenylation factor (CPF) in yeast (Figure 1.1D) (7). The CPSF/CPF complexes host multiple enzymatic subunits including an endonuclease (CPSF73 in humans, Ysh1 in yeast) and a poly(A) polymerase (PAP in humans, Pap1 in yeast), which catalyse the two steps of pre-mRNA 3′ end processing (8, 9). Two protein phosphatases are also part of CPF, but are peripherally associated with CPSF (6, 9). In addition, multiple accessory protein factors have been implicated in the activation and regulation of both cleavage and polyadenylation (Figure 1.1D). For example, cleavage factors activate the CPSF/CPF endonuclease and can be considered part of the active 3′ end processing machinery. These include cleavage stimulatory factor (CStF), and cleavage factors Im and IIm (CFIm and CFIIm) in humans, as well as cleavage factor IA and IB (CF IA and CF IB) in yeast (10–12). Other RNA-binding proteins regulate the length of the poly(A) tail as well as the selection of alternative cleavage sites (known as alternative polyadenylation or APA).

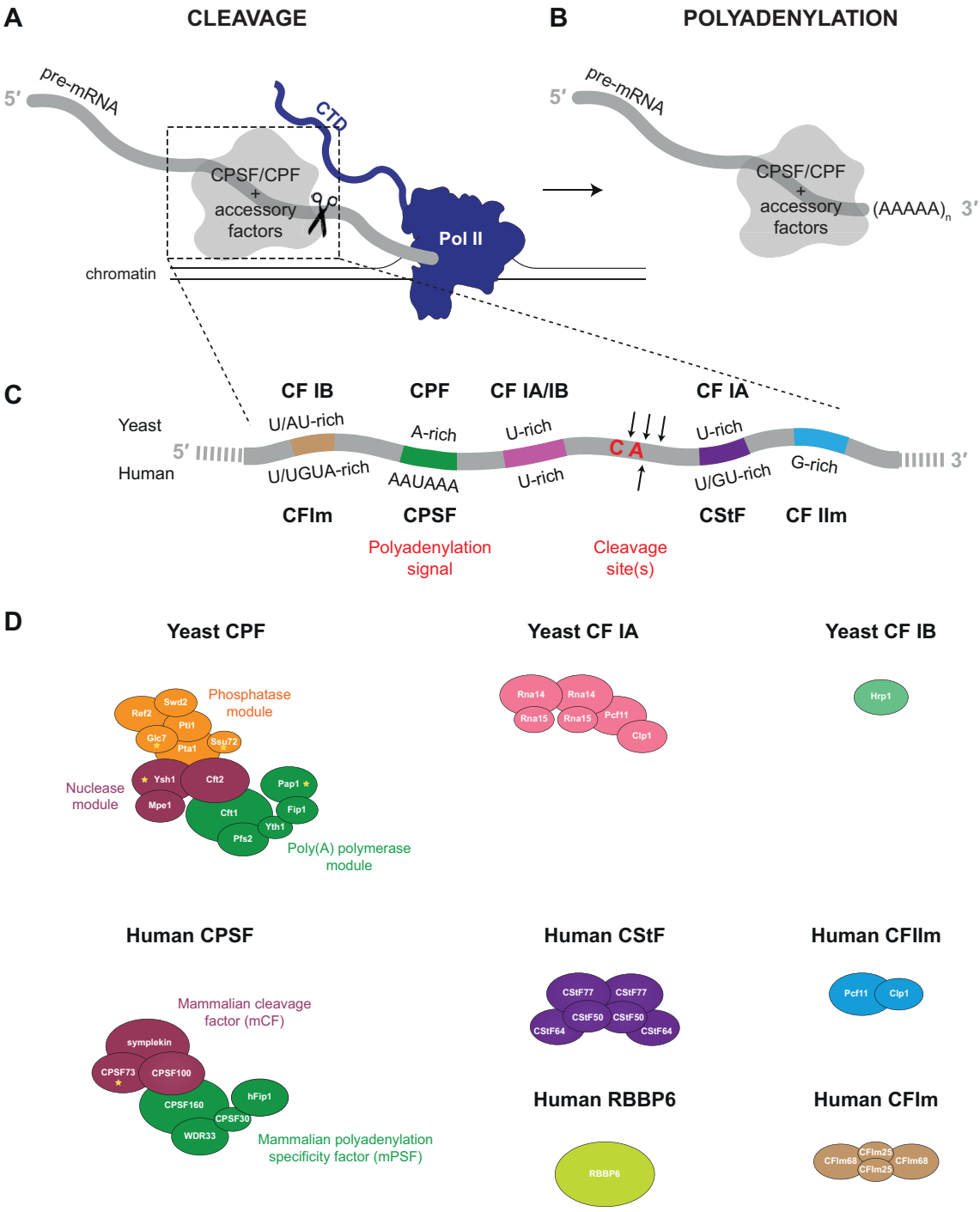


Figure 1.1 3′ end processing depends on multiple protein complexes and cis-regulatory elements. Schematic representation of co-transcriptional cleavage **(A)** and polyadenylation **(B)** reactions catalysed by CPSF and regulated by accessory factors. The pre-mRNA region carrying the cis-regulatory elements required for 3′ end processing are boxed out in **(A)** and schematically depicted in **(C)**. Protein factors binding to each element in yeast (top) and human (bottom) are indicated in **(C)** and schematically depicted in **(D)**.

1.1 3′ end processing machinery

1.1.1 CPSF/CPF performs 3′ end processing in eukaryotes

Early studies aiming to understand the mechanistic basis of cleavage and polyadenylation of eukaryotic pre-mRNAs suggested that 3′ end processing takes place within large protein complexes specifically assembled on the pre-mRNA substrate (13, 14). Fractionation of nuclear extracts enabled the identification of key 3′ end processing factors, and subsequent purification of endogenous protein complexes led to the determination of their subunit compositions (Appendix Table 8.1) (9, 15–19).

The subunits of the pre-mRNA 3′ end processing machinery are highly similar across eukaryotes, highlighting their conserved function. Interestingly, the affinities between some components differ between yeast and humans, potentially allowing more regulation in higher eukaryotes. Human CPSF consists of 7 subunits (8, 17). The same subunits are found in yeast CPF but it also contains 8 additional proteins that perform and regulate phosphatase activities (9, 18). Both CPSF and CPF are comprised of modules, centered around a different enzymatic activity: endonuclease, poly(A) polymerase and, in the case of the yeast CPF complex, protein phosphatase (Figure 1.1D).

1.1.2 mPSF or polymerase module

The polymerase module, known as mammalian polyadenylation specificity factor (mPSF) in humans, is a structural scaffold for the CPSF/CPF complexes (8, 9). It recruits CPSF/CPF to pre-mRNAs and binds the poly(A) polymerase enzyme to mediate polyadenylation after endonucleolytic cleavage. mPSF, along with PAP, is sufficient for specific and efficient polyadenylation *in vitro* (20). Recent structural analyses using X-ray crystallography, electron cryo-microscopy (cryo-EM) and nuclear magnetic resonance spectroscopy (NMR) have provided new insights into the function of mPSF/polymerase module (9, 21–26).

mPSF contains four protein subunits: CPSF160, WDR33, hFip1 and CPSF30 (Cft1, Pfs2, Fip1 and Yth1 in the yeast polymerase module). Human PAP is not constitutively associated with the mPSF complex but, interestingly, Pap1 is a stable subunit in yeast. Structural analyses of both human and yeast complexes have revealed details of their highly conserved

architecture. CPSF160/Cft1 contains three β-propeller domains which are arranged in a trefoil configuration with two of these domains forming a binding cavity for an N-terminal helix domain of WDR33/Pfs2 (Figure 1.2A&B). WDR33/Pfs2 also contains a β-propeller downstream of its helical domain, which sits on top of the CPSF160/Cft1 subunit. CPSF30/Yth1 interacts with the complex by contacting both CPSF160/Cft1 and WDR33/Pfs2 with two of its five zinc fingers (ZnFs): ZnF1 and ZnF2.

The overall structure of mPSF and the polymerase module shares a similar architecture but little sequence similarity to the DDB1-DDB2 complex, which recognises UV-damaged DNA, and the SF3b complex, which is part of the U2 snRNP involved in pre-mRNA splicing (9, 22). This suggests that the protein complexes structured around a β-propeller scaffold may have a common ancestor and are used as a common scaffold to facilitate binding to nucleic acids in eukaryotes.

1.1.3 Recognition of the polyadenylation signal

Sites for 3′ end processing are marked by the hexameric polyadenylation signal (PAS) sequence (27). The PAS is located approximately 10-30 nucleotides upstream of the cleavage site and is often surrounded by auxiliary RNA motifs that bind cleavage factors (Figure 1.1C) (28). CPSF/CPF binds the PAS directly (20, 21, 23), and consequently, the sequence of the PAS determines the efficiency of 3′ end processing at a particular site. The consensus PAS sequence with the highest affinity for the 3′ end processing complex in most eukaryotes is AAUAAA (29, 30). However, only ~50% of PAS sites in humans contain the canonical AAUAAA motif (28, 31). Non-canonical PAS sequences are recognised less efficiently and may contribute to regulation of 3′ end processing.

The mechanism of PAS recognition by CPSF/CPF was first studied by UV cross-linking which demonstrated that PAS RNA contacts the mPSF complex (20, 32). More recently, human mPSF bound to PAS-containing RNA was visualised using cryoEM (21, 23). All six nucleotides of the AAUAAA sequence were visible, revealing that the hexamer interacts with ZnF2 and ZnF3 of CPSF30 and a surface on WDR33 (Figure 1.2C). Amino acid residues of CPSF30 form base-specific contacts with the RNA, facilitating the recognition of the PAS sequence. Additional specificity is provided by bases U3 and A6 of the PAS, which form a Hoogsteen base pair that inserts into a pocket of WDR33 and stabilises the mPSF-bound conformation of the RNA.

The *in vitro* affinity of human mPSF for RNA is approximately two orders of magnitude higher than that of the yeast polymerase module (12, 29). A recent cryo-EM structure of the yeast polymerase module bound to a nuclease module subunit, Mpe1, showed how the 5' part of the PAS is recognised (Figure 1.2D) (33). The first two adenosines of the PAS are located at the same position as observed in the human complex, demonstrating that their recognition is highly conserved between yeast and human proteins. An additional nucleotide upstream of the PAS (U(-1)) was bound in a surface pocket of Yth1, raising the possibility that the nucleotides surrounding the PAS may also contribute to RNA binding and recognition by the polymerase module.

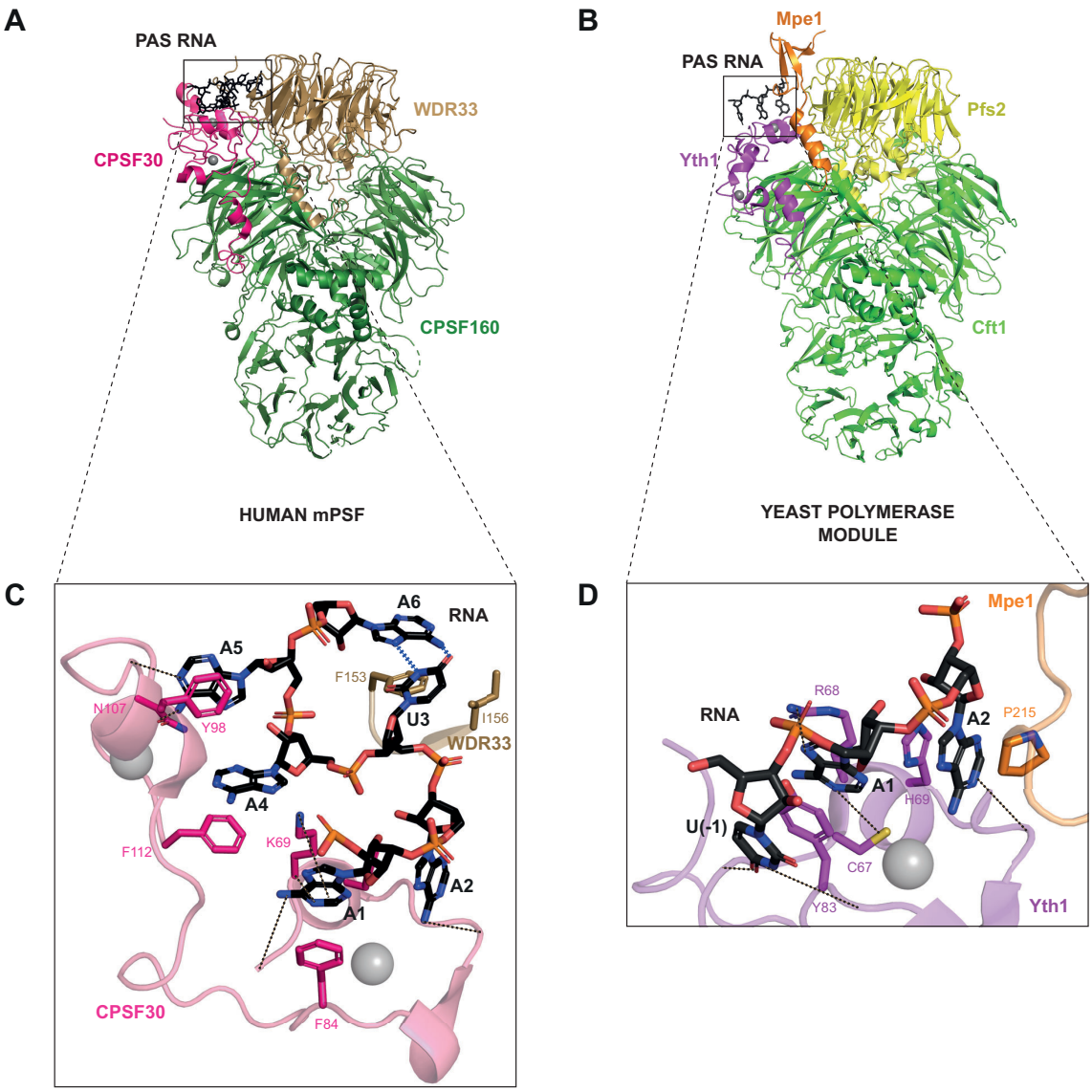


Figure 1.2 mPSF/polymerase module specifically recognises the PAS RNA. Overall architecture of human mPSF bound to PAS RNA **(A)** (PDB 6DNH) and of the budding yeast polymerase module bound to PAS RNA and Mpe1 **(B)** (PDB 7ZGR) (21, 33). Close-up view of the PAS RNA binding site of human mPSF **(C)** and of the yeast polymerase module **(D)**. Some hydrogen bonds between the RNA and the protein subunits are indicated by dashed black lines. Hydrogen bonds that mediate Hoogsteen base-pairing between U3 and A4 in the human complex are depicted in dashed blue lines. Some protein residues that make hydrophobic and stacking interaction with the RNA are also shown in stick representation. Zinc ions bound to ZnF domains of CPSF30/Yth1 are shown in grey.

1.1.4 Poly(A) polymerase

Polyadenylation of cleaved pre-mRNAs is catalysed by PAP/Pap1 using a polymerisation mechanism dependent on two catalytic magnesium ions (34). On its own, PAP displays only weak and distributive PAS-independent activity, likely because of its low affinity for RNA. mPSF stimulates PAP activity, likely by recruiting the enzyme to cleaved pre-mRNAs (20). Both PAP and Pap1 interact with mPSF/polymerase module via the hFip1/Fip1 subunit (Figure 1.1D) (35). A structure of residues 80-105 of yeast Fip1 bound to Pap1 has been determined, but additional residues also contribute to their interaction (Figure 1.3) (24, 36). hFip1/Fip1 also interacts with CPSF30/Yth1 and thereby acts to tether Pap1 to the complex (24–26).

Fip1 is an intrinsically disordered protein in isolation but is known to form some secondary structure upon binding to Yth1 and Pap1 (24, 37). The region of Fip1 connecting the Pap1 and Yth1 interaction sites, known as the central low complexity region (LCR), has been shown to remain dynamic in the context of CPF (24). This may allow Pap1 to be flexibly tethered to the polymerase module and facilitate addition of adenine residues to a growing poly(A) tail (Figure 1.3). The dynamics of Fip1 may also explain why Fip1 and Pap1 were not resolved in cryoEM maps of the polymerase module (9, 33). However, it remains to be determined if Pap1 remains flexibly tethered in the context of the full 3′ end processing machinery assembled on RNA.

Interestingly, native mass spectrometry of endogenous CPF identified a population of the complex bound to two, instead of just one, copies of Pap1 and Fip1 (9). Recent crystal structures of the human CPSF30-hFip1 complex revealed that one copy of hFip1 binds to each of ZnF4 and ZnF5 of CPSF30 (Figure 1.3) (25, 26). Notably, ZnF4 has a significantly higher affinity for hFip1 than ZnF5. The two copies of hFip1 enable tethering of two PAP molecules to a single mPSF complex. Further investigation is required to determine the functional significance of the non-uniform stoichiometry of PAP/Pap1 and hFip1/Fip1 *in vivo*.

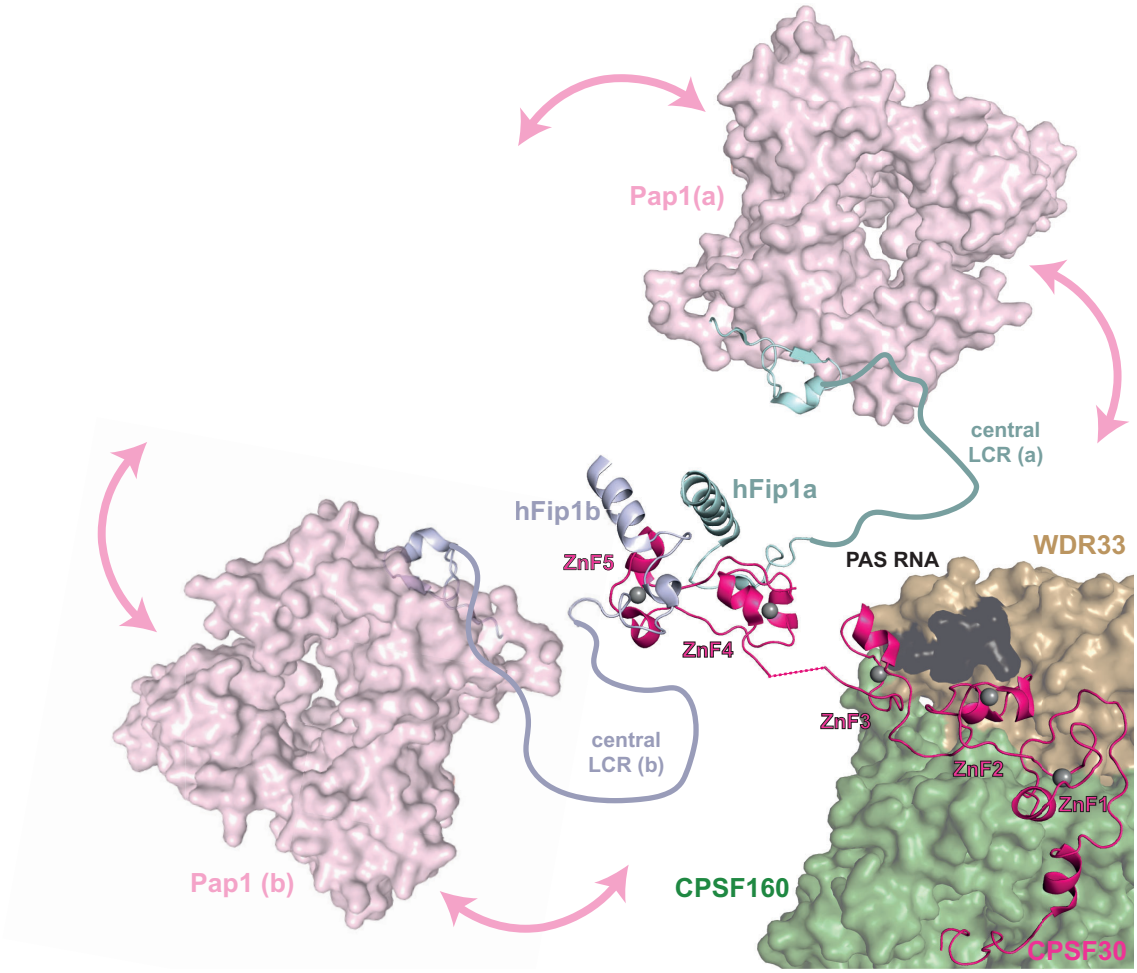


Figure 1.3 hFip1/Fip1 flexibly tethers poly(A) polymerase to mPSF/polymerase module. The zinc finger (ZnF) domains 1-3 of CPSF30 resolved within the structure of mPSF (PDB 6DNH) are separated by 4 residues (pink dashed line) from ZnF4 and ZnF5 whose crystal structure in complex with two copies of hFip1 ("a" and "b") has been determined (PDB 7K95) (21, 25). The dynamic central low complexity region (LCR) of hFip1/Fip1 connects the CPSF30/Yth1 and PAP/Pap1 binding site. Thus, PAP/Pap1 may remain flexible relative to mPSF/polymerase module. The structure of yeast Fip1 bound to PAP is depicted (PDB 3C66) (36). An equivalent structure of human proteins has not yet been determined.

1.1.5 Nuclease module or mCF

The nuclease module, known as mammalian cleavage factor (mCF) in humans, hosts the 3′ endonuclease enzyme, CPSF73/Ysh1, a pseudonuclease CPSF100/Cft2 and a third non-enzymatic subunit (Figure 1D). In humans, this is a scaffold protein, symplekin (8). Its yeast orthologue, Pta1, also interacts with Ysh1 and Cft2 in a conserved manner, but Pta1 has been attributed to the phosphatase module of CPF based on native mass spectrometry data (9). This is in agreement with the two modules being intimately associated. The yeast nuclease module contains Mpe1 as its third subunit.

CPSF73/Ysh1 and CPSF100/Cft2 contain metallo-β-lactamase and β-CASP (metallo-β-lactamase, Artemis, CPSF, Pso2) domains. The active site of CPSF73/Ysh1 is located in a cleft between these two domains where two catalytic zinc ions are coordinated by highly conserved amino acid side chains (38). The zinc binding residues are less well conserved in CPSF100/Cft2 which has therefore been described as an inactive pseudonuclease (39). The endonuclease and pseudonuclease subunits form a constitutive dimer due to an interaction between their C-terminal domains, stabilised by the binding of symplekin/Pta1 (Figure 1.4) (8).

mCF/nuclease module is connected to mPSF/polymerase module through an interaction of a highly conserved peptide of CPSF100/Cft2, called an mPSF-interacting motif (PIM), with a surface on CPSF160/Cft1 and WDR33/Pfs2 (Figure 1.4) (8, 33). The PIM peptide is located in the middle of a ~130 residue long intrinsically disordered region connecting the metallo-β-lactamase and β-CASP domains. Electron microscopy analysis of recombinant complexes has revealed that mCF/nuclease module does not occupy a fixed position relative to mPSF/polymerase module (8, 12). The inherent flexibility of CPSF/CPF may allow for processing of a large variety of eukaryotic pre-mRNA, for example, with variable distances between the PAS and the cleavage site. It is possible that CPSF/CPF becomes more rigid in the presence of the RNA substrate and cleavage factors, forming a structured active complex.

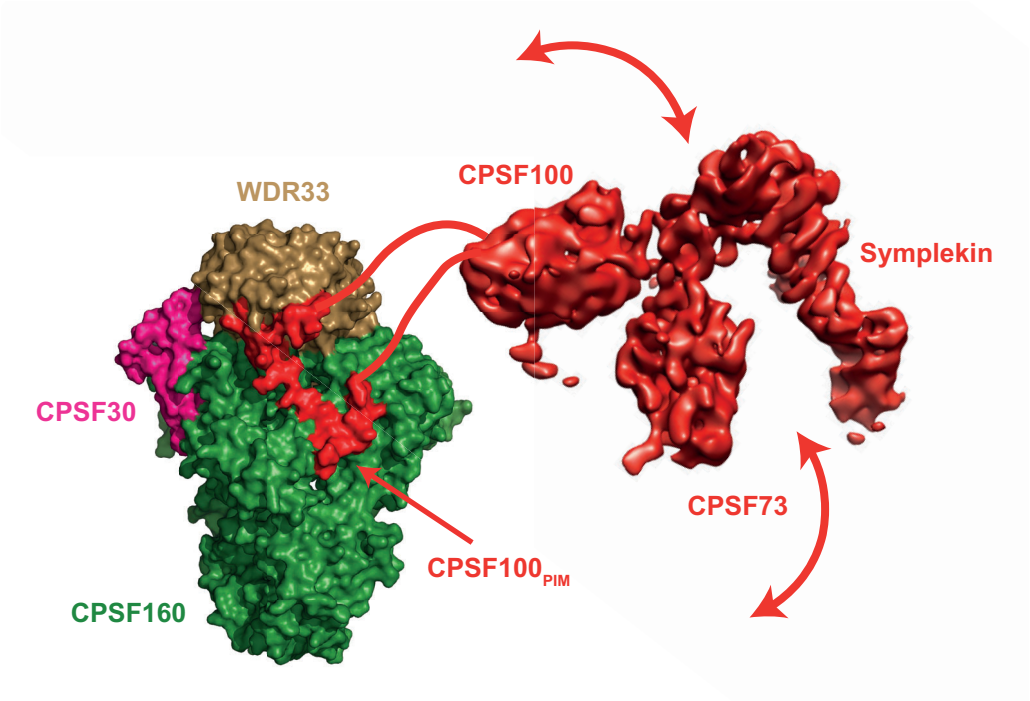


Figure 1.4 mCF/nuclease is flexibly tethered to mPSF/polymerase module. Structural model of the CPSF/CPF complex based on the structure of human mPSF bound to the PIM peptide of CPSF100 shown in surface representation (left; PDB 6URG) and the cryoEM map of human mCF at 7.4 Å resolution (right; EMDB 20859) (8). An intrinsically disordered loop of CPSF100 (red lines) tethers mCF to mPSF in a flexible manner.

1.1.6 RBBP6/Mpe1

A yeast protein called Mpe1 is part of the CPF nuclease module and is required for cleavage activity by the CPF complex (Figure 1.1D) (12). Mpe1 is a multi-domain protein containing a ubiquitin-like domain (UBL), a zinc finger and a RING finger domain. The UBL domain of Mpe1 has been shown to interact with the metallo-β-lactamase domain of the endonuclease subunit. A crystal structure of this dimeric complex from yeast shows that the UBL domain is located adjacent to the active site cleft of the endonuclease (see Figure 1.9B) and that it contributes a positively charged surface which may direct the pre-mRNA substrate into the active site (12). Thus, the UBL domain of Mpe1 may contribute to the activation of the endonuclease enzyme in the context of the complete 3′ end processing machinery.

A recent cryo-EM structure of the yeast polymerase module bound to Mpe1 and PAS RNA revealed that Mpe1 interacts with the Pfs2 subunit and, surprisingly, also contacts the PAS RNA (Figure 1.2B&D) (33). Therefore, the region of Mpe1 that interacts with the polymerase module was referred to as the pre-mRNA sensing region (PSR). Mpe1 contacts RNA through a CH-π interaction between a proline residue and the A2 nucleotide of the PAS. This is a rather weak interaction and therefore, Mpe1 affects neither the affinity of the polymerase module for RNA nor the sequence specificity of RNA binding. Instead, functional studies *in vivo* and *in vitro* showed that mutations of the proline resulted in defective pre-mRNA cleavage by CPF. Thus, Mpe1 may sense correct binding of PAS RNA to the polymerase module, coupling RNA recognition directly to the endonuclease through its UBL domain. In addition, Cft2 and Mpe1 interactions with the polymerase module appear to be antagonistic and this may indicate that binding of either nuclease module subunit may represent alternative conformational states of the complex.

Interestingly, Mpe1 is also computationally predicted to bind to the poly(A) polymerase Pap1 (40). Mutations in the PSR of Mpe1 lead to hyper-polyadenylation of a cleaved pre-mRNA substrate *in vitro*, providing evidence for functional importance of this predicted interaction (33). Given its essential role in activating pre-mRNA cleavage, Mpe1 may coordinate the two 3′ end processing reactions by the CPF complex and may mediate the potential conformational changes between the cleaving and polyadenylating states of CPF.

The human orthologue of Mpe1, RBBP6, has the same domain architecture but also carries a long, largely disordered C-terminal extension which may interact with transcription factors and splicing regulators (41–43). In fact, the name of RBBP6 stems from the fact that

it was originally identified as an Rb-binding protein (41, 42). Compared to Mpe1, the role of RBBP6 in 3′ end processing in humans is far less understood. RBBP6 has been shown to regulate alternative cleavage site selection in human cells and was also detected in native post-cleavage complexes bound to the 5′ product of the cleaved pre-mRNA (6, 44). However, it remained unclear whether RBBP6 interacts with CPSF and whether it is part of the active 3′ end processing machinery in humans.

1.1.7 Protein phosphatases of 3′ end processing complexes

The yeast phosphatase module incorporates two protein phosphatases, Ssu72 and Glc7, which target the highly conserved $Y_1S_2P_3T_4S_5P_6S_7$ repeats on the C-terminal domain (CTD) of Pol II (Figure 1.1D) (45). The phosphatase module also contains four additional non-enzymatic proteins, including the symplekin orthologue Pta1, Swd2, Pti1 and Ref2 (9).

Pol II CTD is primarily phosphorylated at positions 2 and 5, while the frequency and the physiological role of the other phosphorylation sites, especially in yeast, remain controversial (46). Incorporation of Ssu72 and Glc7 into a multi-subunit protein complex likely confers substrate specificity to the phosphatases. Ssu72 dephosphorylates serine 5 and serine 7 residues of the CTD repeats during transcription elongation, while Glc7 promotes transcription termination by dephosphorylating tyrosine 1 as well as a transcription elongation factor SPT5 (47–50). Since CPF associates with actively transcribed protein-coding genes at promoters, CPF may contribute to the coordination of all stages of transcription (51).

Human cells express orthologues of both Ssu72 (SSU72) and Glc7 (protein phosphatase 1, or PP1). SSU72 and PP1 are both pulled down with an active RNA-associated 3' end processing complex, and they perform conserved functions in coordinating transcription with 3′ end processing (6, 45). Both human and yeast Ssu72 interact with symplekin/Pta1 (52). In human cells, PP1 is part of a distinct protein complex that also contains WDR82, Tox4 and PNUTS subunits, the latter being a key regulator of the enzymatic activity by PP1 in transcription termination (53, 54).

A distinct protein complex called associated with Pta1 (APT) mediates transcription termination of non-coding snRNAs and snoRNAs in yeast (55). APT consists of all the subunits of the phosphatase module, but also contains a unique subunit Syc1 which

distinguishes APT from CPF. Syc1 is homologous to the C-terminal domain of Ysh1 and blocks the association of the nuclease and polymerase modules with the APT complex. As a result, snRNAs and snoRNAs are neither cleaved nor polyadenylated, but instead they are released from transcribing Pol II by the helicase activity of the Nrd1-Nab3-Sen1 (NNS) complex (56). The phosphatase activities of APT may regulate transcription of non-coding RNAs in the same way that the phosphatase module functions on protein-coding genes.

In humans, 3′ end processing of snRNAs is executed by the Integrator complex, which does not share homology with APT (57). Integrator has been shown to resolve promoter proximal pausing of Pol II at protein coding genes, preventing the transition of Pol II into productive elongation (58). Thus, Integrator regulates the production of mRNAs as well as snRNAs. It contains at least 14 subunits including both endonuclease and protein phosphatase enzymatic activities (59). The Integrator-PP2A complex (INTAC) can dephosphorylate serine 2, serine 5 and serine 7 of the Pol II CTD *in vitro*, suggesting that a single phosphatase within INTAC could regulate both transcription elongation and transcription termination (59–61). The nuclease module of the Integrator complex has a highly similar architecture to mCF of CPSF, including the endonuclease INTS11, which belongs to the metallo-β-lactamase/β-CASP family, similar to CPSF73 (see Figure 1.9C) (59, 62, 63).

1.2 Accessory factors regulate enzymatic activities of CPSF/CPF

1.2.1 CStF, CFIIm/CF IA

CF IA is essential for the endonuclease activity of yeast CPF (12). It is comprised of two copies of the Rna14 and Rna15 subunits, which form a tetramer, and one copy of each of Pcf11 and Clp1 (Figure 1.1D) (64, 65). In humans the orthologues of CF IA subunits are present within two separate complexes: Human CStF complex contains an orthologue of Rna14, CStF77, an orthologue of Rna15, CStF64, and a human-specific protein CStF50, each in two copies (11). Human Pcf11 and Clp1 are located within a dimeric CFIIm complex (10).

CStF and CF IA bind G/U-rich elements located downstream of the cleavage site (Figure 1.1C). RNA binding is mediated by the RRM domains of CStF64/Rna15, shown to have a preference for G/U-rich sequences *in vitro* (66). The half a tetratricopeptide repeat (HAT) domains of CStF77 dimerise and interact with mPSF, contacting both CPSF160 and WDR33 subunits (8). By binding both CPSF and specific sequence elements of the pre-mRNA substrate, CStF/CF IA may contribute to the specificity of cleavage site selection and also position the RNA for endonucleolytic cleavage.

CFIIm may further increase the sequence specificity of the 3′ processing machinery by binding to the G-rich element found on some pre-mRNAs further downstream of the G/U-rich motif (Figure 1.1C) (10). Pcf11 physically bridges CStF/Rna14-Rna15 and Clp1, and also interacts with the C-terminal domain of Pol II, likely helping to coordinate 3′ end processing with transcription (67, 68). Interestingly, human Clp1 is an active polynucleotide kinase, while the yeast protein lacks the catalytic residues (69).

Overall, CF IA/CStF and CFIIm are required to activate the 3′ end processing endonuclease, but their mechanism of activation remains unknown.

1.2.2 CFIm/CF IB

Additional cleavage factor proteins in both yeast and humans play regulatory roles in pre-mRNA 3′ end processing. These factors tend to be specific to each species, suggesting that, while the basal cleavage and polyadenylation machinery is highly conserved, it may be

regulated differently in yeast and humans (Appendix Table 8.1). Yeast CF IB consists of a single protein, Hrp1, that imparts specificity to CPF endonuclease activity by preventing secondary cleavage of the pre-mRNA substrate (12). CF IB binds to U-rich sequences both upstream and downstream of the PAS, but further details of how it regulates cleavage are unclear (Figure 1.1C).

In humans, the tetrameric CFIm complex contains two copies of the RNA-binding protein CFIm25 and two copies of either CFIm68 or CFIm59 (Figure 1.1D). CFIm favors the usage of poly(A) sites with upstream UGUA motifs by binding to such sequences and recruiting CPSF via an interaction between an RS-like (arginine, serine-rich) domain of CFIm68 and an RE/D (arginine, glutamate/aspartate-rich domain of hFip1 (Figure 1.1C) (70). Thus, CFIm contributes to alternative cleavage site selection and alternative polyadenylation.

1.2.3 Nuclear poly(A)-binding proteins control polyadenylation

The 3' end processing machinery controls polyadenylation, so that all poly(A) tails are synthesised to a relatively uniform length. The median length of a poly(A) tail of a mature mRNA upon nuclear export is species specific, varying between ~60-80 adenosines in budding yeast to ~250 adenosines in humans (71–73). Hyper-polyadenylated mRNAs are degraded by the nuclear exosome, whereas mRNAs containing poly(A) tails that are too short are not exported from the nucleus, highlighting the importance of poly(A) tail length control (74, 75). Poly(A) tail length is likely regulated through controlled processivity of PAP/Pap1 and accessory proteins (Appendix Table 8.1).

Nuclear poly(A)-binding proteins, including Nab2 and PABPN1, contribute to normal control of poly(A) tail length. Once a poly(A) tail reaches ~60 adenosines in yeast and ~250 adenosines in humans, Nab2 and PABPN1, respectively, are thought to bind and inhibit processive polyadenylation (71, 76). In yeast, Pab1, which is less abundant in the nucleus than Nab2, can substitute for Nab2 and restrict the poly(A) tail length to ~90 adenosines if Nab2 is unavailable (71). In a current model, poly(A)-binding proteins promote PAP/Pap1 dissociation from RNA. However, mechanistic understanding of poly(A) tail length control by poly(A) binding proteins is still lacking.

Even in the absence of poly(A) binding proteins, polyadenylation by yeast CPF can be restricted to ~100-200 nucleotides *in vitro* and in cells (71). The mechanistic basis of such

intrinsic length control remains unclear, but physiologically it is thought to act as a fail-safe mechanism in yeast under stress conditions when Nab2 and Pab1 are depleted. A similar restriction of poly(A) tail length is observed when the concentration of PABPN1 in the nucleus is decreased in mammalian cells, suggesting that CPSF also exhibits intrinsic length control of polyadenylation. In the absence of PABPN1, CPSF is less processive and its rate of adenosine addition decreases as the distance between the PAS and the 3′ end of the poly(A) tail increases (76). Thus, the mechanisms of intrinsic poly(A) tail length control may be conserved between yeast and human.

1.3 3′ end processing of histone pre-mRNAs

Not all protein-coding mRNAs in eukaryotes are polyadenylated. 3′ end processing of metazoan replication-dependent histone pre-mRNAs involves endonucleolytic cleavage downstream of a conserved stem loop structure bound by the stem loop-binding protein (SLBP). SLBP is upregulated prior to the onset of the S phase when genes encoding replication-dependent histones are transcribed (77). SLBP promotes 3′ end processing of histone pre-mRNAs as well as their export, translation and stability (78, 79). This allows coordinated production of replication-dependent histone mRNAs specifically during the S phase of the cell cycle when histones are required for chromatin assembly during DNA replication.

3′ end processing of replication-dependent histone pre-mRNAs is executed by a specialised ribonucleoprotein machinery. The histone pre-mRNA 3′ end processing complex contains a seven subunit Sm ring bound to U7 snRNA, subunits Lsm10, Lsm11, FLASH, SLBP and the HCC subcomplex that is comprised of the same subunits as mCF, including endonuclease CPSF73 (Figure 1.5) (80, 81). CStF64 has also been shown to be associated with the histone processing complex, but the functional consequences of this interaction remain unknown. Despite sharing the endonuclease subunit, the histone pre-mRNA 3′ end processing complex and CPSF differ in their mechanisms of both RNA recognition and endonuclease activation (82). For example, unlike protein-mediated PAS recognition by CPSF, the U7 snRNA within the histone complex recognises a conserved histone downstream element (HDE) in the pre-mRNA using canonical base pair interactions. The HDE-U7 duplex is ~15 nucleotides long, which, along with the recognition of the stem loop structure by SLBP, ensures highly specific recognition of histone pre-mRNAs.

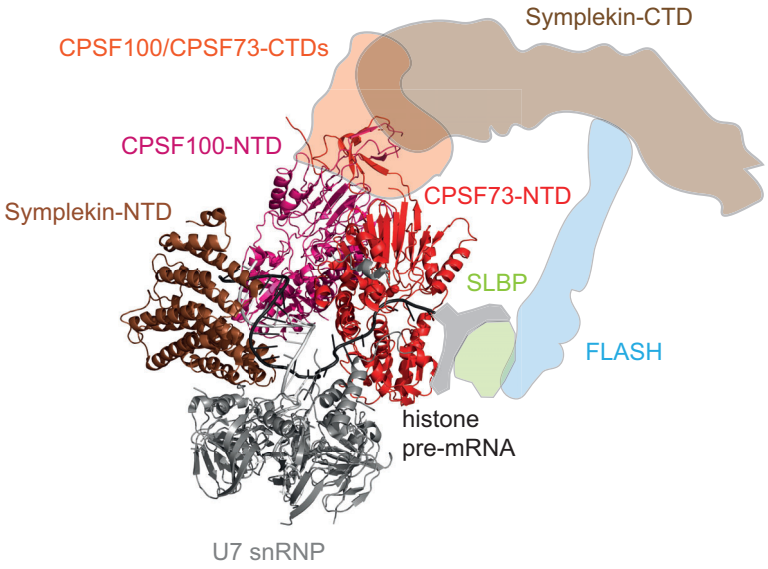


Figure 1.5 3′ ends of histone pre-mRNAs are processed by a specialised ribonucleoprotein machinery. CryoEM structure of the histone pre-mRNA processing machinery (82). Subunits with built atomic models are shown in cartoon representation (PDB 6V4X), while the subunits observed in the cryoEM map but not modelled, including the HDE of histone pre-mRNA, are contoured according to the published map (EMD-21050).

1.4 Regulation of 3′ end processing

1.4.1 Mechanisms of alternative of polyadenylation

About 70% of protein-coding genes in both budding yeast and metazoans produce several mRNA isoforms which differ in the sequence of their 3′ ends (83). Although this has been traditionally termed alternative polyadenylation (APA), the specific selection of alternative cleavage sites is at the heart of generating alternative 3′ ends of the same mRNA (Figure 1.6). APA has a regulatory capacity to tune when, where, how much and what protein is translated from each mRNA and is therefore highly regulated based on cell type, developmental stage and cellular conditions (2).

APA can change the identity of the protein product if cleavage occurs before the stop codon. This can for example produce either a protein lacking its C-terminal regions, often affecting its function, or a truncated non-functional polypeptide. In contrast, APA after the stop codon does not alter the amino acid sequence of the translated polypeptide but instead changes the length and sequence of the 3′ UTR of the mRNA with potential consequences to its

stability and localisation. For instance, alternative mRNA isoforms with altered 3′ UTR lengths often display different decay rates in the cytoplasm.

In a kinetic model of PAS recognition, changes in the cellular concentrations of the 3'end processing machinery can alter the cleavage site. A greater abundance of CPSF/CPF or accessory factors may change the RNA binding landscape to promote the use of proximal PAS sites that emerge early from Pol II but are often suboptimal. In contrast, lower cellular concentrations of the 3' end processing machinery enforces the use of distal, canonical PAS sites. However, it was recently shown that several PAS sites can be used on the same transcript: some pre-mRNAs that are cleaved using a distal PAS first are retained in the nucleus until they are subsequently processed a second time post-transcriptionally using a proximal PAS (84).

The kinetic model is likely to have a global effect on cleavage sites. For example, mRNAs in proliferating cells generally have shorter 3′ UTRs, while in differentiated cells distal PAS sequences tend to be used more often (85). In agreement with this, cell types that show preference for proximal PAS sequences tend to express higher levels of proteins that are part of the core 3′ end processing machinery than cell types that produce isoforms with longer 3′ UTRs (86).

Regulation of APA by varying expression levels of core 3′ end processing factors appears to be widespread, and many specific examples have been documented. For example, during B cell differentiation, elevated levels of CStF64 promote a switch from using the distal PAS to the proximal PAS which leads to the production of IgM antibody missing its transmembrane domain, allowing the antibody to be secreted (87). Interestingly, many 3′ processing factors (including Pcf11, CStF77 and RBBP6) control their own expression by APA, resulting in a negative autoregulatory feedback loop (44, 88, 89).

A second mechanism for APA involves accessory binding proteins that either promote or repress certain cleavage sites. For example, sequence-specific RNA-binding proteins such as NOVA2, ELAV, FUS and SR proteins regulate APA of their target RNAs (2). How these proteins affect the core 3′ end processing machinery to regulate APA is largely unknown. The best studied sequence-specific APA regulator is the CFIm complex. The CFIm25 subunit recognises UGUA motifs that are enriched upstream of distal polyadenylation sites. CFIm promotes production of long 3′ UTR isoforms by recruiting CPSF via an interaction between an RS-like domain of CFIm68 and the hFip1 subunit of CPSF (70). An alternative CFIm

subunit, CFIm59, can partially rescue CFIm68 depletion but appears to be a weaker activator of 3′ end processing, suggesting that the relative expression levels of CFIm68 and CFIm59 may fine-tune APA regulation.

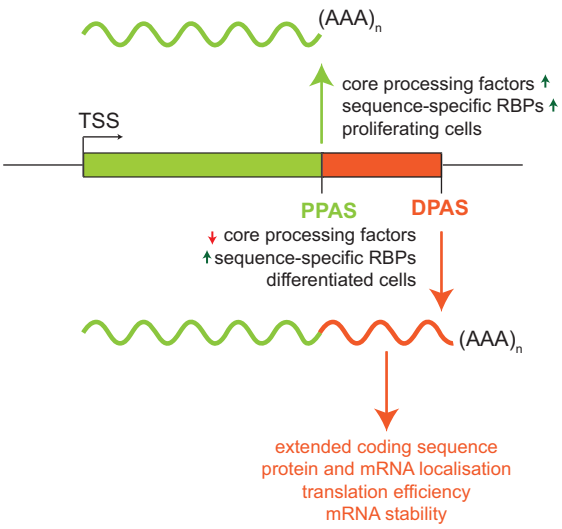


Figure 1.6 Selection of 3′ cleavage sites of eukaryotic pre-mRNAs is highly regulated. Schematic representation of alternative polyadenylation (APA). TSS – transcription start site; PPAS – proximal polyadenylation site; DPAS – distal polyadenylation site.

1.4.2 Regulation of 3′ end processing in disease

3' end processing is deregulated in disease. Mutations in genes encoding components of the 3' end processing machinery are found in cancer, neurological disease and developmental disorders, and their expression is also often misregulated (90). Point mutations in the sequences that specify the cleavage site can also lead to disease. For example, a single point mutation in the PAS of the globin gene alters the cleavage site, leading to destabilisation of the globin transcript and thalassemia.

1.4.2.1 CPSF73 as a therapeutic target

Recent years have seen a growing interest in targeting 3′ end processing pharmacologically, in particular by inhibiting the endonuclease activity of the CPSF complex (91). Compounds with anti-cancer, anti-inflammatory and anti-protozoan properties have been shown to

bind in the active site of CPSF73 and compete with substrate RNA, thereby inhibiting endonucleolytic cleavage (92–96). For instance, benzoxaborole compounds, such as AN3661, specifically inhibit protozoan orthologues of CPSF73 and have a potential to treat malaria and toxoplasmosis (95, 96). Despite the high degree of sequence conservation among CPSF73 orthologues, small differences in the conformation of their active sites enables AN3661 to selectively inhibit protozoan enzymes without causing severe toxicity to the human host (95).

Another compound that targets CPSF73, JTE-607, prevents inflammation and also inhibits growth of Ewing's sarcoma, acute myeloid leukemia and pancreatic ductal adenocarcinoma cell lines (93, 97). JTE-607 induces global read-through transcription and R loop formation in cancer cells (93). Many types of cancers overexpress CPSF73 (97, 98). The elevated transcriptional activity of cancer cells may make their survival more dependent on pre-mRNA processing than that of non-transformed tissues, providing a therapeutic window where cancer cells are susceptible to inhibition of CPSF73. Proximal polyadenylation sites tend to be less optimal and therefore, they may be more sensitive to inhibition of CPSF73. Thus, JTE-607 may also prevent growth of cancer cells by restoring the usage of more optimal distal polyadenylation sites seen in healthy tissues, but this has not yet been investigated. In contrast, in other cancers, such as renal clear cell carcinoma, transcriptional read-through is widespread, explaining why inhibition of the 3′ endonuclease might not be effective against all cancer types (99).

Overall, CPSF73 has emerged as a druggable node both in transformed human cells and eukaryotic parasites. Structural studies have already illuminated how some clinically-important compounds inhibit isolated CPSF73 (93, 95). Use of an *in vitro* reconstituted 3′ end processing reaction with purified proteins for compound screening may allow identification of new drugs that target 3′ end processing. Such a high-throughput system has been recently established for the histone pre-mRNA 3′ end processing reaction (100). It will be exciting to see if any novel CPSF73 inhibitors enter clinical use in the near future.

1.4.2.2 3′ end processing machinery is targeted by viruses

3′ end processing of pre-mRNAs is a vital step in gene expression, and thus, several viruses, including Influenza A and Herpes simplex virus-1, have evolved mechanisms to interfere with 3′ end processing of host transcripts. Infection with Influenza or HSV-1 causes

transcriptome-wide inhibition of 3′ end processing leading to defects in transcription termination and pervasive read-through transcription (101, 102). Such read-through transcription results in global downregulation of host gene expression, allowing the virus to take over the cellular machinery for its own amplification. Single proteins within each virus, non-structural protein 1 (NS1) in Influenza A and ICP27 in HSV-1, are thought to be the primary inhibitors of 3′ end processing, because ectopic overexpression of these proteins is sufficient to replicate most of the 3′ end processing defects observed in respective viral infections (101, 103).

Both NS1 and ICP27 directly interact with the CPSF complex, but their inhibitory mechanisms do not seem to be conserved. NS1 binds ZnF2 and ZnF3 of CPSF30 and may compete with PAS recognition, preventing CPSF recruitment to pre-mRNAs (104). However, 3′ end processing of host pre-mRNAs is inhibited even upon infection with Influenza strains that carry polymorphisms that prevent NS1 binding to CPSF30 (104). It suggests that this interaction is not the sole mechanism of how Influenza virus disrupts cleavage and polyadenylation. ICP27 has been proposed to interact with hFip1 and CPSF73 subunits of CPSF and prevent the assembly of productive 3′ end processing machinery (101). On the other hand, ICP27 stimulates cleavage and polyadenylation of viral transcripts, and certain sequences motifs found upstream of the PAS of viral pre-mRNAs have been implicated in enabling ICP27 to act as an activator of 3′ end processing. Further studies will be required to explain how ICP27 discriminates viral transcripts from host pre-mRNAs and either stimulates or inhibits cleavage and polyadenylation. In-depth understanding of how viruses affect 3′ end processing of their metazoan host may also aid in the development of new therapeutics.

1.5 Impact of 3′ end processing on gene expression

1.5.1 Coordination with transcription termination

3′ end processing of nascent RNAs is closely linked to transcription termination, since defects in 3′ end cleavage caused by a variety of perturbations all result in transcription read-through (105). The dependence of transcription termination on PAS recognition was postulated more than 30 years ago, and we are now beginning to obtain insight into the molecular mechanism of their coordination on protein-coding genes (Figure 1.7) (106).

A prevailing model of transcription termination in eukaryotes posits that Pol II changes into a termination competent complex after the PAS has been transcribed (the allosteric model) (107). This allosteric change likely causes Pol II to slow or pause. Subsequent cleavage of the 5′-capped pre-mRNA by CPSF/CPF and its release from the transcribing Pol II leaves an unprotected 5′ phosphate on the nascent RNA, which is then targeted for degradation by a processive 5′-to-3′ torpedo exonuclease Xrn2. The reduced rate of Pol II progression along the DNA template allows Xrn2 to catch up with the polymerase and displace it from chromatin, terminating transcription (the torpedo model) (108). The 3′ end processing machinery is critical both for the deceleration of the transcribing polymerase and for providing access to Xrn2 (109). The PNUTS-PP1 complex has been demonstrated to reduce the rate of Pol II progression by dephosphorylating the C-terminal region of the transcription elongation factor SPT5 (53). In yeast, CPF-mediated dephosphorylation of tyrosine-1 of the CTD of Pol II at the 3′ end of protein-coding genes, promotes recruitment of Pcf11 and Rtt103, an interactor of the yeast homologue of the Xrn2 exonuclease, Rat1 (47). In addition, recent work shows that dephosphorylation by CPF promotes Pol II dimerisation (110). This dimer is compatible with basal transcription but not with the binding of transcription elongation factors. Therefore, Pol II dimerisation may be the allosteric change that promotes transcription termination. Overall, the phosphatase activity of PNUTS-PP1/Ref2-Glc7 may contribute to transcription termination by acting on several different substrates.

A major outstanding question is how the CPSF/CPF phosphatase activities are triggered after transcription of the PAS. The PNUTS-PP1 complex is strongly enriched on the chromatin regions around annotated 3′ cleavage sites, suggesting that specific recruitment of PNUTS-PP1 by the fully assembled 3′ end processing complex may lead to dephosphorylation of SPT5 around the site of 3′ end cleavage (53). However, the cleavage activity by CPSF/CPF is not strictly required for PAS-dependent transcription termination (111, 112).

The cleavage and polyadenylation machinery may also have roles in transcription termination beyond mediating Pol II deceleration and Xrn2 accessibility (Figure 1.7). Co-immunoprecipitation studies have detected physical association between Pol II and CPSF subunits, and recent work shows that Pol II and CPF interact directly *in vitro* (110, 113). Pcf11 also interacts with the CTD of Pol II (68).

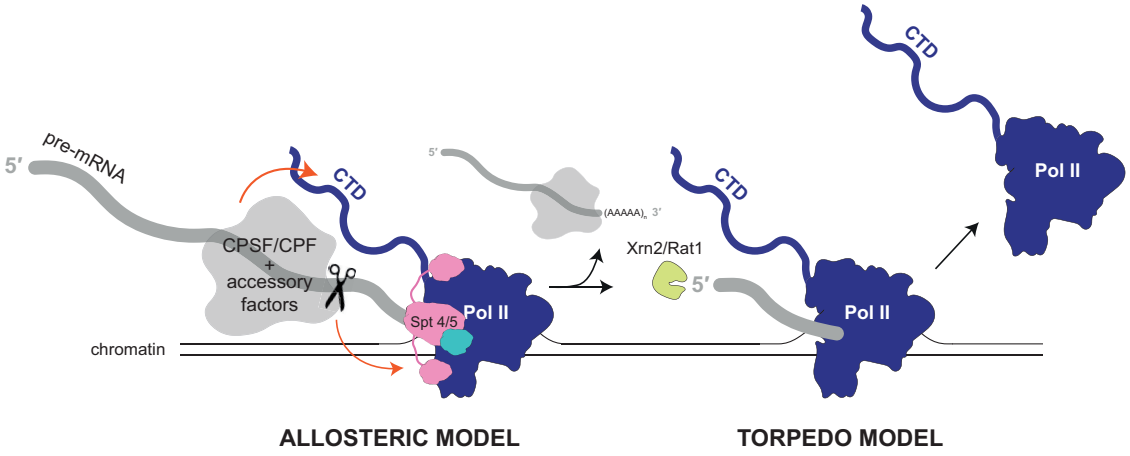


Figure 1.7 Eukaryotic pre-mRNA 3′ end processing is tightly coordinated with transcription termination. Schematic representation of transcription termination. Orange arrows depict phosphatase activity of CPSF/CPF acting on the CTD of Pol II and on transcription elongation factor SPT5. Adapted from (110).

1.5.2 Coordination with splicing

Recent advances in nascent RNA sequencing methods have revealed coordination between co-transcriptional splicing and 3′ end processing of eukaryotic pre-mRNAs. Long-read sequencing of chromatin-associated RNA using either PacBio or Nanopore platforms has enabled the elucidation of full-length pre-mRNA processing intermediates, allowing unambiguous determination of their splicing status and the position of Pol II along the gene at the time of sample preparation (114–116). These studies have uncovered an "all-or-none" nature of pre-mRNA processing: fully spliced pre-mRNAs are successfully cleaved at their 3′ ends and do not display read-through transcription, while transcripts that retain introns tend to be cleaved inefficiently as indicated by transcription continuing far beyond their annotated PAS (Figure 1.8) (114). This "all-or-none" coupling suggests extensive cross-talk between the splicing and the 3′ end processing machineries. Since co-transcriptional splicing precedes 3′ end processing, the removal of introns has been suggested to improve the efficiency of 3′ end processing. The influence of splicing on 3′ end cleavage has been demonstrated by studying the effects of a mutation found in some patients of β-thalassemia. A point mutation located in the intron of the β-globin gene creates a cryptic 3′ splice site, which is used more efficiently than the canonical 3′ splice site. The

increased splicing efficiency of the intron also leads to improved usage of the annotated PAS and reduced transcriptional read-through of the β-globin pre-mRNA (114).

How splicing stimulates 3′ end processing remains unknown. It has been suggested that splicing alleviates constitutive inhibition of 3′ end cleavage. A candidate inhibitor of 3′ end cleavage is the U1 snRNP that is bound to the 5′ splice site of the 3′-most intron of the pre-mRNA. U1 snRNP interacts with transcribing Pol II and also prevents intronic polyadenylation at cryptic PAS sites by inhibiting the assembly of an active 3′ end processing complex, and its removal after successful splicing of the 3′-terminal intron may then allow efficient 3′ end cleavage (117, 118). Other proteins that may inhibit 3′ end processing until splicing is complete include U2AF, PTB and hnRNPC. It is also possible that the proteins that are deposited on pre-mRNAs after successful excision of introns, such as the exon junction complex and SR proteins, activate 3′ end processing. Finally, the coupling between splicing and 3′ end processing could be indirect, for example, mediated by the kinetics of processing events, RNA secondary structures or local chromatin landscape.

Read-through transcripts that remain unspliced are generated at low levels even in unperturbed cells in the absence of external stress (116). Such unprocessed RNAs are typically retained in the nucleus and degraded by the nuclear exosome. Thus, inefficient pre-mRNA processing effectively downregulates mRNA levels independent of transcription initiation and may act as a regulatory mechanism of gene expression.

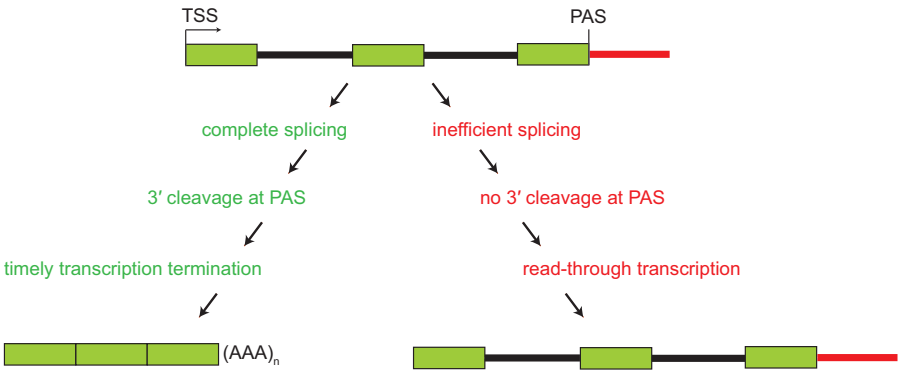


Figure 1.8 Eukaryotic pre-mRNA 3′ end processing is likely coordinated with splicing. Schematic representation of coupling between splicing, 3′ end processing and transcription termination. Green rectangles represent exons, black bars – introns, red bars – regions of read-through transcription. TSS – transcription start site; PAS – polyadenylation site.

1.6 Mechanism of CPSF endonuclease activation

Despite our growing understanding of how 3′ end processing is regulated and coordinated with other processes in the nucleus, the mechanism of the key enzymatic reaction in mRNA processing – the 3′ endonucleolytic cleavage – remains to be elucidated. CPSF73/Ysh1 adopts an inactive closed conformation in most of the structures that have been determined to date (Figure 1.9). The 3′ endonuclease has only weak and nonspecific nuclease activity either in isolation or within the purified CPSF/CPF complex, consistent with it being in an inactive conformation (38). This implies that the 3′ endonuclease requires an activation step which likely involves accessory proteins.

Visualisation of both the Integrator and the histone pre-mRNA 3′ end processing complex in their active state by cryoEM revealed the critical role of non-enzymatic subunits and accessory factors in the activation of cleavage by these large 3′ end processing machineries. The structure of the active histone pre-mRNA 3′ end processing machinery revealed large-scale conformational rearrangements within the HCC complex and the opening of the CPSF73 active site due to the metallo-β-lactamase and β-CASP domains pivoting away from each other (Figure 1.9A) (82). The N-terminal domain of symplekin, a subunit shared with CPSF, is essential for this activation and contacts the HDE-U7 RNA duplex and CPSF100. However, CPSF73 itself is bound and directly activated by the subunits that are unique to the histone pre-mRNA processing complex, Lsm10 and Lsm11. Hence, this structure does not explain how CPSF73 is activated in the context of CPSF. Specifically, Lsm11 interacts with the metallo-β-lactamase domain of the endonuclease. Interestingly, the same surface of Ysh1, the yeast orthologue of CPSF73, is bound by the UBL domain of Mpe1 in the CPF complex (Figure 1.9B). In addition, Lsm10 forms contacts with the β-CASP domain of the endonuclease subunit, which is reminiscent of the transcription elongation factor SPT5 interaction with INTS11 that activates the Integrator endonuclease (Figure 1.9C) (63, 119). Canonical pre-mRNA 3′ end processing can be uncoupled from transcription, suggesting that SPT5 is unlikely to be involved in the activation of the CPSF endonuclease. However, by analogy, CPSF73 within the canonical 3′ end processing machinery is likely activated by combined binding of the RBBP6 UBL to the metallo-β-lactamase domain and binding of another interactor to the β-CASP domain. An active state structure of CPSF bound to its substrate and auxiliary activators will be essential to determine how the active site of CPSF73 is pried open within the canonical 3′ end processing machinery.

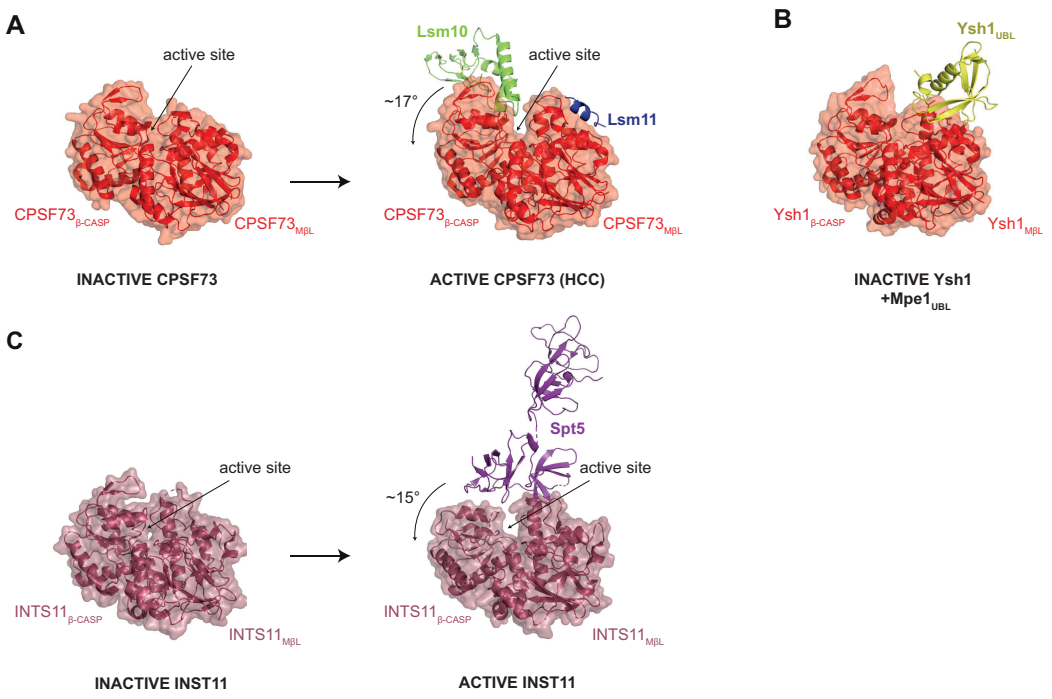


Figure 1.9 Activation of 3′ processing endonucleases requires accessory factors. (A) Structural comparison between the inactive state of CPSF73 (PDB 2I7T) (38) and activated CPSF73 (PDB 6V4X) (82) bound to Lsm10 and Lsm11 within the histone cleavage complex (HCC). Lsm10 acts as a wedge that induces a pivot of the metallo-β-lactamase domain (MβL) relative to the β-CASP, opening the active site of the endonuclease. **(B)** Structure of Ysh1 bound to the UBL domain of Mpe1 (PDB 6I1D) (12). **(C)** Structural comparison between the inactive state of INTS11 (PDB 7BFP) (62) within the Integrator cleavage module and activated INTS11 within Integrator bound to the paused Pol II complex (PDB 7PKS) (63). Similarly to Lsm10, SPT5 promotes opening of the active site.

1.7 Aims of this Study

So far, the activation of human CPSF endonuclease has been mainly studied by *in vitro* experiments in fractionated nuclear extract prepared from cultured human cells. However, the full protein composition of partially purified 3′ end processing machinery from nuclear extract is not known, making it difficult to infer molecular mechanisms. To enable detailed mechanistic studies of CPSF endonuclease activation, an *in vitro* assay containing a well-defined set of highly pure proteins is required. *In vitro* reconstitution of the yeast 3′ end cleavage reaction with purified recombinant proteins revealed that CPF requires cleavage factors CF IA and CF IB for its endonuclease activation and that the phosphatase module of CPF is dispensable for this activity (12). Despite significant progress in the field, the structure of the active yeast machinery has not been yet elucidated. In contrast, the reconstitution of the endonuclease activity of human CPSF has eluded researchers for decades, and exactly what proteins are required for its activation remains a mystery. This also suggests that there could be some differences between yeast and human machineries.

