Architecture of Eukaryotic mRNA 3' End Processing Machinery and Insights into the Mechanism of Polyadenylation

This dissertation is submitted for the degree of Doctor of Philosophy

Churchill College

August 2019

Ananthanarayanan Kumar

UNIVERSITY OF CAMBRIDGE

MRC | Laboratory of Molecular Biology

To Usha and Kumar
Thank you for everything

and

To Mahesh (1992 – 2018)
Wherever you are, you are not alone

Preface

This dissertation is the result of my own work and includes nothing that is the outcome of work done in collaboration except as declared in the Preface and specified in the text. The structure determination of the polymerase module was a collaborative project with Ana Casañal. Gianluca Degliesposti performed the cross-linking mass spectrometry experiments. Sarah Maslen performed the HDX experiments. NMR experiments were collaboration with Conny Yu. The assays described in chapter 4 using native CPF were performed alongside Ana Casañal. Any work from collaborations is specifically acknowledged in the methods and results section.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text.

A major part of this work described in this dissertation has been published:

Architecture of Eukaryotic mRNA 3′-End Processing Machinery. Casañal*, Kumar* et al, *Science*, 358, 1056–1059 (2017). * Equal contribution

Activation of the Endonuclease that Defines mRNA 3′ Ends Requires Incorporation into an 8-Subunit Core Cleavage and Polyadenylation Factor Complex. Hill et al, Molecular Cell, 73, 1217–1231 (2019).

It does not exceed the prescribed word limit for the relevant Degree Committee.

Ananthanarayanan Kumar
August 2019

Summary

Architecture of eukaryotic mRNA 3' end processing machinery and insights into the mechanism of polyadenylation

This dissertation is submitted for the degree of Doctor of Philosophy by Ananthanarayanan Kumar

Almost all eukaryotic messenger RNAs (mRNAs) have a polyadenosine (polyA) tail at their 3′ end that is added by a multi-protein complex known as cleavage and polyadenylation factor (CPF). CPF, along with accessory cleavage factors (CF) IA and IB, cleaves the pre-mRNA within the 3' untranslated region (UTR) and adds a poly(A) tail. Previous work has shown that CPF is a fourteen-subunit complex that is organised into three enzymatic modules: phosphatase, nuclease and polymerase.

The polymerase module consists of the poly(A) polymerase Pap1, three RNA binding proteins (Cft1, Pfs2 and Yth1) and an unstructured protein Fip1. To understand how polymerase module recognizes specific RNA elements and how polyA tail addition is coordinated with other factors, I recombinantly expressed and purified a five-subunit polymerase module. Electron cryomicroscopy analysis resulted in a 3.5 Å resolution structure of Cft1-Pfs2-Yth1, revealing four β propellers in an arrangement reminiscent of other nucleic acid binding complexes. Using biochemical assays, I show that CF IA stimulates polyadenylation activity of CPF by interacting with polymerase module and tethering it to substrate RNA. Thus polymerase module acts as a hub to bring together the RNA, Pap1 and cleavage factors for specific and efficient polyadenylation.

The poly(A) tail length of newly made pre-mRNAs in *S. cerevisiae* is ~ 60 As. The nuclear poly(A) binding protein Nab2 is known to have a role in poly(A) tail length control. The molecular mechanism behind how CPF terminates polyadenylation to regulate uniform poly(A) tail length remains elusive. Using an *in vitro* polyadenylation assay with highly pure protein complexes, I have studied the mechanism of poly(A) tail length control by CPF. The assays highlight the contribution of the cleavage factors and the phosphatase module of CPF towards regulating the poly(A) tail length of a substrate RNA. Taken together, the findings discussed in this dissertation provide new insights into the architecture of eukaryotic mRNA 3' end processing machinery and into the mechanism of polyadenylation by CPF.

Acknowledgements

First and foremost, I am very grateful to my supervisor Lori Passmore for her support and supervision throughout my PhD. I would like to thank Lori for providing me with a lot of freedom to pursue interesting trajectories in the projects. I feel privileged to have been under Lori's tutelage at the LMB for my PhD. Thank you for everything Lori.

Next, I would like to thank my co-worker Ana Casañal. It was a great learning experience working together with Ana on the structure determination of polymerase module. We've had the joy of seeing many things together for the first time. I vividly remember the time in graphics room when we were able to see how Pfs2 inserts into Cft1 once we built the N-terminal alpha helix of Pfs2. Thank you Ana for all the good times. Big thanks to Chris Hill for all his help. I would like to thank Juan Rodríguez for being a solid source of support, for all the late night discussions about CPF and importantly for always being super excited about my new results. Thank you Manuel Carminati for being a wonderful colleague and a friend. We had lots of fun cloning the phosphatase module and APT together. It has been a pleasure working with Vytaute Boreikaite, both during her summer project and now during her PhD. I would also like to acknowledge Jana Wolf, Katrin Wiederhold, Gillian Dornan and Ashley Easter's contribution to the CPF project. Their work forms the foundation for almost all the experiments I present in this dissertation. It was an interesting experience for me working together as a team for the CPF project. I have learnt a lot of important lessons that I will carry with me.

I am glad to have found a friend in James Stowell, Michael Webster and Terence Tang. Thank you Michael for helping me get a head start with my project. I am forever grateful for all your advice. I would not have managed to make recombinant CPF if not for the motivation I received from James. You have had a tremendous influence on me during my PhD; thank you very much comrade! Terence, allow me to express my immense gratitude for everything you've done. I certainly would not have made it through if not for you. Thank you Fanconi boys - Shabih Shakeel and Pablo Alcon for all the good times in the lab. And Shabih, for those late night counsels. I feel privileged to be working alongside Conny Yu on the NMR project. You have not only taught me the basics of NMR but you remain an inspiration for me about how to approach science. I am beholden to Terence, Juan, Ana, Conny, Shabih, Pablo

and Lori for proof reading my thesis. All of you have made the lab a remarkable and an absolutely enjoyable place to work. And you all wonder why I am always in the lab?

I thank all the facilities at the LMB without which none of these experiments would have been possible: the EM facility, mass spec facility, the NMR facility, the biophysics facility, scientific computing, the media kitchen and Jianguo Shi (insect cell facility). I wish to acknowledge the technical help and advice from Chris Russo, Vinoth Kumar, Christos Savva, Alan Brown and Andy Boland. I would also like to thank my second supervisor Kiyoshi and my University supervisor Luca Pellegrini for their guidance. I thank Maud Pilkington, Jennie Lightfoot and Georgia Ntatsiou from the LMB, and Rebecca Sawalmeh from Churchill College for their support. A special thank you to Cristina Rada, Madan Babu and Vish Chandrashekaran for their guidance and wise words during challenging times.

My PhD was possible thanks to the generous funding I received from GATES Cambridge and the MRC. I feel very lucky to be a part of such a diverse GATES community. I would also like to thank Churchill college for making me feel home in Cambridge (and for instilling a love for Brutalism in me). I thank the Addenbrooke's Japanese Society for letting me be their honorary member. I always missed Japan a little less after our weekly Tuesday lunches. I would also like to thank my landlady Sarah Chivers for making me feel like family. Big thanks to my aunt Archana Venkat in London for letting me spend several end of term holidays with her family.

I am extremely grateful to have the love of my friends and family all over the world. Especially Sandhya Krishnan, Shrinidhi Guru, Surya Swaminathan, Arun Mohit, Ramya Chandran, Vijay Ravikumar, Siddarth Narasimhan, Paulina Rowicka, Shridhar Jagannathan, Eric Dawson, Krishna Sharma, Sankar Narayanan and Girish Krishnan. None of this would have been possible without my mother Usha and my father Kumar. I am so glad that you let me have a secret chemistry lab in my cupboard when I was twelve. I cannot thank you enough for everything you have done for me. I am forever indebted. Once again, thank you Lori for believing in my abilities and giving me this opportunity to work in your lab. Four years ago when I interviewed at the LMB on that cold January morning, never did I imagine that I would be spending the next four years of my life trying to answer such extremely interesting biological problems.