Therefore, in this study, I set out to establish a system to study human pre-mRNA 3′ end processing with purified recombinant proteins. In particular, I will focus on the cleavage activity by human CPSF, aiming to provide mechanistic insights into the activation mechanism of the CPSF endonuclease, which represents a major knowledge gap in the field of 3′ end processing.

The specific aims of this Dissertation are:

1. To purify recombinant human CPSF and its accessory protein factors (Chapter 2).
2. To reconstitute specific and efficient CPSF endonuclease activity with purified components and determine the minimal set of factors required for activation of pre-mRNA 3′ end cleavage in humans (Chapter 3).
3. To explore the molecular mechanism of CPSF endonuclease activation (Chapters 3 and 4).
4. To structurally characterise human pre-mRNA 3′ end processing machinery (Chapter 3 and 4).

Chapter 2:

Purification of recombinant protein factors involved in human pre-mRNA 3′ end processing

To establish an *in vitro* system to study human pre-mRNA 3′ end processing, I first aimed to purify the various proteins that have been implicated in this process in human cells. Production of sufficient amounts of highly pure proteins presents a major bottleneck in biochemical studies. Recombinant overexpression in a heterologous expression host enables large-scale protein production. However, human proteins involved in pre-mRNA 3′ end processing exhibit several features that make them particularly difficult to generate using a recombinant approach. First, most of these protein factors are, in fact, protein complexes composed of several different polypeptides, many of which may require co-translational assembly with their binding partners to attain their native conformation and evade degradation in the expression host. For this reason, I chose a biGBac system for baculovirus-mediated protein overexpression in insect cells (see Figure 6.1) (12, 120). The biGBac system allows multiple protein subunits to be simultaneously co-expressed from the same virus, allowing protein complexes to assemble co-translationally with correct stoichiometry. Each protein-coding sequence is inserted into the baculoviral genome with its own promoter and terminator using Gibson assembly, which enables quick and efficient generation of viruses encoding any desired combination of human proteins. The use of a higher eukaryote as an expression system also decreases the chance of human proteins misfolding when over-expressed in an orthogonal host. Second, many constituent subunits of human 3′ end processing complexes contain intrinsically disordered regions (IDRs) absent from their yeast orthologues. IDRs tend to be susceptible to proteolytic degradation, which may reduce the apparent expression level of orthogonally over-expressed proteins (121). In addition, IDRs in purified proteins often cause protein aggregation in the absence of their native interaction partners. Although the precise function of these regions is often unknown, they are likely to mediate protein-protein and protein-RNA interactions that are specific to higher eukaryotes, and hence, are unlikely to be involved in the conserved functions of the complex. Thus, for many of the 3′ end processing factors, I used truncated protein constructs to produce large amounts of recombinant protein complexes.

In this Chapter, I will provide an overview of my strategies to express and purify major human pre-mRNA 3′ end processing factors. I will also discuss some of their properties with regards to polyadenylation activity.

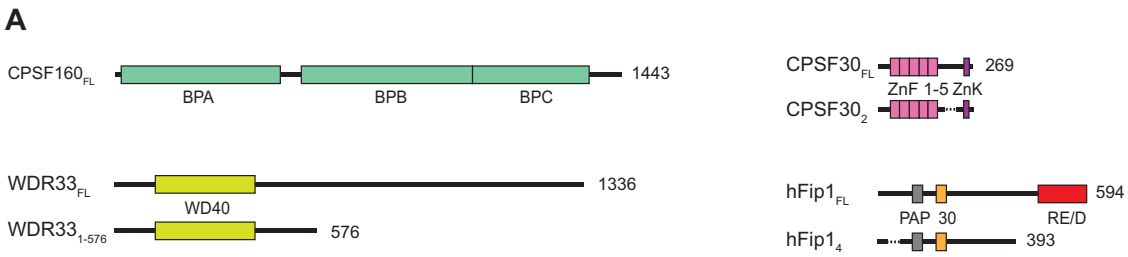
2.1 Purification and characterisation of recombinant human CPSF

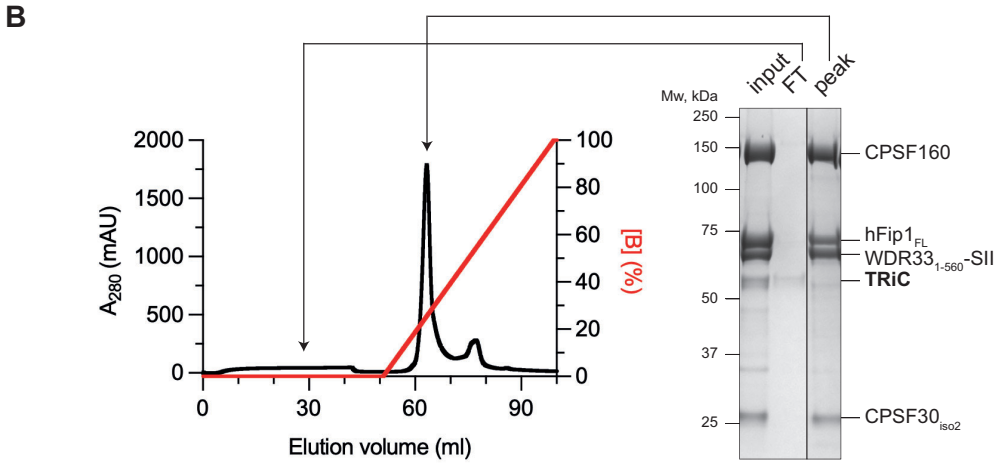
2.1.1 Purification of recombinant mPSF

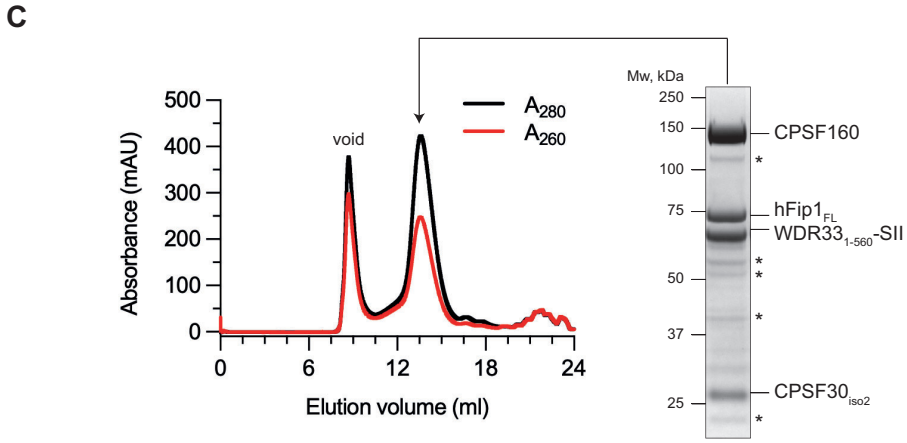
At the start of this project, full human CPSF could only be obtained from endogenous sources in low quantities (17). In contrast, the mPSF module of CPSF had been successfully produced recombinantly by multiple research groups (20–22). Due to the modular nature of protein complexes, I hypothesised that complete CPSF could be assembled from individually purified mPSF and mCF modules, and therefore began the project by purifying recombinant human mPSF.

First, I attempted to purify human mPSF containing all four full-length subunits: CPSF160, WDR33, CPSF30 and hFip1 (Figure 2.1A). I performed affinity purification using a Strep-II tag on the C-terminus of WDR33 subunit followed by ion exchange chromatography. However, low expression levels meant that the yield of the full-length complex was insufficient for extensive biochemical characterisation. Thus, an intrinsically-disordered C-terminal region that lacks sequence conservation was removed from subunit WDR33, as described previously (Appendix Figure 8.1) (21), and the full-length version of CPSF30 was replaced with its shorter isoform 2 (Figure 2.1A). The resulting truncated mPSF complex could be expressed in large quantities in Sf9 insect cells. However, an abundant contaminant of a molecular weight of ~60 kDa co-purified with mPSF after both ion exchange chromatography on a Heparin column and gel filtration chromatography (also see "input" in Figure 2.1B). Mass spectrometry analysis of the contaminant band revealed it to be the T-complex protein Ring Complex (TRiC). TRiC is composed of 16 structurally related subunits of similar molecular weight (~60 kDa). TRiC is a chaperonin that has been shown to assist in the folding of WD40 domains, four of which are present within the mPSF complex (122). Thus, TRiC may be involved in the folding of WDR33 and CPSF160 subunits. I tried several strategies to remove TRiC from the mPSF complex. For instance, I tested the expression of mPSF in Hi5 insect cells and found that, based on qualitative SDS-PAGE analysis, mPSF expressed in Hi5 cells was bound to slightly less TRiC than when expressed in Sf9 cells, which prompted me to switch the cell line for mPSF expression. Also, I found that TRiC complex failed to bind to the Resource Q anion exchange column, while mPSF did. Therefore, replacing a Heparin column for a Resource Q column for an ion exchange step enabled efficient removal of TRiC from the mPSF complex (Figure 2.1B).

After obtaining TRiC-free mPSF by anion exchange chromatography, I attempted to further purify the complex by gel filtration chromatography. The gel filtration trace revealed a significant void peak, which indicated that the complex formed soluble aggregates (Figure 2.1C). In addition, many additional bands, most likely representing protein degradation products, could be detected by SDS-PAGE analysis. Most of the previous studies used a version of mPSF either containing a truncated version of hFip1 or lacking this subunit altogether (21–23), suggesting that hFip1 could be causing the observed aggregation. To test this possibility, I replaced the full-length hFip1 subunit with its shorter isoform 4, which lacks the C-terminal RE/D domain (Figure 2.1A). The mPSF complex containing truncated WDR33, CPSF30 and hFip1 subunits (mPSF-hFip1$_4$) yielded milligram quantities of highly pure protein, which appeared to be monodisperse by gel filtration chromatography (Figure 2.1D). This truncated mPSF complex will be used throughout the experiments described in this Thesis, unless indicated otherwise.


A
CPSF160FL
BPA
BPB
BPC
1443
CPSF30FL
269
ZnF 1-5 ZnK
CPSF302
WDR33FL
WD40
1336
WDR331-576
576
hFip1FL
PAP 30
RE/D
594
hFip14
393


B
input
FT
peak
Mw, kDa
250
150
100
75
50
37
25
CPSF160
hFip1FL
WDR331-560-SII
TRiC
CPSF30iso2
A280 (mAU)
[B] (%)
Elution volume (ml)


C
A280
A260
void
Absorbance (mAU)
Elution volume (ml)
Mw, kDa
CPSF160
hFip1FL
WDR331-560-SII
CPSF30iso2

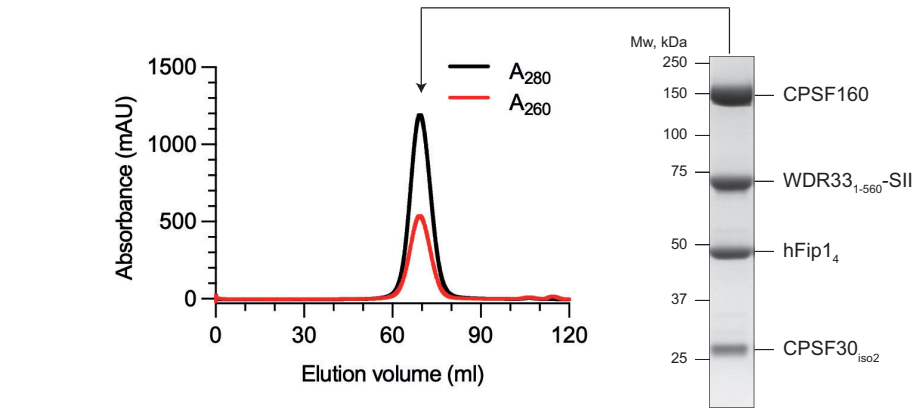

D
A280
A260
Absorbance (mAU)
Elution volume (ml)
Mw, kDa
CPSF160
WDR331-560-SII
hFip14
CPSF30iso2

Figure 2.1 Purification of recombinant human mPSF. (A) Domain diagrams of human mPSF subunits. The number of residues is indicated to the right of each diagram. The residue boundaries of truncated protein variants are marked as black lines. FL – full length; BPA, BPB, BPC – β propeller A, B, C; ZnF – zinc finger; ZnK – zinc knuckle; PAP – PAP-binding region; 30 – CPSF30-binding region. **(B)** Chromatogram of the mPSF complex containing full-length hFip1 run on a Resource Q 1 ml anion exchange column and SDS-PAGE analysis of selected fractions. FT – flow-through. The position of the TRiC contaminant on the gel is indicated. [B] – relative concentration of buffer B containing 1 M NaCl. **(C)** Size exclusion chromatogram of the mPSF complex containing full-length hFip1 run on a Superose 6 10/300 column and SDS-PAGE analysis of the peak fraction. Degradation products are marked with an asterisk. **(D)** Size exclusion chromatogram of the mPSF complex containing isoform 4 of hFip1 run on a Superose 6 XK 16/700 column and SDS-PAGE analysis of the peak fraction.

2.1.2 Recombinant human mPSF is active in polyadenylation

Human mPSF has been shown to be necessary and sufficient for PAS-dependent polyadenylation *in vitro* by recruiting PAP to the RNA substrate via the hFip1 subunit (20, 35). Thus, I investigated if the recombinant mPSF complex I purified was active in polyadenylation, and, specifically, if the truncation of hFip1 required to optimise the preparation of the complex affected its ability to stimulate polyadenylation. To that end, in addition to mPSF, I also cloned and purified recombinant human PAP from insect cells (Figure 2.2A). As a model substrate for polyadenylation assays, I used a 41 nt fragment of the adenoviral L3 RNA, which mimics the pre-mRNA after 3′ end cleavage (Table 6.2). A longer version of the L3 pre-mRNA has been used extensively in experiments using mammalian nuclear extract, and it is known to be cleaved and polyadenylated efficiently *in vivo*. The synthetic substrate carried a 6-FAM fluorescent label on its 5′ end, which enabled convenient detection of the RNA after denaturing gel electrophoresis. mPSF (50 nM), PAP (50 nM) and the substrate RNA (300 nM) were mixed, and ATP (2 mM) was added to start the reaction. Aliquots were taken at various time points, the reaction was stopped by adding buffer that denatures proteins, and the samples were analysed by denaturing gel electrophoresis. The bands of higher molecular weight than the substrate RNA corresponded to polyadenylated products.

I compared the polyadenylation activity of PAP alone and of PAP in the presence of the following mPSF constructs: mPSF with full-length hFip1 (mPSF-hFip1$_{FL}$), mPSF containing truncated hFip1 (mPSF-hFip1$_4$), and mPSF lacking hFip1 altogether (mPSF-ΔhFip1) (Figure 2.2B). Every mPSF construct noticeably stimulated the polyadenylation activity of PAP, as

indicated by higher molecular weight products accumulating more quickly in the presence of mPSF compared with PAP alone. Both mPSF-hFip1$_{FL}$ and mPSF-hFip1$_4$ stimulated polyadenylation to a similar extent, suggesting that isoform 4 of hFip1 retains the same polyadenylation activity. In contrast, mPSF-ΔhFip1 resulted in lower levels of polyadenylation, consistent with the key role of hFip1 in recruiting PAP to mPSF. Overall, these data show that I have successfully purified recombinant human mPSF complex that is active in polyadenylation.

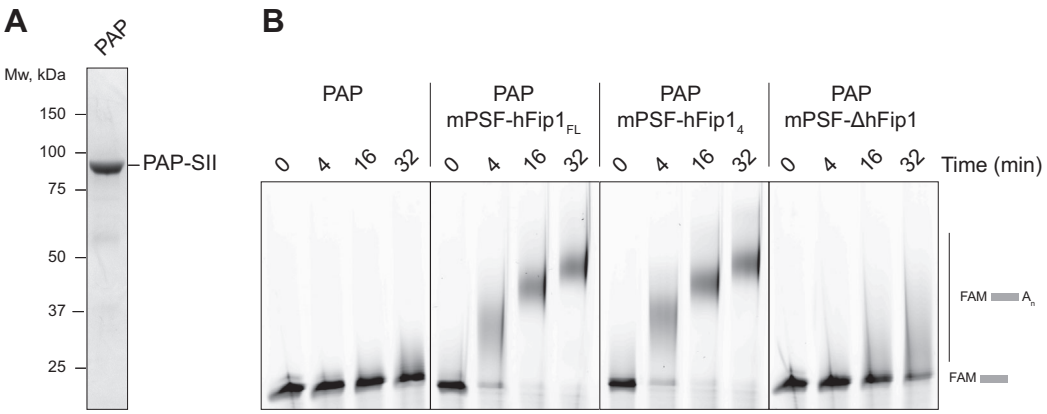


Figure 2.2 Recombinant human mPSF is active in polyadenylation. (A) SDS-PAGE analysis of purified recombinant human PAP. **(B)** Denaturing gel electrophoresis analysis of polyadenylation assays in the presence of PAP and various mPSF constructs. Bands of high molecular weight represent polyadenylated substrate RNAs with poly(A) tails of variable length.

2.1.3 Purification of recombinant mCF

After obtaining purified mPSF active in polyadenylation, I aimed to reconstitute the complete recombinant CPSF complex. To this end, I also cloned, expressed and purified human mCF, containing three subunits: symplekin, CPSF100 and the 3′ RNA endonuclease CPSF73 (Figure 2.3A). A Strep-II tag on the C-terminus of the scaffold protein symplekin was used for affinity purification of the complex, followed by anion exchange and size exclusion chromatography. The full-length mCF yielded milligram quantities of pure protein complex (Figure 2.3B). Therefore, unlike in the case of mPSF, no truncations were made to the subunits of the mCF complex.

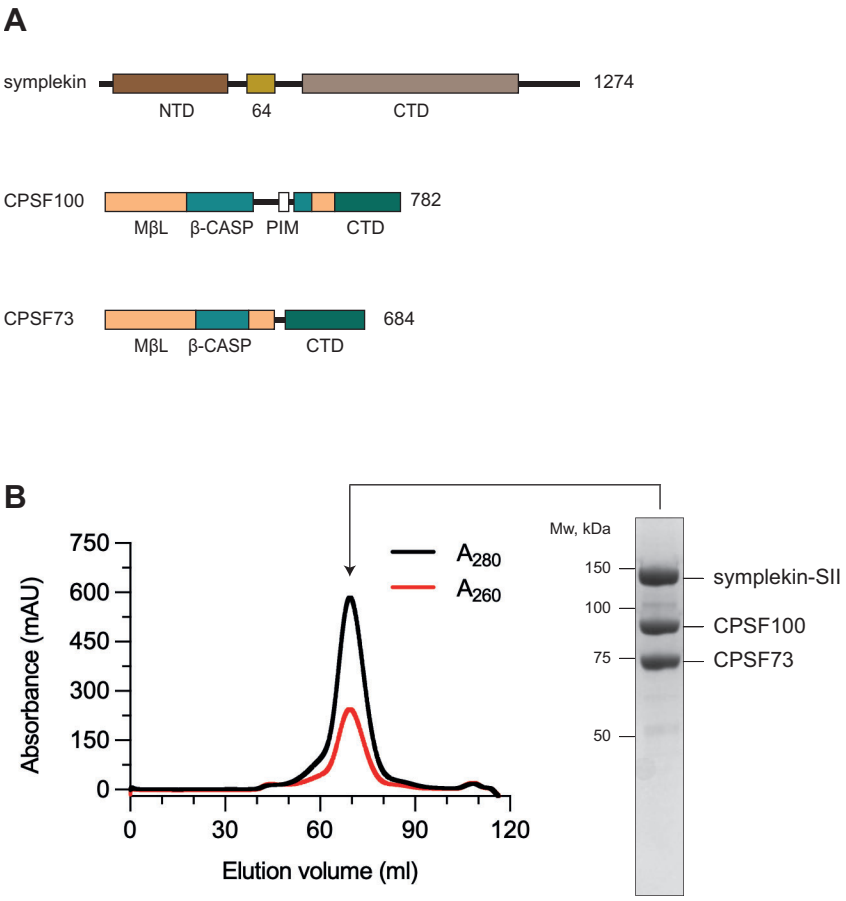


Figure 2.3 Purification of recombinant human mCF. (A) Domain diagrams of human mCF subunits. The number of residues is indicated to the right of each diagram. NTD – N-terminal domain; CTD – C-terminal domain; 64 – CStF64-interacting region; MβL – metallo-β-lactamase domain; PIM – mPSF-interacting motif. **(B)** Size exclusion chromatogram of the mCF complex run on a Superose 6 KX 16/700 column and SDS-PAGE analysis of the peak fraction.

2.1.4 Human mPSF and mCF form a stable CPSF complex *in vitro*

Formation of the CPSF complex by mixing individually purified mPSF and mCF modules was tested by size exclusion chromatography. mPSF and mCF were mixed at equimolar concentrations (2.5 μM each) and loaded onto an analytical size exclusion column. As controls, the two modules were run separately under the same conditions, and the peak fractions were then analysed by SDS-PAGE. When mixed, both mPSF and mCF showed a noticeable leftwards shift in their elution profiles, which was indicative of the formation of a larger protein complex (Figure 2.4). In addition, mPSF and mCF eluted from the column together as a single peak, and all the subunits from both modules were present at comparable stoichiometry (Figure 2.4). Thus, mixing separately purified mPSF and mCF modules at equimolar concentrations of at least 2.5 μM allows efficient formation of the 7-subunit human CPSF complex. In all subsequent assays, CPSF was reconstituted by mixing mPSF and mCF without an additional gel filtration run. This strategy allowed to minimise the loss of purified protein during CPSF reconstitution and, more importantly, enabled me to easily swap various versions of either module (containing mutant or truncated subunits, for example) and test their effects on CPSF activity.

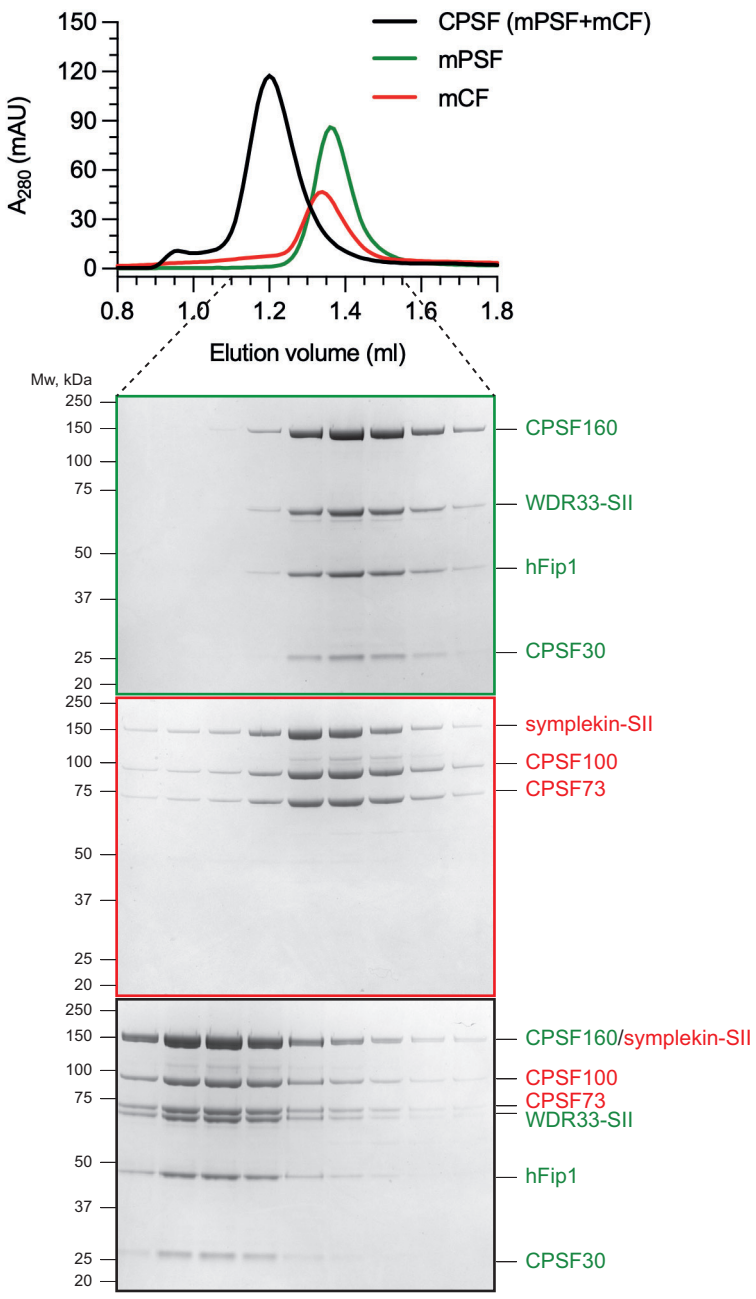


Figure 2.4 mPSF and mCF assemble into the CPSF complex. Size exclusion chromatograms of either mPSF (2.5 μM) or mCF (2.5 μM) alone, or the two modules mixed together run on a Superose 6 3.2/300 column. SDS-PAGE analysis of the corresponding peak fractions is shown below the chromatograms.

2.2 Purification and characterisation of recombinant human cleavage factors

In the previous section, I developed a protocol to produce human CPSF recombinantly and showed that the purified complex is active in polyadenylation. However, it is known that CPSF alone is not sufficient to catalyse the endonucleolytic cleavage of pre-mRNAs (8, 38). This suggests that the CPSF endonuclease needs to be specifically activated to ensure fidelity and specificity of pre-mRNA 3′ end processing. Many protein complexes have been implicated in this activation mechanism, including CStF, CFIIm and CFIm. In this section, I will describe the purification and characterisation of these three cleavage factor complexes.

2.2.1 Purification of recombinant CStF and CFIIm complexes

CStF and CFIIm are the two mammalian cleavage factors that were proposed to be essential for the endonucleolytic RNA cleavage catalysed by CPSF. The CStF complex is a dimer of trimers composed of highly-conserved subunits CStF77 and CStF64 as well as a metazoan-specific protein CStF50 (Figure 2.5A). CFIIm is a dimeric complex containing conserved proteins Pcf11 and Clp1 (Figure 2.6A). CStF and CFIIm complexes were expressed in Sf9 insect cells and purified by affinity chromatography using a Strep-II tag on the C-terminus of CStF77 and Pcf11, followed by anion exchange and size exclusion chromatography. Full-length CStF appeared monodisperse and could be purified in large quantities using this protocol (Figure 2.5B).

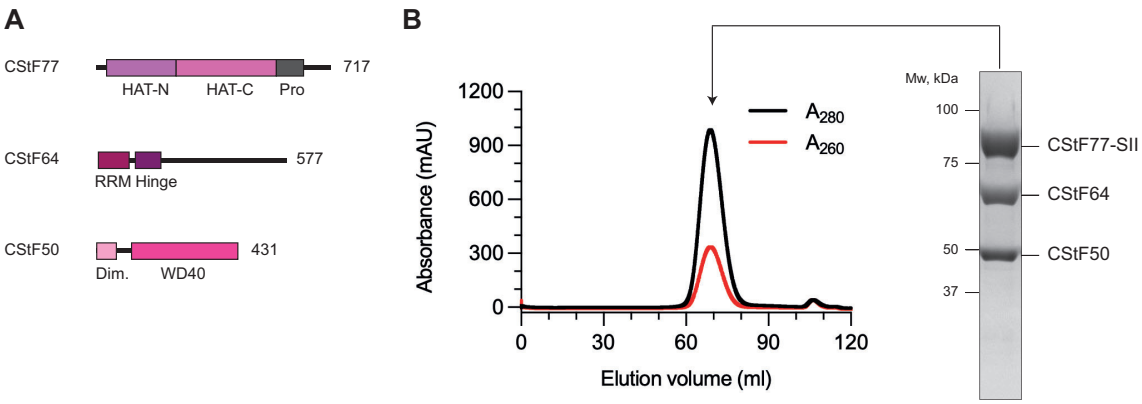


Figure 2.5 Purification of recombinant human CStF. (A) Domain diagrams of human CStF subunits. The number of residues is indicated to the right of each diagram. HAT-N, HAT-C – N-terminal and C-terminal, respectively, half a tetratricopeptide repeat domain; Pro – proline-rich domain; RRM – RNA recognition motif domain; Dim. – dimerisation domain. **(B)** Size exclusion chromatogram of the CStF complex run on a Superose 6 XK 16/70 column and SDS-PAGE analysis of the peak fraction.

Full-length CFIIm, however, eluted in the void volume of the gel filtration column when injected at concentrations above ~1 mg/ml (Figure 2.6B). Formation of soluble aggregates could be prevented by keeping the CFIIm complex at concentrations <1 mg/ml, but this was too dilute for subsequent protein-protein interaction studies by analytical gel filtration chromatography (Figure 2.6B). It is important to note that even when the total amount of CFIIm injected was comparable, less protein was recovered from the column when the sample was more concentrated, as indicated by the total area under the trace (Figure 2.6B). Some protein may have been either retained on the column filter or remained bound to the resin, highlighting the tendency of CFIIm to precipitate at higher concentrations. In addition, full-length Pcf11 noticeably degraded during purification, leading to two overlapping size exclusion peaks: one containing intact CFIIm, and another composed of partially degraded Pcf11 (Figure 2.6B). Therefore, I deleted part of the N-terminal region of Pcf11 (residues 1-769), which was previously reported to be dispensable for CPSF cleavage in nuclear extract but may be susceptible to both proteolysis and aggregation (10) (Figure 2.6A). The resulting $Pcf11_{770\text{-}1555}$-Clp1 complex remained fully soluble at concentrations of >2 mg/ml and was used throughout the experiments described in this Thesis unless indicated otherwise (Figure 2.6C).

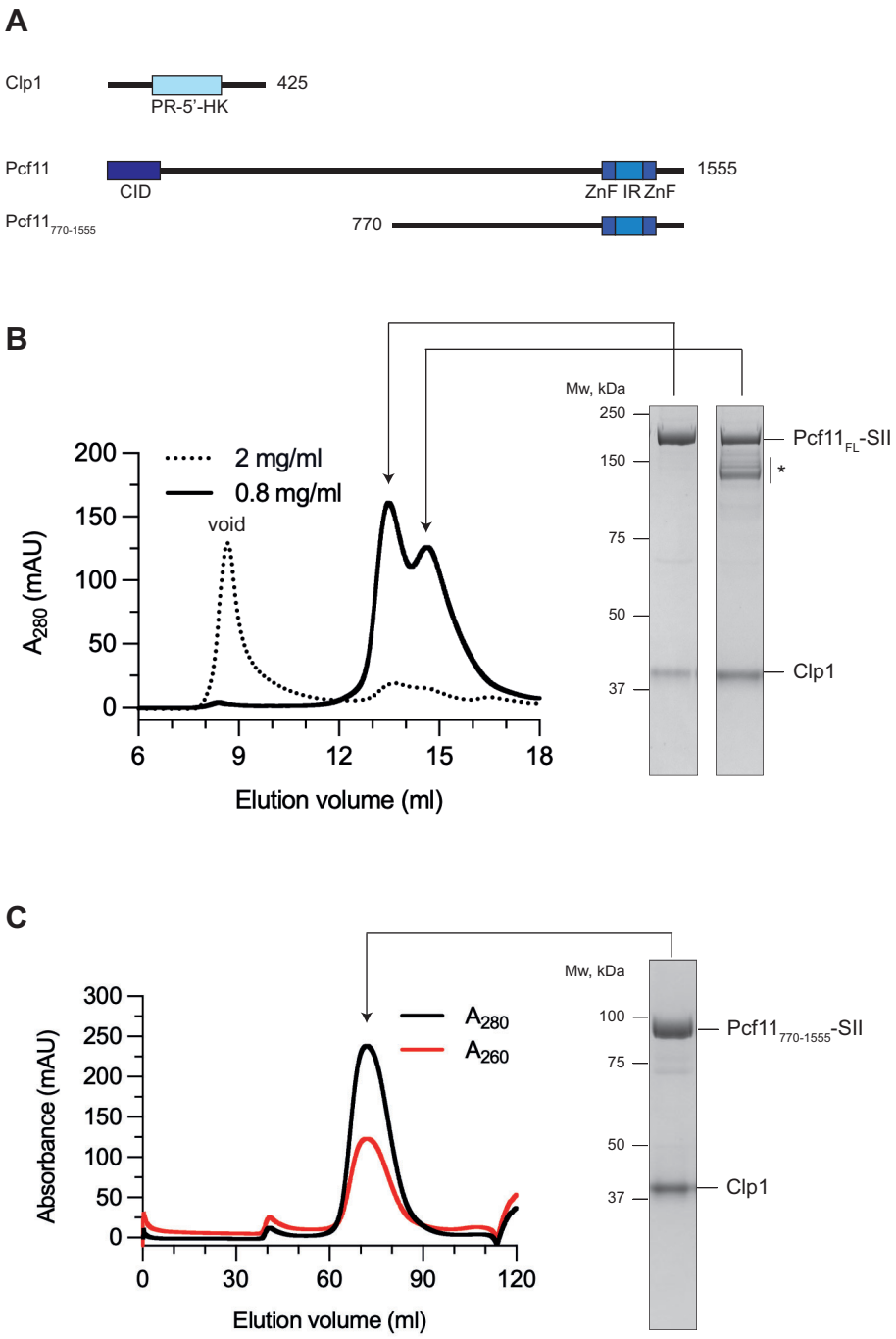

A
Clp1
425
PR-5'-HK
Pcf11
1555
CID
ZnF IR ZnF
Pcf11770-1555
770
B
Mw, kDa
250
150
75
50
37
Pcf11FL-SII
*
Clp1
200
150
100
50
0
2 mg/ml
0.8 mg/ml
void
A280 (mAU)
6
9
12
15
18
Elution volume (ml)
C
Mw, kDa
100
75
50
37
Pcf11770-1555-SII
Clp1
300
250
200
150
100
50
0
A280
A260
Absorbance (mAU)
0
30
60
90
120
Elution volume (ml)

Figure 2.6 Purification of recombinant human CFIIm. (A) Domain diagrams of human CFIIm subunits. The number of residues is indicated to the right of each diagram. The residue boundaries of truncated Pcf11$_{770\text{-}1555}$ are indicated as a black line. PR-5′-HK – polyribonucleotide 5′-hydroxyl-kinase domain; CID – RNA polymerase II CTD-interacting domain; ZnF – zinc finger domain; IR – Pcf11-interacting region. **(B)** Size exclusion chromatograms of full-length CFIIm run on a Superose 6 10/300 column and SDS-PAGE analysis of selected fractions. The samples were injected at the protein concentrations of ~2 mg/ml (dashed line) and ~0.8 mg/ml (solid line), but the total amount of protein injected was roughly the same. The total area under the curve of the ~0.8 mg/ml sample appears to be much larger than that of the ~2 mg/ml sample, indicating a greater protein recovery from the column at a lower protein concentration. Pcf11 degradation products are marked with an asterisk. **(C)** Size exclusion chromatogram of the CFIIm complex containing truncated Pcf11$_{770\text{-}1555}$ run on a Superose 6 XK 16/700 column and SDS-PAGE analysis of the peak fraction.

2.2.2 CStF and CFIIm form a stable complex

Yeast orthologues of CStF (except for CStF50) and CFIIm subunits together form a constitutive cleavage factor complex CF IA. Despite the conserved function of these two protein complexes, it remained unclear whether they interact in humans (10). To test this, I mixed purified CStF (2 μM) with molar excess of CFIIm (2.5 μM) and analysed the sample by analytical gel filtration chromatography (Figure 2.7). The leftwards shift in the elution volume depicted in the chromatogram as well as SDS-PAGE analyses of the peak fractions revealed the formation of a stable CStF-CFIIm complex. Thus, although these two protein factors do not associate constitutively in human cells, CStF and CFIIm have a potential to interact, at least *in vitro*, likely in a similar manner as within yeast CF IA.

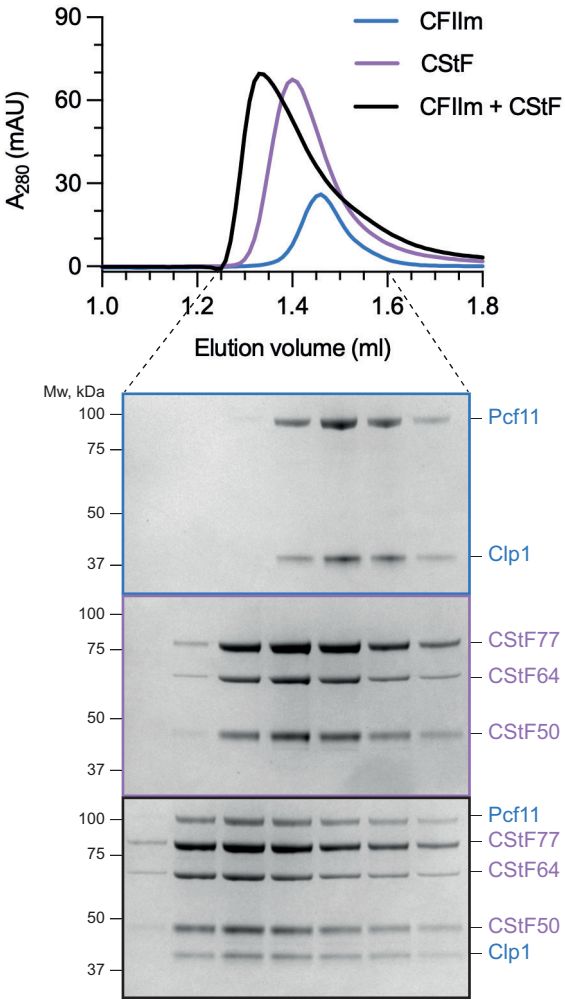


Figure 2.7 Human CStF and CFIIm complexes interact directly. Size exclusion chromatograms (top) and SDS-PAGE analyses (bottom) of the peak fractions of CFIIm, CStF and the two complexes mixed together. The fractions analysed by SDS-PAGE are indicated.

2.2.3 CStF enhances specificity of mPSF-dependent polyadenylation *in vitro*

CStF interacts with both RNA and the CPSF subunit CPSF160, and could therefore stabilise mPSF binding to the substrate (8, 66). The G/U-rich elements recognised by CStF64/Rna15 tend to be located downstream of the cleavage site. Nevertheless, the Rna14-Rna15 complex from yeast has been shown to stimulate polyadenylation of a pre-cleaved RNA substrate catalysed by purified yeast polymerase module, possibly by binding to RNA upstream of the cleavage site (9). I proposed that CStF could also stimulate mPSF-dependent polyadenylation activity by human PAP. To test this, I added a four-fold molar excess of CStF (200 nM) over PAP and mPSF into a polyadenylation reaction (described in Section 2.1.2). Almost no effect was observed upon addition of CStF except for a slightly broader distribution of poly(A) tail lengths, as indicated by more diffuse polyadenylation product bands in the presence of CStF (Figure 2.8A). Since CStF may stabilise mPSF on the substrate RNA, I rationalised that CStF may stimulate polyadenylation of suboptimal pre-mRNAs, for instance, containing a mutant PAS (AACAAA), which has reduced affinity for mPSF (3, 29). Indeed, polyadenylation of the mutant substrate was noticeably more distributive than that of the wild-type RNA, as evidenced by a broader distribution and shorter length of polyadenylation products (Figure 2.8B). Surprisingly, addition of CStF into a polyadenylation reaction of the substrate containing the AACAAA sequence as its PAS resulted in a dose-dependent inhibition of polyadenylation (Figure 2.8B). This suggests that in humans, in addition to activating endonucleolytic cleavage, CStF may also prevent polyadenylation of suboptimal pre-mRNA substrates and therefore enhance specificity of 3′ end processing.

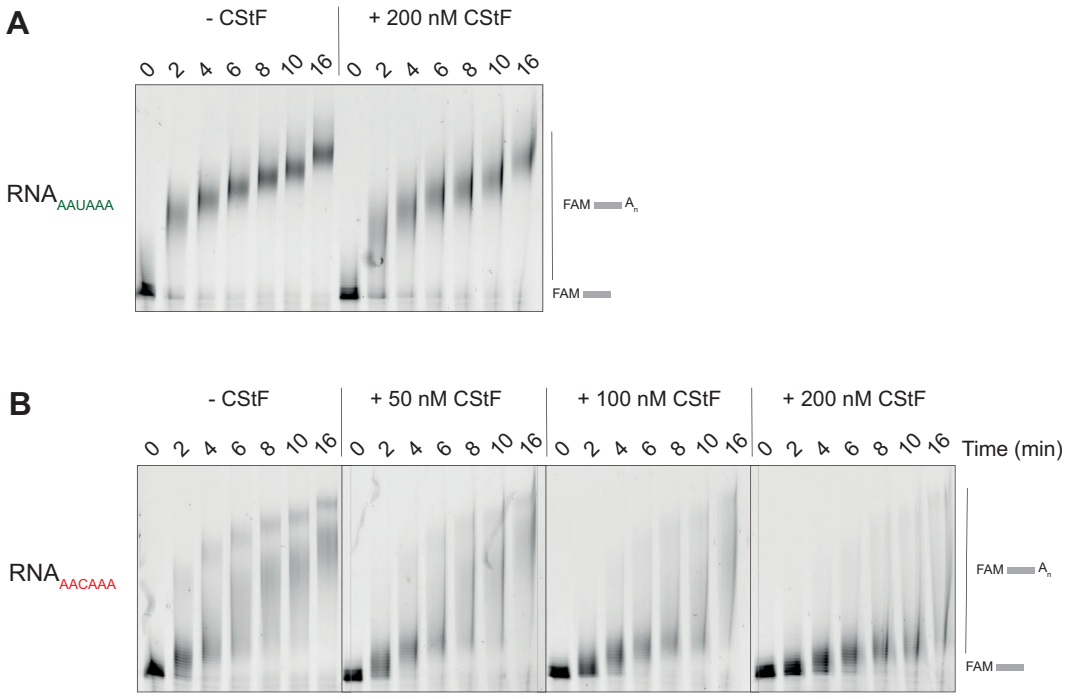


Figure 2.8 CStF inhibits polyadenylation of suboptimal substrate RNAs. (A) Denaturing PAGE analysis of polyadenylation assays performed in the presence of 300 nM 5′-FAM-labelled 41 nt pre-cleaved L3 pre-mRNA substrate containing wild-type PAS, 50 nM mPSF, 50 nM PAP, 2 mM ATP in the presence or absence of CStF. **(B)** Denaturing PAGE analysis of polyadenylation assays performed in the presence of 300 nM 5′-FAM-labelled 41 nt pre-cleaved L3 pre-mRNA substrate containing mutant PAS and variable concentrations of CStF.

2.2.4 Purification of recombinant CFIm

Unlike CStF and CFIIm, CFIm is thought not to be essential for CPSF endonuclease activation (70). Instead, it may recruit the 3′ end processing machinery to specific PAS sites containing an upstream UGUA motif. To purify CFIm, I tagged with a Strep-II tag on the C-terminus of CFIm25 and co-expressed it with CFIm68 in Sf9 insect cells (Figure 2.9A). In human cells, CFIm25 can also associate with an alternative CFIm subunit CFIm59. Although CFIm68 and CFIm59 have partially redundant functions, CFIm68 appears to be a stronger activator of 3′ end processing and was therefore chosen for this study (70). Similar to the other complexes described here, the purification of CFIm was first attempted by Strep-Tactin affinity, anion exchange and gel filtration chromatography. However, the size exclusion trace of CFIm showed a prominent void peak, indicating the presence of soluble aggregates (Figure 2.9B). The aggregation of CFIm was strongly dependent on the concentration of NaCl in the size exclusion buffer, with higher salt concentrations (~300 mM) improving the solubility of the complex (Figure 2.9B). In addition, increasing the salt concentration in the gel filtration buffer noticeably improved protein recovery from the column, suggesting that salt may prevent not only the precipitation of CFIm but also its interactions with the resin.

CFIm68 contains an extensive IDR, which may cause the observed aggregation. The IDR of CFIm68 includes a proline-rich segment as well as a C-terminal RS-like domain which, by interacting with hFip1, recruits CPSF to specific cleavage sites (Figure 2.9A) (70). Therefore, due to the functional significance of this IDR, it could not be removed to improve the solubility of the complex. To minimise CFIm aggregation, I avoided the gel filtration step, which requires the complex to be concentrated, potentially promoting protein precipitation. Instead, I dialysed the pooled peak fractions from the anion exchange step against a buffer containing 400 mM NaCl, which is the approximate concentration of salt that CFIm eluted at from an anion exchange column (Figure 2.9C). The dialysis step was necessary to adjust the salt concentration across all pooled fractions to a known value, since the precise NaCl concentration at which the complex eluted from an anion exchange column cannot be easily measured. This allowed accurate control of the salt concentration in the assays containing CFIm. Nevertheless, some soluble aggregates of CFIm may still form even at these high salt conditions. Overall, the purification of CFIm was the most challenging out of all of the human 3′ end processing factors, likely due to its functionally important IDRs.

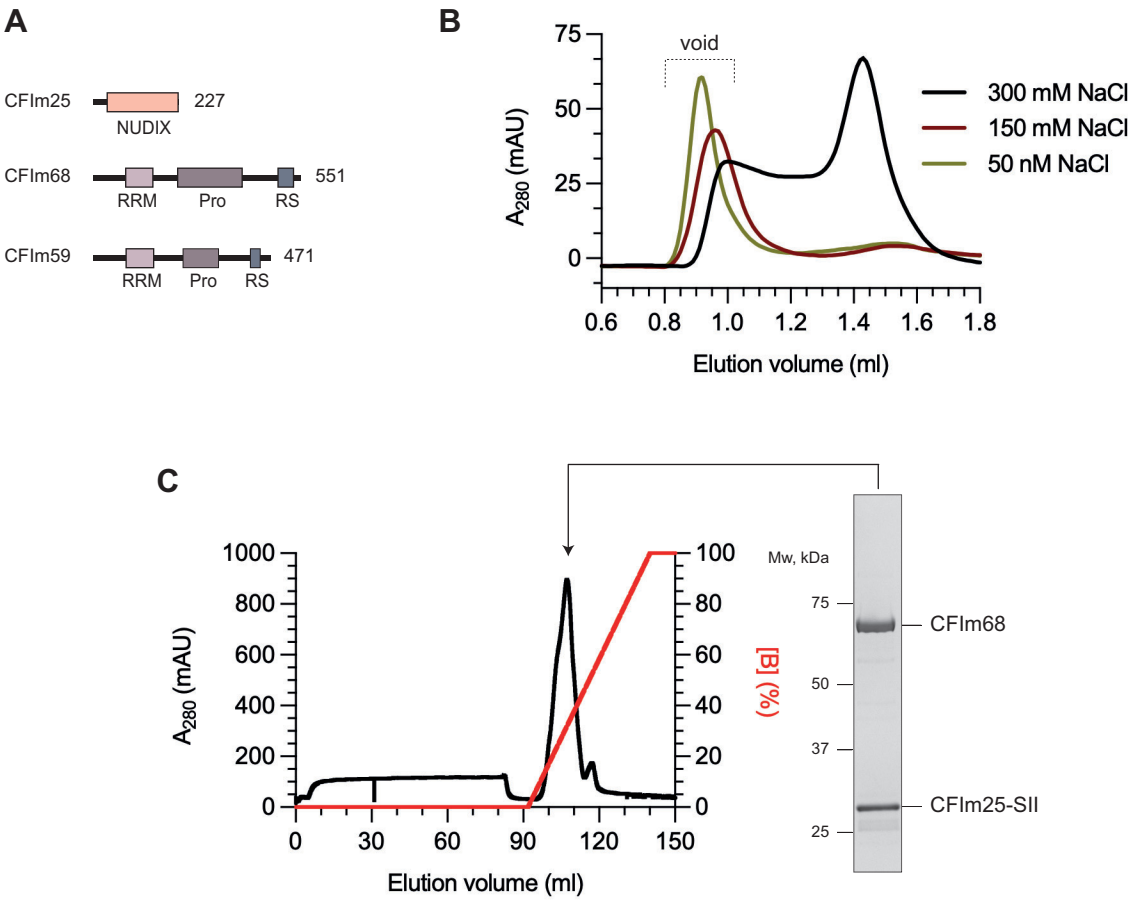


Figure 2.9 Purification of recombinant human CFIm. (A) Domain diagrams of human CFIm subunits. The number of residues is indicated to the right of each diagram. NUDIX – Nudix hydroxylase-like domain; RRM – RNA recognition motif domain; Pro – proline-rich domain; RS – arginine/serine-rich domain. **(B)** Chromatogram traces of purified CFIm complexes run on a Superose 6 3.2/300 gel filtration column equilibrated in buffers containing a variety of salt concentrations. **(C)** Chromatogram of a Resource Q 1 ml anion exchange column run of the CFIm complex and SDS-PAGE analysis of the peak fraction. [B] – relative concentration of buffer B containing 1 M NaCl. The pooled peak fractions were dialysed.

2.2.5 Purification of RBBP6

The multidomain protein RBBP6 has been implicated in 3′ end processing in humans (44). However, the exact role of RBBP6 remained unclear, and I aimed to investigate it using recombinant proteins. Full-length RBBP6 failed to express in insect cells, and hence, I tried expressing truncated versions of the protein. The longest construct that led to protein expression contained the N-terminal 844 amino acids of RBBP6 (Figure 2.10A). However, I decided to focus on RBBP6 construct 1-335, because it encompasses the highly conserved region that is equivalent to its yeast orthologue, contains the folded domains and excludes the C-terminal IDR, which is presumed to interact with protein factors involved in processes other than cleavage and polyadenylation (Appendix Figure 8.2) (41, 42, 123).

The yeast orthologue of RBBP6, Mpe1, is a constitutive subunit of the yeast CPF complex. Thus, I first attempted to purify RBBP6$_{1\text{-}335}$ bound to CPSF. I co-expressed in insect cells Strep-II tagged RBBP6$_{1\text{-}335}$ with untagged CPSF and its modules, and performed a pull-down experiment using Strep-Tactin beads. RBBP6$_{1\text{-}335}$ was expressed in large quantities on its own, confirming that it was indeed the C-terminal IDR that impeded the production of the full-length protein. However, despite high levels of CPSF expression, hardly any CPSF subunits, apart from substoichiometric amounts of CPSF73, co-purified with RBBP6$_{1\text{-}335}$ (Figure 2.10B). Thus, this experiment demonstrated that to study RBBP6, it had to be purified separately in the absence of CPSF.

RBBP6$_{1\text{-}335}$ was purified using Strep-Tactin affinity, anion exchange and size exclusion chromatography. SDS-PAGE analysis revealed that RBBP6 was expressed, albeit as a doublet band running at the expected molecular weight (~41 kDa) (Figure 2.10C). The protein species of a higher apparent molecular weight was consistently more abundant, but both bands remained present throughout the purification. Mass spectrometry confirmed that both bands indeed corresponded to RBBP6. However, attempts to determine by mass spectrometry whether either proteolysis or different post-translational modifications were responsible for the doublet band were inconclusive. Regardless, I successfully produced large quantities of relatively pure RBBP6$_{1\text{-}335}$ protein, which will be referred to as RBBP6 in this Thesis, unless indicated otherwise.

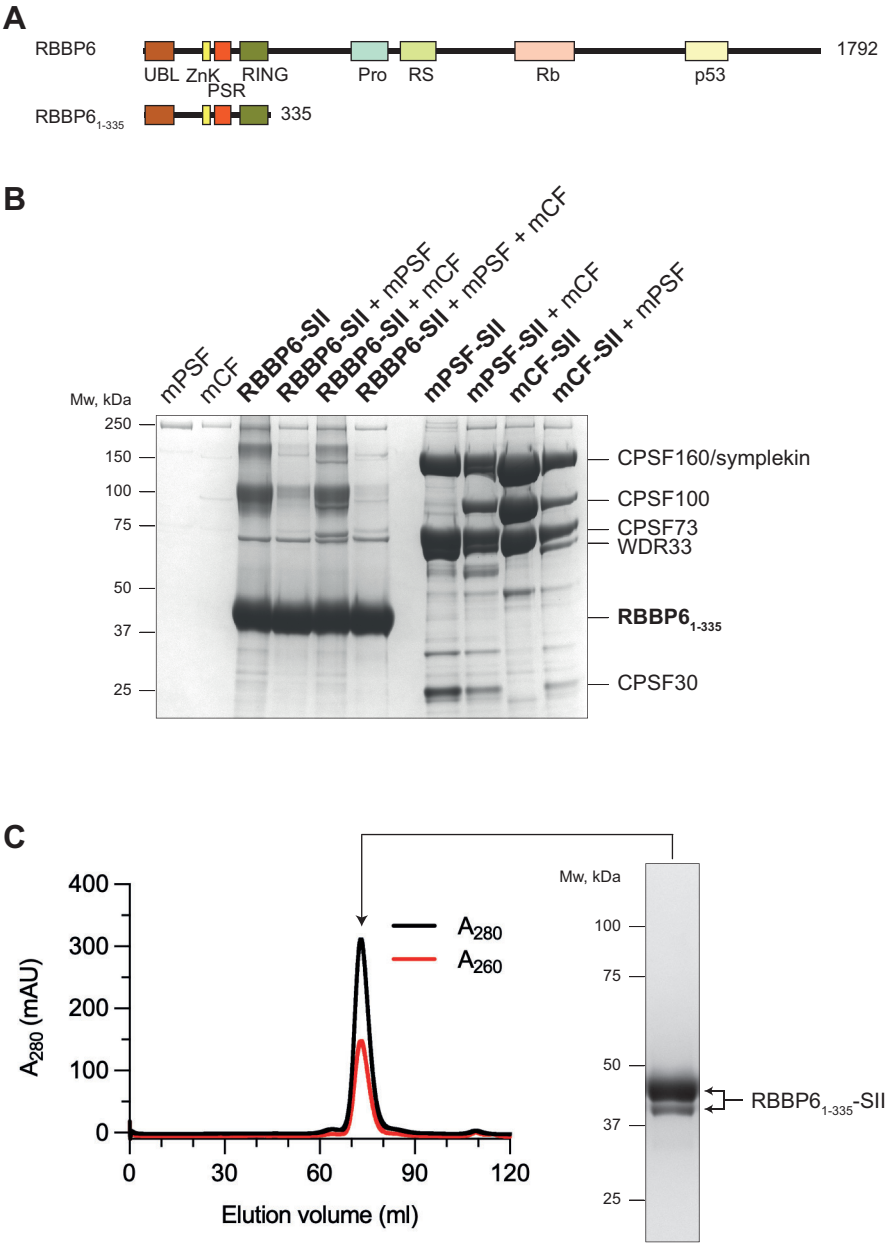


Figure 2.10 Purification of recombinant human RBBP6. (A) Domain diagram of human RBBP6. The number of residues is indicated to the right. The residue boundaries of the RBBP6$_{1\text{-}335}$ construct are marked with a black line. UBL – ubiquitin-like domain; ZnK – zinc knuckle; PSR – pre-mRNA-sensing region; Pro – proline-rich region; RS – arginine-serine-rich domain; Rb – Rb-interacting motif; p53 – p53-interacting motif. **(B)** Pull-downs on Strep-Tactin beads using lysates from Sf9 insect cells co-infected with tagged RBBP6$_{1\text{-}335}$ and untagged modules of CPSF (left); either tagged mCF or mPSF co-expressed with untagged mPSF or mCF, respectively, to show the expression levels of untagged CPSF modules (right). **(C)** Size exclusion chromatogram of RBBP6$_{1\text{-}335}$ run on a Superdex 200 16/600 column and SDS-PAGE analysis of the peak fraction. The two bands of the doublet are indicated with arrows.

2.2.6 RBBP6 stimulates PAP activity independently of CPSF

A recent study that aimed to computationally predict all the binary protein-protein interactions in the budding yeast proteome included a structural model of Mpe1 bound to poly(A) polymerase (33, 40). The residues at this putative binding interface appeared to be highly conserved, and a model of an orthologous human complex between RBBP6 and PAP could be predicted with high confidence using AlphaFold (Figure 2.11A; Appendix Figure 8.3A) (124). Specifically, the predictions of both human and yeast complexes suggest that a helix of RBBP6/Mpe1 located between UBL and zinc knuckle domains (residues 109-147 of RBBP6) may bind PAP. Thus, I aimed to test if RBBP6 had any effect on the polyadenylation activity by recombinant PAP. I titrated increasing concentrations of RBBP6 into polyadenylation reactions containing either PAP alone; PAP and mPSF; or PAP and CPSF. Interestingly, RBBP6 stimulated the activity of the PAP enzyme alone, independent of mPSF or CPSF (Figure 2.11B). When either mPSF or CPSF was present in the assay, two populations of reaction products appeared (Figure 2.11C&D). A highly polyadenylated species likely represented processive polyadenylation by PAP bound to either mPSF or CPSF. A second set of shorter reaction products were polyadenylated at about the same rate as in the absence of mPSF or CPSF, most likely corresponding to polyadenylation by free PAP (Figure 2.11C&D).

A direct physical interaction between RBBP6 and PAP could explain how RBBP6 could stimulate polyadenylation independent of mPSF or CPSF. PAP has a relatively low affinity for RNA (125, 126). However, if PAP binds to RBBP6 which can also interact with the polyadenylation substrate, the combined RNA binding interface of the two proteins may enable more efficient PAP recruitment to its substrate in the absence of CPSF. Alternatively, RBBP6 may activate PAP by an allosteric mechanism. Further studies will be required to test these possibilities, and whether the effect of RBBP6 on polyadenylation is relevant in the context of the full pre-mRNA 3′ end processing machinery.

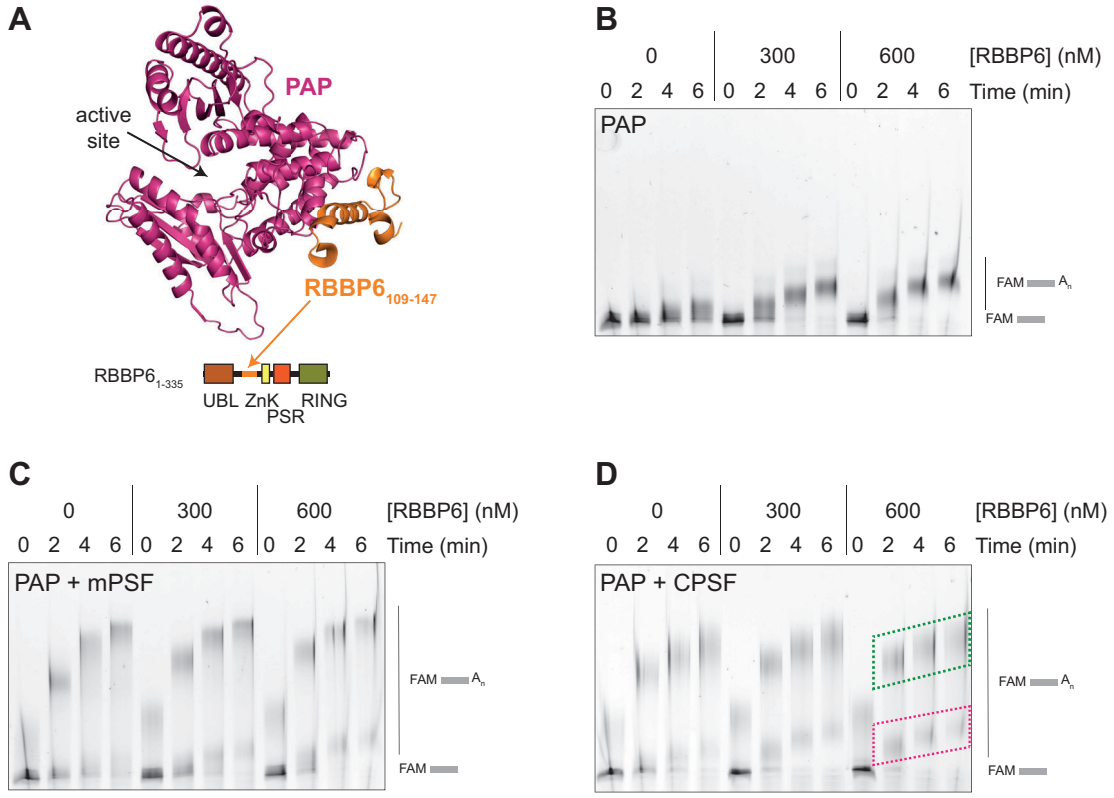


Figure 2.11 RBBP6 stimulates polyadenylation by PAP independent of mPSF or CPSF. (A) Structural model generated using AlphaFold Multimer of the complex between human PAP (pink) and RBBP6 (orange). PAP active site is indicated with an arrow. The position of the helix of RBBP6 that is predicted to interact with the folded domain of PAP is indicated in the domain diagram of RBBP6 with an orange arrow. **(B)** Polyadenylation assays containing 50 nM PAP, 2 mM ATP, 300 nM 41 nt L3 RNA substrate and various concentrations of RBBP6. **(C)** Polyadenylation assays containing 50 nM PAP, 50 nM mPSF, 2 mM ATP, 300 nM 41 nt L3 RNA substrate and various concentrations of RBBP6. **(D)** Polyadenylation assays containing 50 nM PAP, 50 nM CPSF, 2 mM ATP, 300 nM 41 nt L3 RNA substrate and various concentrations of RBBP6. The bands likely corresponding to polyadenylation products by PAP bound to mPSF/CPSF is indicated by a green box, while the products of free PAP are marked with a pink box.

2.3 Conclusions and perspectives

2.3.1 Production of human 3′ end processing factors requires deletions of several IDRs

In this Chapter, I have provided an overview of my strategies to express and purify human protein factors implicated in pre-mRNA 3′ end processing. Removal of IDRs has been a common solution for improving the expression levels and solubility of protein complexes. I rationalised that the deleted IDRs should not affect the activity of the purified protein complexes in cleavage and polyadenylation reactions. The removed IDRs tend to be specific to higher eukaryotes and are often absent from their yeast orthologues. This suggests that such IDRs are not involved in the conserved functions of the protein and are unlikely to be required for the reconstitution of the minimal system to study CPSF function. Indeed, the metazoan-specific complex CFIm, containing functionally important IDRs, was the most challenging to produce.

IDRs are, nevertheless, likely to play important roles *in vivo*, for example, by interacting with other nuclear factors or by promoting condensate formation (127). For instance, the C-terminal domain of RBBP6 contains short linear motifs that may interact with transcription factors and splicing regulators to coordinate 3′ end processing with other events in gene expression. It is also possible that, even though the proteins lacking IDRs are active in cleavage and/or polyadenylation, the IDR could still be needed to modulate the activity by accessory proteins. For instance, although mPSF-hFip1$_4$ is active in polyadenylation, isoform 4 of hFip1 lacks an RD/E domain that interacts with CFIm, which may regulate polyadenylation activity. To study the function of IDRs with purified proteins *in vitro*, strategies to purify intact human proteins should be developed. For instance, adding a folded protein tag to an end of an IDR may both improve the solubility of the protein and protect it from proteolysis.

2.3.2 Covalent modifications may affect cleavage and polyadenylation activities of purified proteins

Many 3′ end processing factors carry covalent modifications in human cells (128). The post-translational modification state of the proteins purified here is largely unknown but may have an impact on the reconstituted cleavage and polyadenylation activity. For instance,

global dephosphorylation of the 3′ end processing reaction in nuclear extract has been shown to inhibit CPSF cleavage activity (129). All the human proteins described in this Thesis were recombinantly expressed in insect cells and may therefore be modified differently than native human proteins. In-depth mass spectrometry analysis will be required to compare the post-translational modification state of recombinantly overexpressed factors with that of the native human proteins in the cell as well as to determine the functional consequences thereof.

2.3.3 Several auxiliary factors regulate polyadenylation by mPSF

In this Chapter, I have also reconstituted the mPSF-dependent polyadenylation activity with purified components and investigated how CStF and RBBP6 may regulate this reaction. Surprisingly, I found that CStF suppresses mPSF-dependent polyadenylation preferentially of a substrate containing a suboptimal PAS sequence. The effect of CStF is strongly dependent on its concentration, and it is possible at high concentrations, CStF may associate with the RNA sequences surrounding the PAS site and compete with CPSF for binding to the substrate, preventing polyadenylation. In addition, while this Thesis was in preparation, a study by Martin Jinek's group showed that CStF directly interacts with an N-terminal α helix of hFip1 (26). This interaction may displace the hFip1-bound PAP away from the 3′ end of the RNA substrate, thereby inhibiting polyadenylation. Substrates containing mutant PAS bind to mPSF with a lower affinity and hence could be more sensitive to the inhibition of polyadenylation by CStF, which may contribute to the fidelity of the second step in 3′ end processing. Unlike CStF, RBBP6 regulates PAP activity independent of mPSF, which is unlikely to represent a physiologically-relevant situation *in vivo*, since PAP does not function independently of CPSF. Nevertheless, if RBBP6 is indeed involved in pre-mRNA 3′ end processing in humans, RBBP6 may still contribute to how polyadenylation is regulated within the CPSF complex.

It is important to emphasise that the polyadenylation activity of the mPSF/CPSF complex has not been studied exhaustively, and many more insights into this reaction could be obtained from the minimal reconstituted system. Nevertheless, in this Study, I was more interested in how the endonuclease of CPSF is activated and used the purified recombinant proteins to study this reaction in detail as described in the subsequent Chapters.

Chapter 3:

In vitro reconstitution of CPSF endonuclease activity with purified proteins

In the previous Chapter, I described the insights into the mechanism of human polyadenylation gained from the experiments using purified recombinant proteins. The *in vitro* polyadenylation reaction with human factors was first reconstituted almost a decade ago and has since proven to be an extremely powerful tool used by many research groups to study polyadenylation (20). A big caveat is that such assays always use a substrate that mimics an already cleaved pre-mRNA substrate. In the cell, however, polyadenylation is always preceded by endonucleolytic cleavage of the pre-mRNA by the CPSF complex.

Reconstituting endonuclease activity by human CPSF has proven to be challenging. Despite many putative activator complexes having been identified, the exact set of proteins required for the activation of CPSF endonuclease has remained elusive. As I described in the previous Chapter, I have purified large amounts of highly pure protein factors that have been implicated in endonuclease activation, including a previously overlooked protein RBBP6. This put me in a great position to attempt to reconstitute the CPSF endonuclease activity *in vitro*. In this Chapter, I will describe how I established an assay with purified proteins to study the cleavage by CPSF and the insights into the mechanism of endonuclease activation gained from this minimal *in vitro* system.

3.1 Determining the set of proteins required for CPSF endonuclease activation

In addition to the recombinant protein factors, I also had to produce a model pre-mRNA substrate in order to study the endonuclease activity by recombinant CPSF. I chose a 218 nt fragment of the simian virus 40 late polyadenylation site (SV40), because it had been shown to be processed efficiently both *in vivo* and in nuclear extract (Table 6.2) (130, 131). The SV40 RNA was prepared by *in vitro* transcription. I also made a 520 nt RNA substrate containing the L3 polyadenylation site from adenovirus 2, which was used for certain experiments (132, 133). The L3 pre-mRNA also had three MS2 loops on its 5′ end, which were employed in pull-down experiments described later in this Dissertation. Which pre-mRNA was used for a particular experiment will be clearly indicated throughout this Thesis. I mixed the model SV40 pre-RNA with CPSF and its putative activators, incubated the mixture for 150 min at 37°C, which is likely to be the optimal temperature for a human enzyme, stopped the reaction and then analysed the samples by denaturing gel electrophoresis. The pre-mRNA substrate was not labelled, and therefore, the gel was stained with SYBR Green. In case of a successful endonuclease reaction, I would observe the appearance of two cleavage product bands, each shorter than the substrate itself (Figure

3.1A). To focus on the endonuclease activity, PAP and ATP were omitted from the reaction, so the 5′ cleavage product could not be polyadenylated.

In activity assays, I mixed the SV40 pre-mRNA with CPSF alone but failed to observe any cleavage activity (Figure 3.1B). This was consistent with the idea that the CPSF endonuclease has to be specifically activated by auxiliary protein factors. CStF and CFIIm were determined to be essential for activation of endonucleolytic cleavage in mammalian nuclear extract, and therefore, I added these two complexes into CPSF cleavage assays (134, 135). Addition of either CStF or CFIIm did not stimulate CPSF and neither did the two cleavage factors combined (Figure 3.1B). This result was rather puzzling, because it contradicted many published studies. It prompted me to consider whether any of the other proteins implicated in 3′ end processing in humans that I had purified – PAP, CFIm or RBBP6 – could be essential for CPSF endonuclease activation. Cleavage and polyadenylation can be uncoupled in nuclear extract, suggesting that the two steps in 3′ end processing are mechanistically independent and that PAP is unlikely to be required for endonuclease activation (136). CFIm may bind the SV40 pre-mRNA, which contains an upstream UGUA motif, but the CFIm complex has been consistently shown to be a regulator rather than an essential activator of CPSF cleavage (70). In contrast, the yeast orthologue of RBBP6, Mpe1, is absolutely essential for endonuclease activity by the yeast machinery, making RBBP6 the most probable candidate for the missing activator of CPSF cleavage. To test this hypothesis, I added recombinant RBBP6 into a cleavage assay (Figure 3.1C). Addition of RBBP6 alone to the CPSF assay already resulted in some weak endonuclease activity. Addition of either CStF or CFIm did not provide any additional stimulation, but combining CStF, CFIIm and RBBP6 lead efficient cleavage of the SV40 pre-mRNA substrate. The product bands did not appear when CPSF was omitted from the complete reaction, which confirmed that CPSF was responsible for the observed cleavage activity (Figure 3.1C). Overall, I concluded that the activation of CPSF endonuclease requires three additional protein factors: CStF, CFIIm and RBBP6 (Figure 3.1D). The role of RBBP6 in activating the endonuclease was particularly intriguing and will be investigated in more detail in Chapter 4. The cleavage assays described in the rest of this Thesis were performed in the presence of all four protein factors, unless indicated otherwise.

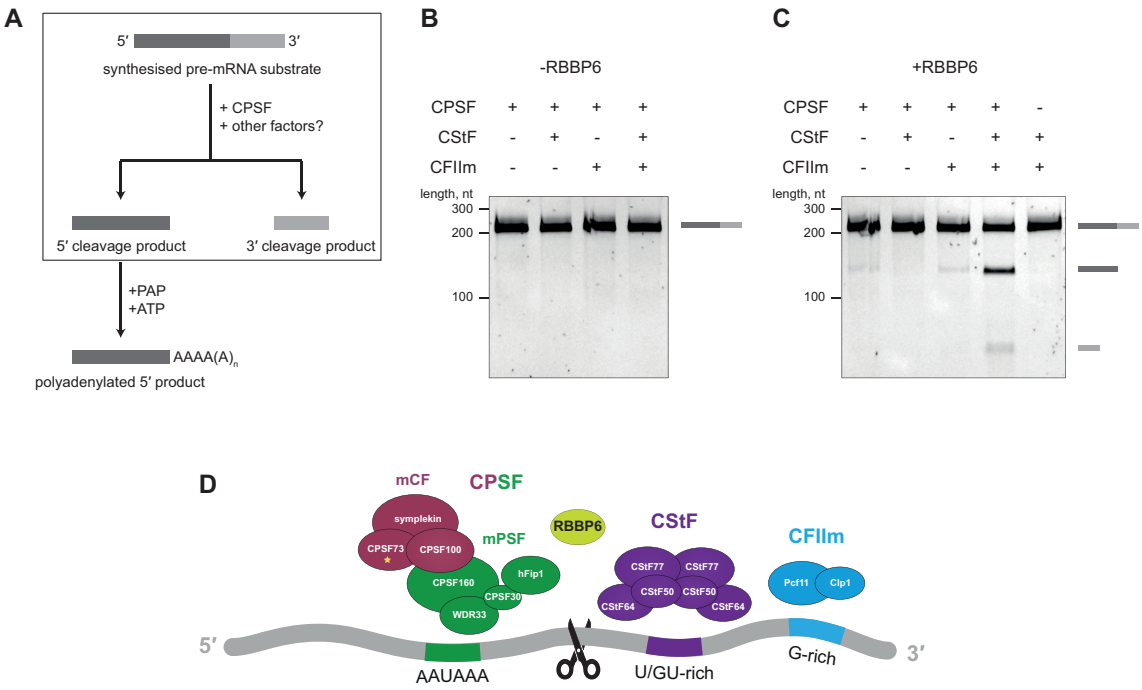


Figure 3.1 CStF, CFIIm and RBBP6 are required to activate CPSF endonuclease. (A) Schematic representation of the pre-mRNA 3′ end processing assay. The cleavage step investigated here is boxed out. **(B)** Denaturing gel electrophoresis analysis of end-point CPSF cleavage assays using the SV40 pre-mRNA substrate with various combinations of CStF and CFIIm in the absence of RBBP6. **(C)** Denaturing gel electrophoresis analysis of end-point CPSF cleavage assays using the SV40 pre-mRNA substrate with various combinations of CStF and CFIIm in the presence of RBBP6. **(D)** Schematic depiction of the proteins and cis-regulatory elements required for CPSF cleavage activity. Cleavage site is marked with a scissor symbol.

3.2 Optimising conditions of recombinant CPSF endonuclease assay

Having determined the proteins required for CPSF endonuclease activation, I aimed to optimise reaction conditions in order to attain as high a rate of substrate cleavage as possible *in vitro*. Determining the optimal conditions of the reaction may also provide some insights into the mechanism of endonucleolytic cleavage. I investigated the effects of both assay buffer conditions and protein concentrations on the endonuclease activity of recombinant CPSF.

3.2.1 Optimising buffer conditions

Early in the project, I tested the endonuclease activity of recombinant CPSF in the presence of CStF, CFIIm and RBBP6 taken from the peak fractions of their corresponding anion exchange chromatography runs but, surprisingly, did not detect any cleavage activity. Both CPSF modules and the three auxiliary factors eluted from the anion exchange column at the salt concentration of >350 mM. In contrast, the proteins that enabled successful reconstitution of endonuclease activity were also purified by gel filtration chromatography and were present in buffers containing 150-200 mM NaCl. Thus, I suspected that the final salt concentration in the assay could affect CPSF cleavage activity. I performed cleavage assays in the presence of various concentrations of NaCl and indeed observed dose-dependent inhibition of endonucleolytic cleavage, as the salt concentration was increased (Figure 3.2A). As a result, all subsequent assays were performed at the final NaCl concentration of 50 mM. I hypothesise that high ionic strength may disrupt certain protein-protein or protein-RNA interactions required to activate the endonuclease.

CPSF73 is a zinc-dependent endonuclease. Thus, I predicted that supplementing the assay buffer with zinc ions would stimulate its activity. However, titrating increasing concentrations of zinc acetate into the cleavage reaction did not increase the cleavage activity by CPSF (Figure 3.2B). This observation is consistent with previous studies showing that the affinity of CPSF73 for zinc ions is high enough for them to bind to the enzyme in the expression host and then remain associated in the active site throughout the purification procedure (38).

Next, I tested the activity of CPSF under various pH conditions. Any buffering agent is sensitive to temperature changes, which is important to consider, since the buffers were

prepared at room temperature (~21°C) but used for cleavage assays at 37°C. I prepared several buffers of different pH values as measured under assay conditions at 37°C: 20 mM HEPES-NaOH at 7.0, 7.5, 8.0, and 20 mM CHES-HCl at pH 8.5. Time-course cleavage assays using the 520 nt L3 pre-mRNA substrate were performed in these buffers of varying pH. CPSF appeared to be most active at a pH value of 7.0 (Figure 3.2C). Noticeable non-specific RNA degradation was observed at pH values >8.0, which may indicate non-enzymatic alkaline hydrolysis of RNA. Thus, HEPES at pH 7.0 at 37°C was used as a buffering agent in all future assays.

Several protein components in the cleavage reaction were purified in the presence of magnesium ions. Magnesium salts are thought to stabilise the native conformation of proteins, but I did not test whether they were essential for successful purification of the human 3′ end processing factors (137). Magnesium ions may affect the three-dimensional structure of the pre-mRNA substrate, which could also influence cleavage efficiency (138). I performed CPSF cleavage assays in the presence of various concentrations of magnesium acetate on two different substrates: the SV40 pre-mRNA and the L3 pre-mRNA. Interestingly, the optimal concentration of magnesium acetate differed for the two substrates: the SV40 pre-mRNA was cleaved best at 0.5 mM magnesium acetate, while the L3 pre-mRNA was cut most efficiently in the presence of 4 mM magnesium acetate (Figure 3.2D). These observations suggest that magnesium ions may stabilise a conformation of the L3 pre-mRNA compatible with 3′ end cleavage and destabilise an equivalent conformation of the SV40 pre-mRNA, highlighting the importance of RNA structure to the efficiency of the CPSF endonuclease.

Overall, optimisation experiments allowed me to determine buffer conditions in which the CPSF endonuclease is the most active. The final assay buffer, after accounting for buffer carry-over from the protein stocks, contained 20 mM HEPES-NaOH (pH 7.0 at 37°C), 50 mM NaCl, 0.5 mM TCEP and 2 mM magnesium acetate.

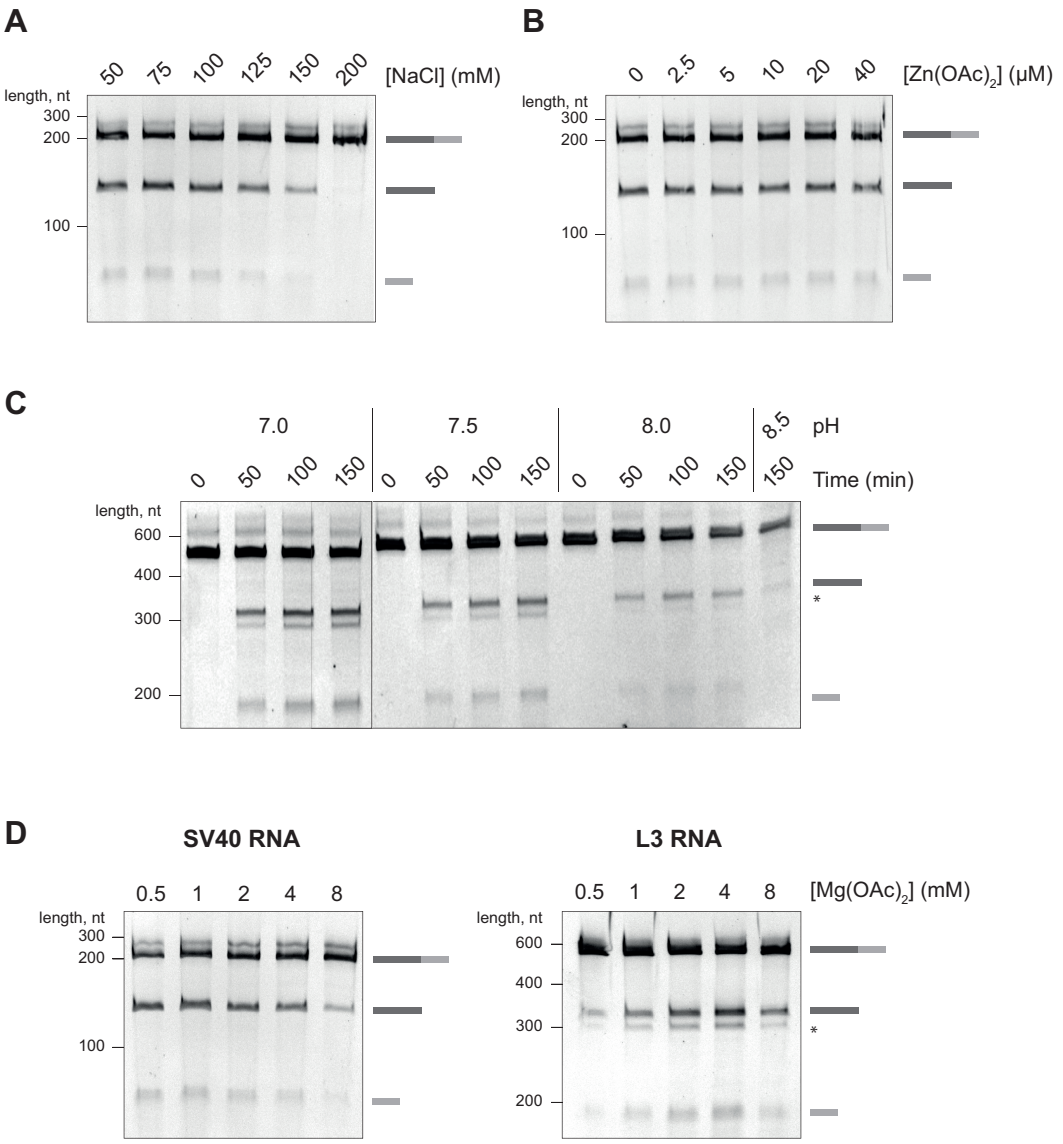


Figure 3.2 Optimising buffer conditions of the recombinant CPSF cleavage assay. (A) Cleavage assays of the SV40 pre-mRNA substrate in the presence of various salt concentrations. **(B)** Cleavage assays of the SV40 pre-mRNA substrate in the presence of various concentrations of zinc acetate. **(C)** Cleavage assays of the L3 pre-mRNA substrate performed in buffers of various pH values. **(D)** Cleavage assays of the SV40 pre-mRNA substrate (left) and of the L3 pre-mRNA (right) in the presence of various concentrations of magnesium acetate. A minor cleavage product of the L3 substrate, which is discussed in more detail later, is marked with an asterisk.

3.2.2 Chemical additives are not required for CPSF endonuclease activity

Prior to this work, the cleavage activity by CPSF has been primarily studied using assays containing partially purified fractions of mammalian nuclear extract. The endonuclease efficiency in such experiments is strongly dependent on the presence of chemical additives, in particular, creatine phosphate and polyvinyl alcohol (PVA) (139, 140). Creatine phosphate is important for cellular metabolism, but its concentration in the nucleus where CPSF acts is unlikely to be significant. In nuclear extract, creatine phosphate facilitates regeneration of ATP and has also been hypothesised to either facilitate certain protein-protein interactions or mimic an unidentified activator phosphoprotein (140). PVA, on the other hand, is a bio-orthogonal polymeric molecule that could act as a crowding agent in the cleavage reaction, promoting assembly of the active 3′ end processing complex and may also affect the conformation of IDRs (141). I tested both creatine phosphate and PVA in a CPSF cleavage assay with purified components (Figure 3.3). Neither compound stimulated the endonuclease activity by CPSF, suggesting that chemical additives are not required for cleavage reconstituted with purified proteins. The dependence on creatine phosphate and PVA in the nuclear extract assay could be linked to the presence of unknown protein factors or small molecules that are carried over during subcellular fractionation. This highlights the advantage of a minimal assay system with highly purified components, in which both protein content and buffer composition are well defined.

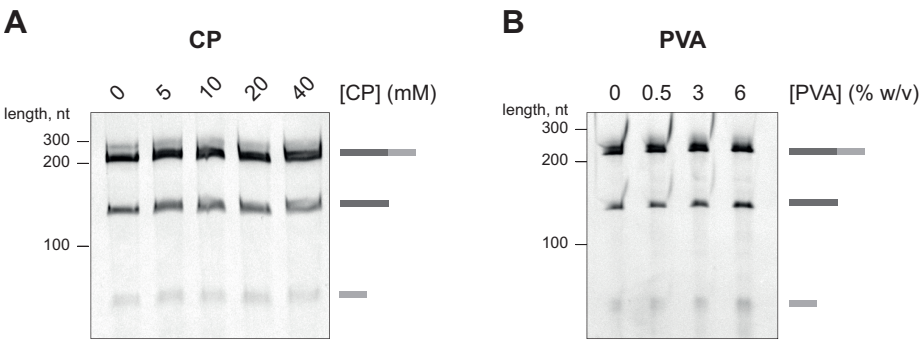


Figure 3.3 Chemical additives are not required for endonuclease activity of recombinant CPSF. Cleavage assays of the SV40 pre-mRNA substrate in the presence of various concentrations of **(A)** creatine phosphate (CP) or **(B)** PVA.

3.2.3 Optimising protein concentrations

The reconstituted CPSF endonuclease reaction contains four different protein components. Their absolute concentrations as well as their relative molar ratios could affect cleavage efficiency. I kept the concentration of the pre-mRNA constant at 100 nM, which enabled clear visualisation by denaturing PAGE, and varied the concentrations of the protein components. Starting from 50 nM CPSF and 100 nM of each CStF and CFIIm, I titrated RBBP6 into a cleavage assay. RBBP6 stimulated the CPSF endonuclease in a dose-dependent manner reaching maximal stimulation at ~300 nM, and this concentration of RBBP6 was used in the subsequent assays (Figure 3.4A). At this point, either CPSF, CStF or CFIIm were rate limiting, and hence, I subsequently tested endonuclease activity in the presence of various concentrations of these three factors. Surprisingly, the starting concentration of 100 nM of CStF and CFIIm lead to the highest cleavage activity, while increasing their concentrations inhibited the reaction (Figure 3.4B). It is likely that excess cleavage factors bind RNA non-specifically and block productive assembly of the 3′ end processing machinery on the substrate. It is also possible that CStF and CFIIm may aggregate at higher concentrations under low salt conditions of the assay. In addition, doubling the concentration of CPSF from 50 nM to 100 nM unexpectedly did not increase endonuclease activity either (Figure 3.4B). In summary, at least six-fold molar excess of RBBP6 over CPSF is required to achieve the optimal cleavage efficiency, and RBBP6 is the rate-limiting component of the reaction under most conditions. Even under the most optimal conditions (100 nM pre-mRNA, 50 nM CPSF, 100 nM CStF, 100 nM CFIIm, 300 nM RBBP6) only ~30% of the substrate gets endonucleolytically cleaved by CPSF, but what limits its rate then is unclear.

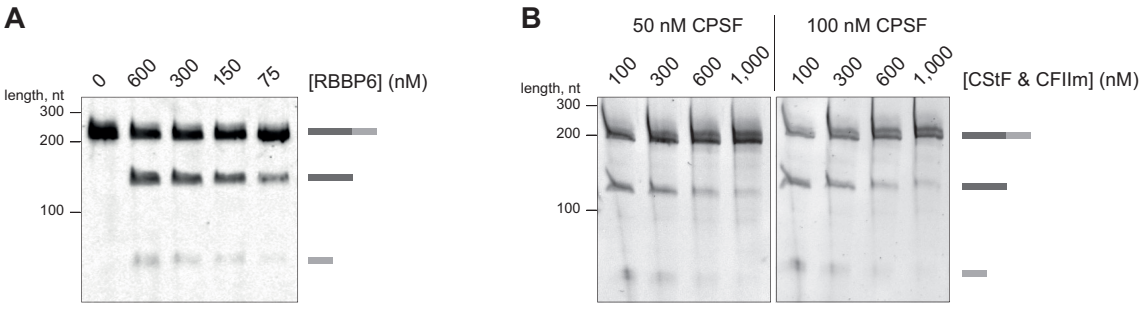


Figure 3.4 Optimising protein concentrations in the recombinant CPSF cleavage assay. (A) Cleavage assays of the SV40 pre-mRNA substrate in the presence of 50 nM CPSF, 100 nM CStF and CFIIm each and a range of concentrations of RBBP6. **(B)** Cleavage assays of the SV40 pre-mRNA substrate in the presence of 300 nM RBBP6 and various concentrations of CPSF, CStF and CFIIm.

3.3 Understanding the specificity of recombinant CPSF endonuclease activity

In the human nucleus, CPSF becomes activated only at specific sites on the pre-mRNA, making 3′ endonucleolytic cleavage highly specific. This both prevents non-specific degradation of RNAs in the nucleus and ensures that mature transcripts of a correct length are produced. The specificity of CPSF could stem from the intrinsic sequence specificity of the endonuclease enzyme as well as the RNA binding specificity of both CPSF and the auxiliary factors. The assay reconstituted with purified proteins is an excellent tool to better understand sequence specificity of CPSF cleavage in the absence of other nuclear proteins that could influence its specificity and whose concentrations are difficult to control *in vivo* or in nuclear extract. Therefore, in this section, I will investigate the specificity of recombinant CPSF using an *in vitro* assay with purified proteins.

3.3.1 Recombinant CPSF endonuclease exhibits sequence specificity

3′ end cleavage sites in the cell are typically marked by a hexameric PAS sequence recognised by the mPSF module. Indeed, purified mCF module failed to cleave RNA in the absence of mPSF, even when CStF, CFIIm and RBBP6 were present (Appendix Figure 8.4). To test whether recombinant CPSF also has specificity for the PAS *in vitro*, I mutated the PAS sequence in the context of the SV40 pre-mRNA substrate from AAUAAA to AACAAA. While the base of the third PAS nucleotide does not contact CPSF, C3 cannot form a Hoogsteen base pair with A6, which results in the AACAAA sequence having an affinity for CPSF several orders of magnitude lower than the wild-type PAS (Figure 1.2C) (29). This mutation reduced the cleavage efficiency by ~80%, confirming that the endonuclease activity of recombinant CPSF is dependent on the presence of PAS (Figure 3.5A).

By recruiting CPSF to the pre-mRNA, PAS may dictate the precise site at which the complex cleaves the substrate. I aimed to determine the precise cleavage site of the SV40 pre-mRNA by recombinant CPSF (Figure 3.5B). To this end, I performed a cleavage assay and purified the 5′ cleavage product. I ligated a DNA adaptor of a known sequence to its 3′ end, carried out reverse transcription and PCR amplification. The resultant PCR products were then sequenced by Sanger sequencing. The last nucleotide corresponding to the sequence of the substrate upstream of the adaptor indicated the cleavage site by CPSF. Out of the 15 cleavage products sequenced, 13 were cleaved 13 nucleotides downstream of the PAS at the CA|A motif, where | corresponds to the cleavage site (Figure 3.5C). This result was

consistent with both the known sequence specificity of 3′ end processing endonucleases and the observation that the cleavage site in human pre-mRNAs *in vivo* is typically located 10-30 nucleotides downstream of the PAS (12, 28). Importantly, the SV40 pre-mRNA was cleaved by CPSF at the exact same site in nuclear extract (131). Therefore, recombinant CPSF recapitulates the same sequence specificity in terms of its cleavage activity as the endogenous complex.

Interestingly, the SV40 pre-mRNA contains three consecutive CAA motifs, but why only one of them is used preferentially by CPSF remains unclear, demonstrating our limited understanding of what determines the site of CPSF cleavage (Figure 3.5C). To begin to address this question, I created a library of pre-mRNAs based on the SV40 substrate, in which the sequence between the PAS and the cleavage site, including all three CAA motifs, was randomised. I performed a cleavage assay using this substrate library and then sequenced the 5′ cleavage products to determine the optimal sequence and position for CPSF cleavage. For preliminary analysis, I sequenced six cleavage products, five of which were cleaved after a CA dinucleotide (Figure 3.5D). In two cases, the pre-mRNA was cleaved at a CA|U motif located 22 nucleotides downstream of the PAS in the region of the substrate that was not randomised. Interestingly, both RNAs did contain two more CAA motifs but they were not used. In the other three cases, however, the substrate was cleaved 13 nucleotides downstream of the PAS after a CA nucleotide, which is highly reminiscent of the features of the original SV40 pre-mRNA. Although these results are only preliminary, the presence of a CA dinucleotide appears to be the key determinant of the CPSF cleavage site, as long as it is located ~13-22 nucleotides downstream of the PAS. Interestingly, this observation is consistent with known genetic polymorphisms associated with human disease caused by either impaired or overly efficient usage of 3′ cleavage sites (142). However, sequencing of many more cleavage products will be required to reach definitive conclusions. This may also reveal if the intervening sequence between the PAS and the cleavage site could be important for CPSF specificity, which could not be decided based on the limited data.

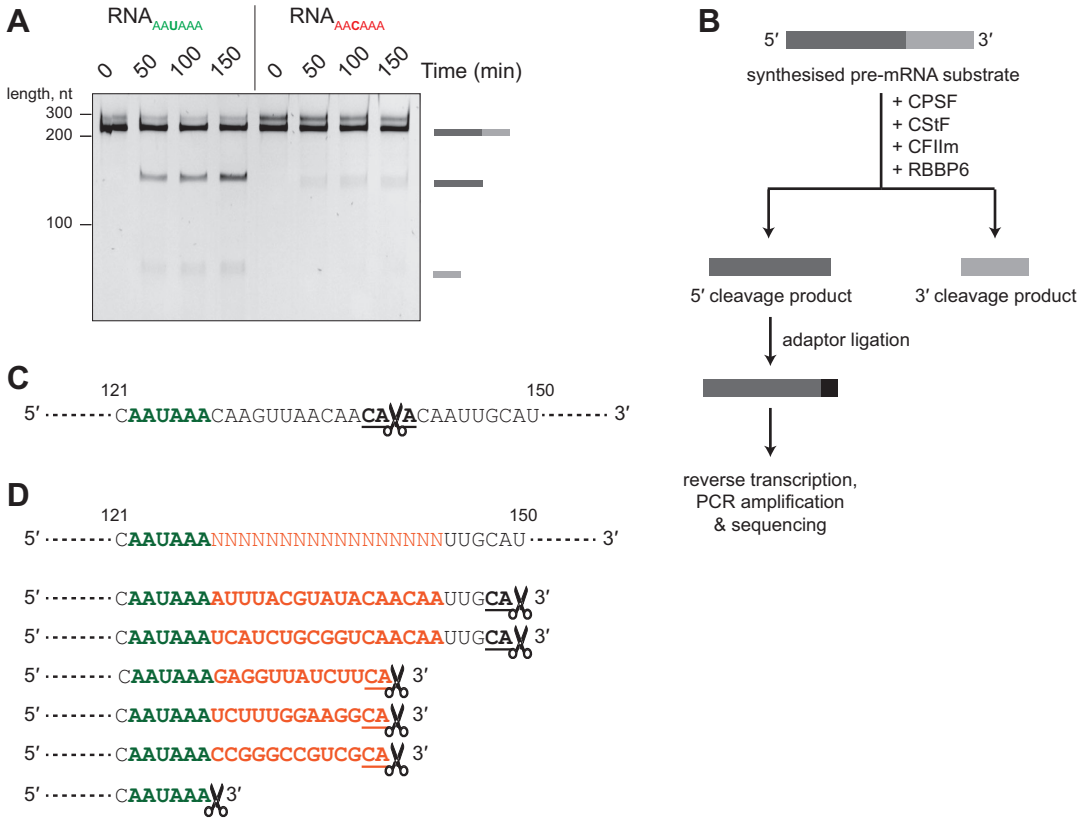


Figure 3.5 Recombinant CPSF endonuclease is sequence specific. (A) Time-course CPSF cleavage assay using either wild-type SV40 pre-mRNA (RNA_{AAUAAA}) or the SV40 pre-mRNA containing a mutant PAS (RNA_{AACAAA}). **(B)** Schematic representation of the protocol used to determine the precise site of CPSF cleavage. **(C)** Part of the SV40 pre-mRNA sequence indicating the determined cleavage site (scissors). PAS is shown in green, CA|A motif – in bold and underlined. **(D)** Sequence of the substrate of the randomised library (top) and of the six 5′ cleavage products sequenced (bottom). Cleavage site is indicated by a scissor symbol, PAS is shown in green, randomised sequence - in orange, CA motif – in bold and underlined. N – random nucleotide.

3.3.2 Recombinant CPSF can cleave multiple substrates

While the CPSF endonuclease activity is highly specific, the 3′ end processing machinery has to be able to accommodate a large variety of pre-mRNA substrates in the human transcriptome. Therefore, I tested whether CPSF could cleave substrates other than the SV40 pre-mRNA. First, I added a 520 nt adenoviral L3 pre-mRNA into a CPSF endonuclease reaction. The L3 substrate was cleaved efficiently, demonstrating that recombinant CPSF can catalyse endonucleolytic cleavage of multiple pre-mRNAs (Figure 3.6A). Notably, two 5′ cleavage products were visible by gel electrophoresis analysis, which suggested that the L3 pre-mRNA was cleaved at two distinct sites. This has not been reported in the literature so far (133). I attempted to determine the 3′ ends of both 5′ cleavage products. The major product appeared to be cleaved 21 nucleotides downstream of the PAS (Figure 3.6B). The cleavage site of the minor product could not be determined unambiguously due to variability in length of the purified cleavage products. Nevertheless, sequencing results suggested that the minor cleavage site was most likely located just upstream of the AAUAAA motif.

The L3 pre-mRNA contains an A/U-rich segment which resembles a PAS ~18 nucleotides upstream of the major PAS (Figure 3.6B). The location of the A/U-rich segment is consistent with the length difference between the major and the minor 5′ cleavage products, and thus, I hypothesised that CPSF binding to this upstream region could be responsible for the minor cleavage event. The AAUAAU sequence has been shown to retain some polyadenylation activity, and I suspected that this particular hexamer may act as the minor PAS (30). To test this, I prepared a series of mutants of the L3 substrate, carrying sequence variations in either the major PAS or the putative minor PAS (Figure 3.6C). Mutating the true PAS to the AACAAA sequence led to the minor cleavage product becoming dominant, suggesting that only the minor PAS was used. Interestingly, when the minor PAS was converted to a canonical AAUAAA sequence, even in the presence of the major wild-type PAS, only the minor PAS appeared to be used. This suggests that the assembly of the 3′ end processing machinery on the minor PAS is more conducive to endonucleolytic cleavage. Since the minor cleavage event has not been reported from *in vivo* studies, it likely represents the usage of a cryptic polyadenylation site. Recombinant CPSF may bind to this cryptic site because of the relatively high protein and RNA concentrations in the assay, which may promote CPSF binding to low affinity sites. In addition, nuclear extract may contain proteins that are missing in the *in vitro* assay but suppress the usage of cryptic polyadenylation sites *in vivo*.

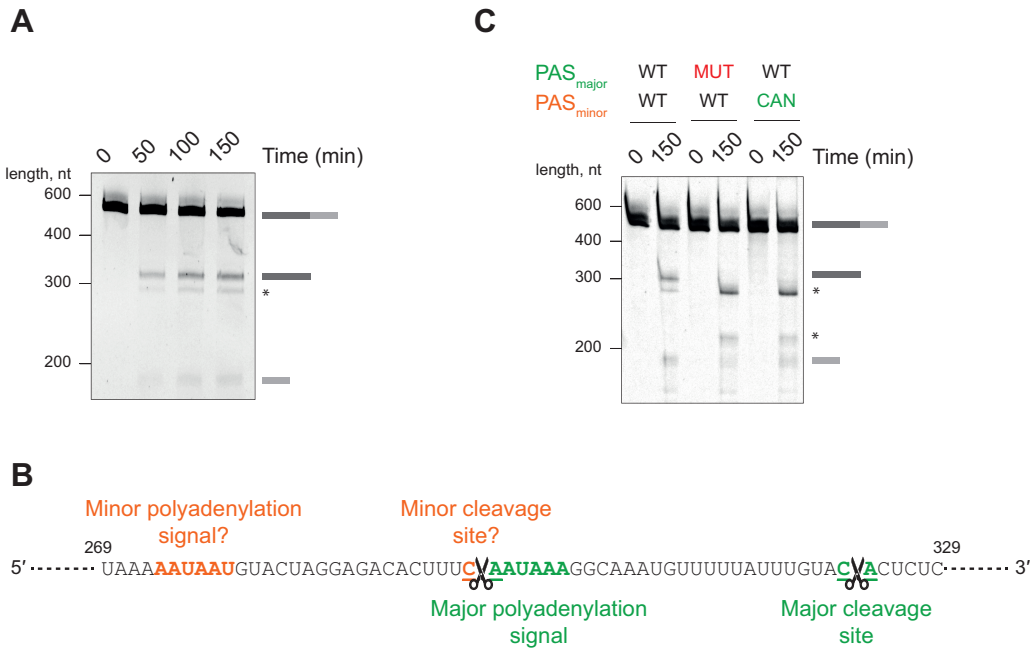


Figure 3.6 Recombinant CPSF can cleave multiple pre-mRNA substrates. (A) Time-course CPSF cleavage assay of the 520 nt L3 pre-mRNA substrate. The two 5′ cleavage products, ~20 nt apart in size, can be observed. Only one 3′ cleavage product can be seen. **(B)** Part of the sequence of the 520 nt L3 pre-mRNA substrate focusing on the two PAS sites and the two cleavage sites. Major PAS is shown in green, an A/U-rich upstream region that may represent the minor PAS – in orange. The cleavage sites are coloured and marked with scissor symbols. **(C)** CPSF cleavage assays of the 520 nt L3 pre-mRNA substrates containing variations of both major and minor PAS sequences. WT – sequence originally found in the L3 pre-mRNA; CAN – canonical AAUAAA sequence; MUT – mutant AACAAA sequence. Key results discussed in the text are boxed out.

I tested several other pre-mRNA substrates in an assay with recombinant CPSF, including a 60 nt synthetic polyadenylation site based on the rabbit β-globin pre-mRNA that is processed with high efficiency in nuclear extract (Table 6.2) (143). Interestingly, recombinant CPSF failed to cleave this substrate (Figure 3.7A). However, I serendipitously discovered that supplementing the assay buffer with 6% w/v PEG 3350 stimulated CPSF and enabled efficient cleavage of the synthetic 60 nt pre-mRNA (Figure 3.7A). Thus, although chemical additives are not strictly required for the endonuclease activity by recombinant CPSF (Figure 3.3), cleavage of some pre-RNAs might be facilitated by crowding agents. Since the 60 nt polyadenylation site contains all the known cis-regulatory elements required for cleavage, it appeared likely that its short length rather than its sequence may confer dependence on PEG 3350. In agreement with this hypothesis, cleavage of the L3 pre-mRNA shortened to 98 nt but retaining the known binding sites for CPSF and the auxiliary factors was also only observed in the presence of a crowding agent (Figure 3.7B).

I predict that a longer substrate (>200 nt) may increase the probability of the 3′ end processing factors encountering the pre-mRNA, likely non-specifically at first, and eventually assembling into a complex competent for cleavage. On the other hand, in case of a short pre-mRNA (<100 nt), a crowding agent may be required to facilitate the encounter between protein factors and the substrate. *In vivo*, 3′ end processing factors associate with a transcribing pol II already at promoters and are likely to be positioned close to the RNA exit channel at all times during transcription elongation (51, 110). Therefore, the assembly of the active endonuclease complex should not depend on the length of the nascent transcript *in vivo*.

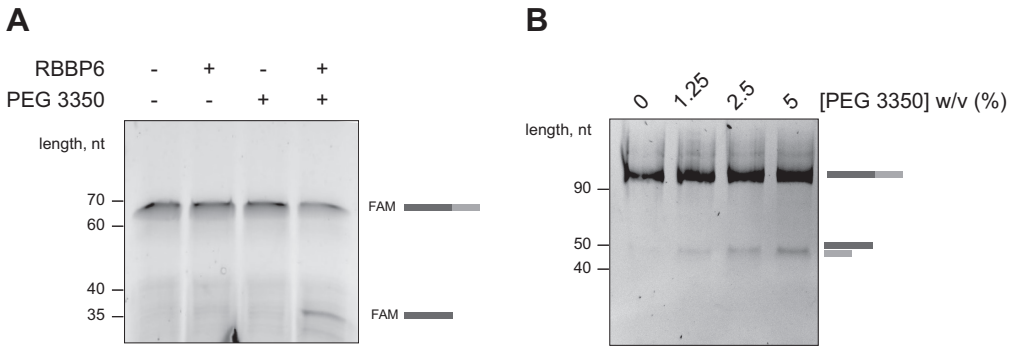


Figure 3.7 Cleavage efficiency of recombinant CPSF is dependent on RNA length. (A) CPSF cleavage assays of the synthetic 60 nt polyadenylation site. The substrate is labelled on its 5′ end with a FAM fluorophore. Only the 5′ product is visible in the fluorescence scan. **(B)** CPSF cleavage assays of the 98 nt L3 pre-mRNA substrate in the presence of various concentrations of PEG 3350. Assuming that the shortened pre-RNA gets cleaved at the same site as the 520 nt L3 substrate, the 98 nt pre-mRNA is cut into two products of the same size.

3.3.3 CPSF73 is the only endonuclease within the CPSF complex

Although eukaryotic pre-mRNA 3′ end processing has been studied since the early 1970s, the actual enzymatic subunit that cleaves the RNA was unambiguously identified only about 15 years ago (38, 144). The search for the endonuclease subunit turned out to be challenging because of the sheer number of protein factors that are required for the 3′ cleavage reaction. The endonuclease CPSF73 was first identified based on its homology to zinc-dependent hydrolase enzymes, and its role was later confirmed by its propensity to cross-link to pre-mRNAs directly at the cleavage site (38, 144). CPSF100 is a homolog of CPSF73 sharing the same domain architecture (Figure 2.3A). Although CPSF100 lacks most of the residues that coordinate the two catalytic zinc ions in the active site of CPSF73, some studies suggested that human CPSF100 could also catalyse endonucleolytic cleavage of RNA (39). Indeed, CPSF100 retains more residues required for catalysis than its yeast orthologue Cft2 (Figure 3.8A). To test if CPSF100 could catalyse endonucleolytic cleavage, I prepared a CPSF complex carrying a mutant version of CPSF73, in which the key zinc-coordinating residues were mutated (CPSF73 D75N H76A) (38, 82). The mutant CPSF complex could be successfully assembled from wild-type mPSF and mCF containing catalytically-dead CPSF73. The mutant CPSF complex was completely inactive in a cleavage assay of the SV40 pre-mRNA substrate (Figure 3.8B). This suggests that CPSF73 is the only active pre-mRNA endonuclease within CPSF, and that CPSF100 cannot catalyse endonucleolytic cleavage, at least in the context of the minimal reconstituted system in the current buffer conditions.

Recent years have witnessed a growing interest in targeting CPSF73 pharmacologically. Various compounds that inhibit CPSF73 and its orthologues have been demonstrated to have anti-inflammatory, anti-cancer and anti-protozoan activities (see Section 1.4.2.1). In particular, human CPSF73 was recently identified as the target of the active form of JTE-607 – a long-known anti-inflammatory drug that has since been shown to be effective at suppressing growth of cell culture models of Ewing's sarcoma, acute myeloid leukaemia and pancreatic ductal adenocarcinoma (93, 97). The crystal structure of the active form of JTE-607 bound to the active site of CPSF73 has been solved, indicating that the compound inhibits the enzyme by competing with binding of the substrate RNA (Figure 3.8C) (93). I aimed to test if JTE-607 could also inhibit the endonuclease activity of recombinant CPSF. The inactive ester form of JTE-607 was purchased from a manufacturer, and the pro-drug was converted into its active acid form by Thomas Elliott from Jason Chin's group (MRC LMB). I titrated increasing concentrations of the acid form of JTE-607 into the cleavage reaction and observed a dose-dependent inhibition of CPSF endonuclease activity (Figure

3.8D). Quantification of the gel-based assay revealed that JTE-607 inhibits CPSF with an IC_{50} value of ~350 nM, which is very similar to the K_d value of the compound for isolated CPSF73 (Figure 3.8E) (93). Recombinant yeast CPF is also inhibited by JTE-607 but only at drug concentrations of ~100 μM, demonstrating the specificity of JTE-607 for the human enzyme. These data not only confirm that CPSF73 is the active endonuclease within CPSF but also suggest that the reconstituted endonuclease assay with purified proteins could be used to develop compounds targeting CPSF73. A high-throughput assay for drug screening was recently developed for the purified histone pre-mRNA 3′ end processing complex (100). A similar assay using, for example, fluorescence read-out to indicate substrate cleavage could also be established for the canonical 3′ end processing complex.

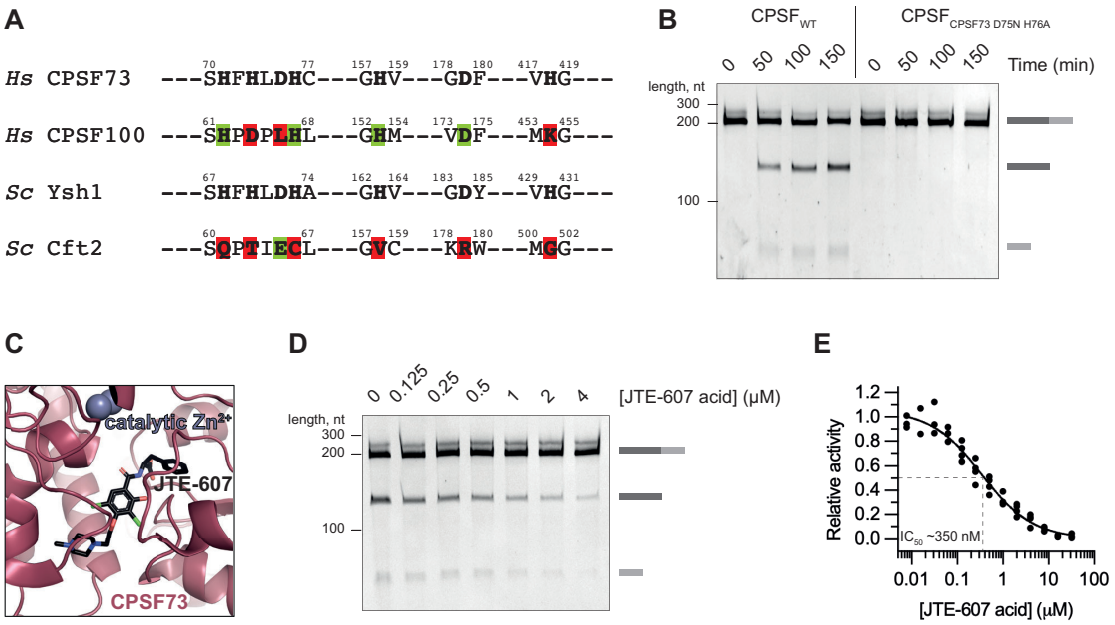


Figure 3.8 CPSF73 is the only active endonuclease within CPSF. (A) Sequence alignments of human CPSF73 and CPSF100 with their yeast orthologues Ysh1 and Cft2, focusing on the regions containing the active site residues that coordinate catalytic zinc ions (indicated in bold). The zinc-coordinating residues that are conserved in the pseudonucleases (CPSF100 and Cft2) are coloured in green, non-conserved catalytic residues – in red. **(B)** Time-course cleavage assays of the SV40 pre-mRNA with either wild-type CPSF ($CPSF_{WT}$) or the CPSF complex containing catalytically-dead CPSF73 ($CPSF_{CPSF73\ D75N\ H76A}$). **(C)** Close-up view of the acid form of JTE-607 bound to the active site of CPSF73 (PDB ID 6M8Q) (93). **(D)** CPSF cleavage assays of the SV40 pre-mRNA in the presence of increasing concentrations of the acid form of JTE-607. **(E)** Dose-response curve demonstrating the relative endonuclease activity of recombinant CPSF in the presence of various concentration of the acid form of JTE-607. Each dot represents a single quantification measurement of a gel-based endonuclease assay. At least three independent measurements were performed for each concentration, but some points overlap.

3.4 NS1 protein from Influenza A virus inhibits pre-mRNA 3′ end processing *in vitro*

3.4.1 NS1 inhibits cleavage activity of recombinant CPSF

The potential clinical relevance of the 3′ endonuclease assay extends beyond testing CPSF73 inhibitors. NS1 protein from Influenza A virus has been shown to inhibit 3′ end processing in infected human cells, and studying the mechanism of this inhibition in a well-controlled system with purified proteins will be useful in developing compounds to disrupt the interactions between the 3′ end processing machinery and NS1 (103, 145). To test whether NS1 could inhibit recombinant CPSF, I aimed to produce the full-length NS1 protein from Influenza A virus H3N2 strain. The plasmid encoding this protein was a kind gift from Loic Carrique and Ervin Fodor (University of Oxford). Full-length NS1 could not be expressed in either *E. coli* or Sf9 insect cells, which was consistent with previous reports, and hence, I introduced two solubility-enhancing mutations in its sequence (146). The resultant full-length NS1 (NS1 R38A K41A) was successfully expressed in *E. coli* and purified in large quantities (Figure 3.9A). I titrated increasing amounts of $NS1_{R38A/K41A}$ into a cleavage assay with recombinant proteins and observed dose-dependent inhibition of CPSF endonuclease activity (Figure 3.9B). Thus, the assay with recombinant proteins recapitulates NS1-mediated inhibition of 3′ end processing.

NS1 from Influenza A contains an RNA-binding domain (RBD), which mediates homodimerisation, and an effector domain (ED), which interacts with a plethora of host proteins and interferes with almost every step in the life of a host mRNA, including 3′ end processing, nuclear export and translation (Figure 3.9C) (147). The ED of NS1 interacts with ZnF2 and ZnF3 of CPSF30 (103, 104). The solubility-enhancing mutations (R38A K41A) are located in the RBD of NS1 and should not affect NS1 binding to CPSF30. The purified ED of NS1 inhibited the endonuclease activity of recombinant CPSF to a similar extent as the full-length NS1 protein, suggesting that NS1-ED is sufficient for inhibition of 3′ end cleavage *in vitro* (Figure 3.9A&B). The crystal structure of the complex between NS1-ED and CPSF30 has been solved, but many questions regarding how NS1 interacts with mPSF remain (Figure 3.9C) (104). First, full-length NS1 is a dimer with two EDs, suggesting that each NS1 could bind to two copies of the mPSF complex. Second, each ED has two interfaces that can bind CPSF30, and it is possible that one ED could interact with two copies of mPSF. Finally, alignment of the crystal structure of the ED-CPSF30 complex and the cryoEM structure of

mPSF shows that NS1 may not only compete with PAS RNA binding to CPSF30 but may also induce a conformational change in the complex, because several loops of WDR33 would clash with the effector domain (Figure 3.9D). To answer these questions, the structure of NS1 bound to the full mPSF module is required.

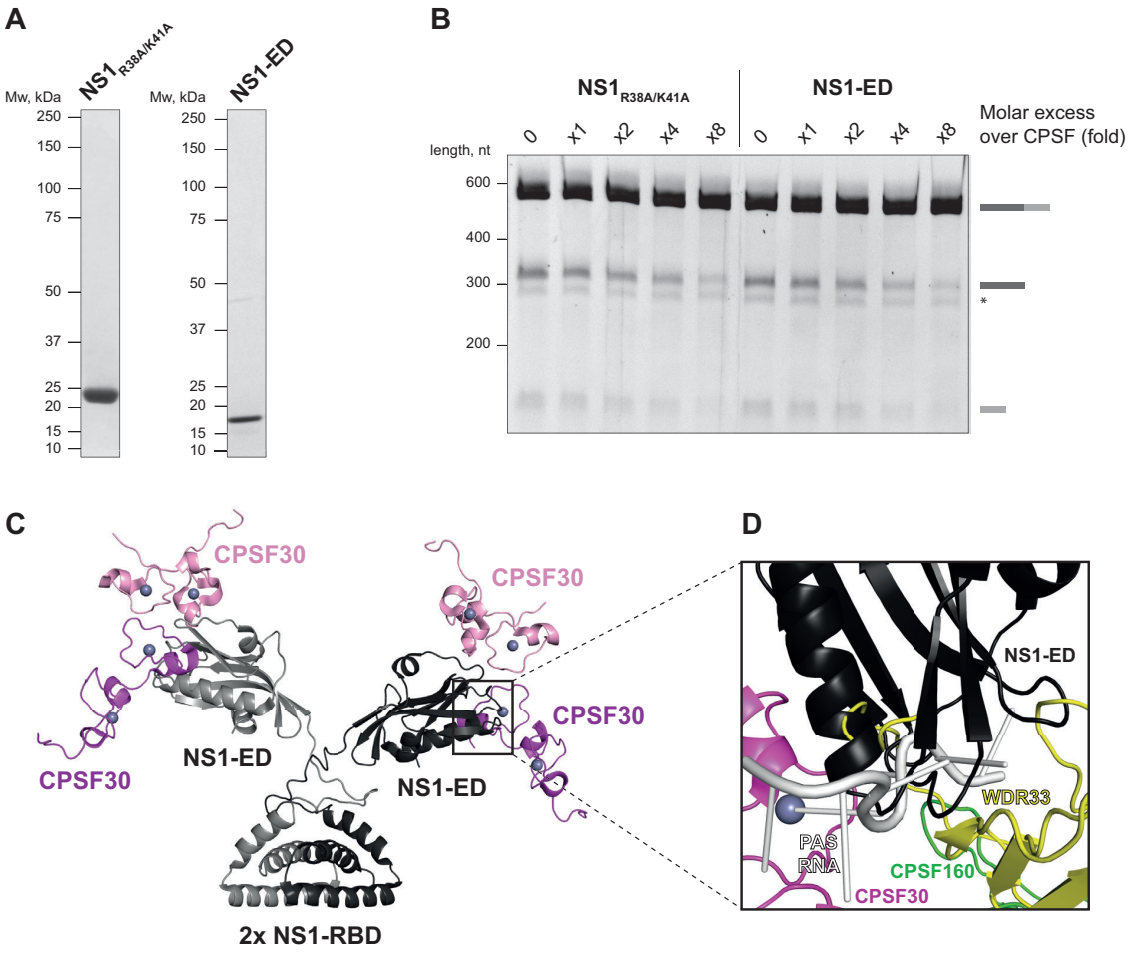


Figure 3.9 NS1 inhibits endonuclease activity of recombinant CPSF. (A) SDS-PAGE analysis of purified recombinant full-length NS1 ($NS1_{R38A/K41A}$) and its effector domain (NS1-ED). **(B)** Cleavage assays of recombinant CPSF (50 nM) using the L3 pre-mRNA substrate in the presence of increasing concentrations of either full-length NS1 or its effector domain. Asterisk denotes the minor 5′ cleavage product. **(C)** Crystal structure of the NS1 effector domain complex with ZnF2 and ZnF3 of CPSF30 (PDB 2RHK) overlayed onto the crystal structure of dimeric full-length NS1 (PDB 5NT2) (104, 148). ED – effector domain; RBD – RNA-binding domain. **(D)** One of the NS1 effector domain-CPSF30 interfaces overlayed onto the cryoEM structure of human mPSF (PDB 6DNH) (21). NS1-ED clashes with both RNA and WDR33.

3.4.2 Structural analysis of the mPSF-NS1 complex

I co-expressed either NS1-ED or full-length NS1 with the human mPSF module in Hi5 insect cells. mPSF bound to either ED or full-length NS1 could be purified using a Strep-II tag on the WDR33 subunit. Both ED and full-length NS1 remained associated with the complex across affinity purification, anion exchange and size exclusion chromatography steps, suggesting that the viral protein was stably bound to mPSF (Figure 3.10A). Interestingly, the mPSF-NS1 complex eluted from the column noticeably earlier than mPSF-ED, suggesting that mPSF-NS1 is much larger and may contain two copies of the mPSF complex. To increase the probability of capturing the NS1-bound mPSF complex in cryoEM, I treated the samples with a chemical cross-linker bis-sulfosuccinimidyl-suberate (BS3) and purified the cross-linked protein complex from aggerates and excess BS3 by gel filtration chromatography (Figure 3.10B). The elution volumes of both complexes were similar before and after cross-linking, suggesting that BS3 treatment did not affect the composition or oligomeric state of either mPSF-ED or mPSF-NS1 complexes. To determine whether ED and NS1 were genuinely cross-linked with mPSF, I performed SDS-PAGE analysis of the peak gel filtration fractions after cross-linking (Figure 3.10C). The gel revealed single discreet bands, demonstrating that the samples were homogenous. mPSF-ED migrated on the gel as a monomeric complex, while the position of the mPSF-NS1 band was consistent with the presence of two mPSF copies. This was in agreement with the gel filtration profiles of both complexes. I proceeded with probing the gel with an anti-His antibody to detect ED or full-length NS1, both of which carried a His_6 tag. The Western blot analysis demonstrated that only full-length NS1 successfully cross-linked with mPSF, and hence, the mPSF-NS1 complex was used for structural analysis by cryoEM (Figure 3.10C).

I prepared unsupported UltrAuFoil® all-gold cryoEM grids of the cross-linked mPSF-NS1 sample, screened them in-house and then imaged the grids using a Titan Krios electron microscope at eBIC (Diamond Light Source) equipped with a K3 direct electron detector in counting mode. I collected 8,226 micrographs revealing well-dispersed particles, which allowed me to pick them manually (Figure 3.10D). I generated 2D classes from the manually picked particles and then used them as templates for automated particle picking in Relion 3.1. ~3 million particles were picked and used for 2D classification with image alignment. The generated 2D classes showed detailed secondary structure features and clearly represented the mPSF complex (Figure 3.10E). NS1 (~40 kDa as a dimer) is likely to be too small to be observed in 2D class averages. However, the 2D classes did not reveal any densities for a second copy of mPSF, suggesting that, despite the evidence of NS1-mediated

mPSF dimerisation in solution, they dimeric complexes could not be observed in vitreous ice.

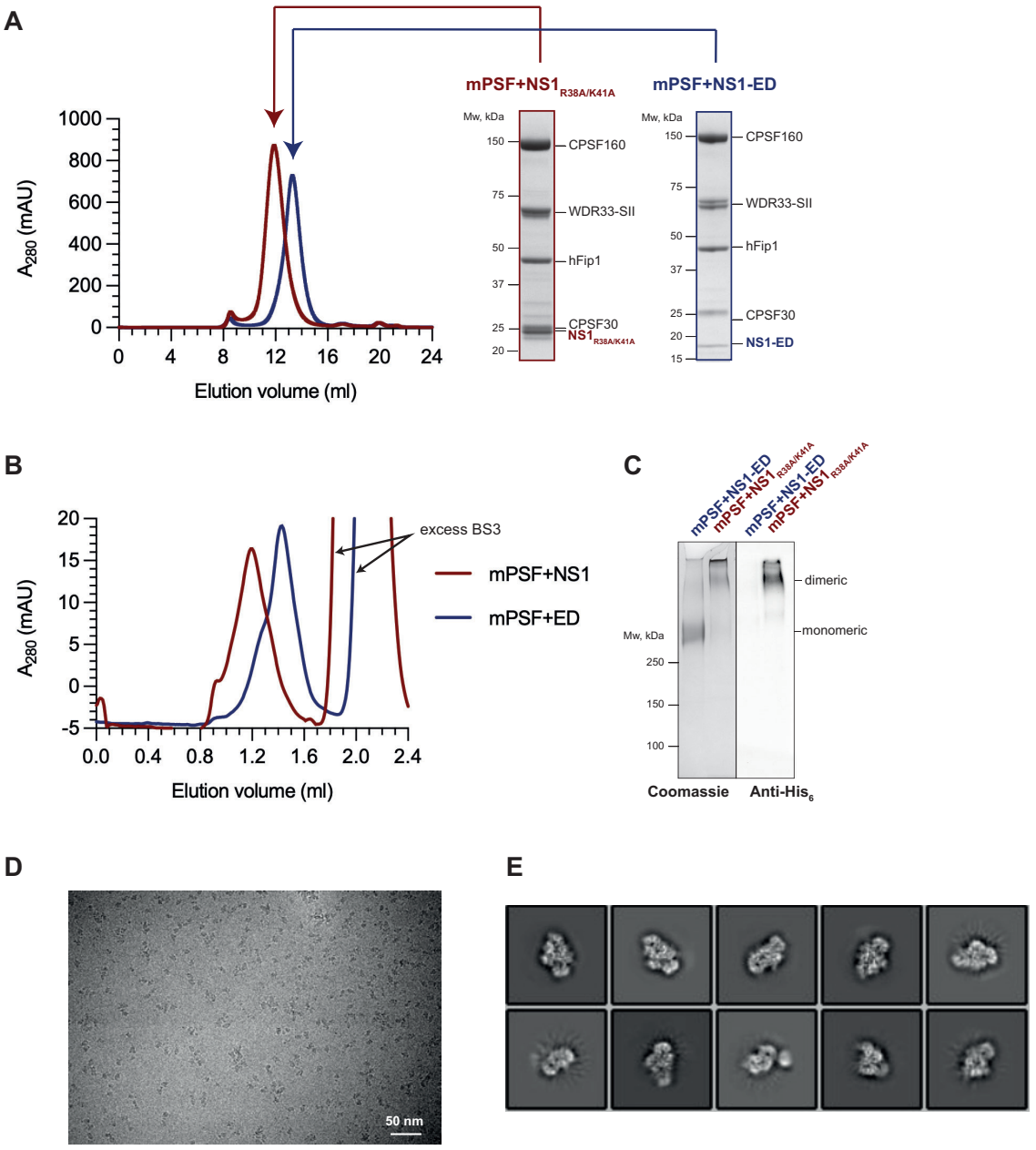


Figure 3.10 Preparation of mPSF-NS1 complex for cryoEM analysis. (A) Size exclusion chromatograms (left) of mPSF bound to either full-length NS1 or its effector domain (ED) and SDS-PAGE analyses (left) of the peak fractions. **(B)** Gel filtrations chromatograms of mPSF bound to either full-length NS1 or its effector domain (ED) after cross-linking with BS3. The large peaks eluting at ~2.2 ml contain excess unreacted BS3. **(C)** SDS-PAGE analyses of the peak fractions from (B) on a gel stained with Coomassie (left) and Western blot analyses of the same samples probed with anti-His_6 tag antibody (right). **(D)** Representative cryoEM micrograph of the mPSF-NS1 complex cross-linked with BS3. **(E)** Representative 2D class averages of the mPSF-NS1 complex.

I performed 3D classification with image alignment using the particles within the most detailed 2D classes, and two high-resolution 3D maps were generated (Figure 3.11). Comparison with the previously determined structures of mPSF suggested that one of the classes contained electron density corresponding to CPSF160, WDR33 and ZnF1 of CPSF30 (class 1), while the other lacked any density for CPSF30 (class 2). 3D refinement of class 2 resulted in a 3.9 Å map of the CPSF160-WDR33 dimer (Figure 3.11). Since NS1 is known to interact with CPSF30, I selected class 1 for further processing. After 3D refinement, I performed 3D classification without image alignment on the particles within class 1 in order to further sort out heterogenous populations of protein particles (Figure 3.11). One of the resultant classes (class 3) appeared to have some weak density for ZnF2 of CPSF30, and I performed focused classification on the region of the map corresponding to WDR33 and CPSF30 subunits, where NS1 should be located. However, no high-resolution density corresponding to NS1 could be resolved (Figure 3.11).

Overall, I showed that NS1 inhibits 3′ end processing *in vitro*, and that a single NS1 protein may interact with two copies of mPSF. However, the structure of mPSF bound to NS1 could not be determined, and further optimisation of the sample is likely to be needed to elucidate how this Influenza A protein interacts with and inhibits CPSF. The NS1 protein was covalently cross-linked with the mPSF module in the cryoEM sample and could not dissociate from the complex during vitrification. Instead, I hypothesise that NS1 may have denatured during grid preparation, so that its electron density could not be reconstructed. To prevent such denaturation, buffers containing detergents or grids coated with a support surface could be used. In addition, despite their conserved fold and function, NS1 proteins from various Influenza A strains differ substantially in their amino acid sequences of the effector domain, which may affect their stability (149). Thus, NS1 from other Influenza A strains should also be tested in this particular cryoEM sample in the future.