Table of Contents

List of Abbreviations

2D two-dimensional

3D three-dimensional

APT associated with Pta1 complex

ATP adenosine triphosphate

BEST–TROSY band selective excitation short transients transverse relaxation optimized spectroscopy

BP beta propeller

BuUrBu disuccinimidyl dibutyric urea

CCB competent cell buffer

CCCH cysteine cysteine cysteine histidine

CF IA cleavage factor IA

CF IB cleavage factor IB

CF II cleavage factor II

cm centimetre

CPF cleavage and polyadenylation factor

CPSF cleavage and polyadenylation specificity factor

Cryo-EM electron cryomicroscopy

CstF cleavage and stimulatory factor

CTD c-terminal domain

CTF contrast transfer function

CYC1 cytochrome c1

DEPC diethyl pyrocarbonate

DNA deoxy ribonucleic acid

dNTP deoxyribonucleotide triphosphate

DQE detective quantum efficiency

EB elution buffer

eBIC electron bio imaging centre

EDTA ethylenediaminetetraaceticacid

EMS electron microscopy sciences

EMSA electrophoretic mobility shift assays

FAM carboxyfluorescein

FBS fetal bovine serum

FEI field electron and ion company

FSC Fourier shell correlation

GAL7 galactose-1-phosphate uridylyltransferase

GCN4 general control protein

GE general electrical

GIF gatan image filter

GOI gene of interest

HAT histone acetyltransferases

HDX hydrogen deuterium exchange

HEAT huntingtin, elongation factor 3, protein phosphatase 2A, yeast kinase TOR1

His histidine

IAB isothermal assembly buffer

IgG immunoglobulin G

IPTG isopropyl β-D-1-thiogalactopyranoside

IVT in vitro transcription

kDa kilo daltons

L litre

LDS lithium dodecyl sulphate

LMB laboratory of molecular biology

M molar

m7G 7-methylguanosine

MFA2 mating hormone A-factor 2

mL milli litre

mM milli molar

MOPS 3-(N-morpholino)propanesulfonic acid

MQ milli-Q water

MRC medical research council

mRNA messenger RNA

MS mass spectrometry

nanoESI-MS nanoelectrospray ionisation mass spectrometry

NEB New England biolabs

NHS N-hydroxysuccinimide

nM nano molar

NMR nuclear magnetic resonance

NRD N-terminal RNA binding domain

PABN1 nuclear poly(A) binding protein 1

PAGE polyacrylamide gel electrophoresis

PAP poly(A) polymerase

PBS phosphate buffer saline

PCR polymerase chain reaction

PDB protein data bank

PF polyadenylation factor

Pol I RNA polymerase I

Pol II RNA polymerase II

Pol III RNA polymerase III

Poly(A) polyadenosine

PTM post translational modification

Py pyrimidine

rcf relative centrifugal force

rCPF recombinant cleavage and polyadenylation factor

RELION regularized likelihood optimization

ResQ resource Q

RMSD root mean square deviation

RNA ribonucleic acid

rpm rotations per minute

RRM RNA recognition motif

rRNA ribosomal RNA

SDS sodium dodecyl sulfate

SEC-MALS size exclusion chromatography with multi-angle light scattering

Sf9 spodoptera frugiperda 9

SnRNP small nuclear ribonucleotide particle

TAE tris-acetate-EDTA

TAPS tandem-affinity purification (protein A and streptactin)

TBE tris-borate-EDTA

TE tris-EDTA

TEMED tetramethylethylenediamine

Tm annealing temperature

tRNA transfer RNA

TY tryptone yeast extract

UBL ubiquitin like

UTR untranslated region

UV ultra violet

WD40 beta-transducin

YFP yellow fluorescent protein

ZF zinc finger

μ M micro molar

1 Introduction

1.1 Messenger RNA processing and regulation

Genes contain the blueprints that instruct the cell how to make proteins (in the case of protein-coding genes), or functional RNAs in the case of ribosomal RNAs (rRNAs) or transfer RNAs (tRNAs) for example. Gene expression is the procesess by which the instructions encoded in genes are utilised by the cellular machinery to synthesise the final gene products. Gene expression broadly involves the following steps: transcription of the genomic DNA into an intermediate RNA, processing of the intermediate RNA into its mature form, export into the cytoplasm and translation into proteins in case of messenger RNAs (mRNAs).