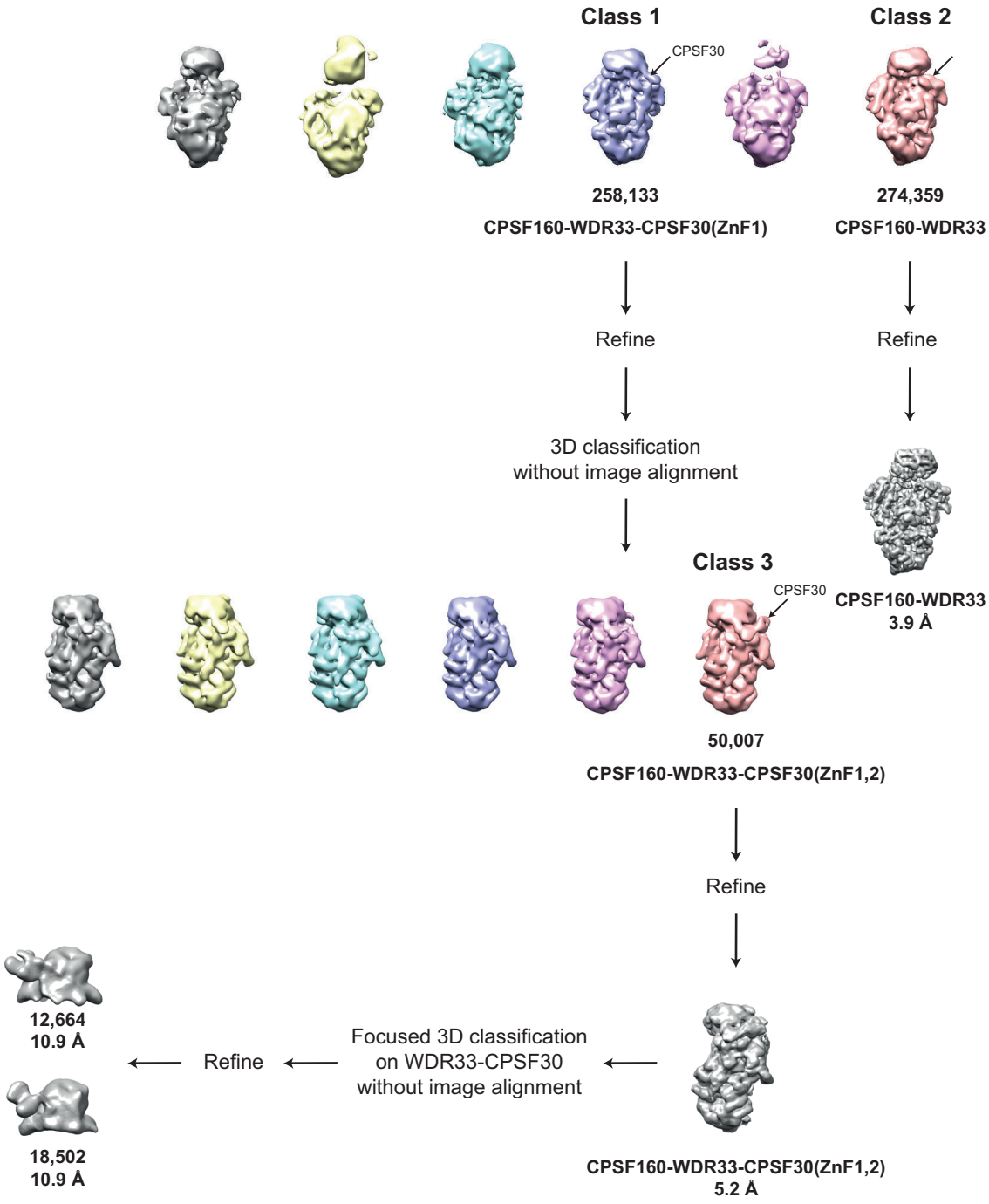


Figure 3.11 Schematic representation of the processing pipeline of the mPSF-NS1 complex in Relion 3.1. The number of particles in specific classes and the resolution of refined maps are indicated. The maps are oriented to clearly show additional densities.

3.5 A direct physical interaction between mPSF and mCF modules is not essential for endonuclease activation

After investigating how the mPSF module is targeted by the Influenza virus, I began to contemplate the role of mPSF in the activation of the CPSF endonuclease. The mCF module alone cannot catalyse endonucleolytic cleavage of the pre-mRNA substrate even in the presence of CStF, CFIIm and RBBP6 (Appendix Figure 8.4). This suggests that the mPSF module is required for the activation of the 3′ endonuclease, possibly by both recruiting CPSF to the substrate and facilitating the conformational change that pries open the active site of CPSF73. mCF and mPSF are physically connected by an interaction between the mPSF subunit CPSF160 and the PIM peptide of CPSF100, a subunit of mCF (Figure 1.4) (8). To further investigate the role of mPSF in CPSF endonuclease activation, I generated a variant of mCF in which the aromatic residues of CPSF100-PIM that mediate its interaction with CPSF160 (F464, W473, Y476) were mutated to alanine. Mixing mPSF and mCF-CPSF100$_{F464A/W473A/Y476A}$ did not result in a leftwards shift of the elution volumes of the two modules in a gel filtration experiment, suggesting that mCF containing mutant PIM failed to bind the mPSF module, in agreement with the published data (Figure 3.12A) (8). The lack of formation of intact CPSF implies that mCF containing the mutations in the PIM peptide should be completely inactive in an *in vitro* cleavage assay. However, surprisingly, mCF-CPSF100$_{F464A/W473A/Y476A}$ mixed with wild-type mPSF, CStF, CFIIm and RBBP6 did cleave the SV40 pre-mRNA substrate, albeit with ~70% lower efficiency than wild-type CPSF (Figure 3.12B). Very similar results were obtained with the mCF variant lacking the PIM peptide entirely (CPSF100 residues 460-486) (Figure 3.12B). These unexpected observations suggest that while physical tethering of mCF to mPSF increases the efficiency of the CPSF endonuclease, it is not strictly required for this enzymatic activity. I hypothesise that in the absence of PIM, the auxiliary factors essential for cleavage may still bridge the two CPSF modules to mediate endonuclease activation. Further studies will be required to address this question.

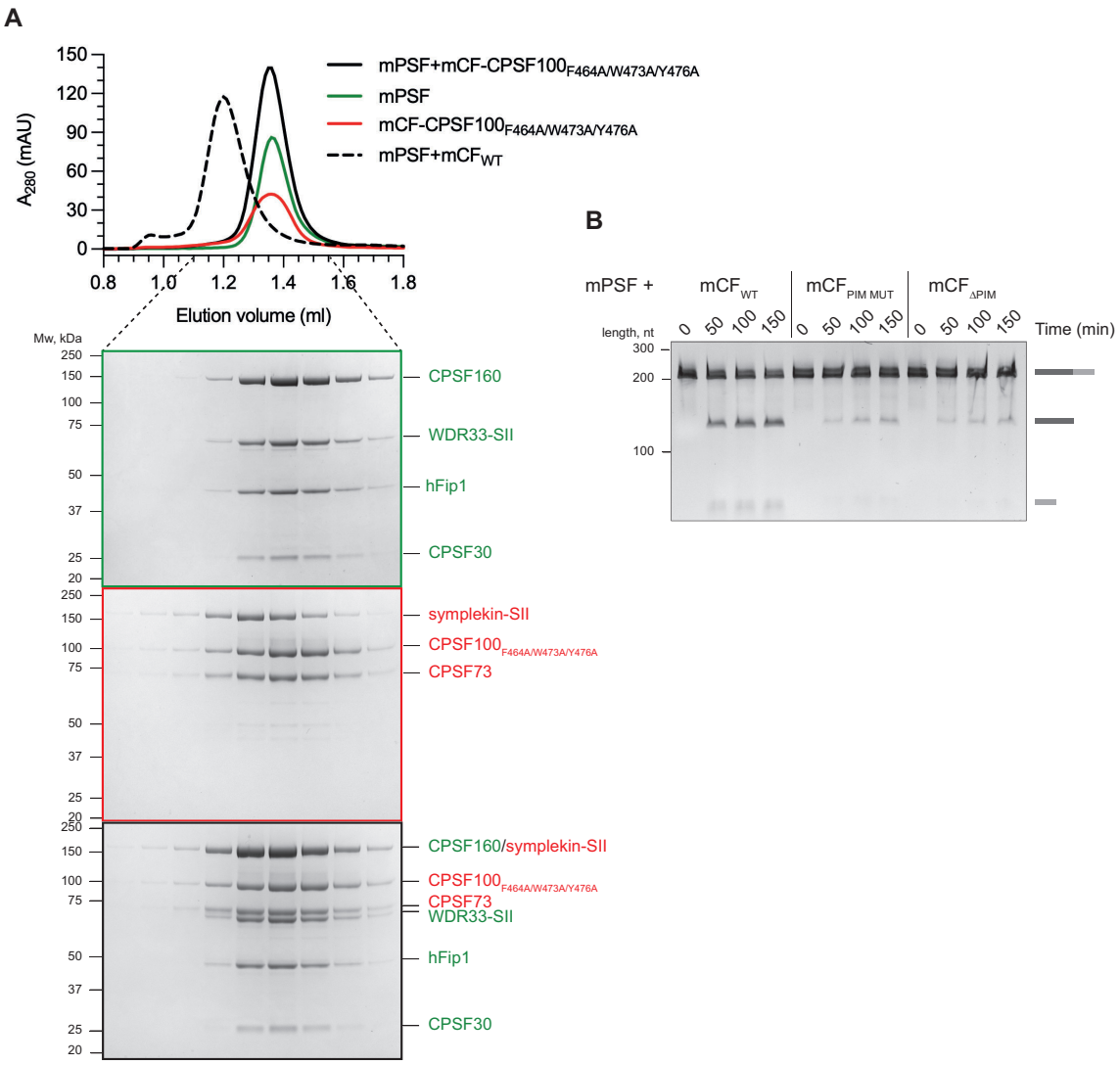


Figure 3.12 Physical tethering of mCF to mPSF is not essential for CPSF endonuclease activation. (A) Size exclusion chromatograms (top) and SDS-PAGE analyses of the peak fractions (bottom) of either mPSF (2.5 μM) or mCF-CPSF100$_{F464A/W473A/Y476A}$ (2.5 μM) alone or the two modules mixed together run on a Superose 6 3.2/300 size exclusion column. The trace of wild-type CPSF (mPSF mixed with wild-type mCF) is shown as a dotted curve for comparison. **(B)** *In vitro* cleavage assays using the SV40 pre-mRNA substrate in the presence of mPSF and either wild-type mCF (mCF$_{WT}$), mCF containing mutations F464A/W473A/Y476A in the PIM peptide (mCF$_{PIM\ MUT}$) or mCF lacking PIM altogether (mCF$_{\Delta PIM}$).

3.6 Truncations of hFip1 and Pcf11 do not affect endonuclease activity of recombinant CPSF

The PIM peptide that physically links mCF and mPSF modules is located within the ~130 residue IDR that divides the β-CASP domain of CPSF100 (Figure 2.3A). It is possible that other IDRs within the 3′ end processing machinery could also mediate functionally important protein-protein interactions. As discussed previously, to optimise the purification of several human 3′ end processing factors, non-conserved IDRs had to be removed from several protein subunits, including WDR33, hFip1, Pcf11 and RBBP6. I have managed to successfully reconstitute the endonuclease activity of recombinant CPSF using these truncated proteins, suggesting that the removed IDRs are not strictly required for endonuclease activation. I also showed that mPSF carrying truncated hFip1 is just as efficient at stimulating polyadenylation as the complex containing full-length hFip1 (Figure 2.2B). Nevertheless, I aimed to test whether using full-length proteins could improve CPSF endonuclease efficiency. Full-length WDR33 and RBBP6 could not be expressed, but I did manage to purify mPSF containing full-length hFip1 and CFIIm carrying full-length Pcf11 subunit. I tested the activity of these complexes in a cleavage assay in order to gain more insight into the mechanism of CPSF endonuclease activation.

A region within the IDR of Pcf11 containing residues 769-1123 consists of a repetitive sequence with a consensus motif FEGP. Although the precise function of FEGP repeats is unknown, they have been shown to be essential for CPSF endonuclease activation in nuclear extract (10) (Figure 3.13A). Thus, I tested the ability of three versions of CFIIm each containing a different Pcf11 variant to activate the CPSF endonuclease: full-length Pcf11 (CFIIm-Pcf11$_{FL}$), Pcf11 lacking 769 N-terminal residues (CFIIm-Pcf11$_{770\text{-}1555}$) used in all other assays throughout this Thesis, and Pcf11 truncated at residue 1123 and hence missing the FEGP repeats (CFIIm-Pcf11$_{1124\text{-}1555}$) (Figure 3.13B). It is important to note that a low concentration stock (~0.5 mg/ml) of CFIIm-Pcf11$_{FL}$ had to be used in the assay to avoid protein aggregation, while the CFIIm complexes containing the truncated variants of Pcf11 were fully soluble even at high (>2 mg/ml) protein concentrations (Figure 2.6B). The cleavage assays revealed that addition of CFIIm-Pcf11$_{FL}$ did not lead to more efficient cleavage by recombinant CPSF compared with CFIIm-Pcf11$_{770\text{-}1555}$, suggesting that the N-terminal region of the Pcf11 IDR, including the CID domain, does not contribute to endonuclease activation (Figure 3.13C). Surprisingly, CFIIm-Pcf11$_{1124\text{-}1555}$ did activate recombinant CPSF, which demonstrated that the FEGP (phenylalanine, glutamate, glycine, proline) repeats are dispensable for endonuclease activation in a minimal reconstituted

system, in contrast to their essentiality in nuclear extract (Figure 3.13C). Similar to other differences between the endonuclease assay with purified proteins and the reaction reconstituted in nuclear extract, protein factors present in the nucleoplasm could cause the dependence of the reaction on the FEGP repeats. Similar peptide motifs have also been shown to form molecular condensates, which, if assembled with other proteins required for endonuclease activation and substrate RNA, could promote cleavage by CPSF in nuclear extract and *in vivo* (150, 151).

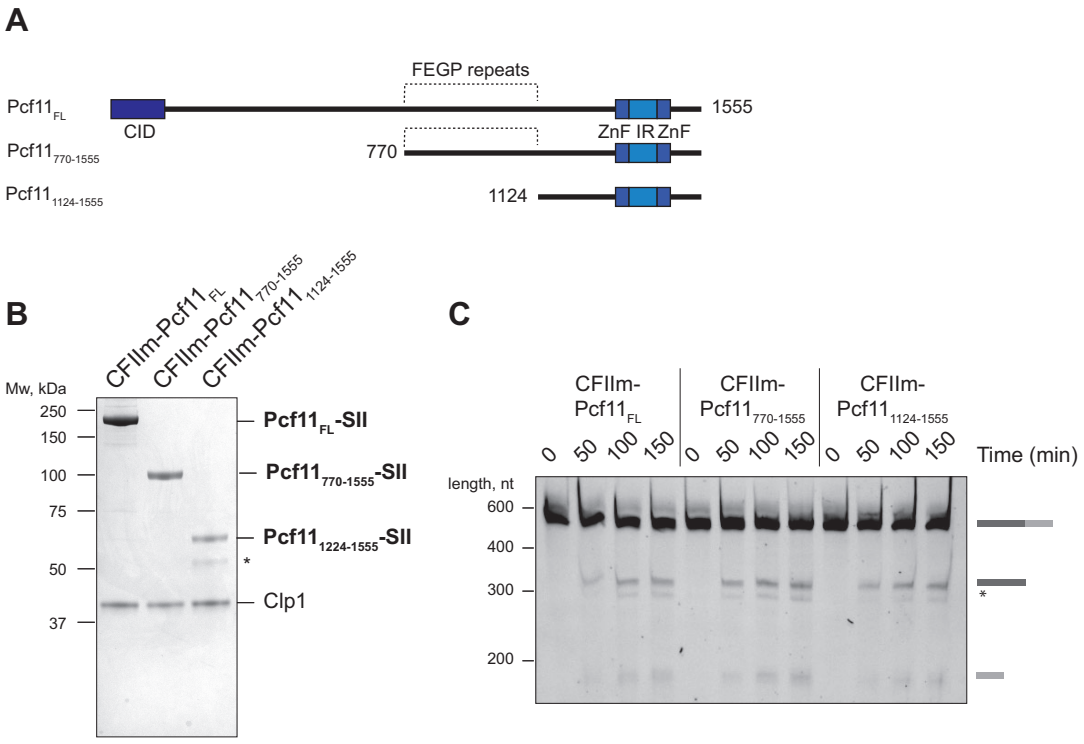


Figure 3.13 FEGP repeats of Pcf11 are not required for endonuclease activity of recombinant CPSF. (A) Domain diagrams of Pcf11 truncations. **(B)** SDS-PAGE analyses of the purified CFIIm complexes containing various versions of the Pcf11 subunit. An asterisk marks a Pcf11 degradation band. **(C)** CPSF cleavage assays using the L3 pre-mRNA substrate in the presence of various version of the CFIIm complex. The minor 5′ cleavage product is marked with an asterisk.

I also tested cleavage activity of CPSF complexes assembled from wild-type mCF and three different version of mPSF: containing full-length hFip1 (CPSF-hFip1$_{FL}$), carrying truncated isoform 4 of hFip1 (CPSF-hFip1$_4$) used in the experiments throughout this Thesis, and lacking hFip1 altogether (CPSF-ΔhFip1) (Figure 3.14A&B). The CPSF-ΔhFip1 complex contained isoform 1, instead of isoform 2, of CPSF30. Both CPSF30 isoforms contain identical binding sites for RNA and all its known protein interactors, and hence, the difference in CPSF30 isoforms should not affect CPSF activity (Figure 2.1A). Similar to CPSF-hFip1$_{FL}$, CPSF-ΔhFip1 was also prone to precipitation, and its peak anion exchange fraction had to be used in an assay to avoid further concentration required for a gel filtration run. This would risk affecting the activity of CPSF, because different amounts of salt would be carried over from the various mPSF complexes. However, the peak anion exchange fractions of both CPSF-hFip1$_{FL}$ and CPSF-ΔhFip1 were highly concentrated, and only a small volume of each complex had to be used in an assay, meaning that the salt carry-over into the assay was negligible.

I tested the cleavage activity of these three complexes and determined that the cleavage efficiency of both CPSF-hFip1$_{FL}$ and CPSF-hFip1$_4$ was almost identical, suggesting that the removed IDRs of hFip1 are not involved in endonuclease activation (Figure 3.14C). The cDNA of isoform 4 of hFip1 has been experimentally detected in the human transcriptome, which suggests that it may be expressed in human cells, although its tissue distribution and expression levels relative to the major hFip1 isoform have not been explored (152). These results imply that the expression of isoform 4 of hFip1 should not affect the basal rate of 3′ end cleavage, but, due to the lack of the RE/D domain, may render CPSF insensitive to CFIm, affecting alternative polyadenylation. In contrast to the other two variants, the endonuclease activity of CPSF-ΔhFip1 was ~40% lower than that of the CPSF complexes containing hFip1 (Figure 3.14C). hFip1 has been shown to interact with CStF and inhibit polyadenylation (26). It is possible that the same interaction has a stimulatory effect on cleavage, which could explain the reduction in CPSF endonuclease activity in the absence of hFip1, but more targeted mutagenesis experiments disrupting the interaction will be required to test this hypothesis.

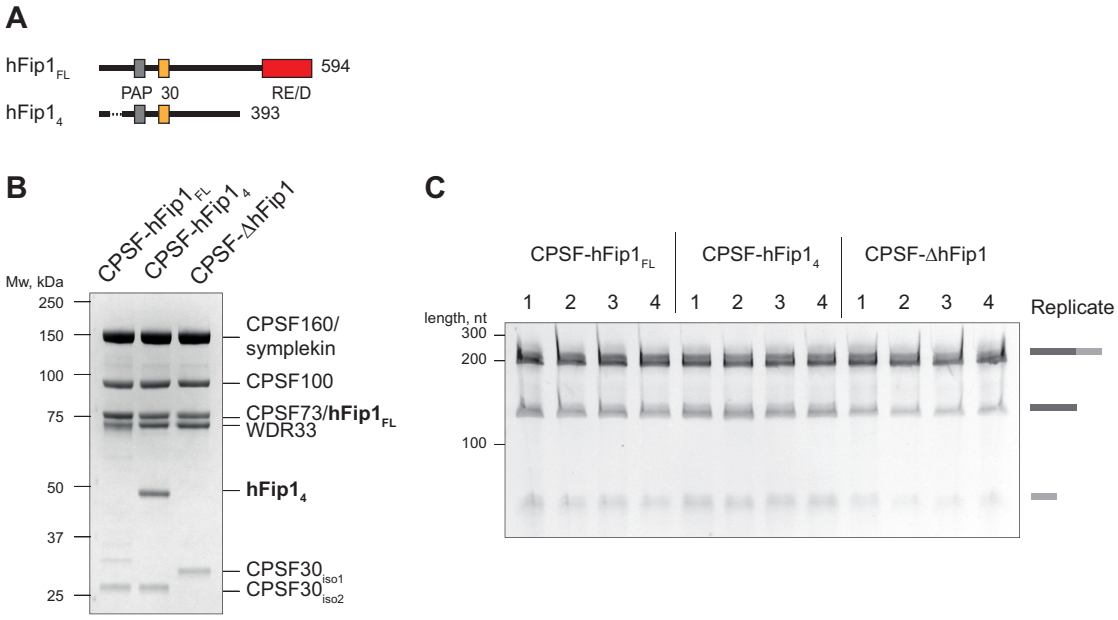


Figure 3.14 IDR of hFip1 is not required for endonuclease activity of recombinant CPSF. (A) Domain diagrams of hFip1 truncations. **(B)** SDS-PAGE analyses of the purified CPSF complexes containing various versions of the hFip1 subunit. **(C)** CPSF cleavage assays using the SV40 pre-mRNA substrate in the presence of various versions of the CPSF complex. Each assay was performed in four replicates.

3.7 CFIm does not stimulate endonuclease activity of recombinant CPSF

After investigating the roles of the protein factors essential for the endonuclease activity of CPSF, I turned my attention to the CFIm complex that I found to be dispensable for the cleavage reaction. CFIm is a major regulator of alternative polyadenylation in human cells. Its subunit CFIm25 selectively binds to UGUA motifs found upstream of certain PAS sites, while CFIm68 recruits the CPSF complex to the pre-mRNA substrate via an interaction between an RS-like domain of CFIm68 and an RE/D domain of CPSF subunit hFip1 (70, 153). The SV40 pre-mRNA substrate does contain an upstream UGUA motif, and hence, its cleavage by CPSF should be stimulated by CFIm (Figure 3.15A). As described in Chapter 2, the CPSF complex used in most of the experiments in this Thesis contains isoform 4 of hFip1, which lacks the C-terminal RE/D domain. Therefore, to test whether CFIm stimulates cleavage of the SV40 pre-mRNA substrate, I first had to purify the CPSF complex containing full-length hFip1 (Figure 2.1A&B). Due to its propensity to aggregate, I avoided concentrating the mPSF-hFip1$_{FL}$ complex and used its peak fraction of anion exchange chromatography directly in a cleavage assay. Various concentrations of CFIm were then titrated into the reaction. To ensure its solubility, CFIm was stored in a buffer containing 400 mM NaCl, but CPSF endonuclease activity is highly sensitive to the salt concentration in the assay buffer (Figure 3.2A). To make sure that salt carry-over upon addition of CFIm does not affect the endonuclease activity, the buffer that CFIm was dialysed against was used to match the salt concentration in every assay regardless of CFIm concentration.

Quantification of three different CFIm titration experiments showed that the efficiency of CPSF endonuclease activity remained unaltered even in the presence of a 20-fold molar excess of CFIm over CPSF (Figure 3.15B&C). The lack of stimulation by CFIm could be explained by the relatively high protein and RNA concentrations used in the assay. The affinity of mPSF for an AAUAAA-containing RNA has been measured to be in a low nanomolar range (~1 nM) (29). Under the current assay conditions (50 nM CPSF and 100 nM RNA substrate), most of the CPSF complex is likely already bound to the substrate, and CFIm may not be able to provide any additional stimulation. It is therefore possible that CFIm may increase the efficiency of endonucleolytic cleavage if the reaction components were diluted. Substrates with multiple UGUA motifs or with multiple potential PAS sites could also be more sensitive to stimulation by CFIm than the SV40 pre-mRNA used here. Finally, the hyper-phosphorylation of the RS-like domain of CFIm has been shown to inhibit its interaction with hFip1, and therefore, treating CFIm with a phosphatase enzyme may allow the complex to stimulate the CPSF endonuclease (70, 154).

Analysis of the L3 pre-mRNA substrate sequence revealed that two UGUA motifs overlap with the A/U-rich segment that may mimic a PAS and be responsible for the minor cleavage event of this substrate (Figures 3.6 & 3.15D). I hypothesised that CFIm could compete with CPSF for binding to the A/U-rich segment and added increasing concentrations of CFIm into a cleavage reaction containing the L3 pre-mRNA substrate. Similar to the SV40 pre-mRNA, CFIm did not seem to stimulate cleavage of the L3 pre-mRNA at the major cleavage site. In contrast, the amount of the minor 5′ cleavage product decreased in the presence of CFIm (Figure 3.15E). Thus, CFIm may compete with CPSF for binding to the A/U-rich region and thereby inhibit the minor cleavage event. This result suggests that CFIm may regulate alternative polyadenylation not only by recruiting CPSF to specific sites but also by blocking the binding of the complex to cryptic PAS sites.

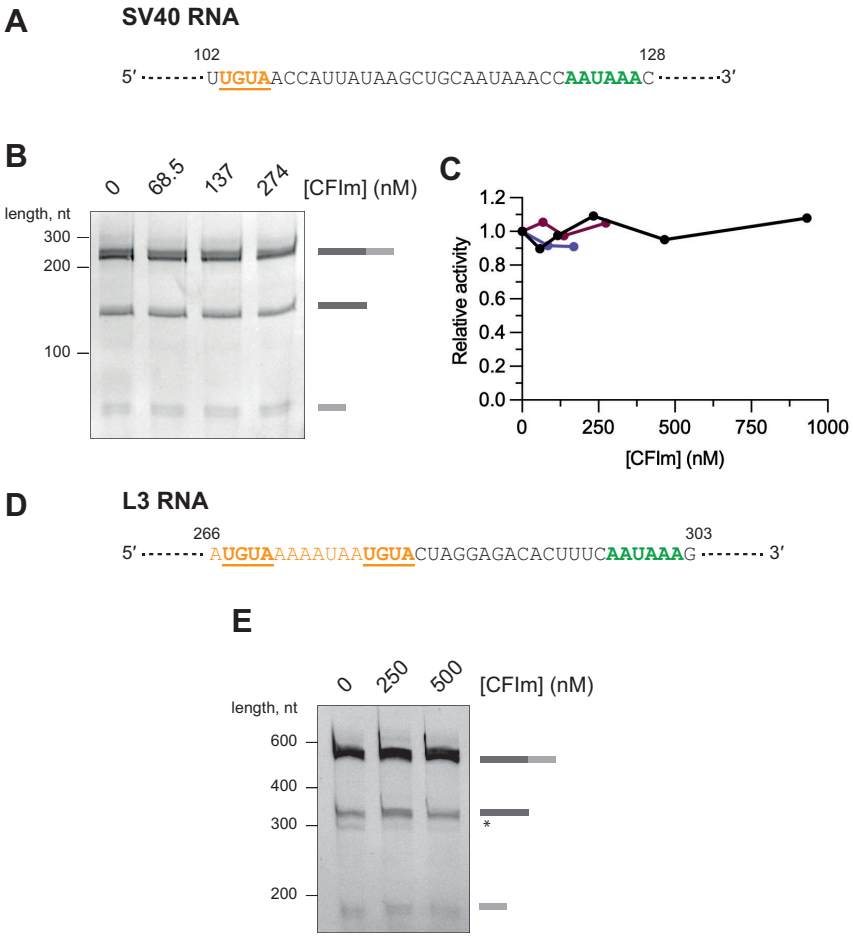


Figure 3.15 CFIm does not stimulate endonuclease activity of recombinant CPSF. (A) Part of the sequence of the SV40 pre-mRNA substrate showing the UGUA motif (orange, underlined) upstream of the PAS (green). **(B)** Cleavage assays of recombinant CPSF using the SV40 pre-mRNA substrate in the presence of various concentrations of CFIm. **(C)** Quantification of gel-based cleavage assays in (B). The data points from three independent replicates are shown in black, purple and maroon. **(D)** Part of the sequence of the L3 pre-mRNA substrate showing that the A/U-rich region (orange) overlaps with two UGUA motifs (orange, bold, underlined). The major PAS is coloured in green. **(E)** Cleavage assays of the L3 pre-mRNA in the presence of various concentrations of CFIm. The minor cleavage product is marked with an asterisk.

3.8 Recombinant CPSF catalyses coupled cleavage and polyadenylation

So far in this Chapter, I described the reconstitution of the endonuclease activity of the human CPSF complex with purified recombinant proteins. I demonstrated that PAP was dispensable for the reconstituted cleavage activity of recombinant CPSF. However, although cleavage and polyadenylation can be uncoupled in nuclear extract, early studies suggested that PAP could be essential for CPSF endonuclease activity (15, 155). Thus, I wondered whether PAP could further stimulate the CPSF endonuclease. I titrated increasing concentrations of PAP into a cleavage reaction, omitting ATP to prevent polyadenylation. Quantification of three different titration experiments did not reveal any clear trends, indicating that PAP does not affect the cleavage efficiency of CPSF in the minimal reconstituted system (Figure 3.16A). What makes an assay in nuclear extract dependent on PAP is difficult to explain, but the effect could be related to either additional proteins present in the extract, the lower concentrations of reaction components *in vitro* or the presence of ATP and other nucleotides in nuclear extract.

Although not required for CPSF cleavage activity, PAP is required to complete 3′ end processing of the cleaved pre-mRNA by adding a poly(A) tail to the majority of protein-coding transcripts in humans. In the previous Chapter, I demonstrated that the recombinant CPSF complex is active in stimulating the polyadenylation activity of PAP. Having subsequently reconstituted its endonuclease activity, I wanted to test if recombinant CPSF could catalyse coupled cleavage and polyadenylation reactions. I added ATP and PAP into a cleavage assay. If the coupled reaction were to take place successfully, the intensity of the 3′ cleavage product band should remain unchanged, while the 5′ cleavage product should disappear and be replaced by a diffuse band of a higher molecular weight corresponding to polyadenylated 5′ cleavage products (Figure 3.16B). However, in the presence of equimolar concentrations of both PAP and CPSF (50 nM each), a band corresponding to RNA longer than the substrate did appear, but almost no 3′ cleavage product was detected (Figure 3.16C). The same pattern was observed in the absence of CStF, CFIIm and RBBP6, and hence, in the absence of cleavage, suggesting that the band of a higher molecular weight than the substrate mostly represented polyadenylated substrate pre-RNAs rather than 5′ cleavage products. Thus, cleavage and polyadenylation reactions were not coupled under these assay conditions. The lack of coupling most likely stems from the two reactions taking place over different time scales: polyadenylation products are discernible after just a few minutes, while it may take hours for a significant proportion of substrate pre-RNAs to be cleaved by CPSF. *In vivo*, the 3′ end of the pre-mRNA is only available after CPSF-catalysed

endonucleolytic cleavage, which releases the transcript from RNA polymerase II. In a test tube, however, the substrate contains a free 3′ end that can be polyadenylated before CPSF may catalyse its cleavage.

I hypothesised that limiting the concentration of PAP in the reaction may increase the probability of substrate cleavage prior to polyadenylation. Indeed, in the presence of substoichiometric amounts of PAP (< 25 nM), the 3′ cleavage product bands persisted, while the 5′ cleavage product was replaced by polyadenylated bands, indicating that 5′ cleavage products were being polyadenylated (Figure 3.16D). Thus, I have managed to reconstitute the complete reaction of pre-mRNA 3′ end processing – coupled cleavage and polyadenylation – with purified recombinant human proteins. Even more efficient coupling of cleavage and polyadenylation in a test tube could be achieved by chemically blocking the 3′ end of the substrate pre-RNA to prevent its polyadenylation prior to cleavage.

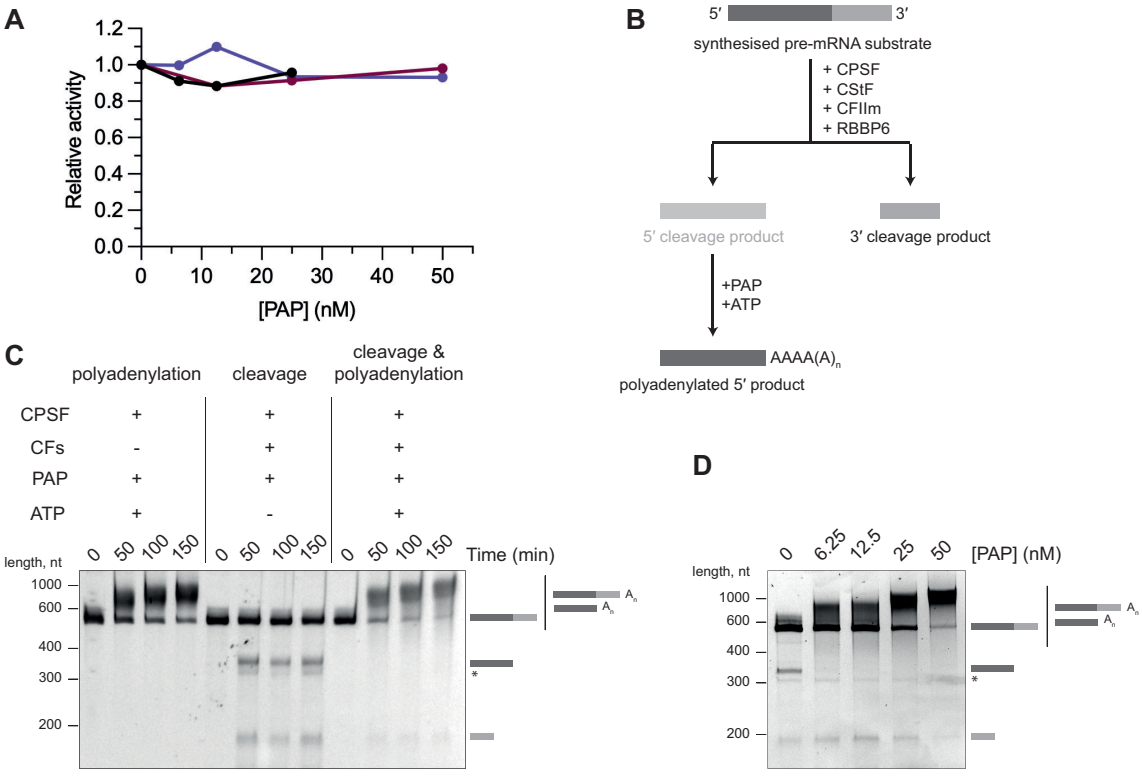


Figure 3.16 Recombinant CPSF can catalyse coupled cleavage and polyadenylation reactions. **(A)** Quantification of gel-based cleavage assays of the SV40 pre-mRNA in the presence of various concentrations of PAP. The data points from three independent replicates are shown in black, purple and maroon. **(B)** Schematic representation of coupled cleavage and polyadenylation reactions. The fading of the 5′ cleavage product indicates that it becomes polyadenylated and disappears from the gel. **(C)** Time-course assays using the L3 pre-mRNA in the presence or absence of cleavage factors (CFs; CStF, CFIIm, RBBP6) and ATP aiming to test polyadenylation, cleavage as well as simultaneous cleavage and polyadenylation activities of recombinant CPSF. Minor cleavage product is marked with an asterisk. **(D)** Cleavage assays of the L3 pre-mRNA in the presence of various concentrations of PAP and 50 nM CPSF. The minor cleavage product does not get polyadenylated and is marked with an asterisk.

3.9 CPSF and histone pre-mRNA 3′ end processing complexes are activated by different mechanisms

The histone pre-mRNA 3′ end processing machinery shares three protein subunits with the canonical CPSF complex: symplekin, CPSF100 and, most importantly, endonuclease CPSF73 (156). The high-resolution cryoEM structure of the histone 3′ end processing complex has been determined in an active state, revealing the mechanism of CPSF73 activation (Figure 1.5) (82). The proteins that directly interact with the endonuclease subunit and mediate the opening of its active site, Lsm10 and Lsm11, are specific to the histone 3′ end processing complex. However, structural and biochemical data suggest that the N-terminal domain (NTD) of symplekin contacts CPSF100 and is essential for endonuclease activation (Figure 3.17A). I investigated if the NTD of symplekin is also involved in activating CPSF73 in the context of the canonical CPSF complex. I prepared a CPSF complex containing a version of symplekin lacking its NTD (CPSF-symplekin$_{\Delta NTD}$). The CPSF-symplekin$_{\Delta NTD}$ complex was still active in an endonuclease assay, suggesting that the NTD of symplekin is dispensable for the cleavage activity of the recombinant CPSF complex (Figure 3.17B). The NTD of symplekin has been shown to interact with a protein phosphatase SSU72, which also inhibits the endonuclease activity of the histone pre-mRNA 3′ end processing complex by sequestering the symplekin-NTD (52, 82). Yeast Ssu72 is also a constitutive component of the CPF complex (9). To determine whether this interaction is also formed in the context of CPSF, I purified human SSU72 from *E. coli* and performed interaction studies using analytical gel filtration chromatography. Indeed, SSU72 did bind to both wild-type CPSF and wild-type mCF but not to the mCF module lacking the NTD of symplekin (mCF-symplekin$_{\Delta NTD}$) (Figure 3.17C&D). Nevertheless, addition of SSU72 did not inhibit endonuclease activity of CPSF (Figure 3.17E). This result shows that symplekin-NTD is not directly involved in the mechanism of CPSF endonuclease activation, and that the same endonuclease subunit CPSF73 is activated by different mechanisms when it is incorporated into CPSF compared with the histone pre-mRNA 3′ end processing complex. In addition to contacting CPSF100, the NTD of symplekin also stabilises the duplex between the substrate RNA and the U7 snRNA in the histone pre-mRNA processing complex (Figure 3.17A) (82). Therefore, a domain shared between the two complexes may have different roles due to the different mechanisms of substrate RNA recognition between the two endonuclease complexes.

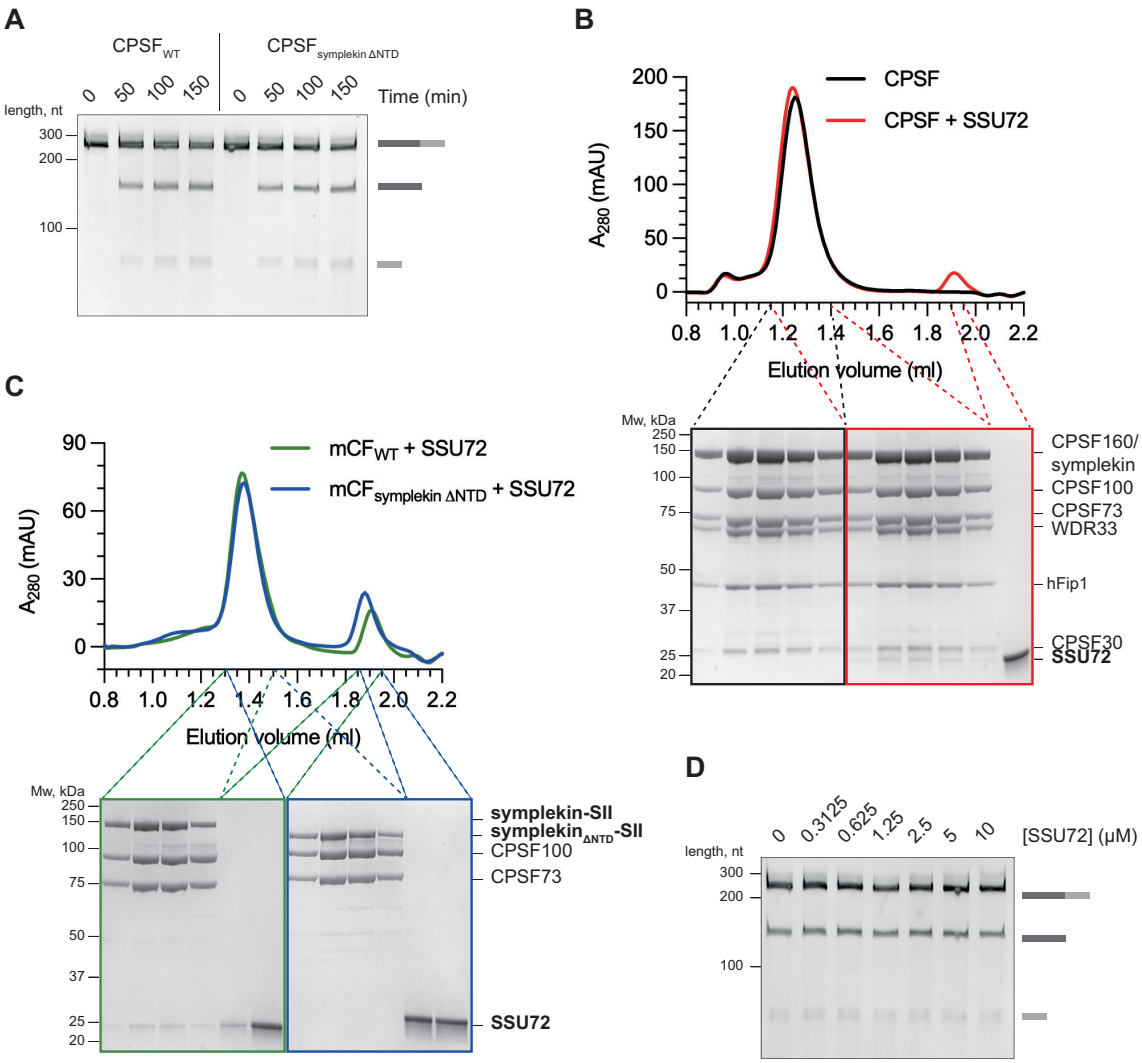


Figure 3.17 CPSF and histone pre-mRNA 3′ end processing complex are activated by different mechanisms. (A) Cleavage assay of the SV40 pre-mRNA substrate in the presence of either wild-type CPSF (CPSF$_{WT}$) or the CPSF complex lacking the NTD of symplekin (CPSF-symplekin$_{\Delta NTD}$). **(B)** Gel filtration chromatograms (top) and SDS-PAGE analyses (bottom) of the corresponding peak fractions of either CPSF alone (2.5 μM) or CPSF (2.5 μM) mixed with SSU72 (10 μM). **(C)** Gel filtration chromatograms (top) and SDS-PAGE analyses (bottom) of the corresponding peak fractions of SSU72 (10 μM) mixed with either wild-type (mCF$_{WT}$) or mCF lacking NTD of symplekin (mCF-symplekin$_{\Delta NTD}$; 2.5 μM). **(D)** CPSF cleavage assay of the SV40 pre-mRNA substrate in the presence of increasing concentrations of SSU72.

3.10 Conclusions and perspectives

3.10.1 CStF, CFIIm, RBBP6 and CPSF are required for pre-mRNA 3′ end cleavage

In Chapter 3, I described how I managed to successfully reconstitute the canonical pre-mRNA 3′ endonuclease activity of human CPSF with purified proteins. I determined that the activation of cleavage requires well-characterised cleavage factors CStF and CFIIm as well as a previously overlooked protein RBBP6. RBBP6 has so far been considered as a regulator rather than an essential activator of CPSF, and therefore, its critical role in activating the endonuclease came as a surprise. RBBP6 was not detected in any of the partially purified fractions of mammalian nuclear extract historically used to reconstitute the 3′ endonuclease activity. However, it must have been co-purified with at least one of these fractions for CPSF endonuclease to be active. Endogenous RBBP6 is a large 200 kDa protein that tends to degrade from its C-terminal end during cell lysis, producing many degradation fragments of variable sizes. I suspect that this may have precluded its identification as an essential component of the CPSF endonuclease assay in nuclear extract.

CPSF, CStF, CFIIm and RBBP6 likely represent the minimal and universal machinery required to carry out the chemistry of 3′ endonucleolytic cleavage. The protein composition of the active human 3′ end processing machinery closely resembles that in budding yeast: CPSF and RBBP6 are equivalent to the CPF complex, while human CStF and CFIIm complexes contain orthologous subunits of the yeast CF IA complex. This demonstrates the high degree of conservation of this fundamental step in eukaryotic gene expression. On the other hand, non-essential factors that regulate 3′ end processing appear to be species specific. For example, the CFIm complex that regulates alternative polyadenylation is only found in metazoans, while CF IB protein that is needed to enforce the specificity of 3′ end cleavage is specific to yeast. Thus, although the general mechanism of pre-mRNA 3′ end endonucleolytic cleavage is highly conserved, mechanisms of its regulation have evolved independently in different species.

3.10.2 Endonuclease activation enforces the specificity of 3′ end processing

Individually purified CPSF73 has only weak and non-specific endonuclease activity (38). Thus, its incorporation into a 7-subunit CPSF complex may ensure that the endonuclease subunit is inhibited until it is specifically activated on transcripts carrying a PAS sequence.

The additional requirement for three accessory RNA-binding factors would further restrict activation, precisely positioning the endonuclease on RNA and preventing premature cleavage. This rather elaborate activation mechanism may ensure a high degree of specificity of 3′ endonucleolytic cleavage. Regulation of CPSF endonuclease is highly reminiscent of the activation of the spliceosome which occurs in a step-wise manner on each individual intron (157).

Despite hours-long incubation of the CPSF endonuclease reaction at 37°C, I never managed to achieve complete cleavage of the model pre-mRNA substrate with recombinant CPSF *in vitro*. This is substantially slower than the rate of cleavage of a model yeast pre-mRNA by yeast CPF under similar conditions (12). This could be due to either variable quality of reaction components, unknown protein factors missing from the reaction reconstituted with human proteins or substrate pre-mRNAs lacking a cleavage-competent tertiary structure *in vitro*. However, it is also possible that human CPSF is an inherently inefficient and potentially more accurate endonuclease that allows more extensive regulation, for example, to enable correct cleavage site selection even on very long 3' UTRs with multiple potential PAS sites. On the other hand, CPF cleavage must be very efficient to prevent transcriptional readthrough into downstream open reading frames in yeast, where genes are closely spaced (33, 158). Thus, the basal 3′ end processing machineries from yeast and humans may have co-evolved along with the genome architecture of the respective species.

3.10.3 CPSF endonuclease could be activated under different conditions

Concurrently with publication of my results, the group of Elmar Wahle published an independent study (Schmidt *et al*) involving a reconstitution of the CPSF endonuclease activity from purified recombinant human proteins (159). The major conclusions of the article, especially regarding the role of RBBP6, matched the observations described here. However, there were some notable difference between the two studies regarding both assay buffer conditions and proteins required for CPSF endonuclease activation. The major similarities and differences between the two studies are summarised in Table 3.1.

The CPSF endonuclease activity described by Schmidt *et al* was dependent on the presence of either PVA or trimethylamine N-oxide (TMAO), while I could reconstitute the cleavage activity of recombinant CPSF in the absence of chemical additives. Both studies used the SV40 pre-mRNA of ~200 nt in length, and hence, the substrate could not account for the

different requirements for chemical additives (Figure 3.3). Schmidt *et al* used significantly lower protein and RNA concentrations, and in their assays, chemical additives may act as crowding agents facilitating the formation of active cleavage complexes. Higher protein and substrate concentrations used here may circumvent this requirement. The use of low concentrations of the reaction components by Schmidt *et al* may also account for their ability to detect stimulation of CPSF endonuclease activity by the CFIm complex, which was not observed here (Figure 3.15).

Surprisingly, the CPSF endonuclease activity reconstituted by Schmidt *et al* strictly required the presence of both the PAP enzyme and ATP in the cleavage assay. Catalytically-dead PAP could also activate CPSF, suggesting that the requirement for ATP is not related to the enzymatic activity of PAP. Instead, Schmidt *et al* proposed that ATP may act as a co-factor of Clp1, a subunit of CFIIm. I performed several experiments and could exclude the possibility of ATP carry-over from either *in vitro* transcription or CFIIm purification in my experiments, suggesting the CPSF endonuclease activity reconstituted here is genuinely independent of ATP (Appendix Figure 8.5). In terms of the role of PAP, I hypothesised that it could stabilise some protein-protein or protein-RNA interactions at low concentrations, but experiments by Schmidt *et al* showed that their assay required PAP even at protein concentrations similar to the ones used here (personal communication).

Notably, most of the proteins used by Schmidt *et al* contained their full-length IDRs, whereas I truncated several proteins to optimise their purification. Therefore, recombinant protein complexes used in the study by Schmidt *et al* more closely resembled the factors that carry out pre-mRNA 3′ end processing in human cells. However, many such full-length factors are likely to suffer from poor solubility and aggregation, which may interfere with the reaction (see Chapter 2). The only truncated protein used by Schmidt *et al* was RBBP6. The construct used in their study contained essentially the same residues as described here (residues 1-340) but was expressed in *E. coli* rather than in insect cells. Although no experiments were performed to address this yet, it is possible that RBBP6 purified from Sf9 insect cells may contain post-translational modifications absent if the same protein is expressed in bacteria. It remains to be determined whether either the behaviour of certain IDRs that I removed from 3′ end processing factors or the possible differences in post-translational modifications of RBBP6 may account for the observed differences between the two studies.

In summary, extensive attempts by myself and by Schmidt *et al* to address the differences in the two sets of assay conditions yielded no conclusive results. Further investigation will

be required to understand the differences in both PAP and ATP dependence. Nevertheless, based on the many control experiments performed here and by Schmidt *et al*, both systems can be used for studying 3′ end processing in humans. In addition, which assay conditions more closely match the environment in the nucleus of the human cell *in vivo* remains to be seen.

Aspect	**This Study**	**Schmidt *et al*** (159)
Proteins required		
CPSF	YES (truncated mPSF)	YES
CStF	YES	YES
mCF	YES (truncated)	YES
RBBP6	YES (from insect cells)	YES (from E. coli)
PAP	NO	YES
Concentrations of components		
Proteins	50-300 nM	3-125 nM
RNA substrate	100 nM	2.5 nM
Buffer conditions		
ATP	NO	YES
Crowding agents	NO	YES

Table 3.1 Comparison between this Study and Schmidt *et al* (159) **of the conditions of the CPSF endonuclease assays reconstituted from purified recombinant human proteins.**

3.10.4 Advantages and limitations of CPSF enzymatic activities reconstituted with purified proteins

The CPSF endonuclease assay reconstituted with purified recombinant proteins does not come without its limitations. Although the exact concentrations of the 3′ end processing factors in the human nucleus are unknown, the purified reaction components in a test tube are likely to be significantly more concentrated. This may mask the effects of certain regulatory proteins (for instance, CFIm) and may also promote the usage of cryptic cleavage sites never used under physiological conditions *in vivo* (Figures 3.6 & 3.15). In general, the

in vitro cleavage assay in its current format is not really suitable for studying cleavage site selection or for comparing processing efficiencies of different pre-mRNAs. *In vivo*, 3′ end cleavage takes place co-transcriptionally, which is not recapitulated in the *in vitro* assay. The order in which cleavage sites emerge from RNA polymerase II influences cleavage site selection, with proximal cleavage sites having a kinetic advantage over distal sites (2). Thus, the choice of cleavage site on a synthetic substrate by recombinant CPSF may not reflect its preference *in vivo*. In addition, buffer conditions such as the concentration of magnesium ions and the presence of crowding agents may not affect the cleavage efficiency of different pre-mRNAs to the same extent (Figures 3.2D & 3.7). Several cell-based reporter assays have been developed to study the cleavage site preference of CPSF *in vivo* (160, 161). However, it would be even more exciting to combine transcription and 3′ end processing assays, both of which have already been reconstituted separately, to study co-transcriptional 3′ end processing with recombinant proteins. In fact, attempts to study 3′ end processing coupled with transcription have been made using the budding yeast machinery (110).

Despite its limitation, the CPSF endonuclease assay with purified recombinant proteins is a powerful tool for mechanistic studies of human pre-mRNA 3′ end processing. Establishing the reaction itself has already revealed the minimal set of proteins required for CPSF endonuclease activation (Figure 3.1). Future improvements in the throughput of the assay may also allow its use in the development of new therapeutics targeting CPSF73, which is overexpressed in many cancer types (97, 98). In addition, recombinant protein complexes are relatively easy to modify by introducing mutations, domain truncations and subunit drop-outs that enable to test hypotheses regarding the mechanism of endonuclease activation. This was already demonstrated in this Chapter by the experiments testing, for example, the cleavage activity of the catalytically-dead CPSF complex or CPSF lacking hFip1 subunit (Figures 3.8 & 3.14). Importantly, the reconstituted reaction has provided information of what proteins need to be assembled for imaging by cryoEM to elucidate the structure of the active human pre-mRNA 3′ end processing machinery. Milligram quantities of the recombinant proteins available also provide a great advantage over endogenous complexes for cryoEM sample optimisation. Overall, the reconstituted reaction will play a critical role in deciphering the precise molecular mechanism of CPSF endonuclease activation using structural biology tools.

Chapter 4:

RBBP6 is a conserved activator of canonical pre-mRNA 3′ end processing

Reconstitution of the CPSF endonuclease activity with purified proteins revealed that a multi-domain protein RBBP6 is essential to activate pre-mRNA 3′ end cleavage. RBBP6 was previously shown to regulate alternative polyadenylation but was not considered to be a key component of the active pre-mRNA 3′ end processing machinery, which meant that its role in activating the endonuclease had been overlooked for many years (44). Therefore, I was particularly intrigued by how RBBP6 may contribute to CPSF endonuclease activation. In this Chapter, I will use biochemical and structural biology techniques, including *in silico* modelling, to investigate the interactions between RBBP6, CPSF and cleavage factors, which may begin to explain how RBBP6 facilitates the transition from an inhibited endonuclease to the active human pre-mRNA 3′ end processing machinery.

4.1 RBBP6 is not a constitutive subunit of human CPSF

The budding yeast orthologue of RBBP6, Mpe1, is a constitutive subunit of the yeast CPF complex, but whether RBBP6 could bind CPSF remained unknown (162). Interestingly, RBBP6 was detected along with CPSF in the post-cleavage complex purified from mammalian nuclear extract bound to the 5′ cleavage product RNA (6). Thus, I wondered whether RBBP6 was also stably associated with CPSF in human cells. To answer this question, I aimed to purify endogenous CPSF and test whether RBBP6 co-purified with the complex. This could also allow me to discover novel interactors of human CPSF.

4.1.1 Over-expression of tagged subunits in mammalian cells leads to the purification of individual CPSF modules

First, I attempted to purify endogenous CPSF by transiently transfecting mammalian Expi293 cells with a plasmid carrying a doxycycline-inducible gene encoding a single tagged subunit of CPSF. The principle behind this approach is that the over-expressed subunit should get incorporated into the complex along with endogenous proteins, and the tag could then be used to purify the endogenous protein complex (Figure 4.1A). This method does not involve genetic manipulation of mammalian cells and hence is relatively quick, which allowed me to tag and test many different CPSF subunits.

To decide which CPSF subunits were the most suitable to tag and over-express, I consulted the published data of the absolute expression levels of the various CPSF subunits measured in the mouse fibroblast cell line (163). CPSF plays a conserved role in gene expression across various cell types and organisms, which suggests that the values measured in this study should be similar to the protein concentrations in Expi293 cells. The lower the abundance of a subunit in the cell, the higher the probability that the tagged version of the protein will be incorporated into CPSF, leading to improved purification yields. Based on the published data, WDR33 is the least abundant protein out of the known CPSF subunits in mammalian cells (Figure 4.1B). RBBP6 itself is also expressed at relatively low levels. Therefore, I transfected the Expi293 cells with mammalian expression plasmids carrying either full-length human WDR33 or RBBP6 genes with a C-terminal 3x-Flag tag. A small-scale pull-down using anti-Flag beads was then performed along with a Western blot with an anti-Flag antibody to test the expression of the tagged proteins. Either protein failed to be produced in Expi293 cells which, based on the attempts to express these proteins recombinantly, was likely caused by their long intrinsically disordered C-terminal extensions.

Next, I aimed to purify the endogenous complex by transiently over-expressing several other known CPSF subunits. I chose CPSF160 and symplekin, which are major scaffold proteins of mPSF and mCF, respectively. These scaffold proteins make physical interactions with multiple subunits, increasing the chances of purifying an intact CPSF complex. I also tagged the endonuclease subunit CPSF73, which must be part of an active and fully-assembled complex. After over-expressing the 3x-Flag-tagged subunits in mammalian cells, a pull-down was performed on anti-Flag beads, and the proteins in the eluate were identified by tandem mass spectrometry analysis (Figure 4.1C). The presence of all known CPSF subunits would indicate that the purification of endogenous CPSF was successful. However, not all known CPSF subunits could be detected in the eluates of the endogenous pull-downs. Instead, proteins belonging to the module of the tagged subunit were preferentially enriched instead of a full CPSF complex. The pull-downs were first performed from whole-cell extracts. However, CPSF performs its function in the nucleus, suggesting that intact functional CPSF should be more abundant in the nuclear fraction. Hence, I attempted to purify endogenous CPSF from the nuclear extract of transiently transfected Expi293 cells. However, pull-downs using, for example, tagged symplekin lead to the purification of endogenous mCF but not the full CPSF complex (Figure 4.1C). It is important to note that no RBBP6 was ever detected in any of the attempted preparations of endogenous CPSF from either whole cells or nuclear extract.

I concluded that transient over-expression of a tagged subunit is not a suitable strategy for purifying endogenous CPSF. I hypothesise that over-expression of a single subunit leads to a mismatch in protein stoichiometry in the cells: not enough endogenous proteins are expressed to assemble into a full complex with the tagged protein, resulting in the purification of assembly intermediates, or, in this case, mPSF or mCF modules. The importance of tagging a protein component that is limiting in copy number is illustrated by how transient over-expression of tagged proteins has in the past enabled the isolation of the Integrator-PP2A complex (INTAC) (59). PP2A may take up to ~1% of the cellular protein content, and hence, enough endogenous PP2A was present in the cell to assemble a stoichiometric complex with the over-expressed Integrator subunits. Thus, it is possible that pull-downs using tagged and over-expressed subunits that are limiting in copy number relative to the rest of CPSF, such as WDR33 and RBBP6, could lead to successful purification of endogenous CPSF, but further optimisation of their expression in mammalian cells would be required.

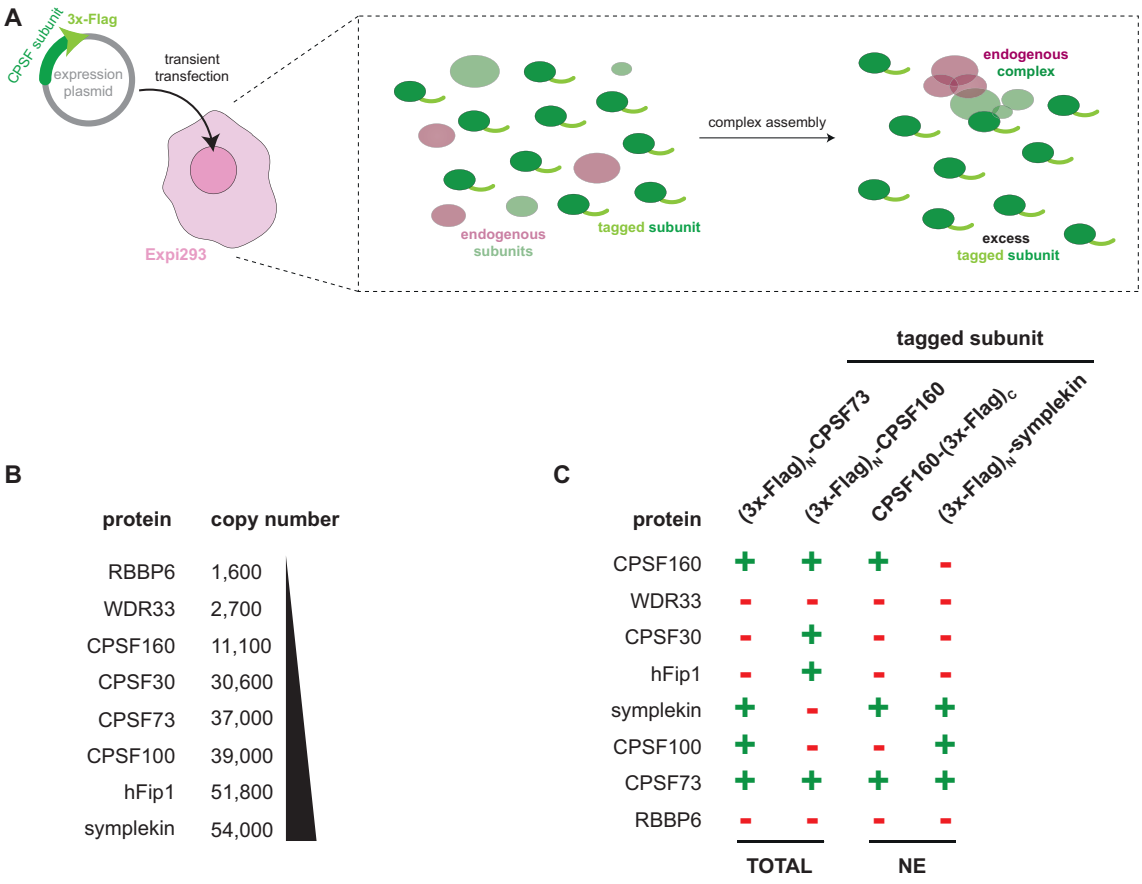


Figure 4.1 Transient over-expression of a tagged subunit does not allow purification of intact CPSF. (A) Schematic representation of the strategy to purify endogenous CPSF using an over-expressed tagged subunit. Purification using anti-Flag resin captures not only an assembled complex but also excess tagged subunit that was not incorporated into an endogenous complex. **(B)** Absolute copy number of each known CPSF subunit and RBBP6 at steady state in the NIH/3T3 mouse fibroblast cell line according to quantitative proteomics data from (163). **(C)** Summary of mass spectrometry analyses of purifications of endogenous CPSF from either whole cell extract (TOTAL) or nuclear extract (NE) using various tagged subunits. "+" indicates that peptides corresponding to a particular subunit were found in the sample, while "-" marks subunits that were not detected.

4.1.2 Purifying endogenous CPSF from a stable cell line carrying a tagged subunit

The CPSF subunit that is limiting in copy number, WDR33, could not be over-expressed by transient transfection. Instead, I aimed to create a stable cell line carrying a tag in the gene encoding WDR33. In this cell line, the tagged protein would be expressed at endogenous levels, ensuring that it is incorporated into an intact CPSF complex with a correct subunit stoichiometry. The tag was added to the 3′ end of the gene (C-terminus of the protein product) (Figure 4.2A). WDR33 contains a long intrinsically-disordered C-terminal domain, which is likely to be susceptible to proteolysis (Figure 2). This may result in the loss of the tag and reduced yield of purification. However, tagging the C-terminus ensures that the purified complex contains full-length WDR33 with an intact C-terminal domain. The role of the C-terminal region of WDR33 has not been studied, but this domain could be important to the function of CPSF. Hence, it was important to make sure that purified endogenous CPSF contained full-length WDR33.

I used a gene targeting strategy using CRISPR-Cas9 to tag the endogenous WDR33 gene in HEK293T cells. I tested two different tags: a tandem affinity purification (TAPS) tag, containing a Strep-II tag, a tobacco etch (TEV) protease cleavage site and protein A; and an HTBH tag, consisting of a His_6 tag, a TEV protease cleavage site followed by a biotin acceptor peptide that is biotinylated by endogenous enzymes in the cell, and another His_6 tag (Figure 4.2A). The cell clones carrying the insertion were selected using hygromycin as an antibiotic selection marker, and the correct addition of the tag to both copies of the WDR33 gene was confirmed by PCR. Endogenous CPSF was purified from the stable WDR33-TAPS and WDR33-HTBH cell lines using Strep-Tactin beads (Figure 4.2A). I followed the presence of the endonuclease subunit as a proxy for CPSF across the various purification steps by Western blot, which indicated that CPSF became enriched on Strep-Tactin beads, and that it was then successfully eluted (Figure 4.2B). The purification from the WDR33-HTBH cell line provided a better protein yield and a lower non-specific protein background than the native CPSF preparation from the WDR33-TAPS cell line, as demonstrated by the SDS-PAGE analysis of the elution fractions (Figure 4.2C). Several features of the HTBH tag could account for this difference (164). Biotin has a significantly higher affinity for Strep-Tactin than the Strep-II tag, which is critical to the yield when purifying endogenous proteins of low abundance. In addition, the HTBH-tagged complex could be eluted from Strep-Tactin beads by specific cleavage with the TEV protease instead of desthiobiotin displacing TAPS-tagged CPSF from the beads. This ensured that hardly any proteins non-specifically-bound to the resin were found in the eluate.

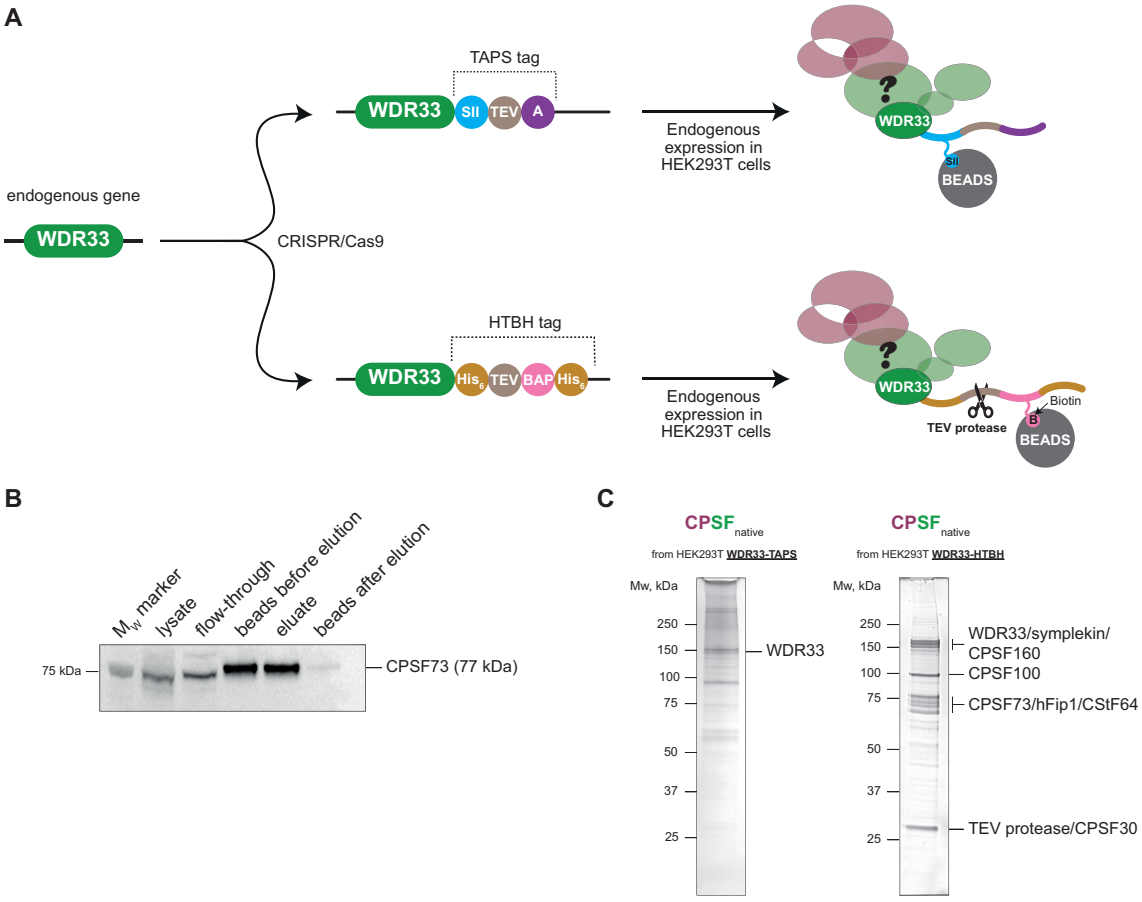


Figure 4.2 Purification of native CPSF from a stable cell line. (A) Schematic representation of the strategy to establish a stable HEK293T cell in which the endogenous genes encoding WDR33 carry either a TAPS or an HTBH tag. SII – Strep-II tag; TEV – cleavage site targeted by the TEV protease; A – protein A; BAP – biotin acceptor peptide; BEADS – Strep-Tactin beads. **(B)** Western blot using an antibody against CPSF73 to test the presence of CPSF at various steps of endogenous purification. This particular blot comes from the experiment in which whole cell extract of the HEK293T WDR33-HTBH cell line was used for purification. **(C)** SDS-PAGE analyses of elution fractions of native CPSF purified either from HEK293T WDR33-TAPS (left; Silver-stained) or HEK293T WDR33-HTBH cell lines (right; stained with SYPRO Ruby). The identities of some prominent protein bands are indicated to the right of each gel. Note that the band representing WDR33 is significantly more prominent relative to the background signal in the sample purified using an HTBH tag than in the preparation of the TAPS-tagged complex.

4.1.3 RBBP6 is not a stable subunit of CPSF in human cells

Due to its advantages in yield and purity, I focused on isolating endogenous CPSF from the stable homozygous WDR33-HTBH HEK293T cell line. I performed three pull-down experiments using slight variations of the purification protocol: I purified native CPSF either from whole cell extract or from nuclear extract prepared either by detergent lysis or mechanical homogenisation. To control for nonspecific protein binding to Strep-Tactin beads, I also carried out the purification procedure with the lysate from wild-type HEK293T cells. After each purification, I analysed the eluate by SDS-PAGE and determined its protein composition by tandem mass spectrometry (Figures 4.2C & 4.3A,B). The eluates from all three experiments contained the same dominant proteins, and so did the preparation of native CPSF purified from the HEK293T WDR33-TAPS cell line (Figure 4.3A). These data demonstrate that the results of the pull-downs were robust and not significantly influenced by variations in the purification protocol that may have affected protein-protein interactions. Mass spectrometry analyses revealed that all seven known CPSF subunits were the top hits in terms of the absolute number of peptides in the eluate in all four pull-down experiments (Figure 4.3A). All seven subunits could also be identified in the SDS-PAGE analysis of the eluate (Figure 4.2C). These results demonstrate the successful purification of native CPSF.

By purifying endogenous CPSF, I expected to identify some previously unknown stable interactors of the complex, but no novel protein was consistently present in every single preparation of native CPSF. However, endogenous CPSF did appear to co-purify under more than one experimental condition with several proteins that have not been previously identified as its interaction partners. These included protein disulfide isomerase (PDI) and SR proteins (SRSF6, SRSF7) that regulate splicing (Figure 4.3B) (165). Interestingly, all four subunits of the SNARP complex – BCLAF1, TRAP150, Pinin and SkIP – were detected in the preparations of native CPSF. BCLAF1 and TRAP150 are both ~100 kDa in molecular weight and may run on SDS-PAGE at the same height as CPSF100, which may explain why the band corresponding to the pseudonuclease appeared super-stoichiometric in SDS-PAGE analysis (Figure 4.2C). The SNARP complex is thought to regulate both transcription and splicing, and could potentially couple these processes to 3′ end processing (166). Some preparations of native CPSF also contained subunits of the TRiC chaperonin, which I previously found associated with recombinant mPSF expressed in insect cells (Figures 2.1B & 4.3B). Since TRiC facilitates protein folding, it was likely associated with native folding and assembly intermediates of CPSF. It will be interesting to determine if any of these proteins actually

interact with CPSF directly. The high absorbance ratio at 260 nm over 280 nm of the eluates suggested that native CPSF co-purified with significant amounts of cellular RNA, and some of the putative binding partners may bind this RNA instead of the CPSF complex itself. Treating the cell lysate with RNase during purification will be required to resolve this issue. In addition, a few peptides belonging to some of the proteins discussed above could be detected in negative control samples, in which the lysate from wild-type HEK293T cells was applied to Strep-Tactin beads in parallel with the extract from the tagged cell line (Figure 4.3B). To distinguish non-specific binding to beads from a genuine interaction with CPSF, quantitative proteomic analysis will be essential.

I also investigated if other proteins required for CPSF endonuclease activation (CStF, CFIIm, RBBP6) were stably associated with the endogenous complex. Neither of the two CFIIm subunits co-purified with endogenous CPSF (Figure 4.3A). In contrast, out of the three CStF subunit, only CStF64 reproducibly co-purified with native CPSF (Figure 4.3A). CStF64 has been shown to interact directly with symplekin (81, 167). Indeed, I could successfully purify a recombinant mCF complex bound to CStF64 upon co-expression of the three mCF subunits with CStF64 in Sf9 insect cells (Figure 4.3C). I tested the activity of the mCF-CStF64 complex in a cleavage assay, but CStF64 did not seem to have a noticeable effect on the endonuclease activity of CPSF (Figure 4.3D). Interestingly, binding of CStF64 to either symplekin or CStF77 has been proposed to be mutually exclusive (167). Thus, these alternative binding events might represent at least two different states of the 3′ end processing machinery. Further investigation will be required to determine the functional relevance of CStF64 binding to CPSF independently of the CStF complex.

Finally, and most importantly, no peptides corresponding to RBBP6 could be detected in any of the preparations of endogenous CPSF (Figure 4.3A). This suggests that, unlike in yeast, RBBP6 is not a stable component of the human CPSF complex. This may explain why the role of RBBP6 in activating the CPSF endonuclease has been overlooked for so long.

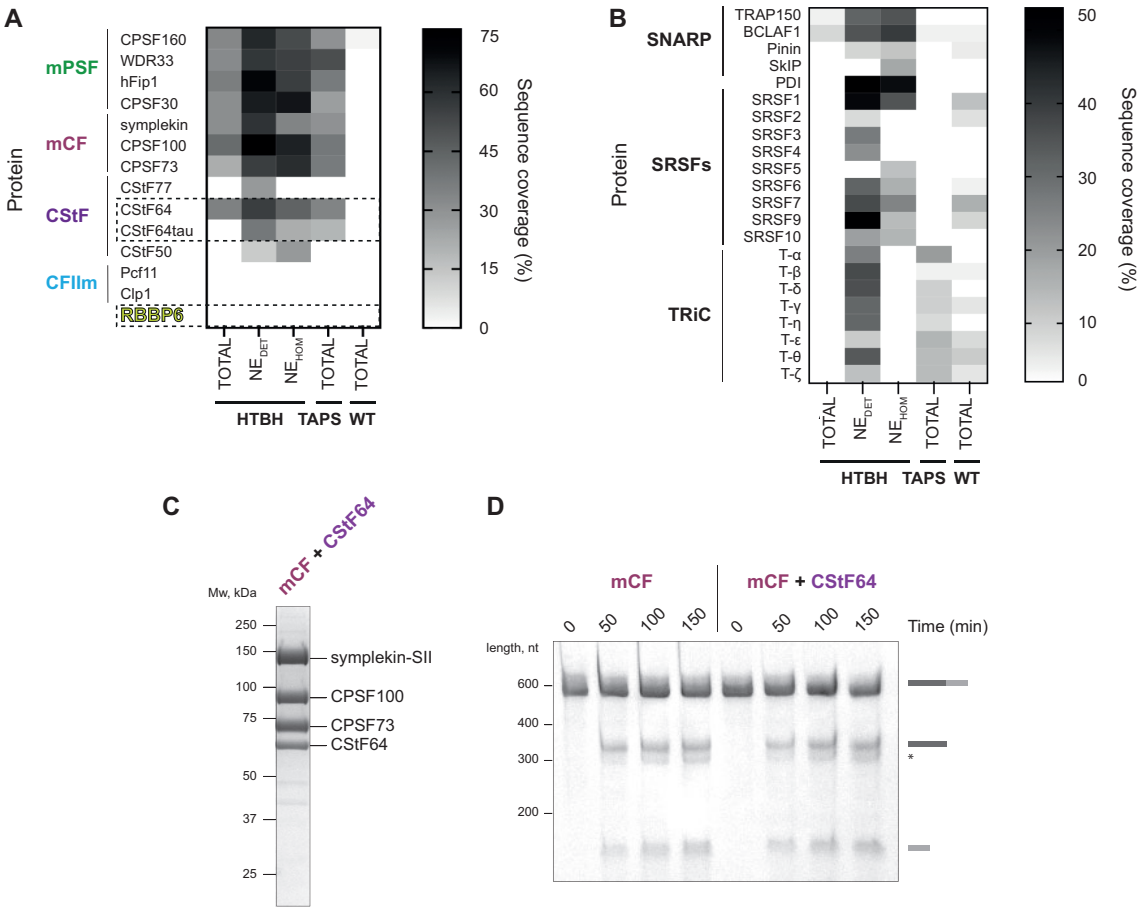


Figure 4.3 Subunit composition of endogenous CPSF. Heatmap showing sequence coverage as determined by mass spectrometry of all proteins required for CPSF endonuclease activity **(A)** and of potential novel interaction partners of CPSF **(B)** across various preparations of the endogenous complex from either HEK293T WDR33-HTBH or HEK293T WDR33-TAPS cells. The experimental conditions are indicated below each column of the heatmap. "TOTAL" indicates a purification from whole cell extract; NE_{DET} – from nuclear extract prepared using detergent lysis; NE_{HOM} – from nuclear extract prepared by mechanical homogenisation; WT – from whole cell extract of wild-type HEK293T cells as a negative control carried out in parallel with the purification from HEK293T WDR33-TAPS cells. PDI – protein disulphide isomerase; T – T-complex protein subunit of TRiC. **(C)** SDS-PAGE analysis of the purified recombinant mCF complex bound to CStF64. **(D)** CPSF cleavage assays of the L3 pre-mRNA substrate in the presence of either mCF complex or mCF co-expressed and co-purified with CStF64, demonstrating no noticeable difference in activity. The minor 5′ cleavage product of the L3 pre-RNA is marked with an asterisk.

4.2 Interactions between RBBP6 and CPSF

Despite its essential role in activating the endonuclease of CPSF, RBBP6 is not stably associated with the CPSF complex in human cells. This poses a question about how RBBP6 could contribute to activation of pre-mRNA 3′ end cleavage. RBBP6 has been implicated in regulating 3′ cleavage site selection, and I hypothesised that the binding of RBBP6 to the 3′ end processing complex could be a regulated step, instead of an obligate interaction, in the pathway towards CPSF endonuclease activation.

4.2.1 RBBP6 is recruited to CPSF in an RNA-dependent manner

I aimed to investigate the interactions between RBBP6 and CPSF with purified recombinant proteins by analytical gel filtration chromatography. Upon mixing mPSF, mCF and RBBP6, the two CPSF modules co-migrated on a size exclusion column, but RBBP6 eluted as a separate peak (Figure 4.4, black trace). This indicated the lack of an interaction between recombinant proteins, mirroring the result obtained from the purification of endogenous CPSF.

RBBP6 was previously detected as part of an RNA-associated 3′ end processing machinery in a post-cleavage state (6). Thus, I predicted that RBBP6 binding to CPSF could be dependent on RNA. To test this possibility, I mixed mPSF, mCF, RBBP6 and the 41 nt L3 RNA substrate, which carries a wild-type PAS site and mimics a cleaved pre-mRNA, and analysed this sample by analytical gel filtration chromatography. In the presence of RNA, a small substoichiometric amount of RBBP6 co-migrated with the RNA-bound CPSF complex, suggesting that RBBP6 may interact with CPSF in an RNA-dependent manner (Figure 4.4, green trace). Interestingly, this interaction was particularly sensitive to salt, and could only be observed in a buffer containing no more than 50 mM NaCl. RNA binding to CPSF was not noticeably affected by increasing the concentration of salt, suggesting that only the protein-protein contacts were disrupted. This observation could explain why the endonuclease activity of CPSF is inhibited in high ionic strength (Figure 3.2A).

CPSF has by far the highest affinity for RNA out of all the proteins required for 3′ endonuclease activation (159). The high affinity and specificity of the CPSF interaction with PAS suggests that this binding event may initiate the assembly of the active 3′ end processing machinery, including recruitment of RBBP6. Therefore, I tested whether the 41

nt L3 RNA in which the PAS sequence was mutated was capable of mediating the interaction between RBBP6 and CPSF. Despite the mutation in the PAS, RNA was bound by the CPSF complex, as indicated by gel electrophoresis and a high A_{260}/A_{280} ratio of the fractions containing CPSF (Figure 4.4, red trace). However, in spite of the presence of bound RNA, virtually no RBBP6 associated with the CPSF complex bound to mutant RNA (Figure 4.4). A difference in elution volume of the CPSF peaks in the presence of wild-type and mutant RNAs indicates that the two RNAs may bind to the complex differently: it is likely that the mutant RNA either binds to CPSF non-specifically or the substrate binds to other components of the complex instead of the PAS recognition site on mPSF. Altogether, it appears that PAS is crucial for the recruitment of RBBP6 to the CPSF complex.

The 41 nt L3 RNA is significantly longer than the six PAS nucleotides bound to mPSF. This means that RBBP6, which interacts with RNA on its own, albeit with a ~10-fold lower affinity than mPSF (159), may associate with the rest of the RNA in the gel filtration experiments and not directly with the CPSF complex. To exclude this possibility, I performed a pull-down experiment using the 520 nt MS2-tagged L3 pre-mRNA substrate that is efficiently cleaved by recombinant CPSF (for example, Figure 4.3D). An MBP-tagged MS2 protein, which specifically interacts with the MS2 loops of RNA, was immobilised on amylose beads, and the beads were then incubated with either RBBP6 alone, CPSF alone or both RBBP6 and CPSF together (Figure 4.5A). In each case, I washed the beads extensively, eluted the MBP-MS2 protein with maltose and analysed the proteins that were pulled down on the L3 pre-mRNA by Western blot using HRP-conjugated streptavidin to detect proteins tagged with a Strep-II tag. Multiple proteins in the sample, including RBBP6 and CPSF subunits symplekin and WDR33, carried a Strep-II tag, but they differed sufficiently in molecular weight to be easily distinguished from one another. I used the signal of the MBP-MS2 protein on the Ponceau-stained membrane as a loading control. On its own, CPSF co-purified with the L3 pre-mRNA, but RBBP6 failed to do so under the same experimental conditions (Figure 4.5B). However, in the presence of CPSF, RBBP6 was pulled-down by the tagged pre-mRNA substrate, suggesting that CPSF facilitates RBBP6 binding to RNA (Figure 4.5B). These results reinforce the previous observations from analytical gel filtration experiments and indicate that CPSF and RBBP6 interact in an RNA-dependent manner. In this way, RBBP6 associates with the 3′ end processing machinery only when CPSF assembles on the pre-mRNA substrate, which could explain how RBBP6 may activate the CPSF endonuclease without being a constitutive component of the human complex. The RNA dependence of this interaction likely stems from RBBP6 binding simultaneously, and

likely cooperatively, with both protein subunits and the RNA substrate, which increases the affinity of RBBP6 for the 3′ end processing complex.

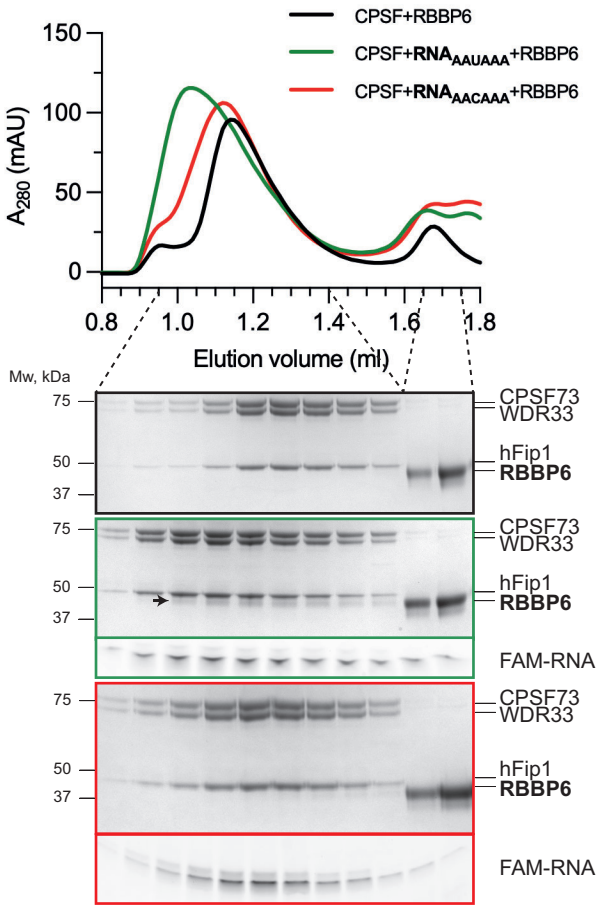


Figure 4.4 CPSF interacts with RBBP6 in an RNA-dependent manner. Chromatograms (top) of size exclusion runs of CPSF (2.5 μM) and RBBP6 (7.5 μM) in the presence and absence of either wild-type 41 nt L3 RNA (RNA_{AAUAAA}; 5μM) or RNA containing a mutant PAS sequence (RNA_{AACAAA}; 5μM), and SDS-PAGE analyses (bottom) of the peak fractions. The band of RBBP6 co-migrating with RNA-bound CPSF is indicated with a black arrow.

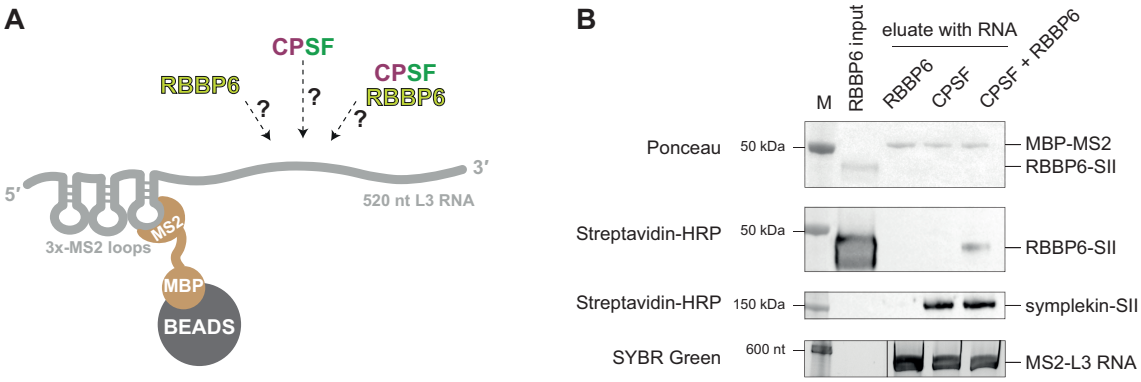


Figure 4.5 CPSF bound to RNA recruits RBBP6. (A) Schematic representation of the pull-down experiment using MBP-MS2 to isolate L3 pre-RNA and associated proteins. **(B)** Results of the pull-down experiment of RBBP6 by the MS2-tagged L3 RNA substrate in the presence and absence of CPSF. The samples were run on SDS-PAGE, transferred onto a membrane and analysed by Ponceau staining as well as Western blot against the Strep-II tag. RNA was analysed by gel electrophoresis separately. M – molecular weight marker.

4.2.2 Structural analysis of the CPSF complex bound to RNA and RBBP6

To understand the structural basis of how RBBP6 could form protein-protein interactions with CPSF in an RNA-dependent manner, I imaged the CPSF-RNA-RBBP6 complex by cryoEM. While the high-resolution structure of mPSF has been solved, only a low-resolution reconstruction of the mCF module of CPSF is available, most likely due to its non-uniform position relative to mPSF within the CPSF complex (8, 21–23). It is possible that the position of mCF may become more rigid in the presence of the RNA substrate and/or RBBP6. Thus, imaging CPSF bound to RNA and RBBP6 may not only reveal how RBBP6 interacts with CPSF subunits but may also improve our structural understanding of the mCF module.

Due to the low stoichiometry of RBBP6 within the target complex, I used chemical cross-linking with sulfo-succinimidyl-diazirine (sulfo-SDA) to increase its occupancy on RNA-bound CPSF. Sulfo-SDA is a highly controllable cross-linking agent containing two reactive chemical groups: a sulfo-succinimide ester and a diazirine group separated by a 3.9 Å-long spacer arm (Figure 4.6A). The presence of two distinct chemical groups within sulfo-SDA increases the specificity of cross-linking and reduces the number of cross-linking events, which prevents aggregation and denaturation artefacts, often caused by other cross-linkers such as glutaraldehyde. I incubated mCF, mPSF, RBBP6 and 41 nt L3 RNA, which mimics a cleaved pre-mRNA, with increasing concentrations of sulfo-SDA on ice, allowing the

succinimidyl ester group to react with amines presents on protein surfaces. The samples were subsequently exposed to UV light of a wavelength of 350 nm, activating the diazirine group, which can react with any amino acid side chain or peptide backbone site within the distance of the spacer arm. SDS-PAGE analysis of the cross-linked samples revealed that, with increasing concentrations of sulfo-SDA, the bands corresponding to uncross-linked proteins started to disappear and were gradually replaced by multiple smeary bands of a higher molecular weight which represented cross-linked complexes (Figure 4.6B). Various proteins displayed different cross-linking efficiency, with CPSF100 and CPSF73 present almost exclusively in cross-linked complexes at 1 mM sulfo-SDA. Importantly, hardly any free RBBP6 remained under these conditions, suggesting that most of RBBP6 was successfully cross-linked. I proceeded with the sample cross-linked with 1 mM sulfo-SDA and purified an intact complex from aggregates and excess cross-linker using size exclusion chromatography. I used the peak fraction of the CPSF-RNA-RBBP6 complex to prepare unsupported UltrAuFoil® cryoEM grids and imaged them using a Titan Krios transmission electron microscope equipped with a K3 detector in counting mode.

Analysis of the micrographs revealed well-dispersed particles which were easy to pick manually (Figure 4.6C). 2D class averages obtained from the manually picked particles were used as templates for automated particle picking in Relion 3.1. 714,444 particles were used for 2D classification, which revealed detailed class averages with clear secondary structure features (Figure 4.6D). The classes closely resembled the 2D projections of the mPSF module solved previously. Some additional fuzzy densities were noticeable around the known PAS binding site, which could either represent the RNA substrate or a protein component. After 3D classification, the class with the highest resolution features was refined, and a map of ~4.1 Å was obtained. Comparison with the known structures demonstrated that the map included electron densities for the mPSF module (three β propellers of CPSF160, one β propeller of WDR33, zinc fingers 1,2 and 3 of CPSF30), PAS RNA and the PIM peptide of CPSF100 (Figure 4.6E). Automated particle picking with the known low-resolution structure of mCF as a template, 2D and 3D classifications with a larger box size, 3D classifications without image alignment as well as focused 3D classifications did not reveal any additional electron densities for mCF, RBBP6 or RNA. Thus, cryoEM analysis of the CPSF-RNA-RBBP6 complex did not provide insights into how RBBP6 may interact with the rest of the CPSF complex.

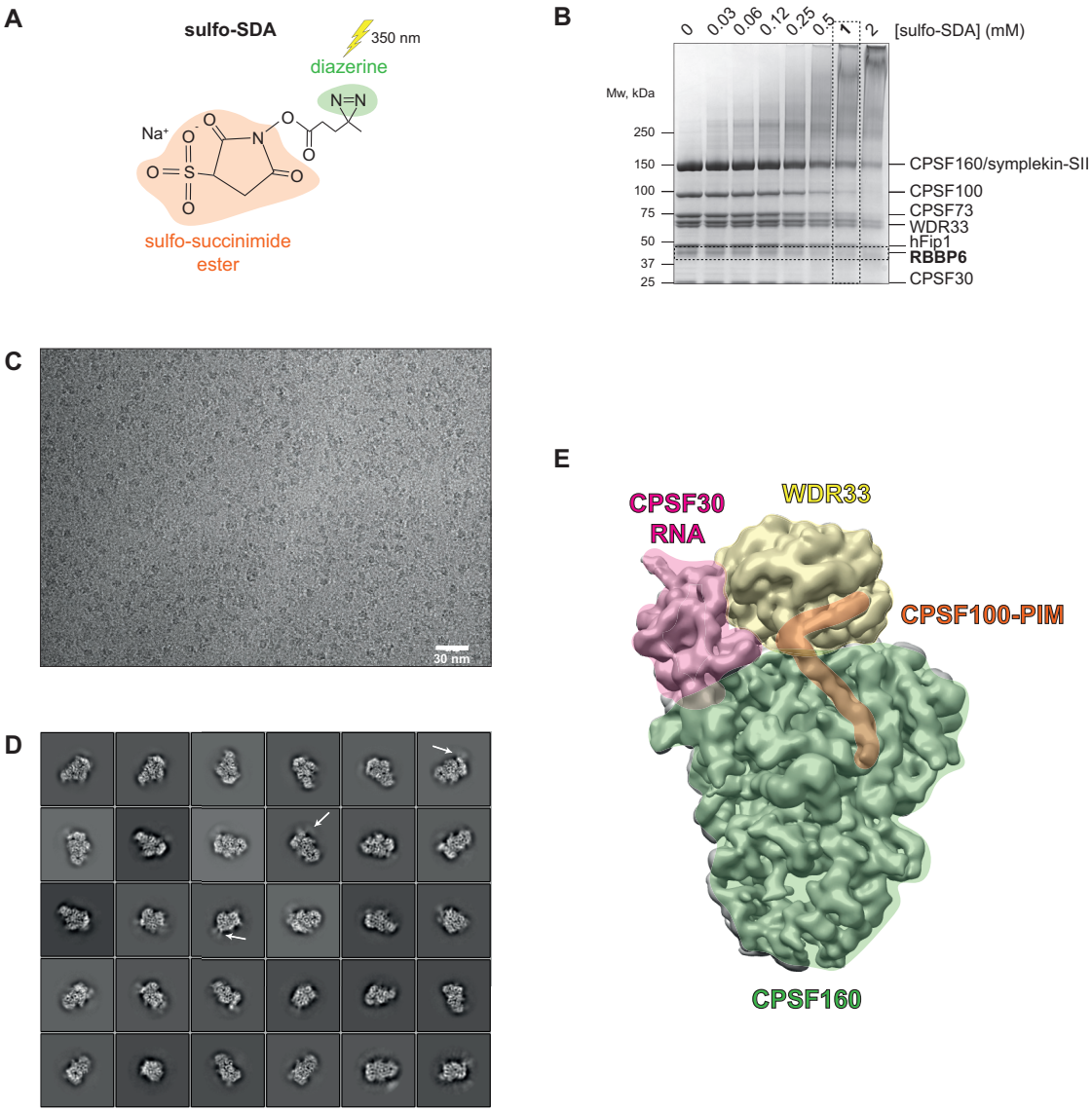


Figure 4.6 Structural characterisation of the CPSF-RBBP6-RNA complex. (A) Structure of the sulfo-SDA cross-linking agent. **(B)** SDS-PAGE analysis on a 3-8% Tris-acetate gel of the CPSF-RBBP6-RNA complex after cross-linking with various concentrations of sulfo-SDA. The concentration of sulfo-SDA used for cryoEM sample preparation (1 mM) as well as the band corresponding to uncross-linked RBBP6 are highlighted with a dashed box. **(C)** Representative micrograph of the sample containing the CPSF-RBBP6-RNA complex cross-linked with 1 mM sulfo-SDA. The sample was applied to the grid at a concentration of ~0.7 mg/ml. **(D)** 2D class averages of the sample. The mask diameter of each class is 18 nm. Fuzzy densities corresponding to RNA upstream and downstream of PAS are marked in some classes with a white arrow. **(E)** 3D reconstruction of mPSF from the sample containing the CPSF-RBBP6-RNA complex. The shading of the map indicates the approximate positions of the various subunits based on the comparison with the published models of the complex bound to RNA (PDB 6DNH) and CPSF100-PIM (PDB 6URG) (8, 21).

4.2.3 RBBP6 may interact with CPSF73

Since my attempts to determine the structure of the complete CPSF-RNA-RBBP6 complex by cryoEM were unsuccessful, I took a biochemical approach combined with *in silico* structural modelling and sequence conservation analysis to decipher how RBBP6 may interact with CPSF73.

I previously determined that the UBL domain of yeast Mpe1 interacts with the endonuclease domain of Ysh1, the yeast orthologue of CPSF73 (Figure 1.9B) (12). The crystal structure of the dimeric yeast complex revealed the residues that are critical for this interaction, and the sequence alignment between the yeast and human proteins demonstrated that these residues were highly conserved in the human orthologues RBBP6 and CPSF73 (Figure 4.7A). The structure of the human complex could be confidently modelled by both HADDOCK and AlphaFold algorithms (Figure 4.7B) (12). Thus, I aimed to test whether RBBP6 and CPSF73 could also interact *in vitro*. I co-expressed in Sf9 insect cells a Strep-II-tagged construct of RBBP6-UBL with either full-length CPSF73, its N-terminal endonuclease domain or its C-terminal domain, and performed a pull-down experiment on Strep-Tactin beads. Tagged RBBP6-UBL successfully co-purified with both full-length CPSF73 and its catalytic domain, suggesting that the UBL domain of RBBP6 and the endonuclease domain of CPSF73 may also interact in humans (Figure 4.7C).

Next, I wanted to investigate whether the interaction between the UBL and the endonuclease could be involved in recruiting RBBP6 to the CPSF complex. To this end, I mutated the residues of the UBL domain that based on sequence alignments and structural modelling should be involved in binding CPSF73 ($RBBP6_{D43K\ R74E}$), and tested whether mutant RBBP6 was able to associate with the RNA-bound CPSF complex (Figures 4.7A & D). I mixed mPSF, mCF, 41 nt L3 RNA and $RBBP6_{D43K\ R74E}$ together, and analysed the sample by analytical gel filtration chromatography. While wild-type RBBP6 co-migrated with the CPSF-RNA complex, the mutations in the UBL noticeably reduced its association, suggesting that disrupting the RBBP6-UBL interaction with the endonuclease abrogated RBBP6 recruitment to CPSF (Figure 4.7E). To test the functional significance of this result, I tested whether $RBBP6_{D43K\ R74E}$ could stimulate the endonuclease activity of CPSF in a cleavage assay. The RBBP6 mutant was completely incapable of activating the endonuclease, suggesting that the interaction between the UBL domain of RBBP6 and the endonuclease domain of CPSF73 is required for both RBBP6 binding to CPSF and its ability to activate 3′ end cleavage (Figure 4.7F).

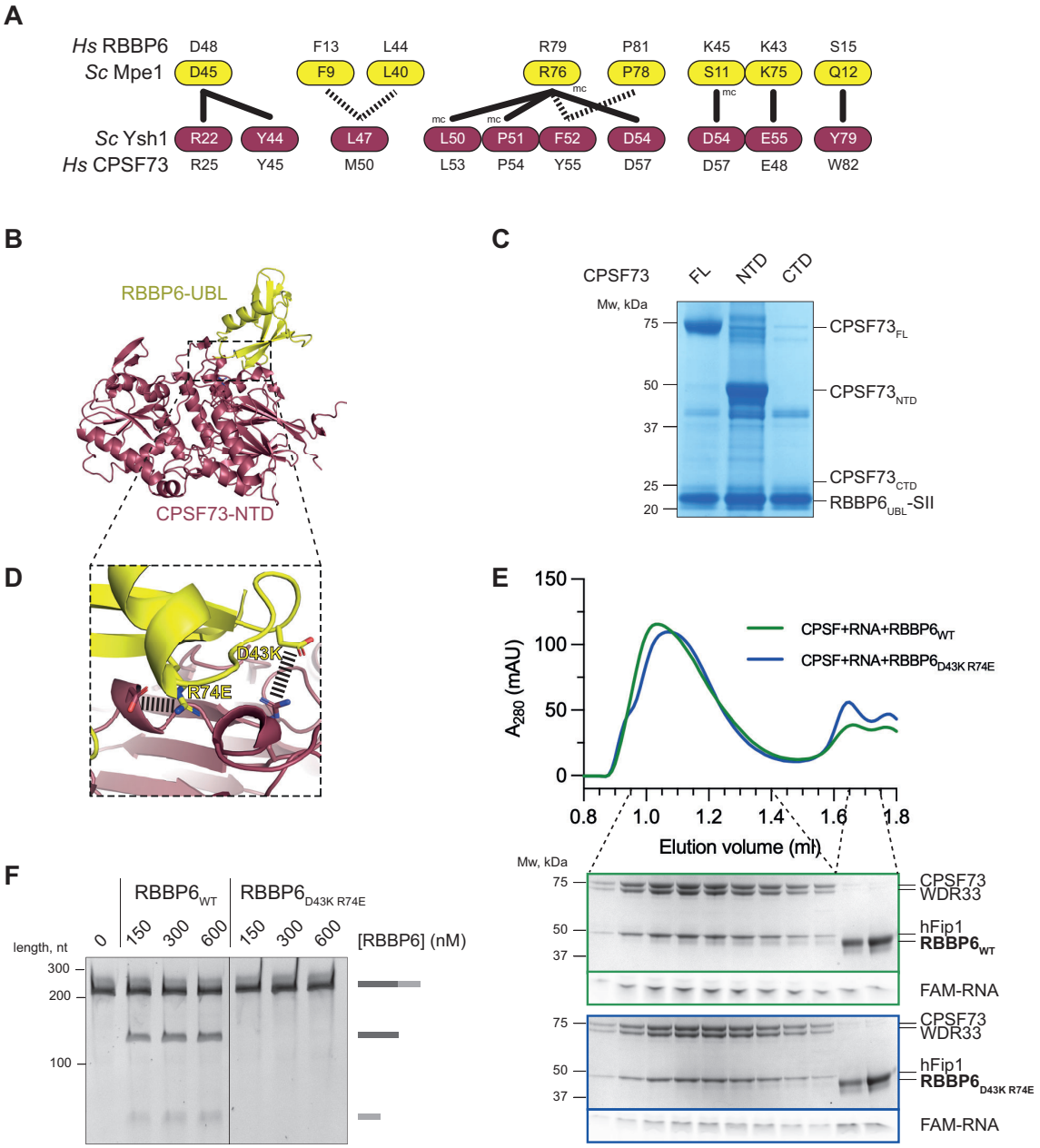

A
Hs RBBP6
Sc Mpe1
Sc Ysh1
Hs CPSF73
D48
D45
R22
Y44
R25
Y45
F13
L44
F9
L40
L47
M50
R79
P81
R76
P78
mc
L50
P51
F52
D54
L53
P54
Y55
D57
K45
K43
S15
S11
K75
Q12
D54
E55
Y79
D57
E48
W82
B
RBBP6-UBL
CPSF73-NTD
C
CPSF73
FL
NTD
CTD
Mw, kDa
75
50
37
25
20
CPSF73FL
CPSF73NTD
CPSF73CTD
RBBP6UBL-SII
D
D43K
R74E
E
150
100
50
0
A280 (mAU)
CPSF+RNA+RBBP6WT
CPSF+RNA+RBBP6D43K R74E
0.8
1.0
1.2
1.4
1.6
1.8
Elution volume (ml)
Mw, kDa
75
50
37
CPSF73
WDR33
hFip1
RBBP6WT
FAM-RNA
CPSF73
WDR33
hFip1
RBBP6D43K R74E
FAM-RNA
F
RBBP6WT
RBBP6D43K R74E
length, nt
0
150
300
600
150
300
600
[RBBP6] (nM)
300
200
100

Figure 4.7 RBBP6 interacts with CPSF73. (A) Schematic representation of the residues that mediate interactions between budding yeast Mpe1-UBL and Ysh1. Orthologous residues of human RBBP6 (top) and CPSF73 (bottom) are indicated. Solid lines represent ionic and hydrogen bonds, dashed lines – hydrophobic interactions. Mc – main chain. Based on (12). **(B)** AlphaFold Multimer model of the human complex between RBBP6-UBL and the N-terminal catalytic domain of CPSF73 (168). Prediction statistics are shown in (Appendix Figure 8.3B). **(C)** Pull-downs of SII-tagged UBL domain of RBBP6 in the presence of various constructs of CPSF73 from Sf9 insect cells. FL - full-length; NTD - N-terminal domain (residues 1-460); CTD - C-terminal domain (residues 461-684). **(D)** Close-up view of the RBBP6-CPSF73 interaction interface. Residues of RBBP6 that may form critical ionic interactions with CPSF73 and were mutated in the following experiments are indicated. **(E)** Size exclusion chromatograms (top) and SDS-PAGE analyses of the peak fractions (bottom) of the samples containing CPSF (2.5 μM) and 41 nt L3 RNA (5 μM) in the presence of either wild-type RBBP6 (WT) or RBBP6-D43K-R74E (5 μM). **(F)** Cleavage assays using the SV40 pre-mRNA substrate in the presence of either wild-type RBBP6 or $RBBP6_{D43K\ R74E}$.

4.2.4 RBBP6 may interact with mPSF

While I was working on the role of RBBP6 in human pre-mRNA 3′ end processing, a cryoEM structure of the yeast polymerase module bound to its yeast orthologue, Mpe1, and PAS RNA was determined (33). It showed how the resolved region of Mpe1 (residues 207-222 and 240-268) contacts Pfs2 (yeast orthologue of WDR33) and Cft1 (CPSF160) subunits as well as the PAS RNA. This region of Mpe1 was therefore termed the mRNA-sensing region (PSR) (Figure 1.2B&D, 2.10A). Specifically, a proline residue (P215) forms a CH-π bond with the RNA, while two aromatic residues (W257 and Y260) insert into a hydrophobic pocket of Pfs2 (Figure 4.8A). These newly identified interactions revealed that, in addition to the PIM peptide of Cft2 (33), Mpe1 may bridge nuclease and polymerase modules within yeast CPF.

It seemed likely that human RBBP6 could interact with the human mPSF module in a similar way. I used analytical gel filtration chromatography to demonstrate that RBBP6 could interact with the mPSF module in the absence of mCF, but that the binding was still dependent on the presence of a PAS-containing RNA (Appendix Figure 8.6). Sequence alignments showed that the proline that may contact PAS RNA and one of the two aromatic residues that may interact with WDR33 were highly conserved from yeast to humans (Figure 4.8B). The possibility that the putative PSR of RBBP6 may interact with mPSF and PAS was especially intriguing, because this interaction could account for the RNA dependence of the RBBP6 binding to CPSF and may also explain why mutating the PAS

sequence abolished RBBP6 binding (Figure 4.4). Thus, I individually mutated the proline (RBBP6$_{P195G}$) and tyrosine (RBBP6$_{Y228G}$) residues of RBBP6, and tested whether these mutants could be recruited to RNA-bound CPSF by analytical gel filtration chromatography. Both RBBP6 mutants failed to co-elute with the CPSF-RNA complex, indicating that contacts with mPSF and, crucially, PAS RNA are functionally conserved and required for RBBP6 recruitment (Figure 4.8C). In addition, an electrophoretic mobility shift assay demonstrated that none of the mutations of RBBP6 tested in gel filtration experiments (in either the UBL domain described above or the PSR) affected the ability of RBBP6 to bind the 41 nt L3 RNA on its own, highlighting that the mutations indeed interfere with direct protein-protein interaction with CPSF subunits (Figure 4.8D). I also tested RBBP6$_{P195G}$ and RBBP6$_{Y228G}$ mutants in a CPSF cleavage assay and observed that both proteins were almost completely incapable of activating the 3′ endonuclease, demonstrating the functional importance of the RBBP6-mPSF-RNA interaction in the mechanism of CPSF endonuclease activation (Figure 4.8E).

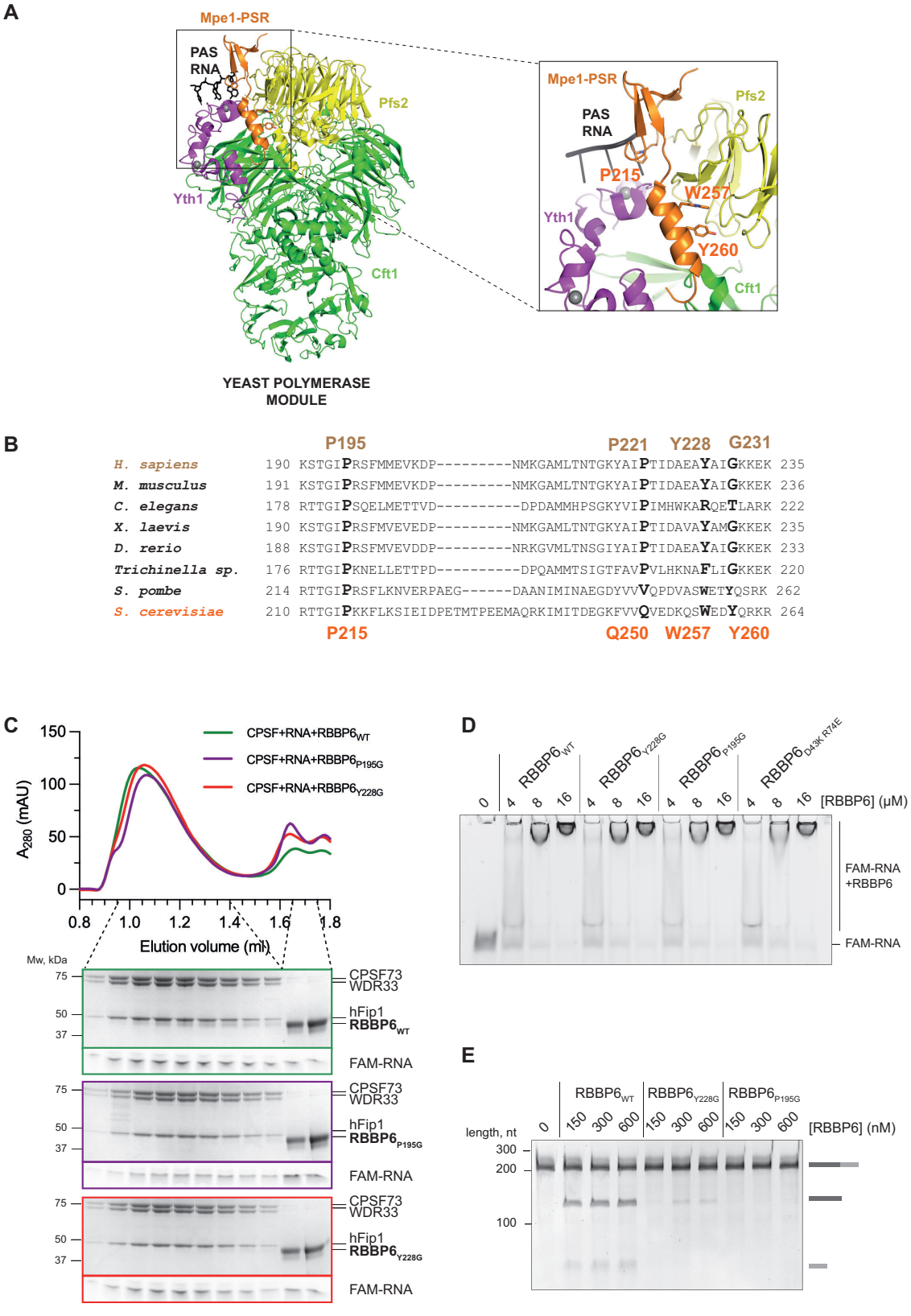

A
Mpe1-PSR
PAS
RNA
Pfs2
Yth1
Cft1
YEAST POLYMERASE
MODULE
Mpe1-PSR
PAS
RNA
Pfs2
P215
W257
Yth1
Y260
Cft1
B
P195
P221
Y228
G231
H. sapiens 190 KSTGIPRSFMMEVKDP---------NMKGAMLTNTGKYAIPTIDAEAYAIGKKEK 235
M. musculus 191 KSTGIPRSFMMEVKDP---------NMKGAMLTNTGKYAIPTIDAEAYAIGKKEK 236
C. elegans 178 RTTGIPSQELMETTVD----------DPDAMMHPSGKYVIPIMHWKARQETLARK 222
X. laevis 190 KSTGIPRSFMVEVEDP---------NMKGAMLTNTGKYAIPTIDAVAYAMGKKEK 235
D. rerio 188 KSTGIPRSFMVEVDDP---------NRKGVMLTNSGIYAIPTIDAEAYAIGKKEK 233
Trichinella sp. 176 RTTGIPKNELLETTPD----------DPQAMMTSIGTFAVPVLHKNAFLIGKKEK 220
S. pombe 214 RTTGIPRSFLKNVERPAEG------DAANIMINAEGDYVVVQPDVASWETYQSRK 262
S. cerevisiae 210 RTTGIPKKFLKSIEIDPETMTPEEMAQRKIMITDEGKFVVQVEDKQSWEDYQRKR 264
P215
Q250
W257
Y260
C
150
100
50
0
A280 (mAU)
CPSF+RNA+RBBP6WT
CPSF+RNA+RBBP6P195G
CPSF+RNA+RBBP6Y228G
0.8
1.0
1.2
1.4
1.6
1.8
Elution volume (ml)
Mw, kDa
75
50
37
CPSF73
WDR33
hFip1
RBBP6WT
FAM-RNA
RBBP6P195G
RBBP6Y228G
D
RBBP6WT
RBBP6Y228G
RBBP6P195G
RBBP6D43K R74E
0
4
8
16
[RBBP6] (μM)
FAM-RNA
+RBBP6
FAM-RNA
E
length, nt
300
200
100
0
150
300
600
[RBBP6] (nM)

Figure 4.8 RBBP6 may interact with the mPSF module. (A) Experimental structure of the yeast polymerase module bound to Mpe1 and PAS RNA (PDB 7ZGR; left) and a close-up view of the Mpe1 binding site (right) (33). **(B)** Sequence alignment of the PSR of RBBP6/Mpe1. The residues discussed in the text are indicated. **(C)** Size exclusion chromatograms (top) and SDS-PAGE analyses of the peak fractions of the samples containing CPSF (2.5 µM), 41 nt L3 RNA (5 µM) and either wild-type RBBP6 (WT), RBBP6$_{P195G}$ or RBBP6$_{Y228G}$ (7.5 µM). **(D)** Electrophoretic mobility shift assays of various concentrations of RBBP6$_{WT}$, RBBP6$_{Y228G}$, RBBP6$_{P195G}$ or RBBP6$_{D43K\ R74E}$ binding to the 41 nt L3 RNA substrate. **(E)** Cleavage assays using the SV40 pre-mRNA substrate in the presence of either RBBP6$_{WT}$, RBBP6$_{Y228G}$ or RBBP6$_{P195G}$.

4.2.5 Interactions between RBBP6/Mpe1-PSR and mPSF/polymerase module may differ in yeast and humans

Next, I investigated what the complex of mPSF bound to RBBP6 and PAS RNA may look like. First, I used AlphaFold Multimer to predict the structure of the mPSF-RBBP6 complex by simultaneously providing the sequences of all four protein subunits as input (168). AlphaFold Multimer confidently predicted the structure of the mPSF module, which closely matched the experimental data, but did not identify an interaction between RBBP6 and mPSF. PAS RNA appears to be essential for RBBP6 binding to CPSF *in vitro*. However, AlphaFold has not been trained to model nucleic acids, which could explain why it failed to predict this interaction. Instead, I analysed the predicted structure of RBBP6 PSR in isolation available in the AlphaFold Protein Structure Database (124). The model shows that the PSR of RBBP6 is indeed likely to adopt a very similar architecture to the one observed in yeast Mpe1 with three antiparallel β strands followed by a short α helix (Figure 4.9A). The high per-residue confidence score of the RBBP6 PSR suggests that the secondary structure elements in the model are predicted with high confidence (Figure 4.9B). I overlayed the structural model of RBBP6-PSR onto the experimental structure of yeast Mpe1 bound to the polymerase module. Surprisingly, the helix of RBBP6 turned out to be positioned at roughly 90° relative to its β strands, which contrasts with the Mpe1 helix that follows the three β strands at an almost straight angle (Figure 4.9A). To ensure that this structural difference was not coincidental, I examined the predicted alignment error (PAE) plot of the model of RBBP6 PSR. PAE provides a pairwise score, indicating how confidently the position of one residue was predicted relative to the position of another residue in the pair. The PAE plot of the RBBP6 PSR model showed that the relative positions of the α helix and the β strands were indeed predicted with high certainty (Figure 4.9C). The predicted

difference in the tertiary structure of the PSR in yeast and human proteins could be explained by Mpe1 PSR changing its conformation upon binding to the polymerase module. However, the AlphaFold prediction of Mpe1 PSR in isolation is in almost perfect agreement with its experimental structure determined in a complex, contradicting this possibility (Appendix Figure 8.3C). Interestingly, a comparison of the PAE plots shows that the relative positions of the β strands and the α helix are predicted with lower confidence in Mpe1 than in RBBP6, which may indicate the higher degree of flexibility of the yeast protein, but this hypothesis will be have to tested experimentally (Appendix Figure 8.3C).

Since a conformational change upon binding does not seem to explain the difference between yeast and human PSR domains, I hypothesised that a proline residue (P221) of RBBP6 located in the linker region that connects the α helix and the β and that is not conserved in yeast Mpe1 could be the cause of this difference in conformation: the proline may be responsible for introducing a kink in the linker, positioning the two secondary structure elements at a right angle in the human protein (Figures 4.8B & 4.9A).

I became interested in how the different positioning of the PSR helix in RBBP6 compared with Mpe1 could affect the binding of RBBP6 to mPSF in humans. Therefore, I compared the putative binding sites of RBBP6 to the human mPSF complex with that of Mpe1 to the yeast polymerase module. I overlayed the experimentally-determined structure of Mpe1 PSR and the AlphaFold model of RBBP6 PSR with either the known structure of human mPSF or the cryoEM structure of the yeast polymerase module (Figure 4.9D&E). Several notable observations could be made from these analyses. A structured loop belonging to CPSF30 (residues 22-38) occupies the conserved pocket between WDR33 and CPSF30 occupied by Mpe1 PSR in the yeast complex (Figure 4.9D). The yeast orthologue of CPSF30, Yth1, does not contain a loop in this position, leaving this space available for Mpe1 binding (Figure 4.9E). The helix in the AlphaFold model of RBBP6 PSR does not clash with the loop of CPSF30 but, based on the structure alignments, may instead contact several β strands on the surface of the WD40 domain of WDR33. Several aromatic and hydrophobic residues of WDR33 are positioned in the proximity of the RBBP6 helix, and could potentially represent an alternative, and likely weaker, binding surface for RBBP6 (Figure 4.9F).

However, it is still possible that RBBP6 could displace the CPSF30 loop to bind to the same site on mPSF as Mpe1. Competition between RBBP6 and the loop of CPSF30 could also account for the lower affinity of RBBP6 for mPSF than of their yeast orthologues. To address this possibility, I prepared a mutant mPSF complex in which the CPSF30 loop was truncated. The complex lacking the loop altogether tended to be degraded during purification, and hence, I chose to only delete a five-residue motif in the middle of the loop (mPSF-CPSF30$_{\Delta PLPFP}$) containing three proline residues that are likely responsible for its rigidity (Figure 4.9G&H). Truncation of the CPSF30 loop did not enhance RBBP6 binding to the complex, and the interaction was still RNA dependent. However, the CPSF-CPSF30$_{\Delta PLPFP}$ complex displayed a reproducibly higher endonuclease activity at a given concentration of RBBP6 than wild-type CPSF in endonuclease assays *in vitro* (Figure 4.9I). These results are consistent with the possibility that RBBP6 may compete with the CPSF30 loop for binding to mPSF. However, for these experiments to be conclusive, it will be necessary to determine whether the five-residue deletion does actually open up the putative RBBP6 binding site. Overall, mapping the precise location of RBBP6 binding to the human mPSF complex will require further investigation.

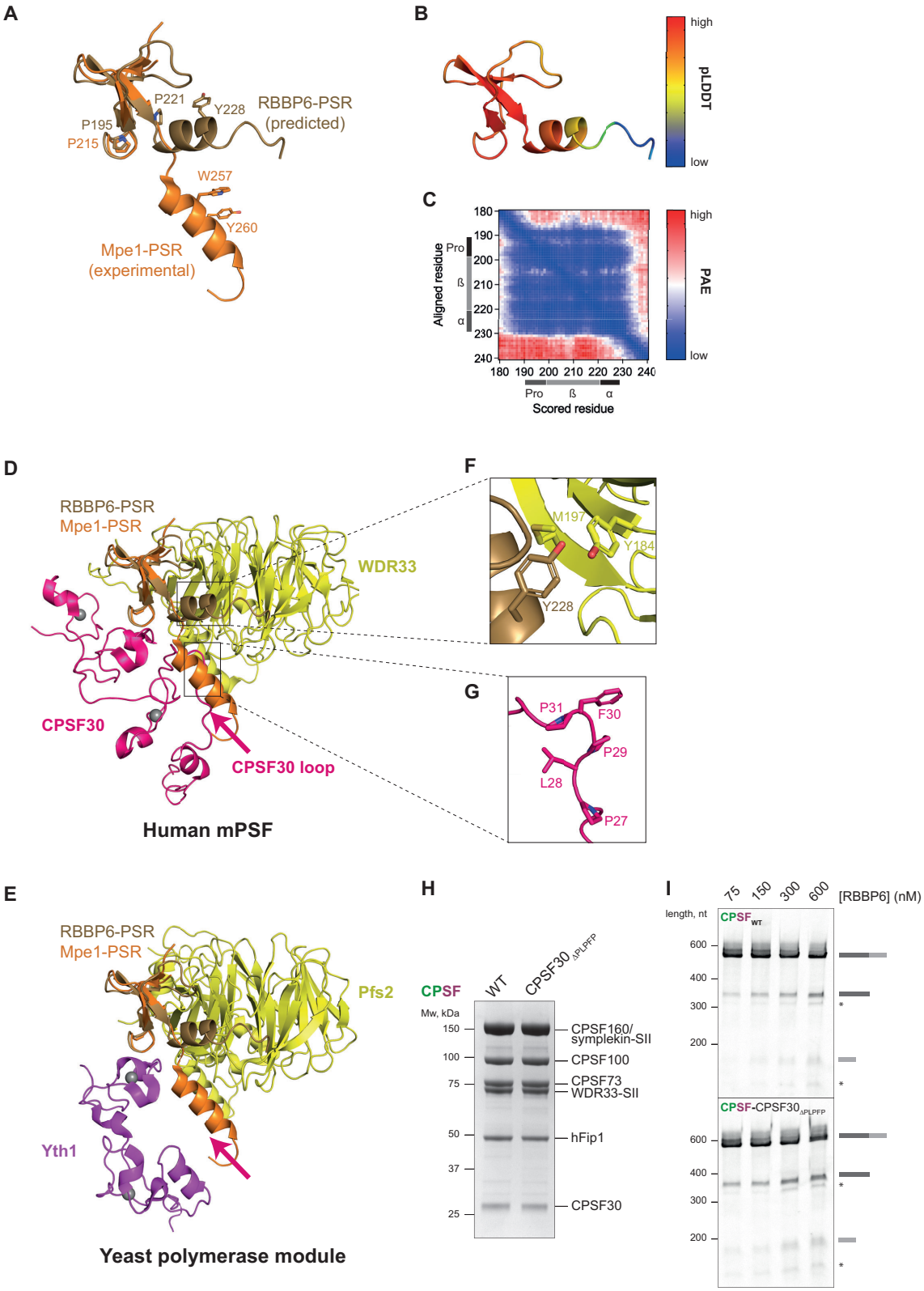

A
P221
Y228
RBBP6-PSR
(predicted)
P195
P215
W257
Y260
Mpe1-PSR
(experimental)
B
high
pLDDT
low
C
Aligned residue
Pro
ß
α
180 190 200 210 220 230 240
Scored residue
high
PAE
low
D
RBBP6-PSR
Mpe1-PSR
WDR33
CPSF30
CPSF30 loop
Human mPSF
F
M197
Y184
Y228
G
P31
F30
P29
L28
P27
E
RBBP6-PSR
Mpe1-PSR
Pfs2
Yth1
Yeast polymerase module
H
CPSF
WT
CPSF30ΔPLPFP
Mw, kDa
150
100
75
50
37
25
CPSF160/
symplekin-SII
CPSF100
CPSF73
WDR33-SII
hFip1
CPSF30
I
75
150
300
600
[RBBP6] (nM)
length, nt
CPSFWT
600
400
300
200
CPSF-CPSF30ΔPLPFP
600
400
300
200

Figure 4.9 Interactions between RBBP6/Mpe1 and mPSF/polymerase module may differ between yeast and humans. (A) Overlay of the experimental structure of Mpe1 PSR (PDB 7ZGR) and the AlphaFold prediction of RBBP6 PSR (33, 124). Residues discussed in the text are indicated. **(B)** Predicted structure of RBBP6 PSR coloured according to the per residue pLDDT score, demonstrating the high confidence of the model. **(C)** Predicted alignment error (PAE) plot of the predicted structure of RBBP6 PSR, showing that the positions of its secondary structure elements relative to each other are predicted with high certainty. Pro – loop containing the proline that may contact PAS RNA; β – β sheets of the PSR; α - C-terminal α helix of the PSR. The experimental Mpe1 PSR structure and the AlphaFold prediction of RBBP6-PSR were overlayed with the experimental structures of the human mPSF complex (PDB 6DNH) **(D)** and of the yeast polymerase module **(E)** (PDB 7ZGR) (21, 33). CPSF160/Cft1 subunits were removed for clarity. **(F)** Close-up view of the potential hydrophobic interactions between the conserved aromatic residue of RBBP6 PSR (Y228) and adjacent residues of WDR33 (M197 and Y184). Y228 of RBBP6 may also form a hydrogen bond with Y184 of WDR33. **(G)** Close-up view of the CPSF30 loop that may block the binding of RBBP6 to mPSF. The residues that were deleted in the CPSF-CPSF30$_{\Delta PLPFP}$ mutant complex are indicated. **(H)** SDS-PAGE analyses of purified wild-type CPSF and of the CPSF-CPSF30$_{\Delta PLPFP}$ mutant complex. Equal molar amounts of each complex were loaded, indicating that the concentrations were estimated correctly. Bands of wild-type CPSF30 and CPSF30$_{\Delta PLPFP}$ run at the same height. **(I)** Cleavage assays of either wild-type CPSF or CPSF-CPSF30$_{\Delta PLPFP}$ using the L3 pre-mRNA substrate in the presence of increasing concentrations of RBBP6. Asterisk denotes minor cleavage products.

4.2.6 mPSF may hinder RBBP6 binding to CPSF73 in the context of the full CPSF complex

So far in this Chapter, I demonstrated that RBBP6 interacts with CPSF in an RNA-dependent manner by contacting both the mPSF subunit WDR33 and the mCF subunit CPSF73 (Figure 4.7 & 4.8). Surprisingly, a subsequent gel filtration experiment showed that mCF alone could interact with RBBP6 in the absence of RNA, while full CPSF failed to do so (Figures 4.4 & 4.10). In addition, I took a look back at my previous results from pull-down experiments, in which I co-expressed in insect cells tagged RBBP6 with either mCF or mPSF individually or both modules together. I noticed that the presence of mPSF noticeably reduced the association of mCF subunits with RBBP6 (Figure 2.10B). Taken these observations together, I hypothesised that binding of mCF to mPSF may sterically hinder the RBBP6 UBL binding site on the endonuclease subunit. Indeed, RBBP6 was able to bind to the mCF variant lacking the PIM peptide in the presence of mPSF, highlighting that a direct interaction between mPSF and mCF prevents RBBP6 binding to CPSF73 in the context of the full CPSF complex (Figure 4.10).

Interestingly, the yeast orthologue of CPSF100, Cft2, was recently proposed to hinder the binding of Mpe1 PSR to the yeast polymerase module (33). This observation together with the data presented here strongly point towards the requirement of a large-scale conformational rearrangement within CPSF to activate the endonuclease. It is likely that in the absence of RNA, CPSF may adopt an inhibited conformation. The recognition of the PAS sequence by mPSF and RNA binding to the complex may alter the conformation CPSF towards its active state that is compatible with RBBP6 simultaneously binding to both mPSF and mCF modules. The precise nature of such conformational changes, however, beg further examination.