Genes coding for rRNAs are transcribed into a pre-rRNA by the RNA polymerase I (Pol I) and the genes coding for tRNAs are transcribed into pre-tRNA by RNA polymerase III (Pol III) (Paule and White 2000). These functional RNA molecules then undergo a series of modifications before being exported into the cytoplasm (Wang and He 2014). In the case of rRNAs, they go on to play important structural and enzymatic roles in the protein synthesis machinery of the cell - the ribosome (Brimacombe and Stiege 1985). rRNAs not only contribute to the assembly of the ribosomes but also function as a ribozyme in catalysing peptide bond formation. On the other hand, tRNAs function in decoding the instructions contained in the mRNA and act as an adaptor molecule bringing amino acids to the ribosome to carryout protein synthesis (Agirrezabala and Valle 2015). There are several other functional RNAs in the cell that perform various other functions e.g. the RNA components of telomerase and spliceosome etc.

Pre-mRNAs are synthesised from protein coding genes by RNA polymerase II (Pol II) machinery. As the pre-mRNA emerges out of Pol II, it undergoes several modifications that render the nascent RNA fit for export into the cytoplasm. These include addition of a 7-methylguanosine (m7G) cap at the 5' end, splicing of the non-coding intronic regions and cleavage at specific sites in the 3' untranslated region (UTR) followed by addition of multiple adenosine monophosphates at the 3' end (polyadenylation) (Figure 1.1). These three modifications transform the pre-mRNA into a mature mRNA that is then used by ribosomes to make proteins (Figure 1.1). It is thought that the components of the mRNA export adaptor interacts efficiently with properly synthesized poly(A) tails bound by poly(A) binding protein Nab2. Nab2 and

the export adaptor components subsequently interact with the nuclear pore complex resulting in efficient export of mRNA into the cytoplasm. In humans and budding yeast, pre-mRNA splicing defects leads to nuclear retention of polyadenylated transcripts. Intriguingly, such polyadenylated transcripts retained in the nucleus can then serve as substrates in mRNA decay pathways (Tudek, Lloret-Llinares, and Jensen 2018). Studies performed in fission yeast and human systems suggest a role for nuclear poly(A) binding proteins and the MTR4 helicase in recruiting poly(A) containing RNAs to the exosome nuclease. In budding yeast, Nab2 has also been implicated in nuclear mRNA decay via its interaction with the spliceosomal proteins. In summary, the various mRNA processing machineries including the spliceosome, the cleavage and polyadenylation factor (CPF), the various poly(A) binding proteins and the exosome work together to regulate the state and levels of mRNA in the nucleus.

By regulating the stability and the chemical properties of the mRNA, cells can fine tune how much of a given protein an mRNA can produce, as well as its spatio-temporal properties. Any changes in the interaction between the mRNA and its binding partners (that are involved in transcription, pre-mRNA processing, export or translation) can alter the stability or the fate of the mRNA. These changes are usually a result of environmental, epigenetic or developmental factors. For example, alternative polyadenylation is a phenomenon in which the mRNA 3' end processing machinery cleaves and polyadenylates at a different site in the 3' UTR rather than the consensus polyadenylation site. This may be caused in part by an increase in the expression levels of cleavage factor proteins that recognize the cleavage site. Changes in the habitual cleavage site in the 3' end of a pre-mRNA results in mRNAs with different UTR lengths, effectively altering the proteome that could interact with the 3' UTR. Such changes play a crucial role in the life of an mRNA and could lead to changes in gene expression. For example in human cells, a normal 3' UTR of the cell surface protein CD47 enables efficient cell surface expression, which protects cells from phagocytosis by macrophages. The uridine-rich elements in the longer form of the CD47 3' UTR interacts with the RNA binding protein HuR, which in turn is responsible for bringing the protein SET to the site of translation of CD47. SET then interacts with the CD47 proteins and recruits it to the plasma membrane via its association with RAC. A shorter form of the 3'

UTR that can no longer interact with HuR, does not recruit SET to the site of translation. This results in localisation of CD47 protein to the site of translation i.e. endoplasmic reticulum (Berkovits and Mayr 2015). Thus, the various mRNA processing steps such as splicing or 3' end processing not only contribute to the normal functioning of an mRNA but also play a crucial role in regulating its functions.

The focal point of this dissertation is to understand the mechanism of mRNA 3' end processing by using a structural and biochemical approach. The introduction chapter begins with a focus on the historical overview of the "mRNA 3' end" field. This is followed by a description of the properties and functions of the components of the mRNA 3' end-processing machinery. Emphasis has been placed on discussing the experiments carried out using *Saccharomyces cerevisiae* as the model organism. The chapter ends by listing the specific aims of this dissertation and how they are addressed.

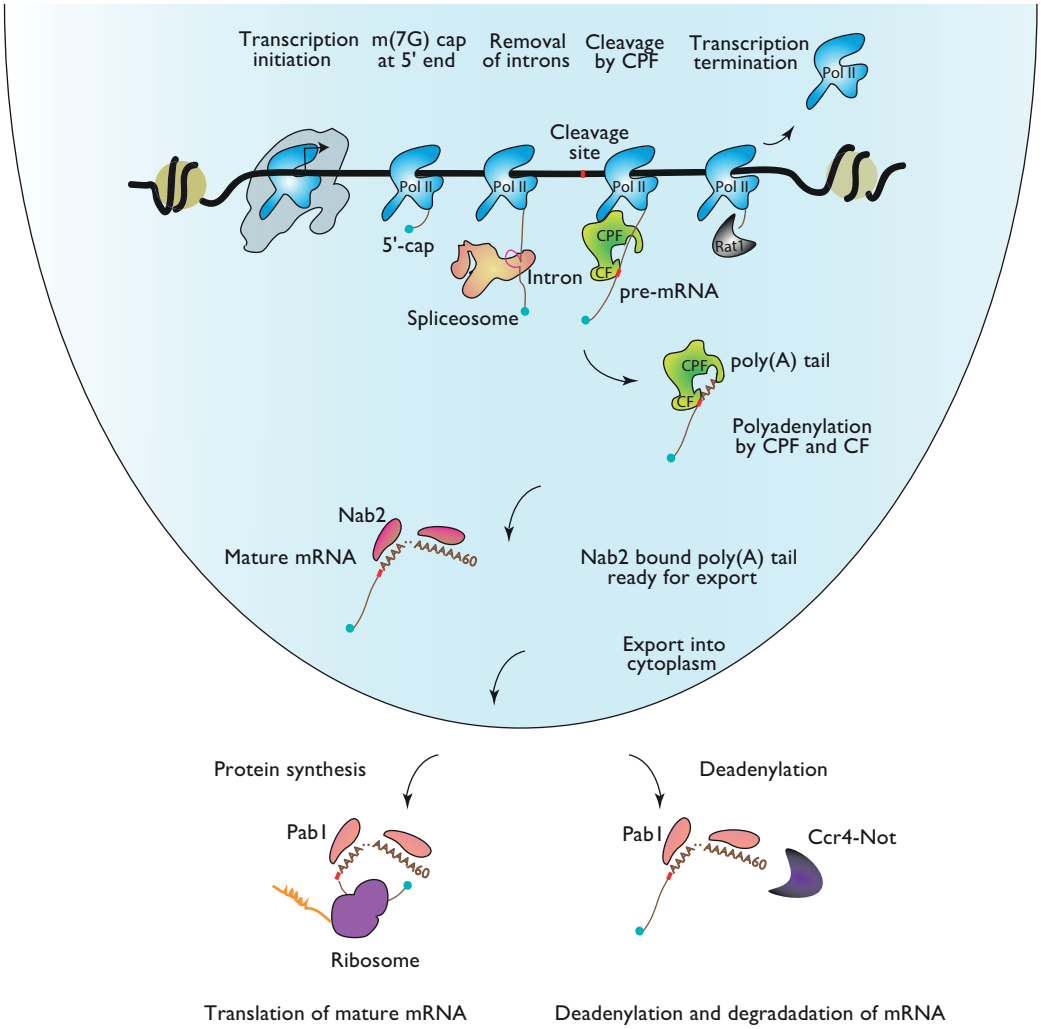


Figure 1.1: The life of an mRNA inside a cell. RNA polymerase II (shown in light blue) transcribes the information in the DNA (in black) into an intermediate pre-mRNA (in brown). As the pre-mRNA emerges out of RNA polymerase II, it undergoes several modifications. The capping enzyme complex carries out the capping at the 5' end, whereas the spliceosome (in peach) removes the introns. The cleavage and polyadenylation factor (CPF) along with the accessory cleavage factors (CFs) cleaves at specific sites in the 3' end, followed by the addition of a poly(A) tail. Nab2 bound mature mRNAs are exported into the cytoplasm by the nuclear pore complex. In the cytoplasm, the information in the mRNA sequence is decoded by the ribosome to synthesize linear chains of amino acids. Alternatively, the mRNA deadenylation complexes including Ccr4-Not targets poly(A) containing RNAs and primes them for subsequent degradation.

1.2 Cleavage and polyadenylation at 3' ends of mRNA

The cleavage and polyadenylation at the 3' end of an mRNA is carried out by a large multi-subunit protein machinery that includes the Cleavage and Polyadenylation Factor (CPF), and accessory cleavage factors (CF) IA and IB. As the pre-mRNA emerges out of pol II, CPF and CF IA/B recognises specific *cis* elements in the nascent RNA, cleaves at a specific site and adds a poly(A) tail to the cleaved 3' end (Figure 1.1) (Hollerer et al. 2014). The cleavage of the mRNA is coupled tightly to the termination of transcription by pol II. In the torpedo model of transcription termination, the cleavage of mRNA by CPF is followed by recruitment of the Rat1 exonuclease to the nascent mRNA. Rat1 degrades the free end of the mRNA