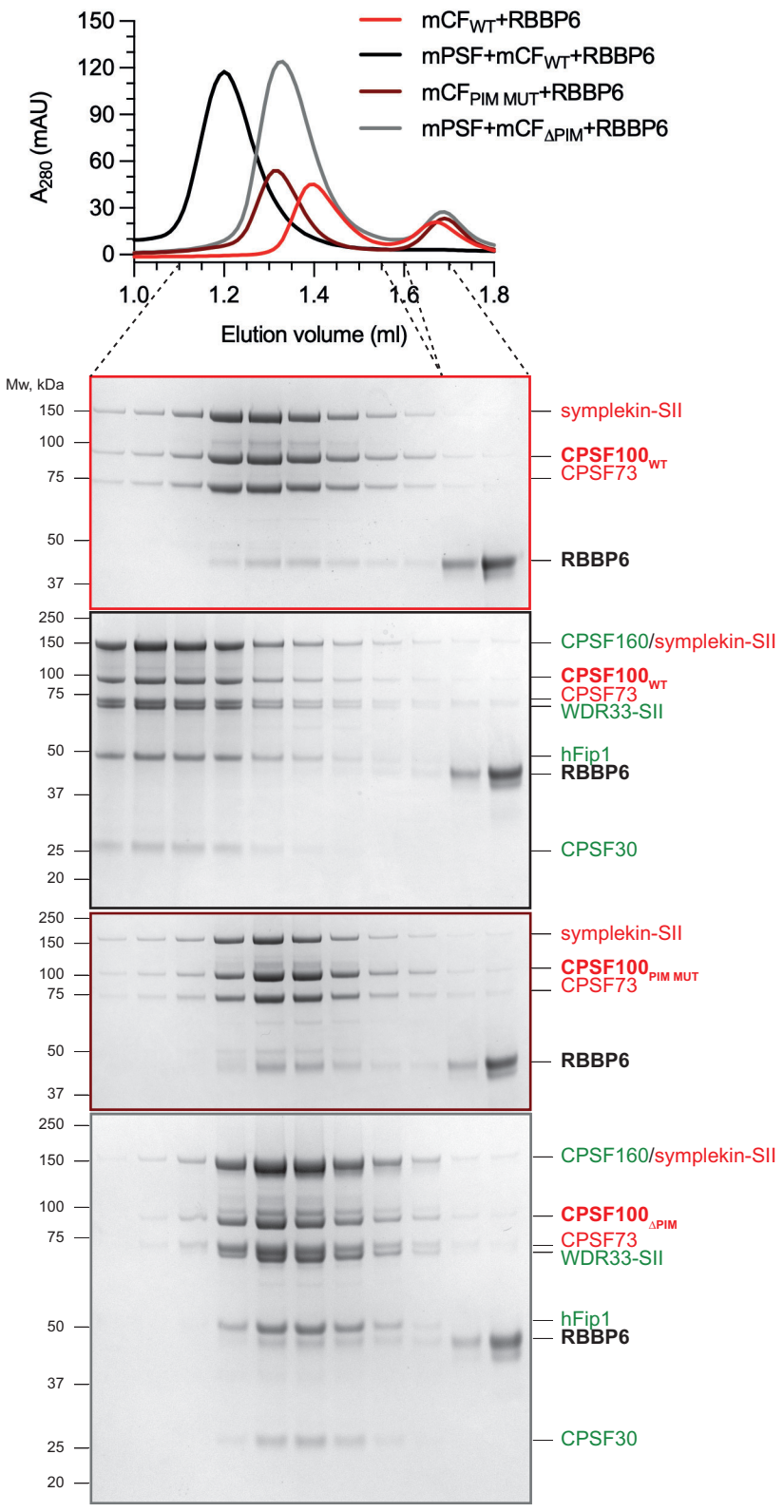


Figure 4.10 mPSF may block RBBP6 binding to mCF in the context of full CPSF. Size exclusion chromatograms (top) and SDS-PAGE analyses of the corresponding peak fractions (bottom) RBBP6 mixed with either wild-type (mCF$_{WT}$), mCF-CPSF100$_{F464A/W473A/Y476A}$ (mCF$_{PIM\ MUT}$) or mCF lacking the PIM peptide (mCF$_{\Delta PIM}$) either in the presence or absence of mPSF.

4.3 Interactions between RBBP6 and cleavage factors

My mutational and sequence analyses as well as structure predictions suggest that RBBP6 may interact with CPSF in an RNA-dependent manner by contacting the CPSF73, WDR33 subunits and RNA. However, accessory protein complexes CStF and CFIIm are also required for CPSF endonuclease activity, and may interact with RBBP6 to collectively facilitate activation of the 3′ endonuclease. Thus, in this section, I investigated potential interactions between RBBP6 and human cleavage factor complexes: CStF and CFIIm.

4.3.1 RBBP6 does not interact with CStF but binds weakly to CFIIm

Previous studies identified only one potential interaction between RBBP6 and the subunits of human cleavage factor complexes: the UBL domain of RBBP6 was proposed to interact with a CStF subunit CStF64 independent of RNA (44). This interaction was demonstrated by a co-immunoprecipitation experiment from mammalian nuclear extract. However, nuclear extract contains other proteins that could bridge RBBP6 and CStF64, and the two proteins may co-precipitate without interacting directly. To test direct binding, I used purified recombinant proteins and analytical gel filtration chromatography. First, I tested the sample containing only the UBL domain of RBBP6 and an intact CStF complex. Mixing CStF and RBBP6 did not change the elution volume of either component compared with the gel filtration runs of each protein individually (Figure 4.11A). SDS-PAGE analyses showed that CStF and RBBP6 did not co-elute from the gel filtration column, suggesting that they did not interact (Figure 4.11A). This experiment was conducted before buffer conditions in a cleavage assay were optimised, and the gel filtration buffer contained 150 mM NaCl, which would inhibit CPSF (Figure 3.2A). Therefore, I repeated the experiment in a buffer containing 50 mM salt. In addition, the RBBP6 construct competent to activate cleavage ($RBBP6_{1\text{-}335}$) was used instead of RBBP6 UBL alone, in case other domains of the protein could also contribute to the interaction with CStF. The analytical gel filtration run of the sample containing $RBBP6_{1\text{-}335}$ and CStF revealed that the two components eluted as separate peaks, suggesting the lack of a stable interaction (Figure 4.11A). Overall, I could not detect direct binding between RBBP6 and the CStF complex, and at the same time CStF64, which contradicts previous observations. However, the interaction of RBBP6 with isolated CStF64 was not tested, and it cannot be excluded that the binding site for RBBP6 could be blocked in the context of the fully assembled CStF complex.

Next, I explored whether RBBP6 could interact with CFIIm, which has not been previously investigated. I mixed RBBP6 and CFIIm and ran the sample on an analytical gel filtration column. While RBBP6 and CFIIm largely eluted in two separate peaks, a small substoichiometric amount of RBBP6 co-migrated with CFIIm (Figure 4.11C). This result was reminiscent of RBBP6 binding to the CPSF-RNA complex (Figure 4.4). However, in the former experiment no RNA was present, suggesting that RBBP6 binding to CFIIm does not require RNA substrate.

Overall, my analytical gel filtration studies revealed that, unlike previously reported, RBBP6 does not form protein-protein interactions with CStF but may instead interact directly, albeit weakly, with CFIIm. It will be interesting to test whether this binding could be important for CPSF endonuclease activation.

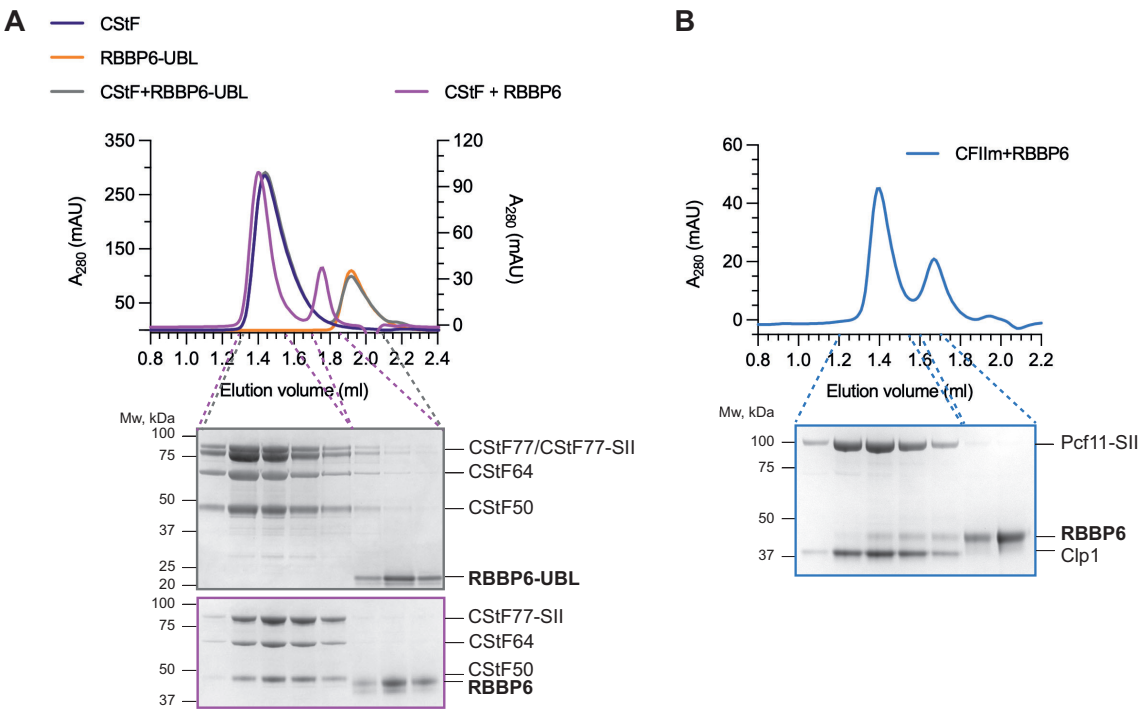


Figure 4.11 RBBP6 does not interact with CStF but may bind CFIIm. Gel filtration chromatograms (top) of samples containing either the UBL domain of RBBP6 (30 µM) mixed with CStF (12 µM) or RBBP6$_{1\text{-}335}$ (7.5 µM) mixed with the CStF complex (2.5 µM) **(A)**, and of RBBP6 (7.5 µM) mixed with CFIIm (4 µM) **(B)**. SDS-PAGE analyses of the corresponding peak fractions are depicted below the chromatograms.

4.3.2 mCF and CFIIm form a stable complex

In addition to the interactions involving RBBP6, understanding how the other two protein complexes required for CPSF endonuclease activation, CStF and CFIIm, may interact with CPSF could also provide significant insights into the activation mechanism of the 3′ endonuclease. CStF has been shown to bind to at least two subunits of CPSF: CStF64 interacts with an mCF subunit symplekin, while the dimer of CStF77 contacts CPSF160 and hFip1 within the mPSF module (8, 26, 167, 169). On the other hand, direct binding of CFIIm to the CPSF complex has never been demonstrated, and hence, I aimed to explore these interactions using analytical gel filtration chromatography.

I tested the binding of CFIIm to both CPSF modules individually. The chromatogram and SDS-PAGE analysis of the sample containing CFIIm and mPSF showed that the two complexes eluted at a very similar elution volume (Figure 4.12A). However, the runs of each complex separately indicated that mixing the two components together did not change the fraction in which the complexes eluted from the column, suggesting that CFIIm and mPSF do not interact. In contrast, mixing CFIIm with mCF caused a noticeable leftward shift in the elution volume of each complex, and SDS-PAGE analyses confirmed that CFIIm and mCF form a stable stoichiometric complex (Figure 4.12B). To my knowledge, this is the first demonstration of a direct interaction between these two protein complexes. mCF carries the endonuclease subunit CPSF73, and it is therefore possible that CFIIm could be directly involved in the conformational change that opens up the endonuclease active site for pre-mRNA 3′ cleavage.

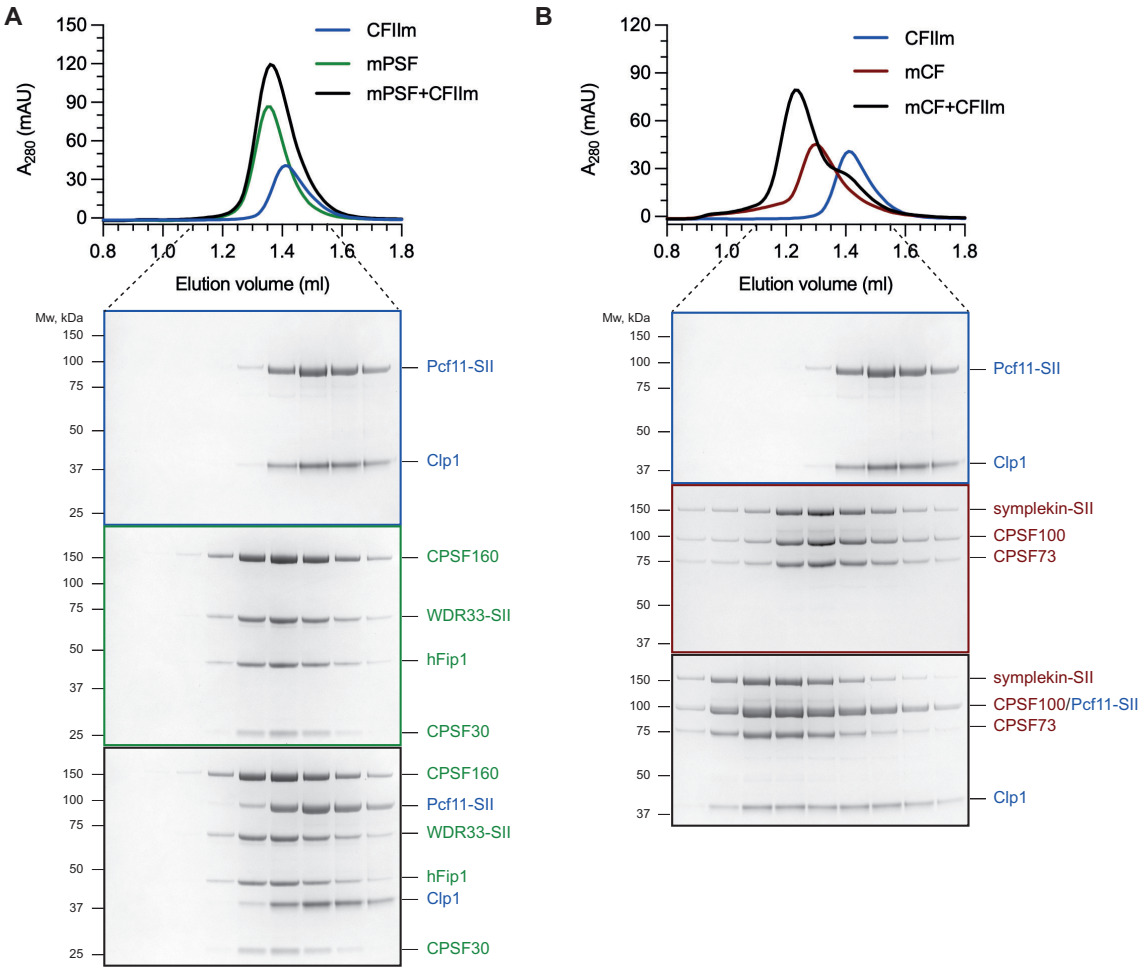


Figure 4.12 CFIIm and mCF form a stable complex. Gel filtration chromatograms (top) of samples containing either CFIIm (4 µM) and mPSF (2.5 µM) **(A)**, or CFIIm (4 µM) and mCF (2.5 µM) **(B)**. SDS-PAGE analyses of the corresponding peak fractions are depicted below the chromatograms.

4.3.3 RBBP6 interacts with the mCF-CFIIm complex in the absence of RNA

I showed that RBBP6 interacts weakly with the CFIIm complex, and that CFIIm in turn forms a stable complex with mCF. RBBP6 also binds the mCF module via an interaction between the UBL domain of RBBP6 and the endonuclease domain of CPSF73 (Figure 4.7). Therefore, I aimed to test whether RBBP6 could also interact with the mCF-CFIIm complex. Indeed, in an analytical gel filtration experiment, the sample containing RBBP6, mCF and CFIIm demonstrated almost stoichiometric binding of RBBP6 to the mCF-CFIIm complex (Figure 4.13A). This result was in contrast with only a small substoichiometric amount of RBBP6 co-eluting with either mCF or CFIIm individually from a gel filtration column (Figure 4.4). It is important to note that the ability of RBBP6 to interact with mCF-CFIIm did not depend on the pre-mRNA substrate.

Based on these observations, I hypothesised that improved RBBP6 stoichiometry was consistent with two possibilities. A single molecule of RBBP6 may simultaneously contact the binding sites on both mCF and CFIIm, cooperatively enhancing the affinity of RBBP6 for the mCF-CFIIm complex. It cannot be excluded that the low affinity binding sites on mCF and CFIIm could be independent of one another, and one copy of RBBP6 could be able to bind to each individual complex resulting in up to two RBBP6 molecules bound per mCF-CFIIm complex. More in-depth structural and biochemical studies will be required to elucidate which case represents the architecture of the mCF-CFIIm-RBBP6 complex.

4.3.4 Structural studies of the mCF-CFIIm-RBBP6 complex

Next, I aimed to determine the architecture of the protein complex containing mCF, CFIIm and RBBP6. The structure of mCF has been determined, and it was also known how Pcf11 interacts with Clp1 within the CFIIm complex (8, 82, 170). I also assumed that RBBP6 UBL should interact with CPSF73 as it does in the context of CPSF. Indeed, RBBP6 mutants that disrupt this interaction showed reduced association with the mCF-CFIIm complex, while mutations in the PSR had no effect on the binding (Appendix Figure 8.7). Despite all this knowledge, it was still unclear how mCF, CFIIm and RBBP6 come together to form a stable complex, and I sought to determine its structure by cryoEM. SDS-PAGE analysis of the peak size exclusion fraction showed that each subunit of the complex was present at almost stoichiometric amounts (Figure 4.13A). Therefore, the peak fraction was used directly to

prepare unsupported UltrAuFoil® cryoEM grids without chemical cross-linking. The grids were then imaged with a Titan Krios electron microscope equipped with a K3 detector.

A relatively small dataset of 320 micrographs was collected, revealing plentiful protein particles, although individual particles were difficult to discern by eye (Figure 4.13B). mCF represents the majority of protein mass in the complex, and its dimensions are known (8, 82). Hence, I manually picked the particles that were at least as big as the mCF module (> 15 nm in diameter), and used 2D class averages reconstructed from these particles as templates to perform automated picking in Relion 3.1. 2D classification was performed on the 47,980 auto-picked particles. The resultant 2D class averages were closely reminiscent of the shape of mCF consisting of three lobes, representing CPSF73, CPSF100 and the C-terminal domain of symplekin (Figure 4.13C). No obvious density that could attributed to either CFIIm or RBBP6 was observable in the 2D class averages. The 2D classes, however, did not exhibit any secondary structure features, and an interpretable 3D electron density map could not be reconstructed. Many factors could account for the poor quality of particle alignment, including inherent flexibility of the complex and denaturation of protein upon vitrification. Overall, the structure of the mCF-CFIIm-RBBP6 complex could not be determined, at least under the current conditions of sample and grid preparation.

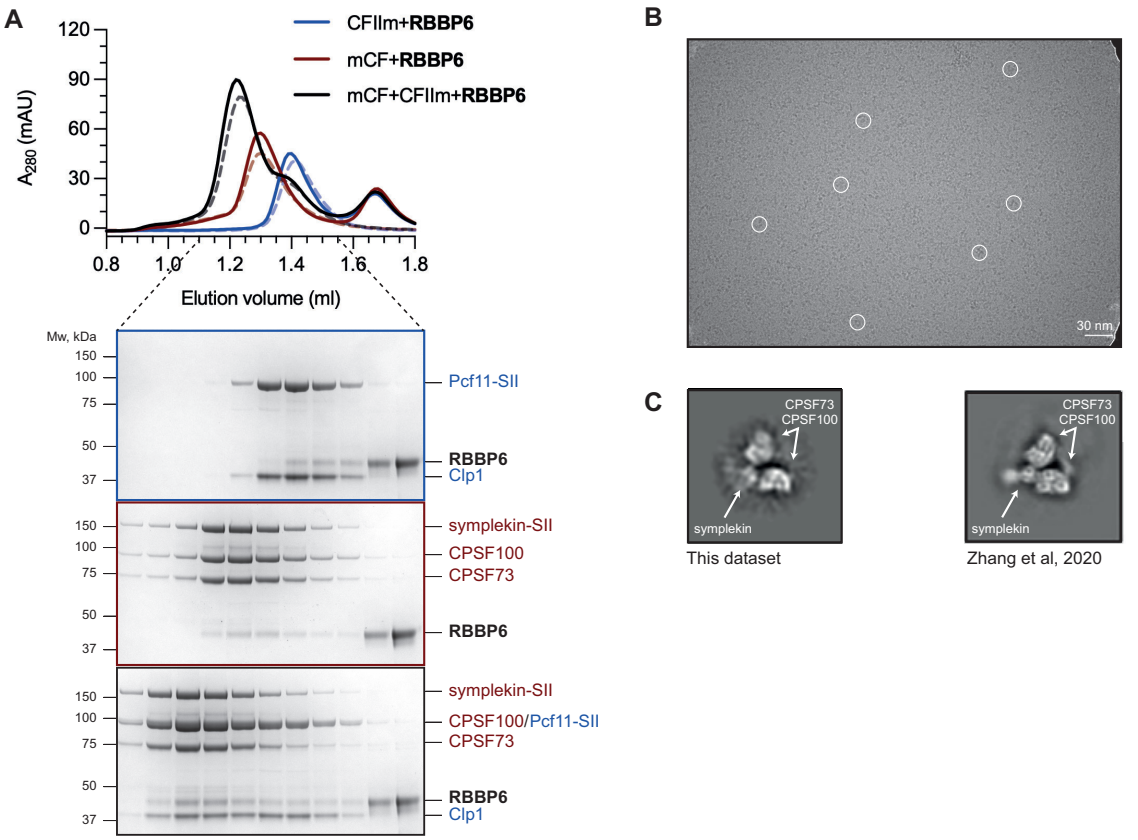


Figure 4.13 RBBP6 stably interacts with the CFIIm-mCF complex. (A) Size exclusion chromatograms (top) of the samples containing RBBP6 (7.5 μM) mixed with either CFIIm (2.5 μM) or mCF (2.5 μM) individually, or both complexes together. The positions of the CFIIm and mCF peaks in the absence of RBBP6 are shown as faded dashed curves. SDS-PAGE analyses of the peak fractions are depicted below the chromatograms. **(B)** Representative cryoEM micrograph of the mCF-CFIIm-RBBP6 complex. Several particles are circled. The diameter of the circle is 16 nm. **(C)** Representative 2D class average from the mCF-CFIIm-RBBP6 dataset (left) compared with published a 2D class of the mCF complex (right) (8). The diameter of the circular mask is 20 nm. It cannot be determined which of the two lobes corresponds to CPSF73 and which to CPSF100 in either image.

4.3.5 Identification of the interaction sites in the mCF-CFIIm-RBBP6 complex

Since my attempts to solve the structure of the mCF-CFIIm-RBBP6 complex experimentally were unsuccessful, I resorted to structure prediction by AlphaFold Multimer to decipher the architecture of this newly identified protein complex. Due to the limit on the number of residues that can be handled by the algorithm, the sequences of all six subunits could not be processed together at once. Instead, I ran AlphaFold Multimer to systematically search for potential pairwise interactions between individual proteins within the complex. I identified two novel binary interactions using AlphaFold Multimer that were predicted with high confidence, as indicated by both high values of the per-residue confidence score (pLDDT) and low values of PAE measured between the predicted interaction domains (Figure 4.14A; Appendix Figure 8.3D&E).

First, Pcf11 appears to interact with the pseudonuclease CPSF100. Residues 1428-1485 of Pcf11 are predicted to adopt a β-sheet-rich fold, which forms extensive interactions along the surface of CPSF100, contacting its C-terminal, metallo-β-lactamase and β-CASP domains (Figure 4.14A). This prediction was corroborated by analytical gel filtration experiments: a CFIIm complex containing a construct of Pcf11 that encompasses the complete putative CPSF100 interaction motif (residues 1340-1555) efficiently co-eluted with mCF and RBBP6. In contrast, CFIIm carrying only the Clp1-binding motif of Pcf11 (residues 1368-1443) did not (Figures 4.15). A closer inspection of the predicted interaction interface revealed that a highly negatively-charged loop of Pcf11 connecting two anti-parallel β-sheets (amino acid sequence WDEEEEEW; residues 1454-1461) is likely to bind to a positively-charged patch formed by combined surfaces of β-CASP and metallo-β-lactamase domains of CPSF100, while the indicated tryptophan residues of Pcf11 may form cation-π interactions (Figure 4.14B&C). The residues of both Pcf11 and CPSF100 that may mediate these interactions appear to be conserved in other eukaryotic species, including budding yeast, which further supports the prediction (Figure 4.14D). Interestingly, the predicted Pcf11 binding site on CPSF100 does not overlap with the binding site of symplekin-NTD in the histone 3′ end processing complex (compare Figures 4.14A & 3.14A). However, to validate the predicted structure, the Pcf11 construct containing only the putative CPSF100 interaction motif as well as point mutants on either side of the interaction interface will have to be tested.

In addition, the statistical parameters of AlphaFold prediction indicated that Clp1 may interact with the UBL domain of RBBP6 (Figure 4.14A & Appendix Figure 8.3E). The likely binding surfaces of CPSF73 and Clp1 do not overlap and are located on the opposite sides of the UBL domain. A closer look at the Clp1-RBBP6 interface suggested that the binding between the two protein could be stabilised primarily by hydrogen bonds (Figure 4.14E). Sequence alignments of the orthologues of RBBP6 and Clp1 showed that the residues implicated in this putative interaction are relatively well conserved across eukaryotes (Appendix Figure 8.8). If this interaction were real and the UBL domain truly interacted with both mCF and CFIIm subunits, the UBL domain would be sufficient for RBBP6 binding to the mCF-CFIIm complex. However, UBL did not co-elute with the mCF-CFIIm complex in an analytical gel filtration experiment (Figure 4.15). These observations suggest that, despite the statistically high-confidence of the prediction, the predicted RBBP6-Clp1 interaction may be weak or not form at all in solution. Further experiments will be required to determine how RBBP6 may interact with the CFIIm complex.

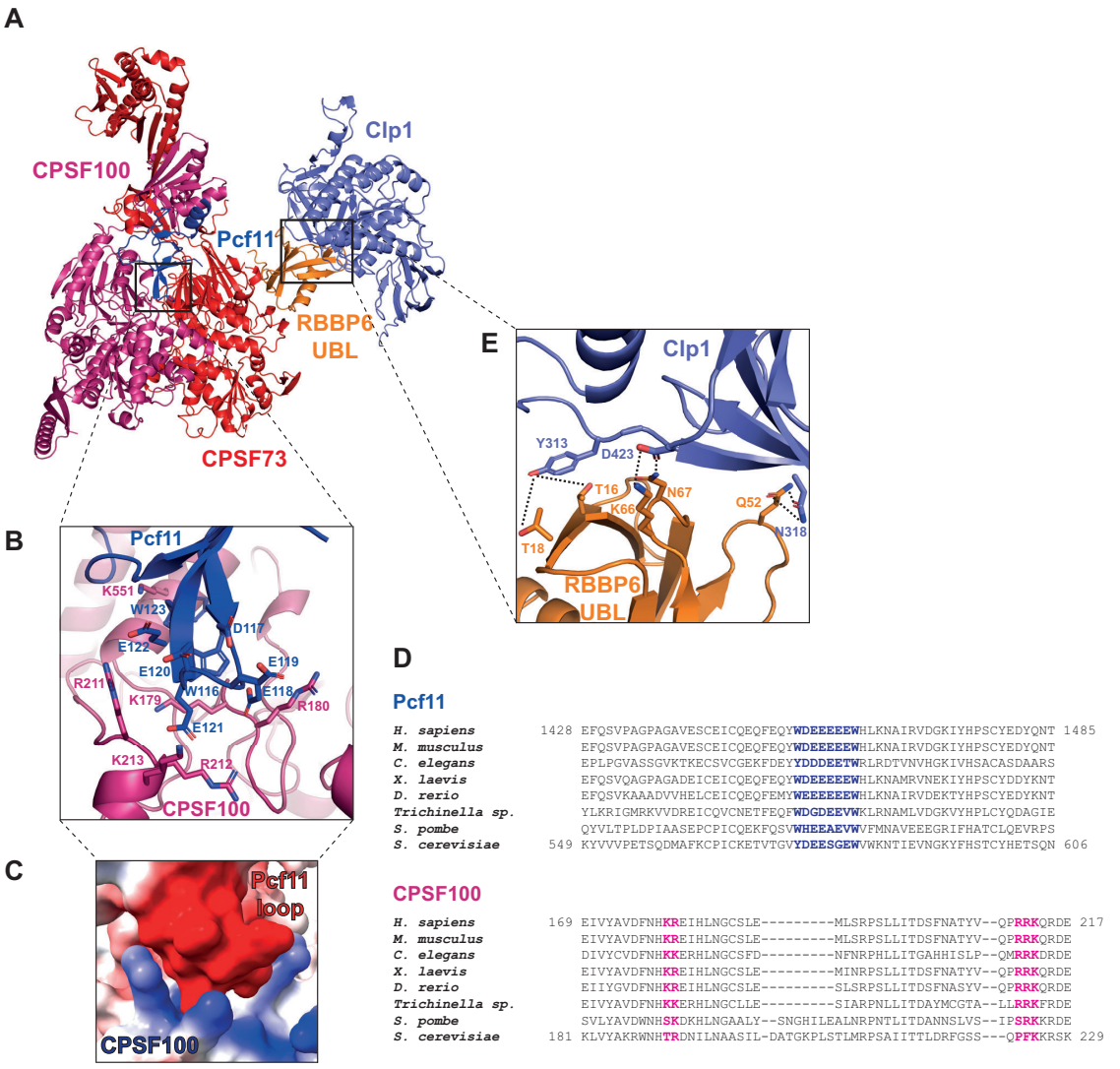


Figure 4.14 Predicted architecture of the mCF-CFIIm-RBBP6 complex. (A) Overall composite structure of the mCF-CFIIm-RBBP6 complex predicted by AlphaFold Multimer. **(B)** Close-up view of the predicted interface between Pcf11 and CPSF100. **(C)** Surface representation of the Pcf11-CPSF100 interface coloured according to electrostatic potential. Red denotes negative and blue – positive charges. **(D)** Sequence alignments of the relevant regions of Pcf11 and CPSF100. The residues that may mediate the interactions between the two proteins are coloured. **(E)** Close-up view of the putative interface between RBBP6-UBL and Clp1. Potential hydrogen bonds are indicated with dashed lines. AlphaFold does not accurately predict the precise position of the side chain, and therefore some of them appear detached from the main chain in the predicted models.

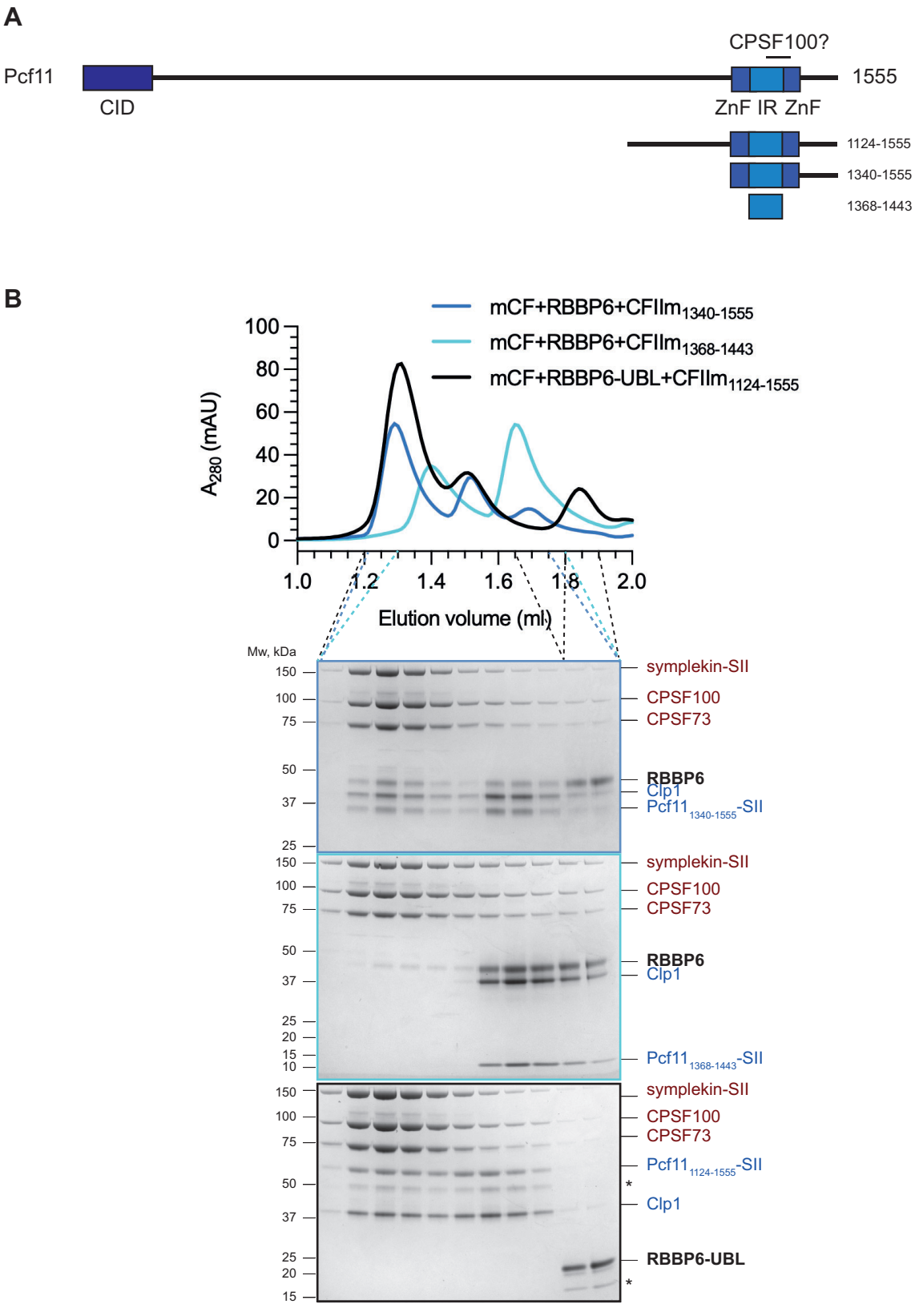

A
CPSF100?
Pcf11
1555
CID
ZnF IR ZnF
1124-1555
1340-1555
1368-1443
B
mCF+RBBP6+CFIIm1340-1555
mCF+RBBP6+CFIIm1368-1443
mCF+RBBP6-UBL+CFIIm1124-1555
A280 (mAU)
Elution volume (ml)
Mw, kDa
symplekin-SII
CPSF100
CPSF73
RBBP6
Clp1
Pcf111340-1555-SII
Pcf111368-1443-SII
Pcf111124-1555-SII
RBBP6-UBL

Figure 4.15 Experimental evidence for direct interactions between CFIIm subunits and either mCF or RBBP6. (A) Domain diagram of Pcf11, highlighting the location of the putative CPSF100 interaction motif as well as the boundaries of the constructs used in gel filtration experiments. **(B)** Size exclusion chromatograms (top) and SDS-PAGE analyses of the corresponding fractions (bottom) of the samples containing mCF (2.5 μM) and various constructs of either CFIIm (4 μM) or RBBP6 (7.5 μM). Asterisks denote degradation products.

4.4 Assembly of the complete active human pre-mRNA 3′ end processing machinery *in vitro*

So far in this study, I have defined the set of proteins that are required for the endonuclease activity by the human CPSF complex and explored the physical interactions amongst these proteins by biochemical and bioinformatics studies to provide mechanistic insights into the activation of cleavage. Subsequently, elucidating the structure of the complete active pre-mRNA 3′ end processing machinery represents the ultimate goal in studying the activity of the 3′ endonuclease: it would simultaneously reveal the architecture of individual protein complexes as well as how these complexes assemble together to facilitate the conformational changes necessary to pry open the active site of CPSF73 for pre-mRNA cleavage.

The 3′ end processing machinery is likely to exist in at least three different conformational, and possibly compositional, states during an endonuclease reaction, including pre-cleavage, cleaving and post-cleavage states. I was particularly interested in the pre-cleavage state, in which the pre-mRNA substrate is already bound to an open active site of CPSF73 poised for cleavage. However, mixing all the reaction components together with RNA may result in instantaneous cleavage of the substrate, leading to a mixture of states, and hence, the endonuclease reaction had to be stalled. To achieve this, I kept the assembled reaction on ice at all times, which inhibited endonuclease activity of CPSF (Figure 4.16A). The pre-cleavage state of the histone pre-mRNA 3′ end processing complex was successfully trapped using the same approach, suggesting that, although cleavage does not take place at low temperatures, the active complex should assemble efficiently (82). In fact, all the protein components required for cleavage assembling into a stable complex with the pre-mRNA substrate is an essential condition for structure determination of an intact active complex by cryoEM. The observation that RBBP6 was not stably associated with CPSF hinted at the possibility that the canonical human pre-mRNA 3′ end processing machinery could be

rather dynamic. To investigate this, I assembled a cleavage reaction containing the synthetic 60 nt polyadenylation site and ran the sample on an analytical gel filtration column at 4°C. Although the 60 nt RNA does not get cleaved efficiently under standard assay conditions, its relatively short length ensures that it can enter the column matrix, whereas a longer RNA may only elute in the void volume. The chromatogram revealed that the elution peak was rather broad, indicating that it may represent a mixture of complexes that may differ slightly in their conformation and/or subunit composition. Nevertheless, SDS-PAGE analysis of the peak fractions suggested that a stable stoichiometric complex containing all the required protein components could assemble on the substrate RNA (Figure 4.16B). Thus, although I observed no endonuclease activity when the reaction components were incubated on ice, the complete active complex assembled successfully. This particular sample will be important in determining the exact molecular mechanism of CPSF endonuclease activation in the future.

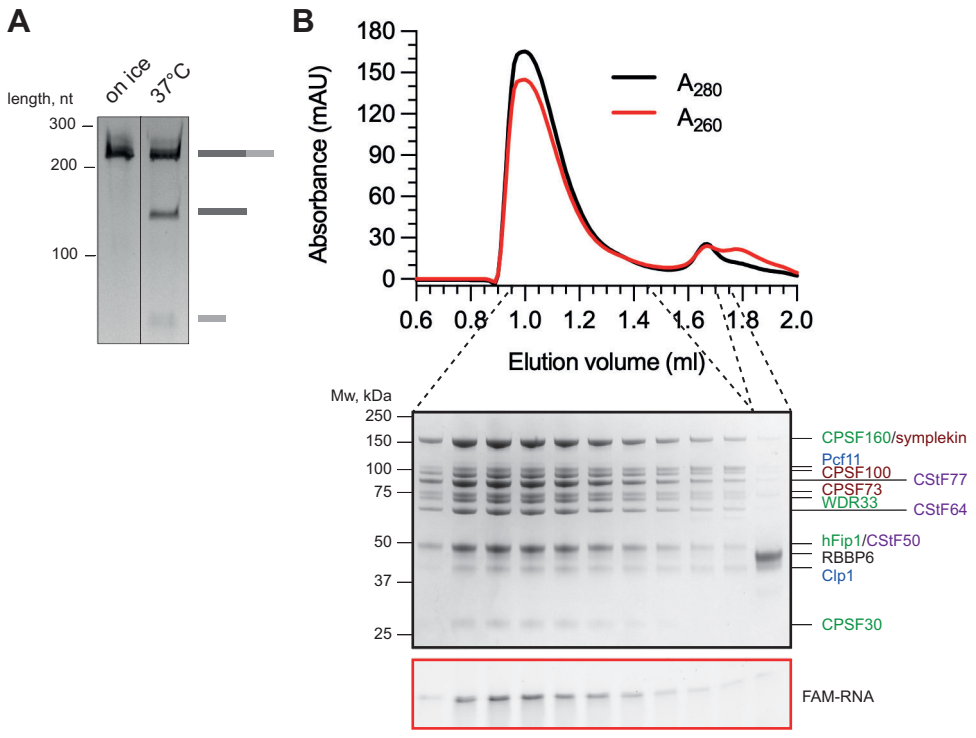


Figure 4.16 Assembly of the complete active pre-mRNA 3′ end processing machinery with human proteins. (A) CPSF cleavage assays using the SV40 pre-mRNA substrate performed either on ice or at 37°C. **(B)** Gel filtration chromatogram (top) and SDS-PAGE analyses of the corresponding peak fractions of the sample containing all protein components required for endonuclease activity of CPSF (2.5 µM mPSF, 2.5 µM mCF, 2.5 µM CStF, 4µM CFIIm and 7.5 µM RBBP6) and 60 nt synthetic pre-mRNA substrate (2.5 µM). The gel was run in MES buffer for better separation of bands of interest.

4.5 Conclusions and perspectives

4.5.1 RBBP6 is not a constitutive subunit of CPSF in humans

Reconstitution of the human CPSF endonuclease activity with purified recombinant proteins revealed that, in addition to CStF and CFIIm complexes, a multi-domain protein RBBP6 was also essential for endonuclease activation. The yeast orthologue of RBBP6, Mpe1, is a constitutive subunit of the yeast CPF complex, and hence, I was puzzled how such an essential factor in human pre-mRNA 3′ end processing had escaped notice for decades. I purified endogenous CPSF from human cells, and determined that RBBP6 was not a stable part of the complex in human cells, which may explain why the role RBBP6 in activating 3′ end processing had been overlooked.

The demonstration that both human RBBP6 and yeast Mpe1 are essential for the activation of the 3′ endonucleases in their respective species highlights the conservation in mechanism of 3′ end processing from yeast to humans. Many protein-protein interactions within the 3′ end processing machinery in humans, including between RBBP6 and CPSF, appear to be of a lower affinity than in budding yeast. For example, the poly(A) polymerase is constitutively associated with Fip1 in yeast but is only transiently recruited to mPSF in humans; CStF and CFIIm are separate complexes in human cells, but their homologues constitute a single CF IA complex in yeast. The weaker interactions amongst human proteins could account for the generally lower rate of endonucleolytic cleavage of CPSF compared with yeast CPF *in vitro* (12). However, the trend towards weaker binding between the components of the human machinery is not universal: human mPSF binds RNA with several orders of magnitude higher affinity than the yeast complex, and the nuclease module of CPF does not seem to interact with the Pcf11-Clp1 dimer *in vitro* (unpublished experiment by Juan Rodriguez, MRC LMB). Overall, the variability in interaction affinities may allow different modes of regulation of alternative polyadenylation in mammalian and yeast cells, while maintaining the same fundamental mechanism of 3′ end processing. In the future, it will be important to study cleavage and polyadenylation machineries in many more species, especially phylogenetically distant ones, to better understand the regulation of 3′ end processing in eukaryotes.

4.5.2 More CPSF interactors may await identification

The purification of endogenous CPSF from HEK293T cells revealed a few novel putative binding partners of the complex, although additional control experiments, including a demonstration of a direct interaction with purified proteins, will be required to validate these candidate interactors. However, the example of RBBP6 illustrates that weak and non-constitutive interactions between proteins are often critical for their function. The enzymatic activities of CPSF are highly regulated in human cells, and therefore, I predict that, similar to RBBP6, many proteins that modulate CPSF function may bind to the complex only transiently and may not co-purify with the endogenous complex. To capture such weak interactions in the native environment, they may need to be stabilised by chemical cross-linking *in situ* using membrane-permeant cross-linking agents. This approach has recently been successful for studying the interactome in bacterial cells, and could also facilitate the study of transient protein-protein interactions in human cells (171). I already performed preliminary experiments, demonstrating that CPSF can be cross-linked in isolated mammalian cell nuclei and that the cross-linked complex can be successfully purified (Appendix Figure 8.9). Increasing the scale of this experiment to obtain sufficient material for analysing and identifying the cross-linked peptides are the challenges that will have to be overcome to identify new transient binding partners of CPSF. In particular, I would expect that the C-terminal intrinsically disordered domains of WDR33 and RBBP6 that contain several highly conserved motifs may interact with proteins that have so far escaped notice (Appendix Figures 8.1 & 8.2). Extending such interaction studies to other cell types may also reveal the mechanistic basis of tissue-specific alternative polyadenylation in human health and disease.

4.5.3 RBBP6 interacts with CPSF and CFIIm

I demonstrated that, although RBBP6 is not a stable component of human CPSF, it gets recruited to the complex once CPSF is bound to PAS RNA. This may delay endonuclease activation until the full 3′ end processing machinery has assembled, enhancing the specificity of 3′ cleavage. RBBP6 likely contacts WDR33 and CPSF73 subunits and the RNA substrate simultaneously, with multiple interaction sites cooperatively stabilising the complex. It is important to note that the observed interactions between RBBP6 and CPSF *in vitro* do not contradict the results of the pull-downs of endogenous CPSF in which RBBP6 was not detected: gel filtration and pull-down experiments with recombinant proteins were

performed at micromolar protein concentrations, which are at least several orders of magnitude higher than the likely concentrations of the same proteins in the cell, allowing to observe relatively weak interactions *in vitro*.

I also identified novel interactions between the mCF and CFIIm complexes and RBBP6. Unlike its interaction with CPSF, the binding of RBBP6 to the mCF-CFIIm complex is independent of RNA. It is unclear which subunits RBBP6 contacts first as the active machinery assembles, but all the interactions involving RBBP6 described in this work are likely to be ultimately required for endonuclease activation.

4.5.4 RBBP6 may mediate cross-talk between various components of the pre-mRNA 3′ end processing machinery

RBBP6 contains multiple domains in its N-terminal region (Figure 4.17). The UBL, PSR, zinc knuckle and RING finger domains are connected by intrinsically disordered regions and do not appear to interact with each other. Such a "beads-on-a-string" arrangement highlights the modular nature of RBBP6, which makes this multi-domain protein ideal for linking the various components of the 3′ end processing machinery both structurally and functionally. Each domain of RBBP6 seems to interact with a different component of the 3′ end processing machinery, connecting both mPSF and mCF modules of CPSF with cleavage factors and poly(A) polymerase (Figure 4.17). The multivalency of its binding to the 3′ end processing machinery may allow RBBP6 to coordinate a conformational transition from an inactive complex to the active endonuclease machinery, and perhaps even facilitate polyadenylation following cleavage.

However, the binding of RBBP6 to the 3′ end processing machinery might be more complex than previously thought. For instance, the sites of RNA binding on RBBP6 as well as its sequence specificity have not been investigated in detail, but both zinc knuckle and RING finger domains may interact with RNA directly (172). The zinc knuckle domain may also contact mPSF (33). The RING domain is not required for the activation of the CPSF endonuclease but could act as an E3 ubiquitin ligase to facilitate ubiquitination of protein substrates (159, 172). For instance, yeast Mpe1 has been suggested to be required for ubiquitination of poly(A) polymerase in yeast cells (172). Since RBBP6 interacts with so many proteins and protein complexes, any of them could be targets for ubiquitination.

However, whether RBBP6 actually regulates pre-mRNA 3′ processing by ubiquitination remains to be determined (173).

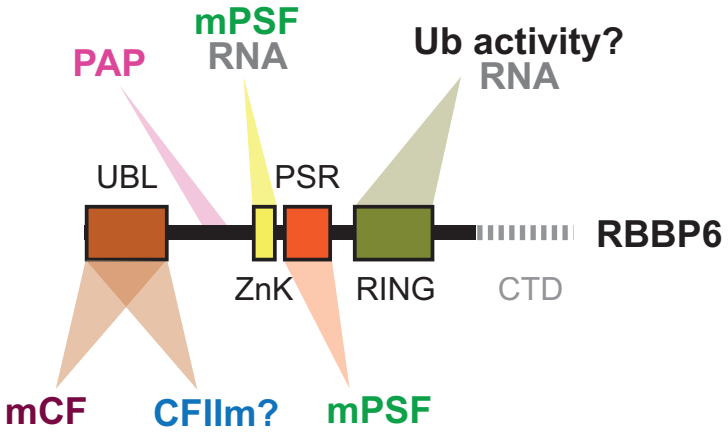


Figure 4.17 RBBP6 interacts with multiple components of the 3′ end processing machinery. Domains on the N-terminal end of RBBP6 are depicted. The 3′ end processing factors that each domain interacts with are indicated either above or below the domain diagram. CTD – C-terminal domain not discussed in this section.

4.5.5 Approaches to improve the preparation of CPSF complexes for cryoEM

In this Dissertation, I described imaging various protein complexes involved in human pre-mRNA 3′ end processing by cryoEM, including the mPSF module bound to NS1 protein from Influenza virus, the CPSF complex bound to RBBP6 and RNA, and the mCF-CFIIm-RBBP6 complex. However, my data collections did not reveal any novel protein densities beyond the previously determined structures of the mPSF and mCF modules of CPSF. mCF is known to be flexible relative to mPSF, which may explain the difficulty in determining the structure of the complete CPSF complex. The failure to reconstruct any densities of additional protein factors suggests that either they do not provide additional stabilisation to the complex, that they dissociate from CPSF upon grid preparation or that they denature in vitreous ice. Various buffer conditions, different types of cross-linkers and grid specimen supports will have to be tested systematically, but at this point it is not possible to rationally design a strategy to improve the quality of the samples.

A variety of other protein samples could also provide significant insights into the mechanism of pre-mRNA 3′ end processing in humans. Pull-downs from mammalian

nuclear extract using a tagged RNA showed that all proteins required for the activation of CPSF remain stably associated with the 5′ cleavage product, suggesting that the active 3′ end processing machinery in its post-cleavage state could be a suitable target for cryoEM (6). In addition, protein factors that are not essential for CPSF endonuclease activation and are not included in the recombinant samples could stabilise the active complex in the nucleus. Thus, imaging the active endogenous machinery could also be informative. Specifically, endogenous pull-downs using tagged RBBP6 could enrich the active endogenous 3′ end processing machinery for cryoEM studies.

In Chapter 3, I showed that recombinant CPSF is inhibited by compounds targeting CPSF73, in particular by JTE-607. In addition to potential therapeutic applications of this finding, JTE-607 could also be used to stabilise CPSF in cryoEM studies. The compound inhibits the endonuclease by competing with the RNA substrate for binding to the active site, and therefore, is unlikely to stabilise the active RNA-bound state of the 3′ end processing machinery. However, binding of JTE-607 to CPSF73 may stabilise the mCF complex bound to CFIIm and RBBP6, which does not contain RNA. Binding of the compound may stabilise the fold of CPSF73, prevent its potential denaturation upon flash-freezing and reduce the movement of its endonuclease domain relative to the rest of the complex. Thus, JTE-607 could improve the quality of the data of the mCF-CFIIm-RBBP6 complex.

4.5.4 AlphaFold is a powerful tool in studying protein complexes

AlphaFold is a recently developed tool that uses a machine learning algorithm trained on the experimentally available data to predict the three-dimensional structure of unknown proteins (124). AlphaFold Multimer predictions of oligomeric protein structures can be incredibly powerful for studying protein complexes, especially for targets that are challenging to solve by experimental methods. Mutational and sequence conservation analyses may be then used to validate the predicted model, as demonstrated here for the RBBP6 interactions with CPSF subunits. In the case of mCF and CFIIm, AlphaFold Multimer revealed how the two complexes may interact and allowed to define the boundaries of the domains that mediate this interaction, which may enable the design of more optimal protein constructs for experimental structure determination. While high-affinity interactions are predicted rather accurately, transient weaker binding events as well as cooperative interactions between different protein surfaces may present a challenge to AlphaFold Multimer. For example, while AlphaFold predicted an interaction between Clp1 and the UBL

domain of RBBP6, experiments showed that UBL alone was not sufficient for binding to the mCF-CFIIm complex. This does not mean that the predicted interaction is necessarily wrong – it is possible that other domains of RBBP6 may also contact subunits of mCF and CFIIm, but these interactions could have been missed by the prediction algorithm. Thus, despite its power in generating hypotheses and assisting with experimental design, at the moment AlphaFold cannot fully replace experimental structure determination of protein complexes.

Chapter 5:

Conclusions and perspectives

5.1 Summary of this Thesis

In this Dissertation, I investigated the mechanisms of 3′ end cleavage and polyadenylation of pre-mRNAs in humans. I employed an *in vitro* approach of reconstituting both processing reactions with highly pure recombinant proteins in a well-controlled minimal system.

In Chapter 2, I developed methods to express and purify large quantities of recombinant human proteins and protein complexes involved in 3′ end processing from insect cells. Initially, with these recombinant proteins I studied the mechanisms of polyadenylation by reconstituting this reaction with a model RNA substrate, recombinant mPSF module and the poly(A) polymerase enzyme. I discovered that the CStF complex may inhibit polyadenylation of suboptimal RNA substrates and thereby confer specificity to the reaction (Figure 2.8). In addition, I showed that a multi-domain protein RBBP6 stimulates the poly(A) polymerase enzyme and may regulate polyadenylation (Figure 2.11).

However, for the vast majority of this Study, I focused on the cleavage activity of the CPSF complex, which had not been reconstituted with recombinant proteins before. In Chapter 3, I established that the activation of the CPSF endonuclease requires three additional protein factors: well-characterised cleavage factor complexes CStF and CFIIm, and a previously overlooked protein RBBP6 (Figure 3.1). The reconstituted cleavage reaction was highly specific, and the model pre-mRNA substrate was cleaved by recombinant CPSF at the same site as *in vivo*, demonstrating that the *in vitro* assay accurately recapitulates the 3′ end processing reaction in human cells (Figure 3.5C). I also demonstrated that the endonuclease of the canonical CPSF complex is activated by a different mechanism than the highly specialised histone pre-mRNA 3′ end processing machinery (Figure 3.17). Thus, explaining the activation mechanism of CPSF demanded further investigation.

The result that RBBP6 was essential for the activation of the CPSF endonuclease was rather surprising. Therefore, in Chapter 4, I explored how this poorly characterised protein may facilitate the activation of the 3′ end processing machinery in humans. I purified endogenous CPSF, and demonstrated that RBBP6 is not constitutively associated with the complex in human cells (Figures 4.2 & 4.3). Instead, RBBP6 appeared to be recruited to CPSF in an RNA-dependent manner (Figures 4.4 & 4.5). RBBP6 is likely to interact both with the endonuclease subunit CPSF73 and with the mPSF module bound to the PAS-containing pre-mRNA substrate (Figures 4.7 & 4.8). Interestingly, I discovered that the mCF module of CPSF and the CFIIm cleavage factor form a stable complex, likely via an interaction between

CPSF100 and Pcf11 (Figures 4.12, 4.14, 4.15). Subsequently, I revealed that, in addition to its interaction with mCF, RBBP6 is also likely to contact the CFIIm complex (Figures 4.13, 4.14, 4.15). All the interactions described here appeared to be important for the mechanism of endonuclease activation.

In the rest of this Chapter, I will discuss the broader implications of this Study to our understanding of gene expression in human cells.

5.2 Coordination of 3′ end processing with other steps of mRNA biogenesis

Most pre-mRNAs that retain introns fail to undergo timely 3′ end processing and transcription termination (Figure 1.8). Such "all-or-none" nature of mRNA biogenesis implies that splicing, 3′ end processing and transcription termination are coordinated. RNA Pol II itself, including its C-terminal domain, has been considered to be a recruitment platform for various pre-mRNA processing factors, thereby coordinating their activities. For instance, recent structural studies revealed that both 3′ end processing complexes, as exemplified by yeast APT and human Integrator, and the U1 snRNP bind Pol II at a similar location adjacent to the nascent RNA exit site (63, 110, 118). Thus, mutually exclusive binding of these machineries to RNA Pol II could be the physical basis for coupling between splicing, 3′ end processing and transcription termination.

Alternatively, having discovered its essential role in the activation of pre-mRNA cleavage, I hypothesise that RBBP6 could be involved in the coordination of transcription, splicing and 3′ end processing. The N-terminal folded domains of RBBP6 interact with a plethora of proteins involved in cleavage and polyadenylation, as described in this Thesis. On the other hand, the C-terminal domain of RBBP6 is primarily intrinsically disordered and contains short linear motifs that interact with transcription regulator VP30 from Ebola virus, endogenous transcription factor p53 and transcriptional co-repressor Rb, from which RBBP6 got its name as a Retinoblastoma-binding protein (Figure 2.10A & Appendix Figure 8.2) (41, 42, 123). The C-terminal domain also has an RS-like domain which could form homotypic interactions with RS domains of splicing regulators, including SR proteins and U1 snRNP (165). Thus, the modular architecture of RBBP6 may not only facilitate cross-talk between various components of the 3′ end processing machinery but may also coordinate cleavage and polyadenylation with transcription and splicing.

My analysis of endogenous CPSF preparations confirmed previous observations that protein phosphatases that regulate transcription termination – PP1 and SSU72 – are not stably associated with the human complex (Figure 4.3). PP1 in human cells is part of a different complex assembled around a scaffold protein PNUTS (53, 54). SSU72, however, is likely recruited to transcription termination sites via a direct interaction with a CPSF subunit symplekin (Figure 3.17B&C). Interestingly, binding of SSU72 is compatible with cleavage activity of CPSF but not with the active state of the histone pre-mRNA 3′ end processing complex (Figure 3.17D) (82). This is also corroborated by *in vivo* studies showing that depletion of SSU72 in cells improved the efficiency of 3′ end processing of histone mRNAs but had an opposite effect on polyadenylated transcripts (174). The effects of SSU72 on 3′ end processing could be independent of its phosphatase activity. Alternatively, these observations suggest that the mechanism of transcription termination, and in particular, the role of various covalent modifications of the Pol II C-terminal domain, may differ between the mRNAs that encode replication-dependent histones and the rest of the protein-coding genes (175). Overall, the biogenesis of replication-dependent histone transcripts appears to have diverged significantly from canonical mechanisms, enabling their highly regulated and timely expression at the onset of the S phase of the cell cycle.

In vitro reconstitution of individual steps in mRNA biogenesis (transcription, splicing, 3′ end processing) with purified proteins have provided important mechanistic insights into each process. It is interesting to note that only 3′ end processing could be reconstituted entirely with recombinant proteins, while transcription and splicing require purification of RNA Pol II and spliceosome components from native sources. I believe that such *in vitro* assays will be instrumental in generating hypotheses regarding how splicing, 3′ end processing and transcription termination are coordinated. Testing these hypotheses will require clever approaches to perturb and study these processes in a more complex environment of the mammalian cell.

5.3 Molecular mechanism of cleavage and polyadenylation

The CPSF complex with the help of several accessory protein factors catalyses both cleavage of the nascent pre-mRNA at its 3′ end and subsequent polyadenylation of the cleaved substrate. Based on the results in this Dissertation and other published studies, I have designed a hypothetical model of how CPSF facilitated by CStF, CFIIm and RBBP6 may perform both cleavage and polyadenylation of human protein-coding transcripts.

3′ end processing is most probably initiated by CPSF binding to the PAS sequence as the nascent pre-mRNA emerges from RNA Pol II (Figure 5.1A). CPSF binds to RNA as an inactive endonuclease and requires recruitment of CStF, CFIIm and RBBP6 for its activation. The exact order in which these other factors subsequently assemble is unclear and will be determined both by the relative affinities of protein-protein and protein-RNA interactions as well as the availability of these factors on transcribed chromatin.

Based on their interactions with CPSF, I hypothesise that CStF, by binding to both the mPSF module and downstream U/G-rich RNA elements, may position the RNA correctly in the active site of CPSF73, while CFIIm and RBBP6, both of which interact with the catalytic mCF module, could be more directly involved in the conformational change of the endonuclease subunit (Figure 5.1B). In particular, the UBL domain of RBBP6 is the only direct interaction partner of the catalytic domain of CPSF73. I predict that the propagation of conformational rearrangements across many protein factors will lead to “pulling” of RBBP6 UBL, causing the metallo-β-lactamase domain of the endonuclease to pivot away from the β-CASP domain, opening the active site of CPSF73 for endonucleolytic cleavage of the pre-mRNA substrate (Figure 5.1B).

Polyadenylation of an RNA *in vitro* requires only mPSF and PAP (Figure 2.2). However, *in vivo* polyadenylation always follows endonucleolytic cleavage and is catalysed by PAP bound to the CPSF complex, which remains associated with the PAS after cleavage. Although CStF and CFIIm complexes preferentially bind to cis-regulatory elements located downstream of the cleavage site, they likely remain associated with CPSF via protein-protein interactions. Hence, addition of a poly(A) tail is likely to take place in the presence of all three accessory factors required for endonuclease activation: CStF, CFIIm and RBBP6. Mounting evidence suggests that these auxiliary factors may also regulate polyadenylation. Each CPSF complex may contain two copies of hFip1 and two copies of PAP (25, 26). In a cleavage state, the N-terminal helices of the two hFip1 subunits may interact with two

equivalent interfaces on the CStF77 dimer, which inhibits polyadenylation (Figures 2.8 & 5.1B) (26). The transition from a cleaving to a polyadenylating complex likely involves dissociation of either one or both hFip1 subunits from CStF77. This may allow PAP to swing into a correct position to accept the 3′ end of the cleaved pre-mRNA into its active site. The interaction with RBBP6 may stabilise this position of the enzyme relative to the rest of the complex (Figure 2.11). Only one copy of RBBP6 is present in the complex, and it is possible that only one copy of PAP is active at any one time. The effect of CFIIm on polyadenylation remains to be explored.

The 3′ end processing machinery remains stably associated with the mature mRNA after cleavage and polyadenylation (6). Thus, it is likely that the disassembly of the machinery is an active process, most likely executed by RNA export factors (Figure 5.1C) (176). Removal of the 3′ end processing machinery allows both export of the mature mRNA out of the nucleus and recycling of protein complexes for subsequent rounds of pre-mRNA processing.

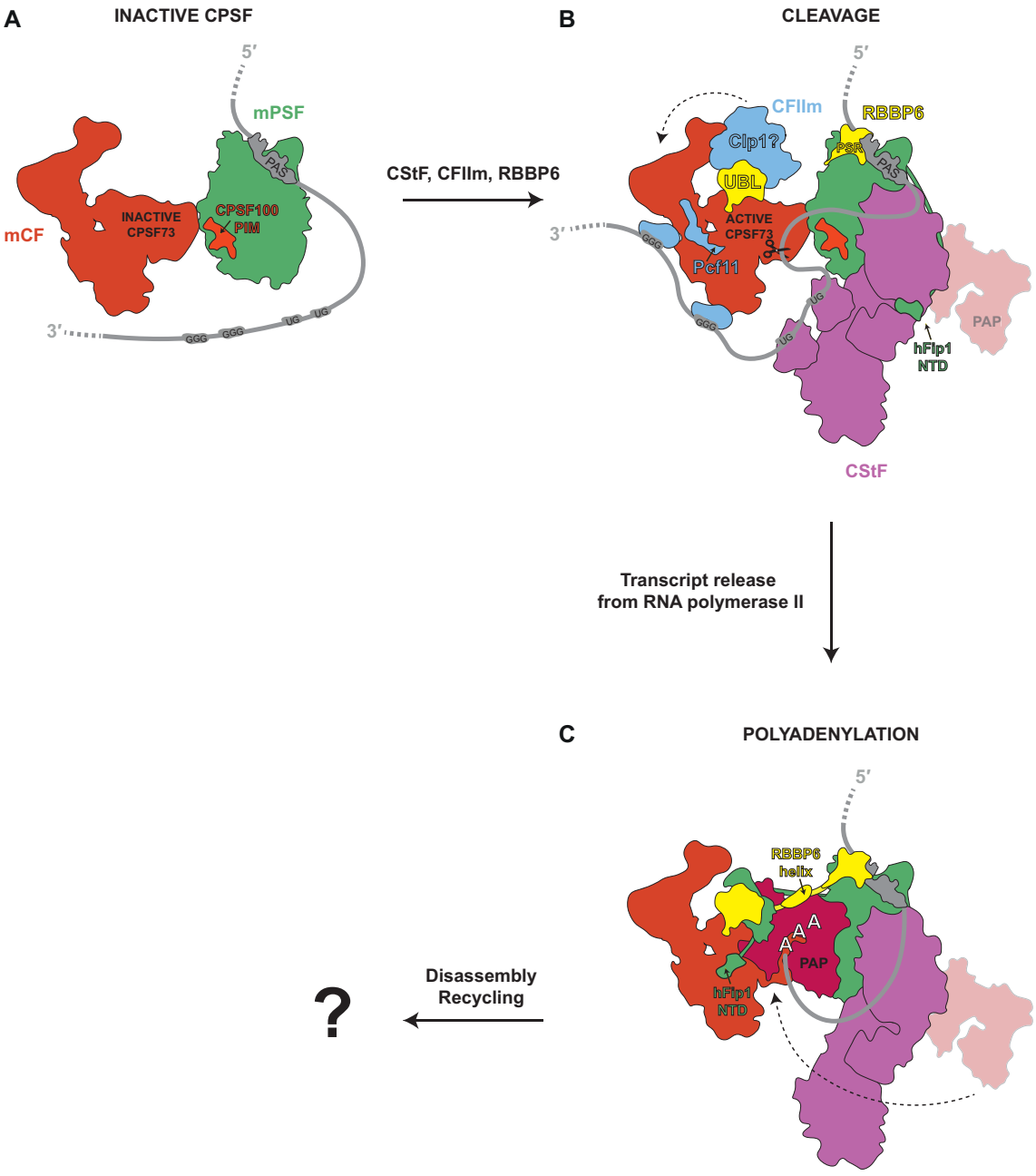


Figure 5.1 Model of molecular mechanisms of cleavage and polyadenylation by human CPSF. Some known interactions as well as connections between certain domains of the same proteins were omitted for clarity. Only one copy of hFip1 and PAP is shown for simplicity. Only proteins and domains relevant to a particular state are labelled. The question mark next to Clp1 indicates the uncertainty of its interaction with RBBP6 UBL. Dashed arrows depict potential conformational changes required for the transition from catalytically-inactive CPSF **(A)** to its cleaving **(B)** and polyadenylating **(C)** states.

5.4 Final conclusions

3′ end processing of eukaryotic pre-mRNAs is essential for regulated expression of protein-coding genes. The relatively simple chemistry of endonucleolytic cleavage requires a complex array of protein factors beyond the CPSF complex that hosts the endonuclease enzyme. Multiple protein-protein and protein-RNA interactions, many of them relatively weak individually, may enable the activation of the CPSF endonuclease and regulate subsequent polyadenylation of the cleaved substrate. Such an elaborate assembly mechanism of the active machinery may ensure fidelity and specificity of 3′ end formation of protein-coding transcripts in eukaryotic cells. I hope that the reagents that were generated in this Study as well as the mechanistic insights gained from my experimental work will enable further in-depth investigation into human pre-mRNA 3′ end processing. From a practical perspective, the assay to study the endonuclease activity of human CPSF could be used to search for new therapeutic compounds that may inhibit the endonuclease enzyme or prevent viral proteins from hijacking the processing machinery in infected cells. Fundamentally, the mystery of how the auxiliary protein factors induce a conformational change in CPSF to activate the endonuclease remains wide open, and structural studies of the purified recombinant machinery I established in this Dissertation will be instrumental in answering this question. Overall, it will be exciting to follow what course the field of eukaryotic pre-mRNA 3′ end processing will take in the future.

Chapter 6:

Materials and Methods

6.1 Cloning

6.1.1 General cloning methods

6.1.1.1 Polymerase chain reaction (PCR) and gel electrophoresis of DNA

Gene fragments for cloning were amplified by PCR in 50 µl reactions using Q5 high-fidelity DNA polymerase (NEB, cat. No. M0491S). Primer annealing temperatures and extension times were determined specifically for each reaction. After PCR was complete, 1 µl Dpn1 (NEB, cat. No. R0176S) was added, and reactions were incubated at 37°C for 1 h to degrade methylated template plasmid DNA. The reactions were then mixed with gel loading dye (NEB, cat. No. B7024S) and run on a 1% agarose gel prepared with low-melting agarose (BioGene, cat. No. 300-600) dissolved in TAE buffer. The gels were stained with SybrSafe (Invitrogen, cat. No. S33102) and visualised using a BioRad Gel Doc XR+ instrument and Image Lab Software. Bands of expected size were cut out, and DNA was extracted using a gel recovery kit (Zymogen, cat. No. D4007).

6.1.1.2 Gibson assembly

Gibson assembly allows construction of entire plasmids from separate fragments containing homologous overlapping sequences of 15-30 nucleotides. The fragments were synthesised by PCR, purified by gel extraction, and their concentrations were measured using a NanoDrop instrument (Thermo Scientific). The purified overlapping fragments were mixed at equimolar concentrations, with the exception of assemblies including a plasmid backbone, in which case the much larger backbone fragment was mixed with gene fragments at a molar ratio of 1:5. The DNA was added to the Gibson reaction mix, containing Taq DNA ligase (NEB, cat. No. M0647S), T5 exonuclease (NEB, cat. No. M0363) and Phusion high-fidelity DNA polymerase (NEB, cat. No. M0530S) in isothermal reaction buffer (100 mM Tris-HCl pH 7.4, 10 mM $MgCl_2$, 0.2 mM dNTPs, 10 µM DTT, 50 mg/ml PEG 8000, 1 mM NAD). The reactions were incubated at 50°C for 90 min before transforming the assembled plasmids into TOP10 *E. coli* cells.

6.1.1.3 Transformation of competent cells

Chemically competent *E. coli* cells (TOP10 for cloning, BL21(DE3) Star for expression in *E. coli*, EmBACY for bacmid preparation) were prepared by various members of Passmore group (MRC LMB). ~100 ng plasmid DNA was added to 30-50 µl of competent cells, and the cells were incubated on ice for 20 min. The cells subsequently underwent heat shock treatment in a 42°C water bath for 45 s and were then chilled on ice for 2 min. The cells were allowed to recover in 400 µl sterile SOB media for 90 min at 37°C. The transformed cells were centrifuged for 1 min at ~1,000 g, and the pellet was plated on agar containing an appropriate antibiotic for the transformed plasmid. The colonies were grown overnight at 37°C.

6.1.1.4 Plasmid amplification and purification

Individual colonies from transformation plates were picked and inoculated in ~5 ml sterile 2X YT media containing an appropriate antibiotic. The cultures were grown overnight in a 37°C shaking incubator. The next morning, the cultures were pelleted for 10 min at 3,000 g. Plasmid DNA was extracted from the cell pellet using a Miniprep kit (either Zymogen, cat. No. D4208T, or Qiagen, cat. No. 271006X4). The concentration of purified plasmid was measured using a NanoDrop instrument (Thermo Scientific).

6.1.1.5 Gene synthesis and Sanger sequencing

E. coli codon-optimised genes encoding full-length proteins of all CPSF, CStF, CFIIm, CFIm subunits, RBBP6 and SSU72, and isoform 2 of CPSF30 (Uniprot O95639-2) in pACEBac vectors for expression in insect cells were synthesised by Epoch Life Science. pcDNA plasmids for transient overexpression in mammalian cells encoding full-length open reading frames of human WDR33 and RBBP6 were synthesised by GenScript.

All cloning done in this Thesis was validated by sequencing (Source Bioscience).

6.1.2 Cloning individual proteins and protein complexes for expression in insect cells

6.1.2.1 Tagging genes in pACEBAC vectors

The full coding regions of CStF77, symplekin, CFIm25 and PAP were amplified by PCR from their original pACEBac vectors and cloned using Gibson assembly into pACEBac vectors containing an in-frame TEV cleavage site followed by a Strep-II tag on its 3′ end. For the following genes, only the sequences encoding the indicated residues were amplified by PCR: 1-576 of WDR33, 770-1555 of Pcf11 (or other regions as indicated in the text), 1-335 of RBBP6, 1-142 of RBBP6 (RBBP6 UBL), 341-1274 of symplekin (symplekin$_{\Delta NTD}$). These fragments were also cloned into pACEBac-TEV-SII vectors as described above.

6.1.2.2 hFip1$_{iso4}$

To generate isoform 4 of hFip1 (Uniprot Q6UN15-4), fragments containing residues 1-28 and 44-393 were amplified by PCR. Substitution F393K was also introduced during the PCR of fragment 44-393. Both fragments were assembled into an empty pACEBac vector using Gibson assembly.

6.1.2.3 Site-directed mutagenesis of CPSF73, CPSF100, CPSF30 and RBBP6

To produce catalytically inactive CPSF73$_{D75N\ H76A}$, the CPSF73 pACEBac plasmid was divided into three overlapping fragments, and these fragments were amplified by PCR. The mutations were located in the overlapping region between two of the three fragments. Point mutations of RBBP6 (D43K and R74E, Y228G, P195G) and CPSF100 (F464A, W473A, Y476A) were introduced into the coding sequence of the respective proteins in a pACEBAC vector using a similar approach, except that only two overlapping fragments were used. To generate point mutants of CPSF73, CPSF100 and RBBP6, the fragments were ligated using Gibson assembly.

To produce CPSF73 NTD and CPSF73 CTD constructs, CPSF73 residues 1-460 and 461-684, respectively, were amplified by PCR and assembled into empty pACEBac vectors using Gibson assembly.

CPSF30$_{\Delta PLPFP}$ and CPSF100$_{\Delta PIM}$ deletion mutants were also generated by amplifying by PCR overlapping fragments of the respective coding sequences. The sequences intended for deletion were omitted from the boundary between two of these fragments. The mutant plasmid were then produced by Gibson assembly.

6.1.2.4 biGBac cloning

A modified biGBac protocol was used to generate pBig1 vectors encoding all subunits of each complex (mPSF, mCF, CStF, CFIIm, CFIm and their variants) containing their own promoters and terminators for simultaneous co-expression of up to five genes in insect cells (12, 24, 120). The sequence encoding the promoter, open reading frame and poly(A) site of each gene was amplified by PCR using primers specific to the position of the gene within the pBig1 vector. The pBig1 vector was linearised by digestion with SwaI (NEB, cat. No. R0604) restriction enzyme. The primers used to amplify genes for insertion contain overhang sequences that allow the PCR products to be inserted into the pBig1 vector in a specific order by Gibson assembly. The multi-gene vectors constructed by Gibson assembly were then validated by diagnostic restriction digest using SwaI enzyme: the primer overhangs contain SwaI recognition sites, and hence, digestion with the enzyme will reveal the number of genes successfully inserted into pBig1 vector (number of bands observed after gel electrophoresis minus one corresponding to the plasmid backbone) as well as their sizes. The whole procedure using the mPSF module as an example is schematically depicted in Figure 6.1. The pBig1 vectors used in this study are listed in Table 6.1.

6.1.2.5 Cloning NS1 protein and its effector domain for co-expression with mPSF insect cells

pFastBac encoding a His$_6$-tagged NS1 protein from Influenza A H3N2 was a kind gift from Loic Carrique and Ervin Fodor (University of Oxford). The sequence encoding His$_6$-NS1 was amplified by PCR and cloned into an empty pACEBAC vector by Gibson assembly. The solubility-enhancing mutations in its RNA-binding domain (R38A and K41A) were subsequently introduced by site-directed mutagenesis. The sequence-encoding the His$_6$-tagged effector domain of NS1 was also subcloned into a pACEBAC vector using PCR and Gibson assembly.

Complex	Gene 1	Gene 2	Gene 3	Gene 4
mPSF-Fip1$_{FL}$	CPSF160	WDR33$_{1-572}$-SII	CPSF30$_{iso2}$	hFip1$_{FL}$
mPSF-Fip1$_{iso4}$	CPSF160	WDR33$_{1-572}$-SII	CPSF30$_{iso2}$	hFip1$_{iso4}$
mPSF-ΔhFip1	CPSF160	WDR33$_{1-572}$-SII	CPSF30$_{iso1}$	
mCF	CPSF100	CPSF73	Symplekin-SII	
mCF-CPSF73$_{D75N/H76A}$	CPSF100	CPSF73$_{D75N/H76A}$	Symplekin-SII	
mCF-symplekin$_{ΔNTD}$	CPSF100	CPSF73	Symplekin$_{ΔNTD}$-SII	
mCF-CPSF100$_{ΔPIM}$	CPSF100$_{ΔPIM}$	CPSF73	Symplekin-SII	
mCF-CPSF100$_{PIM\ MUT}$	CPSF100$_{PIM\ MUT}$	CPSF73	Symplekin-SII	
CStF	CStF77-SII	CStF64	CStF50	
CFIIm-Pcf11$_{FL}$	Pcf11$_{FL}$-SII	Clp1		
CFIIm-Pcf11$_{Δ769}$	Pcf11$_{Δ769}$-SII	Clp1		
CFIIm-Pcf11$_{1124-1555}$	Pcf11$_{1124-1555}$-SII	Clp1		
CFIIm-Pcf11$_{1340-1555}$	Pcf11$_{1340-1555}$-SII	Clp1		
CFIIm-Pcf11$_{1368-1443}$	Pcf11$_{1368-1443}$-SII	Clp1		
CFIm	CFIm25-SII	CFIm68		

Table 6.1 List of pBig1 constructs used in this study. SII – Strep-II tag; FL – full-length; iso – isoform; MUT – mutant.

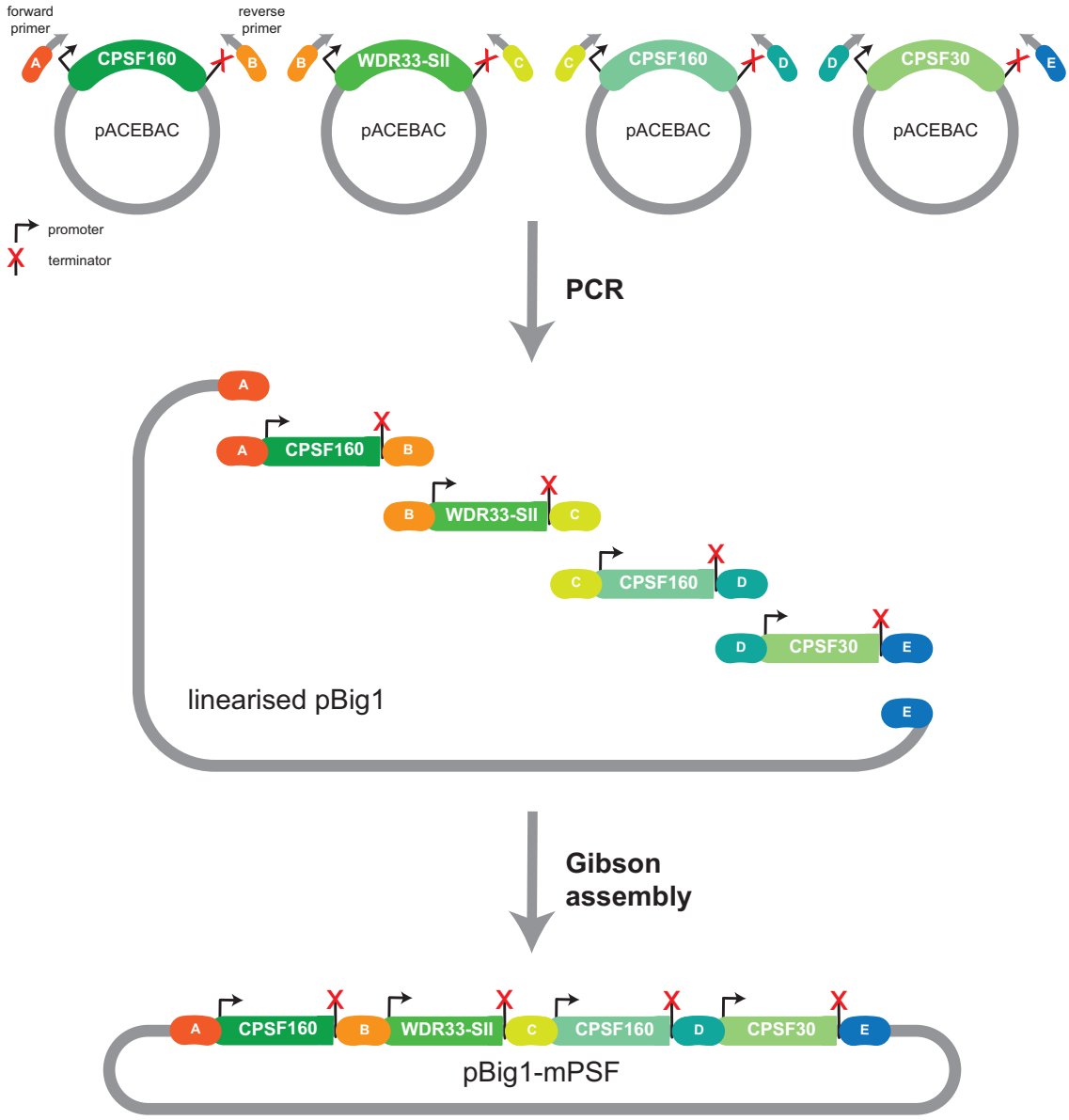


Figure 6.1 Schematic representation of the biGBac cloning protocol. Cloning of the mPSF complex is used as an example. Complementary overhangs are colour coded and marked with letters A-E. Based on (120).

6.1.3 Cloning for expression in *E. coli*

6.1.3.1 Cloning SSU72

To express SSU72 in *E. coli*, the coding region of SSU72 was amplified by PCR from its pACEBac vector. The forward primer contained an NdeI cleavage site, and the reverse primer had a BamHI cleavage site. After digestion with NdeI (NEB, cat. No. R0111) and BamHI-HF (NEB, cat. No. R3136) enzymes, the SSU72 coding region was ligated into an empty pET-28a vector that had been cleaved with the same enzymes using T4 DNA ligase (NEB, cat. No. M0202S). The vector contained an in-frame His_6-tag followed by a 3C protease cleavage site on its 5′ end.

6.1.3.2 Cloning NS1 and its effector domain for expression in *E. coli*

The sequences encoding either full-length $NS1_{R38A\ K41A}$ or the effector domain of NS1 were amplified from their pACEBAC vectors by PCR and inserted by Gibson assembly into a modified pMAL-c5X vector containing a 3C protease cleavage site followed by a maltose-binding protein tag on its 5′ end.

6.1.4 Cloning for expression in mammalian cells

6.1.4.1 CRISPR-Cas9 gene targeting in mammalian cells

Plasmids to target the 3′ end of the endogenous WDR33 gene were a kind gift from Steven West (University of Exeter). The sequence of the HTBH tag was purchased as a gBlock from IDT and inserted into a homology-directed repair plasmid by Gibson assembly (164).

6.1.4.2 Mammalian vectors for transient overexpression

To prepare pcDNA5 vectors encoding 3x-Flag-tagged subunits of CPSF, I first inserted the 3x-Flag sequence into an empty pcDNA5 plasmid. The sequence encoding the tag could not be purchased as a gBlock because of its repetitive sequences causing problems during

synthesis. Instead, the sequence was purchased from Sigma Aldrich as two complimentary oligonucleotides. The oligonucleotides were mixed, heated at 98°C for 5 min on a heat block. The heat block was then switched off, and the oligonucleotides were allowed to anneal overnight as the sample gradually cooled down. The assembled double-stranded DNA encoding the 3x-Flag tag also contained sites for selected restriction enzymes and was inserted into the pcDNA5 vector by restriction-ligation cloning. Specifically, for an N-terminal tag, both the DNA encoding the tag and the plasmid were digested with HindIII (NEB, cat. No. R0104) and BamHI-HF (NEB, cat. No. R3136) restriction endonucleases, while to insert a C-terminal tag BamHI-HF (NEB, cat. No. R3136) and XhoI (NEB, cat. No. R0146) enzymes were used. Subsequently, I inserted the open reading frames of CPSF subunits into the newly established pcDNA5 plasmids carrying the 3x-Flag tag on either the 3′ (for an N-terminal tag) or the 5′ (for a C-terminal tag) end. The CPSF genes were amplified from their pACEBAC vectors by PCR using primers that contain sites for specific restriction enzymes on either end: BamHI and XhoI for ligation into the vector with the tag on the 3′ end, and HindIII and BamHI – for insertion of the tag on the 5′ end of the gene. The genes were then ligated into the appropriate pcDNA5 vector digested with the same enzymes. Milligram quantities of the pcDNA5 plasmids were obtained by large-scale amplification in ~2 l TOP10 *E. coli* cells. The plasmid DNA was purified using a Giga plus plasmid prep kit (Qiagen, cat. No. 12991).

6.2 Protein expression

6.2.1 Baculovirus-mediated protein expression in insect cells

6.2.1.1 Bacmid preparation

pBig1 (mPSF, mCF, CStF, CFIIm, CFIm) or pACEBac (RBBP6, PAP, NS1-R28A-K41A, NS1-ED) vectors were transformed into *E. coli* EMBacY cells. EMBacY cells express a recombinase enzyme that inserts the genes from pBig1 and pACEBAC vectors into the backbone of a bacmid that encodes the genome of an insect baculovirus. The recombination event disrupts the bacmid gene coding for the β-galactosidase enzyme. The transformed EMBacY cells were plated onto agar plates containing X-Gal, and thus, colonies carrying successful insertions in the bacmid appeared white. White colonies were picked and small-scale cultures were prepared in sterile 2X YT media containing gentamicin and kanamycin for

bacmid amplification. After pelleting, the cells were lysed using the reagents and instructions of the Miniprep kit (Qiagen, cat. No. 271006X4). The lysate was mixed with two volumes of isopropanol, and the bacmid was allowed to precipitate for 2 h on ice. The precipitated DNA was pelleted at 20,000 g for 45 min and washed twice with 500 µl 70% ethanol. After the final wash, the supernatant was removed and the pellet was left to air-dry in a fume hood before resuspending in 50 µl sterile elution buffer from the Miniprep kit (Qiagen, cat. No. 271006X4). DNA concentration was measured with the NanoDrop instrument (Thermo Scientific).

6.2.1.2 Propagation of baculovirus

Extracted bacmids (~10 µg/well) were transfected into Sf9 insect cells grown on a 6-well plate (1 million cells/well) using FuGENE (Promega, cat. No. E2311) as a transfection reagent (~7.5 µl/well). The cells were grown at 27°C for 4-7 days until the fluorescence signal from the yellow fluorescent protein expressed constitutively by the baculovirus was clearly visible under a light microscope. The media containing the first passage (P1) of the virus was collected, mixed with an equal volume of FBS (Gibco, cat. No. A4766801) and stored for several years at 4°C.

To produce the P2 virus, Sf9 cells at a density of ~2 million cells/ml were infected with the P1 virus at a volume ratio 100:1 cell culture to virus. The cultures were grown for 3-4 days until the cells were brightly fluorescent. The P2 virus was harvested by collecting and filtering the media of infected cell. To overexpress proteins and protein complexes for purification, large-scale (>2 l) cultures of Sf9 cells (except for mPSF, which was overexpressed in Hi5 insect cells) were infected with the P2 virus (1% cell culture volume) at ~2 million cells/ml. The cells were harvested by centrifugation (1,000 g for 10 min at 4°C), when the cell viability fell below ~90% (after 3-4 days) and the cells were brightly fluorescent. The cell pellets were washed with ice-cold PBS, flash-frozen in liquid N_2 and stored at -80°C.

6.2.2 Protein expression in *E. coli*

>2 l of LB media containing ampicillin was inoculated with overnight small-scale starter cultures (1 ml per 1 l LB) of *E. coli* BL21(DE3) Star cells transformed with selected vectors.

The cultures were grown at 37°C until induced with 0.5 mM IPTG at OD_{600} ~0.6 and grown overnight at 20°C (SSU72), 25°C (NS1 ED) or 18°C ($NS1_{R38A\ K41A}$). The cells were harvested by centrifugation ~3,000 g for 10 min at 4°C, washed with ice-cold PBS, flash-frozen in liquid N_2 and stored at -80°C.

6.3 Protein purification

6.3.1 Protein analysis and quantification

The purity of all recombinant proteins across various steps of purification was assessed by SDS-PAGE. Samples were mixed with NuPAGE sample loading buffer (Invitrogen, cat. No. NP0007), boiled at 98°C for 5 min and loaded onto a NuPAGE 4-12% Bis-Tris 1.0 mm Mini Protein Gel (Invitrogen, cat. No. NP0321). The gels were run in MOPS buffer at 180 V for 50 min, stained with Instant Blue (Abcam, cat. No. 119211) and visualised using a BioRad Gel Doc XR+ instrument equipped with Image Lab Software. Variations of this protocol used in specific experiments are indicated in the text.

The concentration of pure protein samples, as assessed by SDS-PAGE analysis, was estimated using a NanoDrop (Thermo Scientific) instrument. The readings were adjusted based on the extinction coefficient of each protein or protein complex calculated using Expasy ProtParam web tool.

6.3.2 mPSF-hFip1$_{iso4}$-SII: on its own or co-expressed with either NS1-R38A-K41A or NS1-ED

A frozen cell pellet of Hi5 cells was thawed in lysis buffer (50 mM HEPES-NaOH pH 8.0, 300 mM NaCl, 1 mM TCEP, 2 mM $Mg(OAc)_2$), supplemented with 50 µg/ml DNaseI, 3 protease inhibitor tablets (Roche, cat. No. 11836153001) and 1 ml BioLock (IBA, cat. No. 2-0205-050) per 1 l cell culture. The cells were lysed by sonication, and the lysate was cleared by centrifugation. The cleared lysate was filtered through a 0.65 µm filter and incubated with Strep-Tactin beads (IBA, cat. No. 2-1201-025) for 2-3 h. The beads were washed with lysis buffer, and the complex was eluted with 2.5 mg/ml desthiobiotin (IBA, cat. No. 2-1000-005) in lysis buffer. The eluate was diluted to reduce the NaCl concentration to 75 mM, filtered through a 0.45 µm filter and applied to a 1 ml Resource Q column (Cytiva, cat. No. 17117701)

equilibrated in buffer A (20 mM HEPES-NaOH pH 8.0, 75 mM NaCl, 0.5 mM TCEP, 2 mM $Mg(OAc)_2$). The complex was eluted using a linear gradient of buffer B (20 mM HEPES-NaOH pH 8.0, 1 M NaCl, 0.5 mM TCEP, 2mM $Mg(OAc)_2$) over 50 column volumes. The peak fractions were pooled, concentrated and injected onto a Superose 6 XK 17/600 pg column (Cytiva, cat No. 71501695) equilibrated in size exclusion buffer (20 mM HEPES-NaOH pH 8.0, 150 M NaCl, 0.5 mM TCEP, 2 mM $Mg(OAc)_2$). Selected fractions were pooled and concentrated. The concentrated protein was aliquoted, flash-frozen in liquid N_2 and stored at -80°C.

6.3.3 mPSF-hFip1$_{FL}$-SII and mPSF-ΔhFip1-SII

mPSF-hFip1$_{FL}$ and mPSF-ΔhFip1 were purified from Hi5 cells by Strep-Tactin affinity chromatography and anion exchange chromatography as described for mPSF-hFip1$_{iso4}$, but the size exclusion step was omitted. The peak fractions of mPSF-hFip1$_{FL}$ and mPSF-ΔhFip1 from a 1 ml Resource Q column were pooled, aliquoted, flash-frozen in liquid N_2 and stored at -80°C.

6.3.4 mCF-SII, mCF-CPSF73$_{D75N/H76A}$-SII, mCF-symplekin$_{ΔNTD}$-SII, mCF-CPSF100$_{PIM\ MUT}$, mCF-CPSF100$_{ΔPIM}$, mCF-SII bound to CStF64

mCF and its variants were purified from Sf9 cells using the same protocol as mPSF but: 1) 50 μg/ml RNaseA was added to lysis buffer; 2) buffers were supplemented with 5% v/v glycerol before each concentration step; 3) size exclusion buffer contained 20 mM HEPES-NaOH pH 8.0, 150 M NaCl, 1 mM TCEP.

6.3.5 CStF-SII

CStF was purified from Sf9 cells using the same protocol as mCF, except that the size exclusion buffer contained 20 mM HEPES-NaOH pH 8.0, 200 mM NaCl, 1 mM TCEP.

6.3.6 CFIIm-SII

The various constructs of CFIIm were purified from Sf9 cells using the same protocol as mCF with a few modifications. In the lysis buffer, DNaseI and RNaseA were replaced by 50 U/ml benzonase (Merck, cat. No. E1014), and 100 µM PMSF (Merck, cat. No. 93482) was also added. The size exclusion buffer of CFIIm contained 20 mM Tris-HCl pH 8.5, 150 M NaCl, 0.5 mM TCEP, 5% v/v glycerol.

6.3.7 RBBP6-SII, RBBP6$_{Y228G}$-SII, RBBP6$_{P195G}$-SII, RBBP6$_{D43K\ R74E}$-SII

RBBP6 was purified from Sf9 cells using the same protocol as mPSF but with different buffers: Lysis buffer - 50 mM HEPES-NaOH pH 8.0, 400 mM NaCl, 1 mM TCEP; buffer A - 20 mM HEPES-NaOH pH 8.0, 40 mM NaCl, 0.5 mM TCEP; buffer B - 20 mM HEPES-NaOH pH 8.0, 1 M NaCl, 0.5 mM TCEP; size exclusion buffer - 20 mM HEPES-NaOH pH 8.0, 200 mM NaCl, 0.5 mM TCEP, 2 mM $Mg(OAc)_2$. Also, HiLoad 16/600 Superdex 200 pg column (Cytiva, cat. No. 28989335) was used for the size exclusion step.

6.3.8 RBBP6-UBL-SII

RBBP6-UBL was purified using a similar protocol as RBBP6. However, UBL domain did not bind to the Resource Q anion exchange column. The protein was already relatively pure, and therefore, the sample was concentrated and directly injected onto a Superdex 75 10/300 size exclusion column.

6.3.9 CFIm-SII

CFIm was purified from Sf9 cells by Strep-Tactin affinity chromatography and anion exchange chromatography as described for mPSF-hFip1$_{iso4}$ but using different buffers: lysis buffer - 50 mM bicine-NaOH pH 9.0, 400 mM NaCl, 0.5 mM TCEP, 2 mM $Mg(OAc)_2$, 10% v/v glycerol; buffer A - 20 mM bicine-NaOH pH 9.0, 150 mM NaCl, 0.5 mM TCEP, 2 mM $Mg(OAc)_2$, 10% v/v glycerol; buffer B - 20 mM bicine-NaOH pH 9.0, 1 M NaCl, 0.5 mM TCEP, 2 mM $Mg(OAc)_2$, 10% v/v glycerol. The peak fractions of CFIm from a 1 ml Resource Q column

were pooled, aliquoted, flash-frozen in liquid N_2 and stored at -80°C. Before running assays, ~ 100 µl CFIm was thawed and dialysed overnight against 500 ml dialysis buffer (20 mM bicine-NaOH pH 9.0, 400 mM NaCl, 0.5 mM TCEP, 2 mM $Mg(OAc)_2$, 10% v/v glycerol).

6.3.10 PAP-SII

PAP was purified from Sf9 cells by Strep-Tactin affinity chromatography as described for mPSF-hFip1$_{iso4}$. The eluate was incubated overnight at 4°C with 20 µg/ml TEV protease to remove the Strep-II tag. The protein was further purified using a 1 ml HiTrap Q column (Cytiva, cat. No. 29051325) equilibrated in buffer A (50 mM HEPES-NaOH pH 8.0, 100 mM NaCl, 1 mM TCEP) and eluted with a linear gradient of buffer B (50 mM HEPES-NaOH pH 8.0, 1 M NaCl, 1 mM TCEP). The peak fractions were concentrated and loaded onto a HiLoad 26/600 Superdex 200 pg column (Cytiva, cat. No. 28989336) equilibrated in buffer containing 50 mM HEPES-NaOH pH 8.0, 150 mM NaCl, 1 mM TCEP. The peak fractions were pooled, concentrated, and aliquoted. The aliquots were flash-frozen in liquid N_2 and stored at -80°C.

6.3.11 His_6-SSU72

Cells were lysed by sonication in buffer A (50 mM HEPES-NaOH pH 8.0, 500 mM NaCl, 1 mM TCEP, 20 mM imidazole) supplemented with 2 protease inhibitor tablets and 50 µg/ml DNaseI. The lysate was cleared by centrifugation and loaded onto a HisTrap HP 5 ml column (Cytiva, cat. No. 17524701) equilibrated in buffer A. The protein was eluted with a linear gradient of buffer B (50 mM HEPES-NaOH pH 8.0, 500 mM NaCl, 1 mM TCEP, 500 mM imidazole) over 20 column volumes. 43 µg/ml 3C protease was added to the pooled peak fractions to remove the His_6-tag, and the protein was dialysed overnight using a 7 kDa-cut-off membrane against dialysis buffer (50 mM HEPES-NaOH pH 8.0, 500 mM NaCl, 1 mM DTT). The dialysed sample was concentrated with 5% v/v glycerol and loaded onto a HiLoad Superdex 75 26/600 column (Cytiva, cat. No. 28989334) equilibrated in size exclusion buffer (20 mM HEPES-NaOH pH 8.0, 200 mM NaCl, 1 mM TCEP). The peak fractions were concentrated in the presence of 5% v/v glycerol, aliquoted and flash-frozen in liquid nitrogen. The protein was stored at -80°C.

6.3.12 $NS1_{R38A/K41A}$-MBP

Cells were lysed by sonication in lysis buffer (50 mM HEPES-NaOH pH 8.0, 500 mM NaCl) supplemented with 2 protease inhibitor tablets and 50 µg/ml DNaseI and RNaseA per 1 l culture. The lysate was cleared by centrifugation and applied to amylose resin (NEB, cat. No. E8021) equilibrated in lysis buffer. The beads were incubated for 2 h at 4°C and washed with lysis buffer. The protein was eluted in buffer containing 20 mM PIPES-NaOH pH 6.8, 150 mM NaCl, 1 mM TCEP and 20 mM maltose. The eluate was then incubated overnight with 3C protease (1:100 w/w protease to NS1) at 4°C. Visible precipitate formed after cleavage, and hence, the solution had to be cleared by centrifugation before loading onto a HiTrap Heparin HP 1 ml column (Cytiva, cat. No. 17040701) equilibrated in buffer A (20 mM PIPES-NaOH pH 6.8, 75 mM NaCl, 0.5 mM TCEP). The protein was eluted with a linear gradient of buffer B (20 mM PIPES-NaOH pH 6.8, 1 M NaCl, 0.5 mM TCEP), concentrated and loaded onto a HiLoad Superdex 200 pg 26/600 column (Cytiva, cat. No. 28989336). The peak fractions were concentrated in the presence of 5% v/v glycerol, aliquoted and flash-frozen in liquid nitrogen. The protein was stored at -80°C.

6.3.13 NS1-ED-MBP

NS1 ED was purified using amylose beads as described for full-length NS1. The protein eluted from amylose resin was concentrated and loaded onto a HiLoad Superdex 200 pg 26/600 column (Cytiva, cat. No. 28989336). The peak fractions were pooled, concentrated and incubated with 3C protease (1:100 w/w protease to NS1-ED-MBP) overnight at 4°C. NS1 ED was separated from free MBP using the HiLoad Superdex 75 pg 16/600 size exclusion chromatography column (Cytiva, cat. No. 28989333).

6.4 Preparation of RNA substrates

Sequences of the RNAs used in this study are listed in Table 6.2.

5′-FAM fluorescently-labelled 41 nt L3 RNA with either wild-type (AAUAAA) or mutant PAS (AACAAA) and 5′-FAM fluorescently-labelled 60 nt synthetic PAS (143) were synthesised by Integrated DNA Technologies (IDT).

The DNA sequences encoding fragments of the 218 nt SV40 pre-mRNA both containing wild-type and mutant PAS sequences were purchased as gBlocks from IDT. The sequence of the T7 RNA polymerase promoter was added to the 5′ end of the gBlock by PCR amplification.

The template of the 520 nt L3 pre-mRNA was purchased from IDT as a gBlock. The fragment had a KpnI (NEB, cat. No. R0142) cleavage site on its 5′ end and a BamHI site on its 3′ end. After restriction digest with both enzymes, the L3 fragment was ligated into a linearised pUCIDT plasmid encoding the T7 RNA polymerase promoter followed by three MS2 loops upstream of the insert. Mutations in the main PAS and secondary PAS were introduced by overlap extension PCR, using primers carrying the desired mutations. The sequence encoding the 98 nt fragment of the L3 RNA substrate was subcloned by PCR.

DNA encoding the 218 nt SV40 pre-mRNA with the randomised sequence between its PAS and the cleavage site, including the three CAA motifs, was synthesised by IDT. The gBlock was then processed as described above.

All pre-mRNA substrates were transcribed using HiScribe T7 High Yield RNA Synthesis Kit (NEB, cat. No. E2040) and subsequently purified with Monarch RNA Cleanup Kit (NEB, cat. No. T2040).

Name	Sequence
5′-FAM-L3-WT (41nt)	AUGAUCUAGGAGACACAAUAAAGGCAAUGUUUUUAUUUGUA
5′-FAM-L3-MUT (41nt)	AUGAUCUAGGAGACACAACAAAGGCAAUGUUUUUAUUUGUA
MS2-L3-WT (520 nt)	GGGGUCUAGACCUCGAGAAGCUUCGUACACCAUCAGGGUACGAG CUUGCCCUUGGCGUACACCAUCAGGGUACGACUAGUAUAUCUCG UACACCAUCAGGGUACGGAAUUCGGUACCCAACUCCAUGCUUAA CAGUCCCCAGGUACAGCCCACCCUGCGUCGCAACCAGGAACAGCU CUACAGCUUCCUGGAGCGUCACUCGCCCUACUUCCGCAGCCACAG UGCGCAGAUUAGGAGCGCCACUUCUUUUUGUCACUUGAAAAACA UGUAAAAAUAAUGUACUAGGAGACACUUUCAAUAAAGGCAAAU GUUUUUAUUUGUACACUCUCGGGUGAUUAUUUACCCCCCACCCU UGCCGUCUGCGCCGUUUAAAAAUCAAAGGGGUUCUGCCGCGCAU CGCUAUGCGCCACUGGCAGGGACACGUUGCGAUACUGGUGUUUA GUGCUCCACUUAAACUCAGGCACAACCAUCCGCGGCAGCUCGGUG AAGUUUUCACUCCACAGGCUGCGCACCAUGACGG
MS2-L3-MUT (520 nt)	GGGGUCUAGACCUCGAGAAGCUUCGUACACCAUCAGGGUACGAG CUUGCCCUUGGCGUACACCAUCAGGGUACGACUAGUAUAUCUCG UACACCAUCAGGGUACGGAAUUCGGUACCCAACUCCAUGCUUAA CAGUCCCCAGGUACAGCCCACCCUGCGUCGCAACCAGGAACAGCU CUACAGCUUCCUGGAGCGUCACUCGCCCUACUUCCGCAGCCACAG UGCGCAGAUUAGGAGCGCCACUUCUUUUUGUCACUUGAAAAACA UGUAAAAAUAAUGUACUAGGAGACACUUUCAACAAAGGCAAAU GUUUUUAUUUGUACACUCUCGGGUGAUUAUUUACCCCCCACCCU UGCCGUCUGCGCCGUUUAAAAAUCAAAGGGGUUCUGCCGCGCAU CGCUAUGCGCCACUGGCAGGGACACGUUGCGAUACUGGUGUUUA GUGCUCCACUUAAACUCAGGCACAACCAUCCGCGGCAGCUCGGUG AAGUUUUCACUCCACAGGCUGCGCACCAUGACGG
SV40-WT (218 nt)	GGGAGACAUGAUAAGAUACAUUGAUGAGUUUGGACAAACCACA ACUAGAAUGCAGUGAAAAAAAUGCUUUAUUUGUGAAAUUUGUG AUGCUAUUGCUUUAUUUGUAACCAUUAUAAGCUGCAAUAAACA AGUUAACAACAACAAUUGCAUUCAUUUUAUGUUUCAGGUUCAG GGGGAGGUGUGGGAGGUUUUUUAAAGCAAGUAAAACCUCGACG GAU

SV40-MUT (218 nt)	GGGAGACAUGAUAAGAUACAUUGAUGAGUUUGGACAAACCACA ACUAGAAUGCAGUGAAAAAAAUGCUUUAUUUGUGAAAUUUGUG AUGCUAUUGCUUUAUUUGUAACCAUUAUAAGCUGCAACAAACA AGUUAACAACAACAAUUGCAUUCAUUUUAUGUUUCAGGUUCAG GGGGAGGUGUGGGAGGUUUUUUAAAGCAAGUAAAACCUCGACG GAU
L3-WT (98 nt)	GGGUGUACUAGGAGACACUUUCAAUAAAGGCAAAUGUUUUUAU UUGUACACUCUCGGGUGAUUAUUUACCCCCCACCCUUGCCGUCU GCGCCGUUUAA
5′-FAM-Synthetic PAS (60 nt)	AGACACAAUAAAAGAUCUUUAUUUUCAUUAGAUCUGUGUGUUG GUUUUUUGUGUGCCUGG

Table 6.2 List RNA substrates used in this study. WT – wild-type PAS sequence (highlighted in green); MUT – mutant PAS sequence (highlighted in red). MS2 loops are coloured in purple. Other variants of RNAs listed here are described in the main text and Figure 3.5 and Figure 3.6.

6.5 Assays with recombinant CPSF

6.5.1 Polyadenylation assays

Each protein factor was first diluted in polyadenylation assay buffer (20 mM HEPES-NaOH pH 8.0 (measured at room temperature), 150 mM NaCl, 0.5 mM TCEP, 2 mM $Mg(OAc)_2$). The diluted proteins were then mixed on ice in a total volume 19 µl per condition and/or time point at the final concentration of 50 nM PAP and 50 nM mPSF/CPSF, unless indicated otherwise in the text. The volume was adjusted by adding assay buffer supplemented with 3 mM DTT and 1 U/µl RiboLock (Thermo, cat. No. EO0381). 41 nt L3 RNA substrate was mixed with ATP (Thermo Scientific, cat. No. R0441) and both components were added to initiate the reaction at the final concentrations of 2 mM ATP and 300 nM RNA. The assays were stopped at selected times by adding 5 µl stop buffer (130 mM EDTA, 5% v/v SDS, 12 mg/ml proteinase K in polyadenylation assay buffer) and incubating them at 37°C for a further 15 min. The samples were mixed with RNA Gel Loading Dye (Thermo Scientific, cat. No. R0641) and loaded on a pre-run (30 W for 30 min) denaturing 15% polyacrylamide gel containing 7 M urea in TBE buffer. The gels were run for ~25 min at 400 V. RNA was visualised using a FAM channel on a Typhoon FLA 7000 instrument (GE Healthcare).

6.5.2 Cleavage assays

Each protein factor was first diluted in protein dilution buffer (20 mM HEPES-NaOH pH 7.25 (measured at room temperature), 150 mM NaCl, 0.5 mM TCEP, 2 mM $Mg(OAc)_2$). Individually purified mPSF and mCF complexes at 2.5 µM each were mixed in protein dilution buffer and incubated on ice for 30 min. All protein components were then mixed on ice in 19 µl per condition and/or time point at the final concentrations of 50 nM CPSF, 100 nM CStF, 100 nM CFIIm and 300 nM RBBP6 in a buffer containing 20 mM HEPES-NaOH pH 7.25 (measured at room temperature; 7.0 at 37°C calculated using a web calculator https://www.liverpool.ac.uk/pfg/Tools/BuffferCalc/Buffer.html), 50 mM NaCl, 0.5 mM TCEP, 2 $Mg(OAc)_2$ and 1 U/µl RiboLock (Thermo, cat. No. EO0381). The tubes were transferred to 37°C, and the reaction was initiated by addition of the RNA substrate to a final concentration of 100 nM. Unless indicated otherwise, the reactions were stopped after 150 min by adding 5 µl stop buffer (130 mM EDTA, 5% v/v SDS, 12 mg/ml proteinase K in protein dilution buffer) and incubating them at 37°C for a further 15 min. The samples were mixed with RNA Gel Loading Dye (Thermo Scientific, cat. No. R0641) and loaded on a pre-

run (30 W for 30 min) denaturing 10% (218 nt SV40), 6% (520 nt L3) or 15% (60 nt synthetic PAS and 98 nt SV40) polyacrylamide gel containing 7 M urea in TBE buffer. The gels were run for ~25 min at 400 V, stained with SYBR Green (Invitrogen, cat. No. S7564) and imaged using a ChemiDoc XRS+ (BioRad).

6.5.3 Coupled cleavage and polyadenylation assays

Cleavage reactions were set up as described above. To test polyadenylation, PAP was added to the cleavage reaction at a final concentration of 50 nM, unless indicated otherwise. ATP (Thermo Scientific, cat. No. R0441) was mixed with the RNA substrate and both components were added simultaneously to the protein mix, containing CPSF and cleavage factors, to start the reaction. The final concentration of ATP in the reaction was 2 mM. The assays were run and analysed as described above for cleavage-only assays.

6.5.4 Sequencing of 5′ cleavage products

A standard cleavage reaction of the SV40 substrate was analysed on a denaturing gel as described above. The band corresponding to the 5′ cleavage product was excised and submerged in 50 µl crush and soak buffer (3 M Na(OAc) pH 5.2, 0.1 M EDTA pH 7.4, 20% v/v SDS). The gel band was crushed with a sterile pipette tip and incubated overnight at 37°C. After taking off the supernatant, the same steps were repeated with 50 µl fresh crush and soak buffer for 2 h. The two supernatants were combined, and the extracted RNA was precipitated at -20°C for 2 h in 300 µl absolute ethanol with 1µl Glycoblue (Invitrogen, cat. No. AM9516). The RNA was pelleted in a chilled microcentrifuge at maximum speed for 10 min and washed with 500 µl 70% ethanol. The RNA pellet was resuspended in 20 µl DEPC water. An adenylated adaptor of a known sequence was ligated to the 3′ end of the extracted 5′-cleavage product using T4 RNA ligase 2, truncated (NEB, cat. No. M0242). The RNA was purified from the ligation reaction components using Monarch RNA Cleanup Kit. The 5′ cleavage products that contained the adaptor were converted into cDNA using SuperScript IV First-Strand Synthesis System (Invitrogen, cat. No. 18091050) with a forward primer specific to a 5′-region of the SV40 RNA and a reverse primer that anneals to the adaptor. The cDNA was further amplified by PCR and ligated into a bacterial vector using Zero Blunt PCR Cloning Kit (Invitrogen, cat. No. K270040). After transformation into TOP10 *E. coli* cells,

15 colonies were picked, and the isolated plasmids were sequenced using the M13R primer (Source Bioscience) to determine the 3′ end of the 5′ cleavage product.

The same protocol was employed for preliminary experiments investigating cleavage specificity using 218 nt SV40 pre-mRNA libraries, in which the sequence between its PAS and cleavage site was randomised.

6.5.5 Assays with JTE-607 acid compound and assay quantification

The prodrug of JTE-607 was purchased from Tocris and hydrolysed to JTE-607 acid analog by Thomas Elliott from Jason Chin's group (MRC LMB) as previously described (93, 100).

Standard cleavage assays were set up in the presence of various concentrations of the acid form of JTE-607, and the samples were analysed by denaturing polyacrylamide gel electrophoresis as described above. The relative activity of CPSF at each concentration of the drug was calculated as the relative intensity of the cleavage product bands in each lane relative to this ratio in the absence of the drug (x - JTE-607 concentration):

The intensity values were measured in Fiji. The data were plotted in Prism 9 and fitted to the equation of "[Inhibitor] vs. response - Variable slope (four parameters)" with an R^2 value of 0.9656.

The same formula of "relative activity" depicted above was used to quantify the effects of PAP and CFIm on the cleavage activity of CPSF. In these cases, x – concentration of either PAP or CFIm.

6.6 Pull-downs of endogenous CPSF from mammalian cells

6.6.1 Pull-downs using transient transfection of a tagged subunit

3.3 mg of pcDNA5 plasmid carrying a 3x-Flag-tagged CPSF subunit was mixed with 9 mg polyethylenimine transfection reagent in Expi293 Expression Medium (Thermo Scientific, cat. No. A1435102) in the final volume of 200 ml. The resultant transfection solution was incubated at room temperature for 15 min and then added to 3 l of Expi293 suspension cells at a cell density of 1.5-2 million cells/ml grown in Expi293 Expression Medium. The expression of the tagged subunit was induced by addition of doxycycline (Sigma Aldrich, cat. No. D9891) at a final concentration of 30 ng/ml. The Expi293 cells were grown for 3-4 days until reaching cell density of ~6 million cells/ml. The cells were pelleted by centrifugation to 1,000 g for 10 min at 4°C and washed with ice-cold PBS.

The washed cell pellet was resuspended in purification buffer (50 mM Tris-HCl pH 8.0, 300 mM NaCl, 2 mM $Mg(OAc)_2$) and lysed by sonication. The lysate was cleared by centrifugation and subsequent filtration through a PVDF membrane with a pore diameter of 0.65 μm. The cleared lysate was then incubated with anti-Flag magnetic agarose (Thermo Scientific, cat. No. A36797) for 2h at 4°C. The beads were pelleted using a magnetic stand and washed three times with purification buffer. The bound protein was eluted with 0.25 mg/ml Flag peptide in purification buffer. The eluates were analysed by SDS-PAGE, and the constituent proteins were determined by tandem mass spectrometry. The mass spectrometry data was analysed using Scaffold 4 software.

6.6.2 Pull-down using tagged endogenous WDR33 subunit

Two stable HEK293T cell lines in which the endogenous WDR33 subunit carried either a C-terminal HTBH tag or a C-terminal TAPS tag were generated using an established protocol for CRISPR-Cas9-based gene targeting (109, 164). The correct clones were identified by sequencing and Western blotting.

HEK293T cells were grown on 150 mm dishes in high glucose GlutaMAX DMEM media (Gibco, cat. No 10566016) supplemented with 10% foetal bovine serum and penicillin-streptomycin, until the cells were ~90% confluent. Native CPSF was purified either from

total cell extract (replicate HTBH-1 and TAPS-1) or from nuclear extract (replicates HTBH-2 and HTBH-3).

In experiments HTBH-1 and TAPS-1, either HEK293T-WDR33-HTBH or HEK293T-WDR33-TAPS cells were harvested using a cell scraper, washed in PBS and resuspended in hypotonic lysis buffer (20 mM HEPES-NaOH pH 8.0, 2 mM $Mg(OAc)_2$, 2 mM EDTA, 1 mM EGTA, 1 mM DTT, 10% glycerol) supplemented with protease inhibitor tablets and 100 μM PMSF. Total cell extract was prepared by freeze-thaw lysis before adjusting the NaCl concentration to 300 mM. The lysate was clarified by centrifugation and incubated with Strep-Tactin beads. The beads were washed in buffer containing 50 mM HEPES-NaOH pH 8.0, 300 mM NaCl, 1 mM DTT, 2 mM $Mg(OAc)_2$, 10% v/v glycerol. The complex containing WDR33-HTBH was eluted from the beads in the same buffer by cleavage with the TEV protease, while the complex carrying WDR33-TAPS was eluted in wash buffer supplemented with 12 mM desthiobiotin (IBA, cat. No. 2-1000-005). The TEV protease remained in the eluted sample purified from HEK293T-WDR33-HTBH cells.

In experiment HTBH-2, nuclear extract of the HEK293T-WDR33-HTBH cell line was prepared using homogenisation. The cell pellet was resuspended in hypotonic lysis buffer (10 mM HEPES-KOH, pH 7.9, 10 mM KCl, 1 m DTT, 1.5 mM $MgCl_2$) supplemented with protease inhibitor tablets and 100 μM PMSF. The cells were incubated on ice, and the intact nuclei were isolated by centrifugation. The pellet containing the nuclei was resuspended in extraction buffer (20 mM HEPES-KOH pH 7.9, 420 mM KCl, 1 m DTT, 1.5 mM $MgCl_2$, 0.2 mM EDTA, 25% v/v glycerol). The nuclei were lysed by homogenization, and the nuclear extract was clarified by centrifugation. The breakdown of the nuclei was checked by Trypan blue staining. The extract was diluted to the final KCl concentration of 300 mM before applying the sample to Strep-Tactin beads. CPSF was purified as described in experiment 1.

In experiment HTBH-3, nuclear extract of the HEK293T-WDR33-HTBH cell line was prepared using detergent lysis. The cell pellet was resuspended in lysis buffer (10 mM HEPES-KOH pH 8.0, 100 mM KCl, 2 mM $Mg(OAc)_2$, 0.3 M sucrose, 0.2% v/v Igepal (Merck, cat. No. I3021), 1 mM TCEP). The cells were incubated on ice, and the intact nuclei were isolated by centrifugation. The pellet containing the nuclei was resuspended in extraction buffer (20 mM HEPES-KOH pH 8.0, 300 mM KCl, 2 mM $Mg(OAc)_2$, 10% v/v glycerol, 0.2% v/v Igepal, 1 mM TCEP). The breakdown of the nuclei was checked by Trypan blue staining. The nuclear extract was clarified by centrifugation, and the sample was applied to Strep-Tactin beads. CPSF was purified as described in experiment 1.

The eluate from each experiment was analysed by SDS-PAGE as described above. An anti-CPSF73 antibody (Bethyl, cat. No. A301-090A) was used to follow the enrichment of CPSF73 across various purification steps by Western blotting. The gels were stained either with SYPRO Ruby (Invitrogen, cat. No. S12000) for preparations from HEK293T-WDR33-HTBH cells or with Silver stain (Thermo Scientific, cat. No. 24612) in experiments that used HEK293T-WDR33-TAPS cells. The samples were also subjected to protein identification by tandem mass spectrometry. Mass spectrometry data were analysed using Scaffold4 software.

6.7 Pull-downs from insect cells

A P2 virus encoding a tagged protein and a P2 virus carrying a gene or genes of untagged proteins, the binding of which to the tagged protein was being investigated, were used to co-infect Sf9 cells at ~2 million cells/ml. The cultures were harvested after 3 days by centrifugation and washed in ice-cold PBS. The cell pellets were lysed using glass beads (Merck, cat. No. G8772) in lysis buffer (50 mM HEPES-NaOH pH 8.0, 300 mM NaCl, 1 mM TCEP, 2 mM $Mg(OAc)_2$) supplemented with 2 protease inhibitor tablets per 50 ml buffer. The lysates were cleared by centrifugation and applied to Strep-Tactin beads. After a 2 h incubation, the beads were washed in lysis buffer, and the bound proteins were eluted by incubating the samples in NuPAGE LDS Sample Buffer (Invitrogen, cat. No. NP0007) at 98°C for 2 min. The eluted proteins were analysed by SDS-PAGE.

6.8 Analytical gel filtration chromatography

Individual protein components were mixed on ice at concentrations indicated in figure legends and incubated for at least 30 min. The samples were loaded onto a Superose 6 Increase 3.2/300 column (Cytiva, cat No. 29091598) equilibrated in a buffer containing HEPES-NaOH pH 8.0, 50 mM NaCl, 0.5 mM TCEP, which closely matched the conditions in cleavage assays. However, experiments to test SSU72 binding to CPSF and mCF variants as well as assembly of CPSF from mCF and mPSF modules were performed in a buffer containing 20 mM HEPES-NaOH pH 8.0, 150 mM NaCl, 0.5 mM TCEP. The protein content of the peak fractions was analysed by SDS-PAGE. To detect the 41 nt L3 RNA in some

experiments, stop buffer was added to an aliquot of each fraction. After incubation at 37°C for 10 min, RNA loading dye was added, and the samples were loaded onto 15% Novex TBE-Urea gels (300 V, 50 min). The gels were scanned using a FAM channel on a Typhoon FLA 7000 instrument (GE Healthcare).

6.9 Electromobility shift assay (EMSA)

Indicated concentrations of various point mutants of RBBP6 were mixed with 100 nM 41-nt 5′-FAM-labelled L3 RNA and orange G loading dye (0.4% w/v orange G, 50% v/v glycerol, 1 mM EDTA). The protein-RNA mixtures were incubated on ice for 15 min and then loaded onto a 10% native polyacrylamide gel. The gel was run for 50 min at 100 V at 4°C. The RNA was visualised using a FAM channel on a Typhoon FLA 7000 instrument (GE Healthcare).

6.10 *In vitro* pull-downs on M2-L3 pre-mRNA

The pull-downs were performed in pull-down buffer containing 20 mM HEPES-NaOH pH 8.0, 50 mM NaCl, 0.5 mM TCEP, 2 mM $Mg(OAc)_2$. First, 520-nt MS-L3 pre-mRNA was incubated with MBP-tagged MS2 protein at molar ratio 1:3 for 45 min on ice. 3 μM RBBP6, 1 μM CPSF or 3 μM RBBP6 + 1μM CPSF were then added and incubated for 1.5 h. The mixture containing RBBP6/CPSF/RBBP6+CPSF and MBP-MS2-bound L3 pre-mRNA was mixed with amylose beads (NEB, cat. No. E8021) equilibrated in pull-down buffer and incubated rotating at 4°C for 1.5 h. The beads were washed with pull-down buffer. Protein-RNA complexes were eluted in pull-down buffer supplemented with 20 mM maltose (Merck, cat. No. 63418). The eluates were loaded on a NuPAGE 4-12% Bis-Tris 1.0 mm Mini Protein Gel. The proteins were transferred onto a nitrocellulose membrane using Trans-Blot Turbo Transfer System (Bio-Rad, cat. No. 1704158). Strep-II-tagged proteins (RBBP6, symplekin and WDR33) were detected using streptavidin-HRP conjugate (Merck Millipore, cat. No. 18152) and Amershan ECL Detection Reagents (Cytiva, cat. No. RPN2106). The blots were visualised using a ChemiDoc XRS+ (BioRad).

6.11 Cryo-electron microscopy (cryoEM)

6.11.1 mPSF-NS1$_{R38A/K41A}$

The purified mPSF-NS1 complex at a concentration of 3 µM was cross-linked with 2 mM BS3 for 30 min on ice. The reaction was quenched with 180 mM Tris-HCl pH 8.0. The cross-linked sample was then loaded on a Superose 6 Increase 3.2/300 column (Cytiva, cat No. 29091598) equilibrated in a buffer containing 20 mM HEPES-NaOH pH 8.0, 150 mM NaCl, 0.5 mM TCEP, 2 mM $Mg(OAc)_2$. The cross-linked samples were analysed by SDS-PAGE on a NuPAGE 3-8% Tris-acetate gel (Invitrogen, cat. No. EA0375PK2) run for ~2 h at 180V. The peak fractions were concentrated to ~0.25 mg/ml.

UltrAuFoil® 1.2/1.3 grids (Quantifoil, cat. No. N1-A14nAu30-50) (177) were plasma-treated using a Fiscione NanoClean plasma cleaner. 3 µl of the cross-linked sample was applied onto the grid, excess sample was blotted away with a filter paper for 4 s with a blot force of -10, and the grid was flash-frozen in liquid ethane using a Vitrobot IV instrument (Thermo Fisher). Data was collected on a Titan Krios I electron microscope (FEI) at eBIC (Diamond Light Source) equipped with a K3 direct electron detector in counting mode. 8,226 multi-frame movies were collected at a pixel size of 1.06 Å/px in a defocus range of -1.7 µm to -3.3 µm. The data were processed using Relion 3.1 (178). Beam-induced motion was corrected using MotionCor2, and CTF was estimated using Ctffind (179, 180). After 2D classification, class averages with secondary structure features were used to generate an initial model for 3D classification. The strategy for 3D classification is described in the main text and Figure 3.9.

6.11.2 CPSF-RBBP6-RNA

mPSF (2 µM), mCF (2 µM), RBBP6 (6 µM) and 41 nt L3 RNA (2.5 µM) were incubated with 1 mM sulfoSDA (Thermo Scientific, cat. No. 26173) for 2 h on ice. The sample was then illuminated with UV light of a wavelength of 350 nM for 15 min. The cross-linked sample was run on a Superose 6 Increase 3.2/300 column (Cytiva, cat No. 29091598) equilibrated in a buffer containing 20 mM HEPES-NaOH pH 8.0, 150 mM NaCl, 0.5 mM TCEP, 2 mM $Mg(OAc)_2$. The peak fractions were analysed by SDS-PAGE using a NuPAGE 3-8% Tris-

acetate gel (Invitrogen, cat. No. EA0376BOX). The peak fractions were concentrated to ~0.8 mg/ml.

UltrAuFoil® 1.2/1.3 grids (Quantifoil, cat. No. N1-A14nAu30-50) (177) were glow-discharged for 90 s using an Edwards Sputter Coater S150B configured to setting 8. 3 µl of the cross-linked sample was applied onto the grid, excess sample was blotted away with a filter paper for 5.5 s with a blot force of -10, and the grid was flash-frozen in liquid ethane using a Vitrobot IV instrument (Thermo Fisher). Data was collected on a Titan Krios I electron microscope (FEI) at MRC LMB equipped with a K3 direct electron detector in counting mode. 4,096 multi-frame movies were collected at a pixel size of 0.73 Å/px in a defocus range of -1.0 to -3.0. The data were processed using Relion 3.1, as described in the main text and above, except that Gctf was used for CTF estimation (178, 181).

6.11.3 mCF-CFIIm-RBBP

mCF (2.5 µM), CFIIm (4 µM) and RBBP6 (7.5 µM) were mixed in buffer containing 20 mM HEPES-NaOH pH 8.0, 50 mM NaCl, 0.5 mM TCEP, and run on a Superose 6 Increase 3.2/300 column (Cytiva, cat No. 29091598) equilibrated in the same buffer. The peak fractions were concentrated to ~0.8 mg/ml.

UltrAuFoil® 0.6/1 grids (177) were glow-discharged for 90 s using an Edwards Sputter Coater S150B configured to setting 8. 3 µl of the cross-linked sample was applied onto the grid, excess sample was blotted away with a filter paper for 2 s with a blot force of -15, and the grid was flash-frozen in liquid ethane using a Vitrobot IV instrument (Thermo Fisher). Data was collected on a Titan Krios I electron microscope (FEI) at MRC LMB equipped with a K3 direct electron detector in counting mode. 320 multi-frame movies were collected at a pixel size of 1.17 Å/px in a defocus range of -1.0 to -3.0. The data were processed using Relion 3.1, as described in the main text and above (178).

Chapter 7: References

1. Hocine S, Singer RH, Grünwald D. 2010. RNA processing and export. *Cold Spring Harb. Perspect. Biol.* 2:a000752

2. Tian B, Manley JL. 2016. Alternative polyadenylation of mRNA precursors. *Nat. Rev. Mol. Cell Biol.* 18:18–30

3. Gruber AJ, Zavolan M. 2019. Alternative cleavage and polyadenylation in health and disease. *Nat. Rev. Genet.* 20:599–614

4. Passmore LA, Coller J. 2021. Roles of mRNA poly(A) tails in regulation of eukaryotic gene expression. *Nat. Rev. Mol. Cell Biol.* 23:93–106

5. Buratowski S. 2005. Connections between mRNA 3′ end processing and transcription termination. *Curr. Opin. Cell Biol.* 17:257–61

6. Shi Y, Di Giammartino DC, Taylor D, Sarkeshik A, Rice WJ, et al. 2009. Molecular architecture of the human pre-mRNA 3′ processing complex. *Mol. Cell*. 33:365–76

7. Kumar A, Clerici M, Muckenfuss LM, Passmore LA, Jinek M. 2019. Mechanistic insights into mRNA 3′-end processing. *Curr. Opin. Struct. Biol.* 59:143–50

8. Zhang Y, Sun Y, Shi Y, Walz T, Tong L. 2020. Structural insights into the human pre-mRNA 3′-end processing machinery. *Mol. Cell*. 77:800-9

9. Casañal A, Kumar A, Hill CH, Easter AD, Emsley P, et al. 2017. Architecture of eukaryotic mRNA 3′-end processing machinery. *Science.* 358:1056–59

10. Schäfer P, Tüting C, Schönemann L, Kühn U, Treiber T, et al. 2018. Reconstitution of mammalian cleavage factor II involved in 3′ processing of mRNA precursors. *Rna*. 24:1721–37

11. Yang W, Hsu PL, Yang F, Song JE, Varani G. 2018. Reconstitution of the CstF complex unveils a regulatory rolefor CstF-50 in recognition of 3-end processing signals. *Nucleic Acids Res.* 46:493–503

12. Hill CH, Boreikaitė V, Kumar A, Casañal A, Kubík P, et al. 2019. Activation of the endonuclease that defines mRNA 3′ ends requires incorporation into an 8-subunit core cleavage and polyadenylation factor complex. *Mol. Cell*. 73:1217–31

13. Skolnik-David H, Moore CL, Sharp PA. 1987. Electrophoretic separation of polyadenylation-specific complexes. *Genes Dev.* 3:672–82

14. Humphrey T, Christofori G, Lucijanic V, Keller W. 1987. Cleavage and polyadenylation of messenger RNA precursors in vitro occurs within large and specific 3′ processing complexes. *EMBO J.* 6:4159–68

15. Takagaki Y, Ryner LC, Manley JL. 1988. Separation and characterization of a poly(A) polymerase and a cleavage/specificity factor required for pre-mRNA polyadenylation. *Cell*. 52:731–42

16. Bienroth S, Wahle E, Suter-Crazzolara C, Keller W. 1991. Purification and characterization of the cleavage and polyadenylation specificity factor in involved in the 3′ processing of messenger RNA precursors. *J. Biol. Chem.* 266:19768–76

17. Chan SL, Huppertz I, Yao C, Weng L, Moresco JJ, et al. 2014. CPSF30 and Wdr33 directly bind to AAUAAA in mammalian mRNA 3′ processing. *Genes Dev.* 28:2370–80

18. Gavin A-C, Krause R, Grandi P, Marzioch M, Bauer A, et al. 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*. 415:141–47

19. Preker PJ, Ohnacker M, Minvielle-Sebastia L, Keller W. 1997. A multisubunit 3′ end processing factor from yeast containing poly(A) polymerase and homologues of the subunits of mammalian cleavage and polyadenylation specificity factor. *EMBO J.* 16:4727–37

20. Schönemann L, Kühn U, Martin G, Schäfer P, Gruber AR, et al. 2014. Reconstitution of CPSF active in polyadenylation: Recognition of the polyadenylation signal by WDR33. *Genes Dev.* 28:2381–93

21. Sun Y, Zhang Y, Hamilton K, Manley JL, Shi Y, et al. 2018. Molecular basis for the recognition of the human AAUAAA polyadenylation signal. *Proc. Natl. Acad. Sci. U. S. A.* 115:1419-28

22. Clerici M, Faini M, Aebersold R, Jinek M. 2017. Structural insights into the assembly and polyA signal recognition mechanism of the human CPSF complex. *eLife*. 6:1–20

23. Clerici M, Faini M, Muckenfuss LM, Aebersold R, Jinek M. 2018. Structural basis of AAUAAA polyadenylation signal recognition by the human CPSF complex. *Nat. Struct. Mol. Biol.* 25:135-38

24. Kumar A, Yu CWH, Rodríguez-Molina JB, Li X-H, Freund SMV, Passmore LA. 2021. Dynamics in Fip1 regulate eukaryotic mRNA 3′ end processing. *Genes Dev*. 35:1510-26

25. Hamilton K, Tong L. 2020. Molecular mechanism for the interaction between human CPSF30 and hFip1. *Genes Dev.* 34:1753-61

26. Muckenfuss LM, Herranz ACM, Boneberg FM, Clerici M, Jinek M. 2022. Fip1 is a multivalent

interaction scaffold for processing factors in human mRNA 3′ end biogenesis. *eLife* 11:e8033

27. Proudfoot NJ, Brownlee GG. 1976. 3′ Non-coding region sequences in eukaryotic messenger RNA. *Nature*. 263:211–14

28. Tian B, Graber JH. 2012. Signals for pre-mRNA cleavage and polyadenylation. *Wiley Interdiscip. Rev. RNA*. 3:385–96

29. Hamilton K, Sun Y, Tong L. 2019. Biophysical characterizations of the recognition of the AAUAAA polyadenylation signal. *Rna*. 25:1673–80

30. Sheets MD, Ogg SC, Wickens MP. 1990. Point mutations in AAUAAA and the poly(A) addition site: Effects on the accuracy and efficiency of cleavage and polyadenylation in vitro. *Nucleic Acids Res.* 18:5799–805

31. Beaudoing E, Freier S, Wyatt JR, Claverie JM, Gautheret D. 2000. Patterns of variant polyadenylation signal usage in human genes. *Genome Res.* 10:1001–10

32. Keller W, Bienroth S, Lang KM, Christofori G. 1991. Cleavage and polyadenylation factor CPF specifically interacts with the pre-mRNA 3′ processing signal AAUAAA. *EMBO J.* 10:4241–49

33. Rodriguez JB, O'Reilly FJ, Fagarasan H, Sheekey E, Maslen S, et al. 2022. Mpe1 senses the binding of pre-mRNA and controls 3′ end processing by CPF. *Mol. Cell*. 82:2490-504

34. Balbo PB, Bohm A. 2007. Mechanism of poly(A) polymerase: Structure of the enzyme-MgATP-RNA ternary complex and kinetic analysis. *Structure*. 15:1117–31

35. Kaufmann I, Martin G, Friedlein A, Langen H, Keller W. 2004. Human Fip1 is a subunit of CPSF that binds to U-rich RNA elements and stimulates poly(A) polymerase. *EMBO J.* 23:616–26

36. Meinke G, Ezeokonkwo C, Balbo P, Stafford W, Moore C, Bohm A. 2008. Structure of yeast poly(A) polymerase in complex with a peptide from Fip1, an intrinsically-disordered protein. *Biochemistry*. 148:825–32

37. Ezeokonkwo C, Zhelkovsky A, Lee R, Bohm A, Moore CL. 2011. A flexible linker region in Fip1 is needed for efficient mRNA polyadenylation. *Rna*. 17:652–64

38. Mandel CR, Kaneko S, Zhang H, Gebauer D, Vethantham V, et al. 2006. Polyadenylation factor CPSF-73 is the pre-mRNA 3′-end-processing endonuclease. *Nature*. 444:953–56

39. Kolev NG, Yario TA, Benson E, Steitz JA. 2008. Conserved motifs in both CPSF73 and CPSF100 are required to assemble the active endonuclease for histone mRNA 3′-end maturation. *EMBO Rep.* 9:1013–18

40. Humphreys I, Pei J, Baek M, Krishnakumar A, Anishchenko I, et al. 2021. Computed structures of core eukaryotic protein complexes. *Science.* 374:eabm4805

41. Saijo M, Sakai Y, Kishino T, Niikawa N, Matsuura Y, et al. 1995. Molecular cloning of a human protein that binds to the retinoblastoma proteins and chromosomal mapping. *Genomics*. 27:511–19

42. Sakai Y, Saijo M, Coelho K, Kishino T, Nikawa N, Taya Y. 1995. cDNA sequence and chromosomal localization of a novel human protein, RBQ-1 (RBBP6), that binds to the retinoblastoma gene product. *Genomics*. 30:98–101

43. Simons A, Melamed-Bessudo C, Wolkowicz R, Sperling J, Sperling R, et al. 1997. PACT: cloning and characterization of a cellular p53 binding protein that interacts with Rb. *Oncogene*. 14:145–55

44. Di Giammartino DC, Li W, Ogami K, Yashinskie JJ, Hoque M, et al. 2014. RBBP6 isoforms regulate the human polyadenylation machinery and modulate expression of mRNAs with AU-rich 3' UTRs. *Genes Dev.* 28:2248-60

45. Cossa G, Parua PK, Eilers M, Fisher RP. 2021. Protein phosphatases in the RNAPII transcription cycle: erasers, sculptors, gatekeepers, and potential drug targets. *Genes Dev.* 35:658-76

46. Suh H, Ficarro SB, Kang UB, Chun Y, Marto JA, Buratowski S. 2016. Direct analysis of phosphorylation sites on the Rpb1 C-terminal domain of RNA polymerase II. *Mol. Cell*. 61:297–304

47. Schreieck A, Easter AD, Etzold S, Wiederhold K, Lidschreiber M, et al. 2014. RNA polymerase II termination involves C-terminal-domain tyrosine dephosphorylation by CPF subunit Glc7. *Nat. Struct. Mol. Biol.* 21:175–79

48. Larame L, Forest A, Bataille AR, Bergeron M, Hanes SD. 2012. A universal RNA Polymerase II CTD cycle is orchestrated by complex interplays between kinase, phosphatase, and isomerase enzymes along genes. *Mol. Cell*. 45:158–70

49. Parua PK, Booth GT, Sansó M, Benjamin B, Tanny JC, et al. 2018. A Cdk9-PP1 switch regulates the elongation-termination transition of RNA polymerase II. *Nature*. 558:460–64

50. Krishnamurthy S, He X, Reyes-Reyes M, Moore C, Hampsey M. 2004. Ssu72 is an RNA polymerase II CTD phosphatase. *Mol. Cell*. 14:387–94

51. Glover-Cutter K, Kim S, Espinosa J, Bentley DL. 2008. RNA polymerase II pauses and associates with pre-mRNA processing factors at both ends of genes. 15:71–78

52. Xiang K, Nagaike T, Xiang S, Kilic T, Beh MM, et al. 2010. Crystal structure of the human symplekin–Ssu72–CTD phosphopeptide complex. *Nature*. 467:729–33

53. Cortazar MA, Sheridan RM, Erickson B, Fong N, Glover-cutter K, et al. 2019. Control of RNA Pol II speed by PNUTS-PP1 and Spt5 dephosphorylation facilitates termination by a "sitting duck torpedo" mechanism. *Mol. Cell*. 76:896-908

54. Lee JH, You J, Dobrota E, Skalnik DG. 2010. Identification and characterization of a novel human PP1 phosphatase complex. *J. Biol. Chem.* 285:24466–76

55. Lidschreiber M, Easter AD, Battaglia S, Rodríguez-Molina JB, Casã Nal A, et al. 2018. The APT complex is involved in non-coding RNA transcription and is distinct from CPF. *Nucleic Acids Res.* 46:11528-38

56. Porrua O, Libri D. 2015. Transcription termination and the control of the transcriptome: why, where and how to stop. *Nat. Rev. Mol. Cell Biol.* 16:190-202

57. Baillat D, Hakimi M, Näär AM, Shilatifard A, Cooch N, et al. 2005. Integrator, a multiprotein nediator of small nuclear RNA processing, associates with the C-terminal repeat of RNA polymerase II. *Cell*. 123:265–76

58. Elrod ND, Henriques T, Huang KL, Tatomer DC, Wilusz JE, et al. 2019. The integrator complex attenuates promoter-proximal transcription at protein-coding genes. *Mol. Cell*. 76:738-52

59. Zheng H, Qi Y, Hu S, Cao X, Xu C, et al. 2020. Identification of Integrator-PP2A complex (INTAC), an RNA polymerase II phosphatase. *Science.* 370:eabb5872

60. Huang K, Jee D, Stein CB, Elrod ND, Henriques T, et al. 2020. Integrator recruits protein phosphatase 2A to prevent pause release and facilitate transcription termination. *Mol. Cell*. 80:345-58

61. Vervoort SJ, Welsh SA, Devlin JR, Barbieri E, Knight DA, et al. 2021. The PP2A-Integrator-CDK9 axis fine-tunes transcription and can be targeted therapeutically in cancer. *Cell*. 184:3143-62

62. Pfleiderer MM, Galej WP. 2021. Structure of the catalytic core of the Integrator complex. *Mol. Cell*. 81:1246-59

63. Fianu I, Chen Y, Dienemann C, Dybkov O, Linden A, et al. 2021. Structural basis of Integrator-mediated transcription regulation. *Science.* 374:883–87

64. Gordon JMB, Shikov S, Kuehner JN, Liriano M, Lee E, et al. 2011. Reconstitution of CF IA from overexpressed subunits reveals stoichiometry and provides insights into molecular topology. *Biochemistry*. 50:10203–14

65. Gross S, Moore C. 2001. Five subunits are required for reconstitution of the cleavage and polyadenylation activities of *Saccharomyces cerevisiae* cleavage factor I. *Proc. Natl. Acad. Sci.* 98:6080–85

66. Pérez Cañadillas JM, Varani G. 2003. Recognition of GU-rich polyadenylation regulatory elements by human CStF64 protein. *EMBO J.* 22:2821–30

67. Ghazy MA, Gordon JMB, Lee SD, Singh BN, Bohm A, et al. 2012. The interaction of Pcf11 and Clp1 is needed for mRNA 3′-end formation and is modulated by amino acids in the ATP-binding site. *Nucleic Acids Res.* 40:1214–25

68. Zhang Z, Fu J, Gilmour DS. 2005. CTD-dependent dismantling of the RNA polymerase II elongation complex by the pre-mRNA 3′-end processing factor, Pcf11. *Genes Dev.* 19:1572–80

69. Ramirez A, Shuman S, Schwer B. 2008. Human RNA 5′-kinase (hClp1) can function as a tRNA splicing enzyme in vivo. *Rna*. 14:1737–45

70. Zhu Y, Wang X, Forouzmand E, Jeong J, Qiao F, et al. 2018. Molecular mechanisms for CFIm-mediated regulation of mRNA alternative polyadenylation. *Mol. Cell*. 69:62–74

71. Turtola M, Manav CM, Kumar A, Tudek A, Mroczek S, et al. 2021. Three-layered control of mRNA poly(A) tail synthesis in *Saccharomyces cerevisiae*. *Genes Dev.* 35:1290–1303

72. Wahle E. 1995. PolyA tail length control is caused by termination of processive synthesis. *J. Biol. Chem.* 270:2800–08

73. Stewart M. 2019. Polyadenylation and nuclear export of mRNAs. *J. Biol. Chem.* 294:2977–87

74. Bresson SM, Hunter O V., Hunter AC, Conrad NK. 2015. Canonical poly(A) polymerase activity promotes the decay of a wide variety of mammalian nuclear RNAs. *PLoS Genet.* 11:1–25

75. Bresson SM, Conrad NK. 2013. The human nuclear poly(A)-binding protein promotes RNA hyperadenylation and decay. *PLoS Genet.* 9:e1003893

76. Kühn U, Gündel M, Knoth A, Kerwitz Y, Rüdel S, Wahle E. 2009. Poly(A) tail length is controlled by the nuclear poly(A)-binding protein regulating the interaction between poly(A) polymerase and the cleavage and polyadenylation specificity factor. *J. Biol. Chem.* 284:22803–14

77. Whitfield ML, Zheng L, Baldwin AMY, Ohta T, Hurt MM, Marzluff WF. 2000. Stem-loop binding protein, the protein that binds the 3′ end of histone mRNA, is cell cycle regulated by both translational and posttranslational mechanisms. *Mol. Cell. Biol.* 20:4188–98

78. Sullivan KD, Mullen TE, Marzluff WF, Wagner EJ. 2009. Knockdown of SLBP results in nuclear retention of histone mRNA. *Rna*. 73:459–72

79. Sanchez R, Marzluff WF. 2002. The stem-loop binding protein is required for efficient translation of histone mRNA in vivo and in vitro. *Mol. Cell. Biol.* 22:7093–104

80. Yang X-C, Sabath I, Debski J, Kaus-Drobek M, Dadlez M, et al. 2013. A complex containing the CPSF73 endonuclease and other polyadenylation factors associates with U7 snRNP and is recruited to histone pre-mRNA for 3′-end processing. *Mol. Cell. Biol.* 33:28–37

81. Skrajna A, Yang X-C, Dadlez M, Marzluff WF, Dominski Z. 2018. Protein composition of catalytically active U7-dependent processing complexes assembled on histone pre-mRNA containing biotin and a photo-cleavable linker. *Nucleic Acids Res.* 46:4752–70

82. Sun Y, Zhang Y, Aik WS, Yang XC, Marzluff WF, et al. 2020. Structure of an active human histone pre-mRNA 3′-end processing machinery. *Science.* 367:700–3

83. Ozsolak F, Kapranov P, Foissac S, Kim SW, Fishilevich E, et al. 2010. Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. *Cell*. 143:1018–29

84. Tang P, Yang Y, Li G, Huang L, Wen M, et al. 2022. Alternative polyadenylation by sequential activation of distal and proximal PolyA sites. *Nat. Struct. Mol. Biol.* 29:21–31

85. Ransom B, Goldman SA, Meldolesi J, Zhou L, Murai KK, et al. 2008. Proliferating cells express mRNAs with shortened 3′ untranslated regions and fewer microRNA target sites. *Science.* 20:1643–48

86. Ji Z, Tian B. 2009. Reprogramming of 3′ untranslated regions of mRNAs by alternative polyadenylation in generation of pluripotent stem cells from different cell types. *PLoS One*. 4:e8419

87. Takagaki Y, Manley JL. 1998. Levels of polyadenylation factor CstF-64 control IgM heavy chain mRNA accumulation and other events associated with B cell differentiation. *Mol. Cell*. 2:761–71

88. Audibert A, Simonelig M. 1998. Autoregulation at the level of mRNA 3′ end formation of the suppressor of forked gene of Drosophila melanogaster is conserved in Drosophila virilis. *Proc. Natl. Acad. Sci.* 95:14302–07

89. Kamieniarz-Gdula K, Gdula MR, Panser K, Nojima T, Monks J, et al. 2019. Selective roles of vertebrate PCF11 in premature and full-length transcript termination. *Mol. Cell*. 74:158-72

90. Curinha A, Braz SO, Pereira-Castro I, Cruz A, Moreira A. 2014. Implications of polyadenylation in health and disease. *Nucleus*. 5:508–19

91. Liu H, Moore CL. 2021. On the cutting edge: regulation and therapeutic potential of the mRNA 3′ end nuclease. *Trends Biochem. Sci.* 46:772-84

92. Kakegawa J, Sakane N, Suzuki K, Yoshida T. 2019. JTE-607, a multiple cytokine production inhibitor, targets CPSF3 and inhibits pre-mRNA processing. *Biochem. Biophys. Res. Commun.* 518:32–7

93. Ross NT, Lohmann F, Carbonneau S, Fazal A, Weihofen WA, et al. 2020. CPSF3-dependent pre-mRNA processing as a druggable node in AML and Ewing's sarcoma. *Nat. Chem. Biol.* 16:50–9

94. Palencia A, Bougdour A, Brenier-Pinchart M, Touquet B, Bertini R, et al. 2017. Targeting Toxoplasma gondii CPSF3 as a new approach to control toxoplasmosis . *EMBO Mol. Med.* 9:385–94

95. Swale C, Bougdour A, Gnahoui-David A, Tottey J, Georgeault S, et al. 2019. Metal-captured inhibition of pre-mRNA processing activity by CPSF3 controls Cryptosporidium infection. *Sci. Transl. Med.* 11:eaax7161

96. Sonoiki E, Ng CL, Lee MCS, Guo D, Zhang YK, et al. 2017. A potent antimalarial benzoxaborole targets a Plasmodium falciparum cleavage and polyadenylation specificity factor homologue. *Nat. Commun.* 8:14574

97. Alahmari AA, Chaubey AH, Tisdale AA, Schwarz CD, Cornwell AC, et al. 2022. CPSF3 inhibition blocks pancreatic cancer cell proliferation through disruption of core histone processing. *bioRxiv:2022.05.09.491230*

98. Ning YUE, Liu W, Guan X, Xie X, Zhang Y. 2019. CPSF3 is a promising prognostic biomarker and predicts recurrence of non - small cell lung cancer. *Oncol. Lett.* 18:2835–44

99. Grosso AR, Leite AP, Matos MR, Martins FB, Desterro JMP, Carmo-Fonseca M. 2015. Pervasive transcription read-through promotes aberrant expression of oncogenes and RNA chimeras in renal carcinoma. *eLife.* 4:e09214

100. Gutierrez PA, Baughman K, Sun Y, Tong L. 2021. A real-time fluorescence assay for CPSF73, the nuclease for pre-mRNA 3′-end processing. *Rna.* 27:1148-54

101. Wang X, Hennig T, Whisnant AW, Erhard F, Prusty BK, et al. 2020. Herpes simplex virus blocks host transcription termination via the bimodal activities of ICP27. *Nat. Commun.* 11:293

102. Zhao N, Sebastiano V, Moshkina N, Mena N, Hultquist J, et al. 2018. Influenza virus infection

causes global RNAPII termination defects. *Nat. Struct. Mol. Biol.* 25:885–93

103. Nemeroff ME, Barabino SML, Li Y, Keller W, Krug RM. 1998. Influenza virus NS1 protein interacts with the cellular 30 kDa subunit of CPSF and inhibits 3′ end formation of cellular pre-mRNAs. *Mol. Cell*. 1:991–1000

104. Das K, Ma L-C, Xiao R, Radvansky B, Aramini J, et al. 2008. Structural basis for suppression of a host antiviral response by influenza A virus. *Proc. Natl. Acad. Sci.* 105:13093–98

105. Vilborg A, Sabath N, Wiesel Y, Nathans J, Levy-Adam F, et al. 2017. Comparative analysis reveals genomic features of stress-induced transcriptional readthrough. *Proc. Natl. Acad. Sci. U. S. A.* 114:8362–71

106. Eaton JD, West S. 2020. Termination of transcription by RNA Polymerase II: BOOM! *Trends Genet.* 36:664–75

107. Logan J, Falck-Pedersen E, Darnell JE, Shenk T. 1987. A poly(A) addition site and a downstream termination region are required for efficient cessation of transcription by RNA polymerase II in the mouse β^{maj}-globin gene. *Proc. Natl. Acad. Sci. U. S. A.* 84:8306–10

108. West S, Gromak N, Proudfoot N. 2004. Human 5′ → 3′ exonuclease Xrn2 promotes transcription termination at co-transcriptional cleavage sites. *Nature*. 432:522–525

109. Eaton JD, Davidson L, Bauer DLV, Natsume T, Kanemaki MT, West S. 2018. Xrn2 accelerates termination by RNA polymerase II, which is underpinned by CPSF73 activity. *Genes Dev.* 32:127–39

110. Carminati M, Manav MC, Bellini D, Passmore LA. 2022. A direct interaction between CPF and Pol II links RNA 3'-end processing to transcription. *bioRxiv:2022.07.28.501803*

111. Zhang H, Rigo F, Martinson HG. 2015. Poly(A) signal-dependent transcription termination occurs through a conformational change mechanism that does not require cleavage at the article poly (A) site. *Mol. Cell*. 59:437–48

112. West S, Proudfoot NJ, Dye MJ. 2008. Article molecular dissection of mammalian RNA polymerase II transcriptional termination. *Mol Cell*. 29:600–10

113. McCracken S, Fong N, Yankulov K, Ballantyre S, Pan G, et al. 1997. The C-terminal domain of RNA polymerase II couples mRNA processing to transcription. *Nature*. 385:357–61

114. Reimer KA, Mimoso CA, Adelman K, Neugebauer KM, Reimer KA, et al. 2021. Co-transcriptional splicing regulates 3′-end cleavage during mammalian erythropoiesis. *Mol. Cell*. 81:998-1012

115. Drexler HL, Choquet K, Churchman LS, Drexler HL, Choquet K, Churchman LS. 2020. Splicing kinetics and coordination revealed by direct nascent RNA sequencing through Nanopores. *Mol. Cell*. 77:985–98

116. Herzel L, Straube K, Neugebauer KM. 2018. Long-read sequencing of nascent RNA reveals coupling among RNA processing events. *Genome Res.* 28:1008–19

117. Cai Z, Oh J, So BR, Di C, Cai Z, et al. 2019. A complex of U1 snRNP with cleavage and polyadenylation factors controls telescripting, regulating mRNA transcription in human cells. *Mol. Cell*. 76:590-9

118. Zhang S, Aibara S, Vos SM, Agafonov DE. 2021. Structure of a transcribing RNA polymerase II–U1 snRNP complex. *Science.* 371:305-09

119. Zheng H, Jin Q, Qi Y, Liu W, Ren Y, et al. 2021. Structural basis of INTAC-regulated transcription. *bioRxiv:2021.11.29.470345*

120. Weissmann F, Petzold G, VanderLinden R, Huis in 't Veld PJ, Brown NG, et al. 2016. biGBac enables rapid gene assembly for the expression of large multisubunit protein complexes. *Proc. Natl. Acad. Sci.* 113:2564–69

121. Suskiewicz MJ, Sussman JL, Silman I, Shaul Y. 2011. Context-dependent resistance to proteolysis of intrinsically disordered proteins. *Protein Sci.* 20:1285–97

122. Kubota S, Kubota H, Nagata K. 2006. Cytosolic chaperonin protects folding intermediates of Gβ from aggregation by recognizing hydrophobic β-strands. *Proc. Natl. Acad. Sci.* 103:8360–65

123. Batra J, Hultquist JF, Liu D, Shtanko O, Von Dollen J, et al. 2018. Protein interaction mapping identifies RBBP6 as a negative regulator of Ebola virus replication. *Cell*. 175:1917–30

124. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature*. 596:583–89

125. Wahle E. 1991. Purification and characterization of a mammalian polyadenylate polymerase involved in the 3' end processing of messenger RNA precursors. *J. Biol. Chem.* 266:3131–39

126. Martin G, Keller W. 1996. Mutational analysis of mammalian poly(A) polymerase identifies a region for primer binding and a catalytic domain, homologous to the family X polymerases, and to other nucleotidyltransferases. *EMBO J.* 15:2593–603

127. Babu MM. 2016. The contribution of intrinsically disordered regions to protein function, cellular complexity, and human disease. *Biochem. Soc. Trans.* 44:1185–200

128. Ryan K, Bauer DL V. 2008. Post-translational modification of protein factors involved in mammalian pre-mRNA 3′ end formation. *Int. J. Biochem. Cell Biol.* 40:2384–96

129. Ryan K. 2007. Pre-mRNA 3′ cleavage is reversibly inhibited in vitro by cleavage factor dephosphorylation. *RNA Biol.* 4:26–33

130. Zarkower D, Wickens M. 1987. Specific pre-cleavage and post-cleavage complexes involved in the formation of SV40 late mRNA 3' termini in vitro. *EMBO J.* 6:4185–92

131. Sheets MD, Stephenson P, Wickens MP. 1987. Products of in vitro cleavage and polyadenylation of simian virus 40 late pre-mRNAs. *Mol. Cell. Biol.* 7:1518–29

132. Moore CL, Sharp PA. 1985. Accurate cleavage and polyadenylation of exogenous RNA substrate. *Cell*. 41:845–55

133. Moore CL, Skolnik-David H, Sharp PA. 1986. Analysis of RNA cleavage at the adenovirus-2 L3 polyadenylation site. *EMBO J.* 5:1929–38

134. Takagaki Y, Manley JL, MacDonald CC, Wilusz J, Shenk T. 1990. A multisubunit factor, CstF, is required for polyadenylation of mammalian pre-mRNAs. *Genes Dev.* 4:2112–20

135. De Vries H, Rüegsegger U, Hübner W, Friedlein A, Langen H, Keller W. 2000. Human pre-mRNA cleavage factor II(m) contains homologs of yeast proteins and bridges two other cleavage factors. *EMBO J.* 19:5895–904

136. Bentley DL. 2005. Rules of engagement: Co-transcriptional recruitment of pre-mRNA processing factors. *Curr. Opin. Cell Biol.* 17:251–56

137. Piovesan D, Profiti G, Martelli PL, Casadio R. 2012. The human "magnesome": Detecting magnesium binding sites on human proteins. *BMC Bioinformatics*. 13(Suppl 14), S10

138. Misra VK, Draper DE. 1998. On the role of magnesium ions in RNA stability. *Biopolymers*. 48:113–35

139. Adamson TE, Shutt DC, Price DH. 2005. Functional coupling of cleavage and polyadenylation with transcription of mRNA. *J. Biol. Chem.* 280:32262–71

140. Hirose Y, Manley JL. 1997. Creatine phosphate, not ATP, is required for 3′ end cleavage of mammalian pre-mRNA *in vitro*. *J. Biol. Chem.* 272:29636–42

141. Soranno A, Koenig I, Borgia MB, Hofmann H, Zosel F, et al. 2014. Single-molecule spectroscopy reveals polymer effects of disordered proteins in crowded environments. *Proc. Natl. Acad. Sci. U. S. A.* 111:4874–79

142. Sachchithananthan M, Stasinopoulos SJ, Wilusz J, Medcalf RL. 2005. The relationship between the prothrombin upstream sequence element and the G20210A polymorphism: The influence of a competitive environment for mRNA 3′-end formation. *Nucleic Acids Res.* 33:1010–20

143. Levitt N, Briggs D, Gil A, Proudfoot NJ. 1989. Definition of an efficient synthetic poly(A) site. *Genes Dev.* 3:1019–25

144. Ryan K, Calvo O, Manley JL. 2004. Evidence that polyadenylation factor CPSF-73 is the mRNA 3′ processing endonuclease. *Rna*. 10:565–73

145. Twu KY, Noah DL, Rao P, Kuo R-L, Krug RM. 2006. The CPSF30 binding site on the NS1A protein of Influenza A virus is a potential antiviral target. *J. Virol.* 80:3957–65

146. Bornholdt ZA, Prasad BVV. 2008. X-ray structure of NS1 from a highly pathogenic H5N1 influenza virus. *Nature*. 456:958–989

147. Lin D, Lan J, Zhang Z. 2007. Structure and function of the NS1 protein of influenza A virus. *Acta Biochim. Biophys. Sin. (Shanghai).* 39:135–62

148. Koliopoulos MG, Lethier M, Van Der Veen AG, Haubrich K, Hennig J, et al. 2018. Molecular mechanism of Influenza A NS1-mediated TRIM25 recognition and inhibition. *Nat. Commun.* 9:1820

149. Yu R, Jing X, Li W, Xu J, Xu Y, et al. 2018. Non-structural protein 1 from avian Influenza virus H9N2 is an efficient RNA silencing suppressor with characteristics that differ from those of Tomato bushy stunt virus p19. *Virus Genes*. 54:368–75

150. Schmidt HB rode., Görlich D. 2015. Nup98 FG domains from diverse species spontaneously phase-separate into particles with nuclear pore-like permselectivity. *Elife*. 4:1–30

151. Boeynaems S, Bogaert E, Kovacs D, Konijnenberg A, Timmerman E, et al. 2017. Phase separation of C9orf72 dipeptide repeats perturbs stress granule dynamics. *Mol. Cell*. 65:1044-1055

152. Wiemann S, Weil B, Wellenreuther R, Gassenhuber J, Glassl S, et al. 2001. Toward a catalog of human genes and proteins: Sequencing and analysis of 500 novel complete protein coding human cDNAs. *Genome Res.* 11:422–35

153. Kubo T, Wada T, Yamaguchi Y, Shimizu A, Handa H. 2006. Knock-down of 25 kDa subunit of cleavage factor Im in HeLa cells alters alternative polyadenylation within 3′-UTRs. *Nucleic Acids Res.* 34:6264–71

154. Schwich OD, Blümel N, Keller M, Wegener M, Setty ST, et al. 2021. SRSF3 and SRSF7 modulate

3′UTR length through suppression or activation of proximal polyadenylation sites and regulation of CFIm levels. *Genome Biol.* 22:82

155. Christofori G, Keller W. 1988. 3′ cleavage and polyadenylation of mRNA precursors *in vitro* requires a poly(A) polymerase, a cleavage factor, and a snRNP. *Cell*. 54:875–89

156. Sullivan KD, Steiniger M, Marzluff WF. 2009. A core complex of CPSF73, CPSF100, and symplekin may form two different cleavage factors for processing of poly(A) and histone mRNAs. *Mol. Cell*. 34:322–32

157. Wilkinson ME, Charenton C, Nagai K. 2020. RNA splicing by the spliceosome. *Annu. Rev. Biochem.* 89:359–88

158. David L, Huber W, Granovskaia M, Toedling J, Palm CJ, et al. 2006. A high-resolution map of transcription in the yeast genome. *Proc. Natl. Acad. Sci. U. S. A.* 103:5320–25

159. Schmidt M, Kluge F, Sandmeir F, Schäfer P, Tüting C, et al. 2022. Reconstitution of 3' end processing of mammalian pre-mRNA reveals a central role of RBBP6. *Genes Dev.* 36:195–209

160. Hockert JA, Yeh HJ, MacDonald CC. 2010. The hinge domain of the cleavage stimulation factor protein CstF-64 is essential for CstF-77 interaction, nuclear localization, and polyadenylation. *J. Biol. Chem.* 285:695–704

161. Kwon B, Fansler MM, Patel ND, Lee J, Ma W, Mayr C. 2022. Enhancers regulate 3′ end processing activity to control expression of alternative 3′UTR isoforms. *Nat. Commun.* 13:2709

162. Vo LTA, Minet M, Schmitter J-M, Lacroute F, Wyers F. 2001. Mpe1, a zinc knuckle protein, is an essential component of yeast cleavage and polyadenylation factor required for the cleavage and polyadenylation of mRNA. *Mol. Cell. Biol.* 21:8346–56

163. Schwanhüusser B, Busse D, Li N, Dittmar G, Schuchhardt J, et al. 2011. Global quantification of mammalian gene expression control. *Nature*. 473:337–42

164. Wang X, Chen CF, Baker PR, Chen PL, Kaiser P, Huang L. 2007. Mass spectrometric characterization of the affinity-purified human 26S proteasome complex. *Biochemistry*. 46:3553–65

165. Graveley BR, Maniatis T. 1998. Arginine/serine-rich domains of SR proteins can function as activators of pre-mRNA splicing splicing of all metazoan introns and therefore play a central role in the basic splicing reaction. *Mol. Cell*. 1:765–71

166. Bracken CP, Wall SJ, Barre B, Panov KI, Ajuh PM, Perkins ND. 2008. Regulation of cyclin D1 RNA stability by SNIP1. *Cancer Res.* 68:7621–7628

167. Ruepp M-D, Schweingruber C, Kleinschmidt N, Schumperli D. 2011. Interactions of CstF-64, CstF-77, and symplekin: Implications on localisation and function. *Mol. Biol. Cell*. 22:91–104

168. Evans R, O'Neill M, Pritzel A, Antropova N, Senior A, et al. 2022. Protein complex prediction with AlphaFold-Multimer. *bioRxiv:2021.10.04.463034*

169. Bai Y, Auperin TC, Chou CY, Chang GG, Manley JL, Tong L. 2007. Crystal structure of murine CstF-77: Dimeric association and implications for polyadenylation of mRNA precursors. *Mol. Cell*. 25:863–75

170. Dupin AF, Fribourg S. 2014. Structural basis for ATP loss by Clp1p in a G135R mutant protein. *Biochimie*. 101:203–7

171. O'Reilly FJ, Graziadei A, Forbrig C, Bremenkamp R, Charles K, et al. 2022. Protein complexes in Bacillus subtilis by AI-assisted structural proteomics. *bioRxiv:2022.07.26.501605*

172. Lee SD, Moore CL. 2014. Efficient mRNA polyadenylation requires a ubiquitin-like domain, a zinc knuckle, and a RING finger domain, all contained in the Mpe1 protein. *Mol. Cell. Biol.* 34:3955–67

173. Deshaies RJ, Joazeiro CAP. 2009. RING domain E3 ubiquitin ligases. *Annu. Rev. Biochem.* 78:399–434

174. Wani S, Yuda M, Fujiwara Y, Yamamoto M, Harada F, et al. 2014. Vertebrate Ssu72 regulates and coordinates 3′-end formation of RNAs transcribed by RNA polymerase II. *PLoS One*. 9:e106040

175. Hsin J-P, Sheth A, Manley JL. 2011. RNAP II CTD phosphorylated on threonine-4 is required for histone mRNA 3′ end processing. *Science.* 334:683–686

176. Qu X, Lykke-Andersen S, Nasser T, Saguez C, Bertrand E, et al. 2009. Assembly of an export-competent mRNP is needed for efficient release of the 3′-end processing complex after polyadenylation. *Mol. Cell. Biol.* 29:5327–38

177. Russo CJ, Passmore LA. 2014. Ultrastable gold substrates for electron cryomicroscopy. *Science.* 346:1377–81

178. Zivanov J, Nakane T, Forsberg BO, Kimanius D, Hagen WJ, et al. 2018. New tools for automated high-resolution cryo-EM structure determination in RELION-3. *elife*. 7:e42166

179. Rohou A, Grigorieff N. 2015. CTFFIND4 : Fast and accurate defocus estimation from electron micrographs. *J. Struct. Biol.* 192:216–21

180. Zheng SQ, Palovcak E, Armache J-P, Verba KA, Cheng Y, et al. 2017. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods.* 14:331–332

181. Zhang K. 2016. Gctf : Real-time CTF determination and correction. *J. Struct. Biol.* 193:1–12

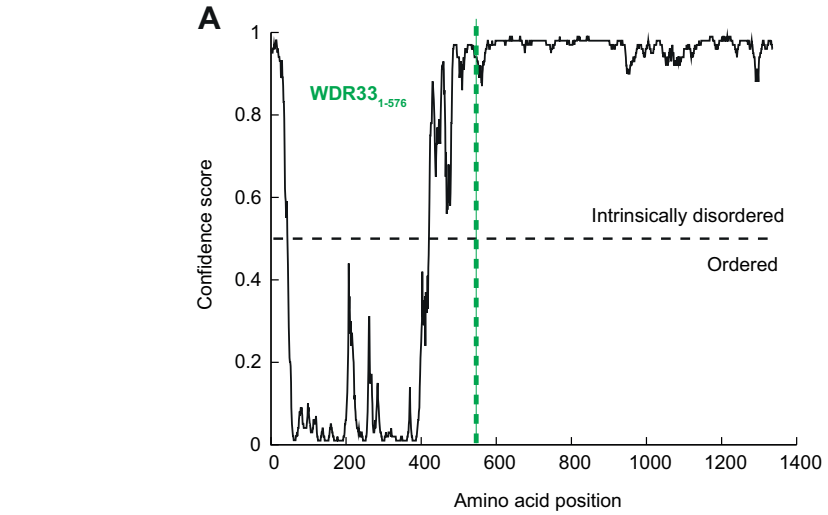
182. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. 2004. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.* 337:635–45

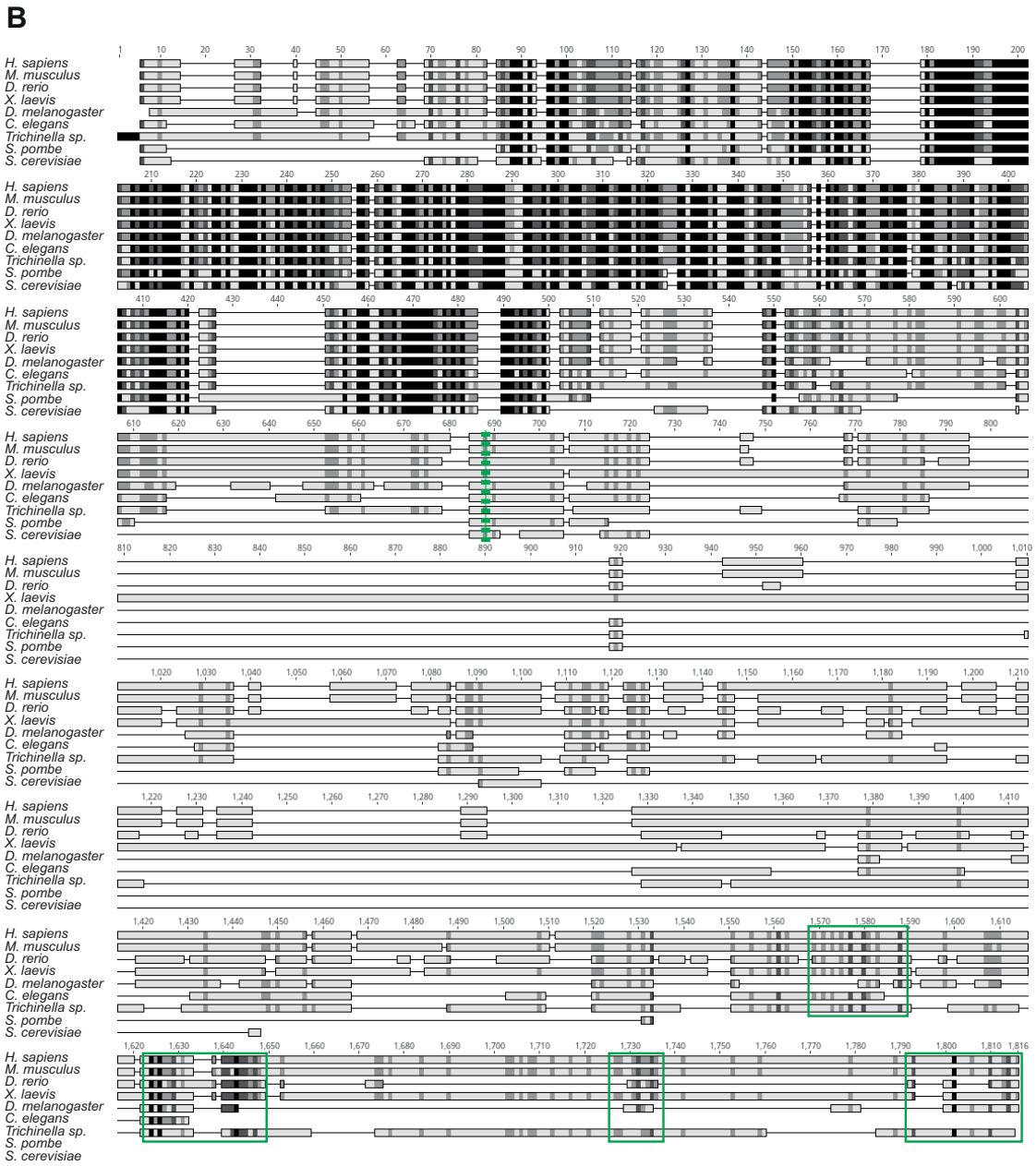
Chapter 8:

Appendices

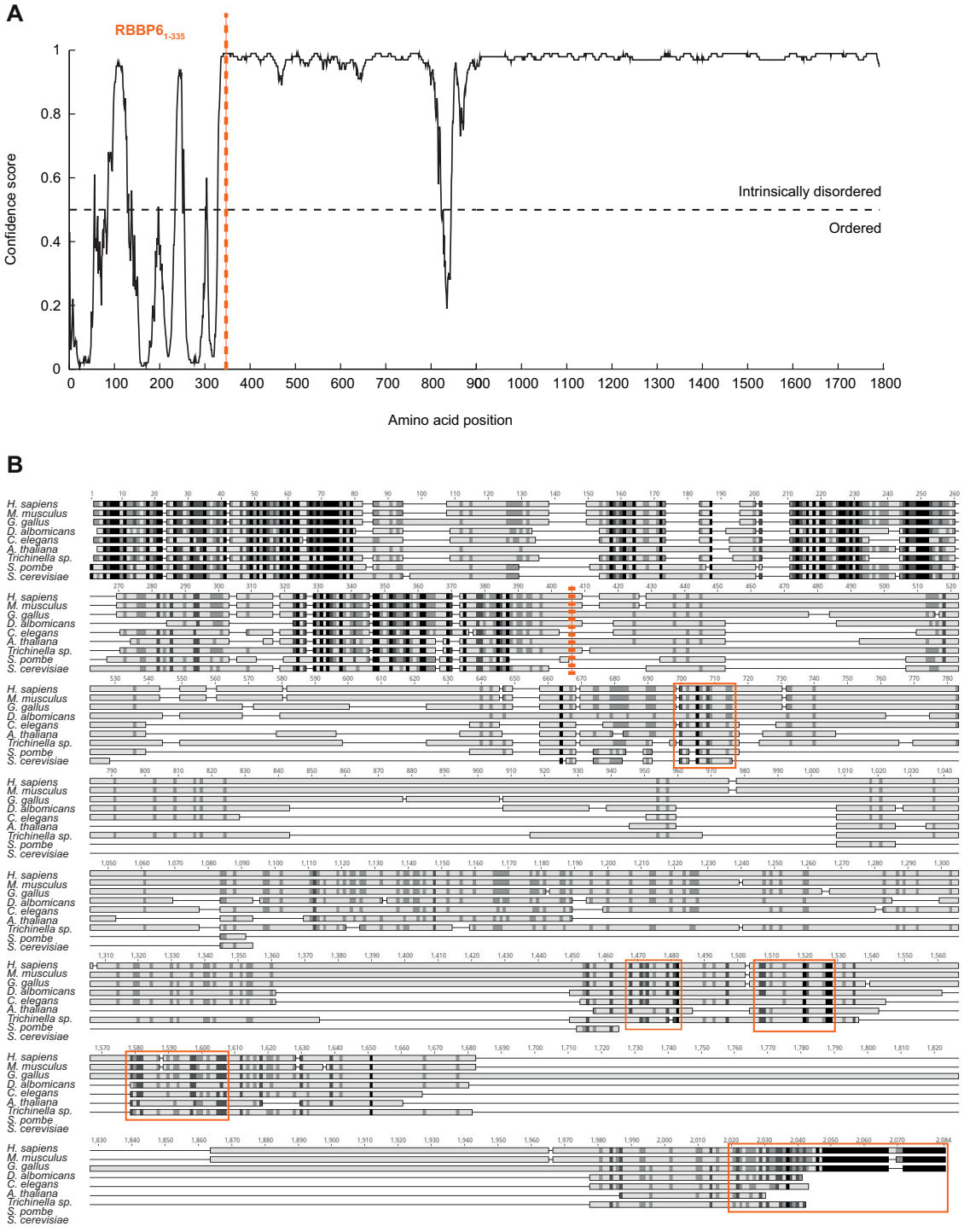
CPSF	mPSF	CPSF160 (1)	Scaffold	Cft1	Poly(A) polymerase module		CPF
		WDR33	Scaffold, RNA binding	Pfs2			
		CPSF30 (4)	RNA binding	Yth1			
		hFip1	PAP/Pap1 recruitment	Fip1			
		PAP	Poly(A) polymerase	Pap1			
		RBBP6	Endonuclease activation	Mpe1	Nuclease module		
	mCF	CPSF100 (2)	Pseudonuclease	Cft2			
		CPSF73 (3)	Endonuclease	Ysh1			
		symplekin	Scaffold	Pta1	Phosphatase module	APT	
		SSU72	Protein phosphatase	Ssu72			
	Phosphatase complex	WDR82	Transcription termination	Swd2			
		PP1	Protein phosphatase	Glc7			
		PNUTS	Scaffold	Ref2			
			Scaffold	Pti1			
		Tox4	DNA binding				
			Antagonising Ysh1	Syc1			
	CStF	CStF50 (1)	Complex stabilisation				
		CStF64 (2)	RNA binding	Rna15	CF IA		
		CStF77 (3)	Binding to CPSF160/Cft1	Rna14			
	CFIIm	Pcf11	Binding to RNA pol II CTD	Pcf11			
		Clp1	Polynucleotide kinase	Clp1			
			Cleavage fidelity	Hrp1	CF IB		
	CFIm	CFIm25 (CPSF5)	RNA binding				
		CFIm68 (CPSF6)	Binding to hFip1				
		CFIm59 (CPSF7)	Binding to hFip1				
		PABPN1	Binding to poly(A) tail				
			Binding to poly(A) tail	Pab1			
		ZC3H14	Binding to poly(A) tail	Nab2			

Appendix Table 8.1 Canonical pre-mRNA 3′ end processing factors in humans and budding yeast, their functions and the multi-subunit protein complexes they belong to. Alternative names are in brackets.

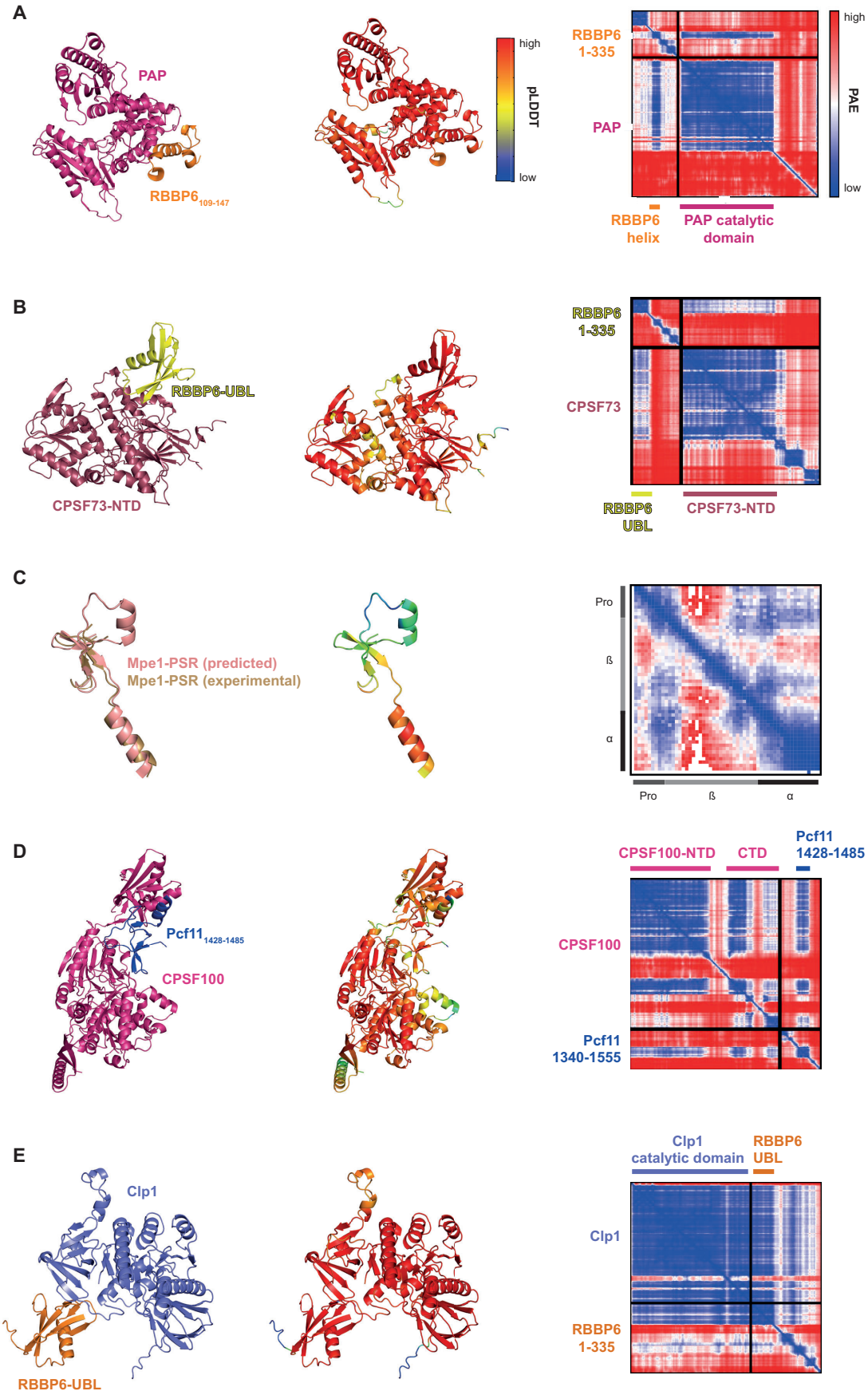

A
WDR33 1-576
Confidence score
Intrinsically disordered
Ordered
Amino acid position
0
0.2
0.4
0.6
0.8
1
0
200
400
600
800
1000
1200
1400


B
H. sapiens
M. musculus
D. rerio
X. laevis
D. melanogaster
C. elegans
Trichinella sp.
S. pombe
S. cerevisiae

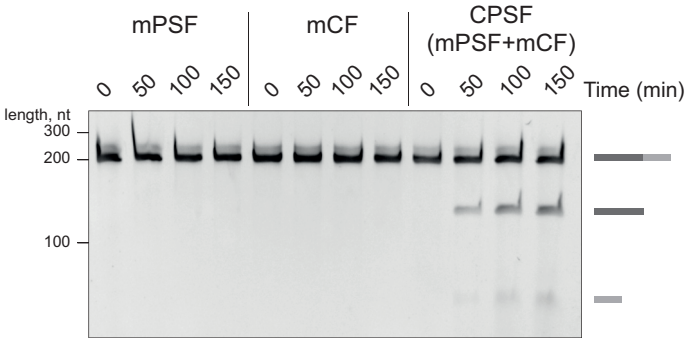
Appendix Figure 8.1 WDR33 contains a long C-terminal IDR. (A) DISOPRED (182) disorder prediction plot of WDR33. The boundaries of the construct containing residues 1-576 are marked with a green dotted line. **(B)** Multiple sequence alignment of WDR33 orthologues. The darker the box, the higher the degree of conservation of that particular residue. The boundaries of the construct containing residues 1-576 are marked with a green dotted line. Highly conserved regions of the C-terminal IDR are indicated by green boxes. The numbers above the alignment do not correspond to the residue number of any particular orthologue.



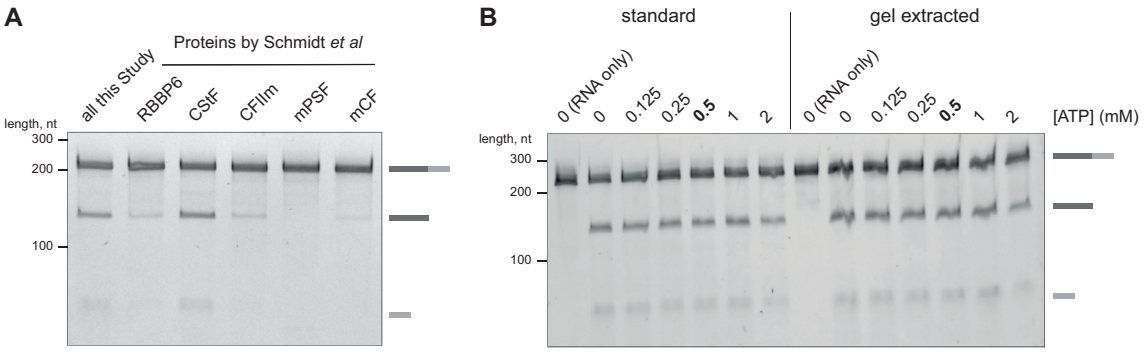
Appendix Figure 8.2 RBBP6 contains a long C-terminal IDR. (A) DISOPRED (182) disorder prediction plot of RBBP6. The boundaries of the construct containing residues 1-335 are marked with an orange dotted line. **(B)** Multiple sequence alignment of RBBP6 orthologues. The darker the box, the higher the degree of conservation of that particular residue. The boundaries of the construct containing residues 1-335 are marked with an orange dotted line. Highly conserved regions of the C-terminal IDR are indicated by orange boxes. The numbers above the alignment do not correspond to the residue number of any particular orthologue.


A
PAP
RBBP6109-147
high
pLDDT
low
RBBP6 1-335
PAP
high
PAE
low
RBBP6 helix
PAP catalytic domain
B
RBBP6-UBL
CPSF73-NTD
RBBP6 1-335
CPSF73
RBBP6 UBL
CPSF73-NTD
C
Mpe1-PSR (predicted)
Mpe1-PSR (experimental)
Pro
ß
α
Pro
ß
α
D
Pcf11 1428-1485
CPSF100
CPSF100-NTD
CTD
Pcf11 1428-1485
CPSF100
Pcf11 1340-1555
E
Clp1
RBBP6-UBL
Clp1 catalytic domain
RBBP6 UBL
Clp1
RBBP6 1-335

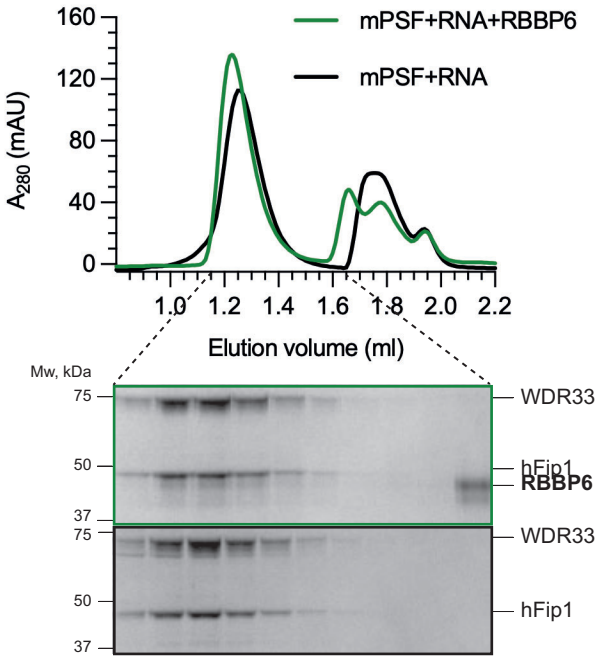
Appendix Figure 8.3 Statistics of the AlphaFold structure predictions discussed in this Thesis. Structure predictions (left), predicted structures coloured according to pLDDT values of each residue (middle) and the heatmaps showing pairwise predicted alignment error (PAE) values in Å (left) of the following: **(A)** complex of PAP and RBBP6 (Figure 2.11); **(B)** complex of RBBP6-UBL and CPSF73 (Figure 4.7); **(C)** Mpe1 PSR also overlayed onto its experimental structure (PDB 7ZGR) (Pro – loop containing the proline that may contact PAS RNA; β – β sheets of the PSR; α - C-terminal α helix of the PSR); **(D)** complex of CPSF100 and Pcf11; **(E)** complex of RBBP6-UBL and Clp1. Colour scales of pLDDT and PAE are shown in (A). Vertical axis – aligned residue number, horizontal axis – scored residue number.



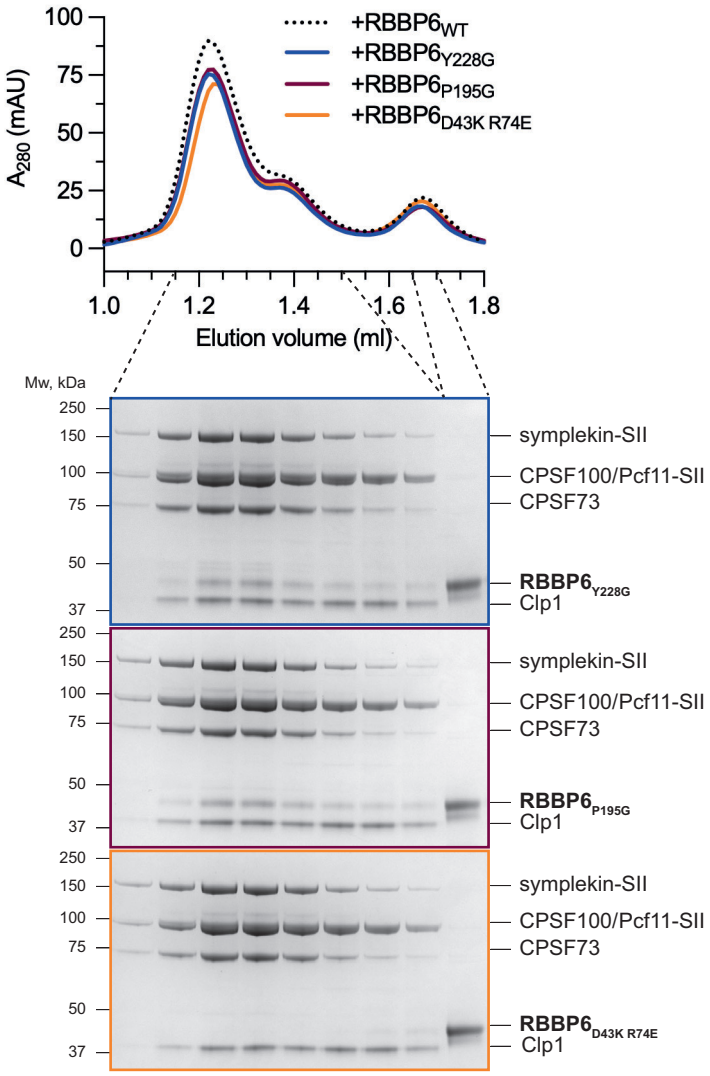
Appendix Figure 8.4 mCF does not catalyse endonucleolytic cleavage in the absence of mPSF. Cleavage assay using the SV40 pre-mRNA substrate in the presence of CStF, CFIIm, RBBP6 and either mPSF, mCF or full CPSF (mPSF and mCF combined).



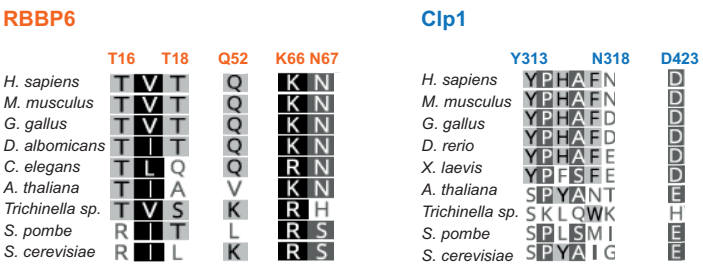
Appendix Figure 8.5 ATP contamination in CPSF cleavage assays does not explain the differences between this Study and Schmidt et al (159)**. (A)** CPSF cleavage assays in which protein factors described in this Study were exchanged individually for the proteins prepared by Schmidt *et al* (which were a kind gift from Elmar Wahle's group). Most of the protein factors purified by Schmidt *et al* were active under reaction conditions used in this Dissertation. **(B)** CPSF cleavage assays under conditions used in this Study of either the SV40 pre-mRNA substrate used throughout this Dissertation or the same RNA additionally purified by gel extraction to eliminate the possibility of ATP carry-over from the *in vitro* transcription reaction. The assays were performed at various concentrations of ATP. The concentration of ATP in the assays by Schmidt *et al* (0.5 mM) is highlighted. The RNA purified by gel extraction is cleaved efficiently, and endonuclease activity is not affected by ATP.



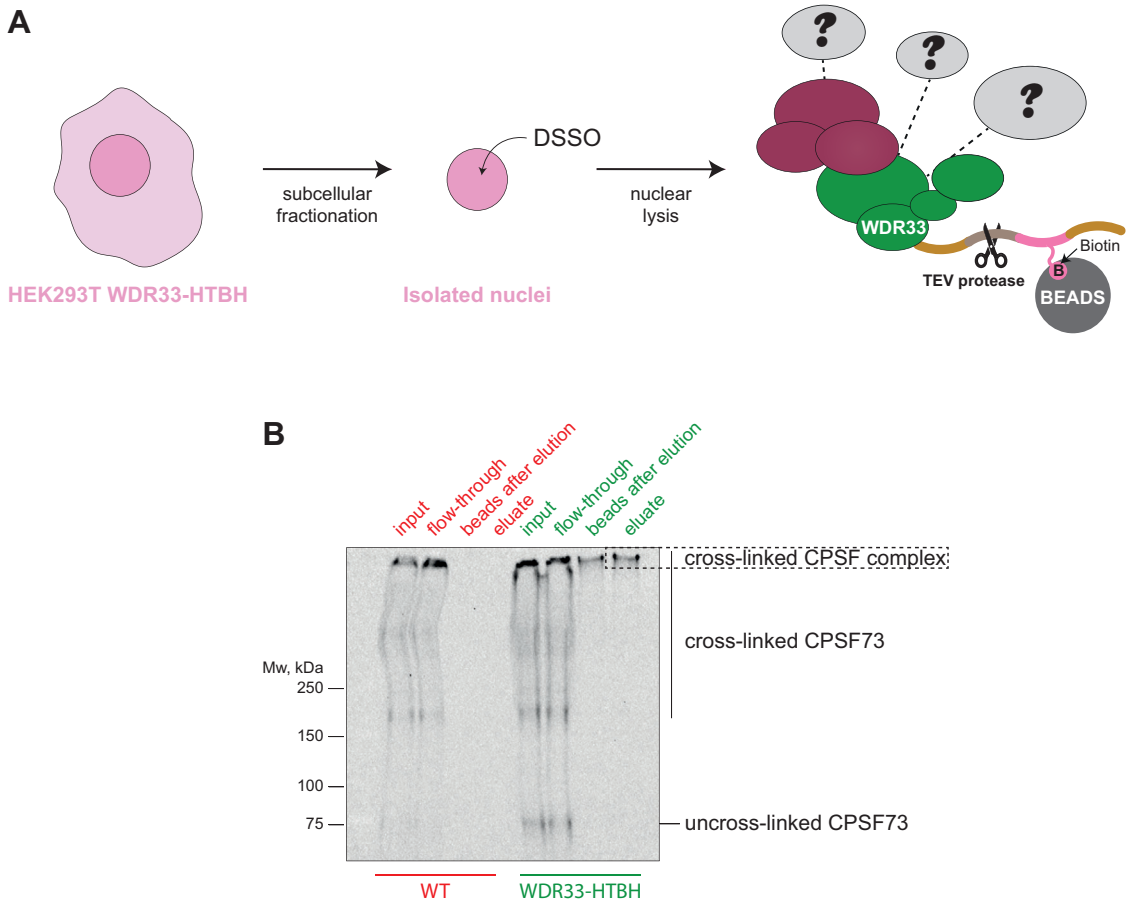
Appendix Figure 8.6 mPSF interacts with RBBP6 in an RNA-dependent manner. Size exclusion chromatogram (top) and SDS-PAGE analyses of the corresponding peak fractions (bottom) of the samples containing mPSF (3.4 μM), 41 nt L3 RNA (6.8 μM) either with our without RBBP6 (10.3 μM).



Appendix Figure 8.7 Disrupting the UBL-CPSF73 interface abolished RBBP6 binding to the mCF-CFIIm complex. Size exclusion chromatograms (top) and SDS-PAGE analyses of the corresponding peak fractions (bottom) of the samples containing mCF (2.5 μM), CFIIm (4 μM) and various mutants of RBBP6 (7.5 μM).



Appendix Figure 8.8 Parts of sequence alignments of RBBP6 and Clp1 orthologues. The darker the shading, the higher the degree of conservation. The residues implicated in the putative interaction between RBBP6 and Clp1 are marked.



Appendix Figure 8.9 CPSF can be cross-linked in situ. (A) Schematic representation of the *in situ* cross-linking experiment. DSSO (disuccinimidyl sulfoxide) is a membrane-permeant cross-linker. Question marks represent unknown interactors that might get cross-linked with the CPSF complex in isolated nuclei. **(B)** Western blot analysis of a denaturing SDS-PAGE gel (3-8% Tris-acetate) using an antibody against CPSF73 of the various stages of purification of native CPSF from isolated nuclei treated with DSSO. CPSF73 might get cross-linked with a variety of nuclear proteins, but pull-downs on Strep-Tactin beads should specifically enrich CPSF. The scale of this experiment needs to be increased significantly to obtain enough material for mass spectrometry analysis of cross-linked peptides.