and stops transcription when it reaches pol II. The length of the poly(A) tails of newly made pre-mRNAs in humans is ~250 As and in *Saccharomyces cerevisiae* is ~ 60 As (M Edmonds, Vaughan, and Nakazato 1971; Mclaughlin et al. 1973; Groner and Phillips 1975). Defects in cleavage and polyadenylation are associated with diseases including cancer and β- thalassemia (Curinha et al. 2014).

In the past 59 years, there has been tremendous progress in our understanding of how mRNA 3' ends are formed. However, there are still many fundamental questions that remain to be addressed. It is interesting to note that some of the major advancements in the field were always accompanied by important technical developments in molecular biology. Similarly, the results published in this thesis are also a result of major technical progress made in electron cryomicroscopy (cryo-EM) and recombinant protein production. Discussed below are some of the important discoveries that resulted in the birth and evolution of the mRNA 3' end-processing field, thus sparking an interest in the study of mRNA 3' ends.

1.2.1 Historical overview

An enzyme that catalyses the ATP-dependent synthesis of a linear sequence of adenosines was discovered in 1960 (Mary Edmonds and Abrams 1960). The linear poly(A) chain was linked via the 3' to 5'- phosphodiester linkage. An elegant study by Nakazato and co-workers demonstrated that the poly(A) tails were indeed covalently linked to the end of an mRNA (M Edmonds, Vaughan, and Nakazato 1971). This was one of the earliest studies also to report the observation of a homogenous poly(A) chain length (~200As) in nuclear RNA from humans. The functional importance for poly(A) tails were soon recognised. Injection of polyadenylated mRNAs into *Xenopus* oocytes conferred stability *in vivo* (Georges Huez et al. 1975). It was also shown that polyadenylation functionally stabilizes histone RNAs (G. Huez et al. 1978). Much later, *in vitro* and genetic studies revealed an involvement of poly(A) tails in the synthesis of proteins (Sachs and Davis 1989; Tarun and Sachs 1995). The discovery of poly(A) tails and emerging understanding of it's properties led to development of mRNA purification techniques and cDNA generation protocols (Venkatesan, Elango, and Chanock 1983).

In the 1970s, there were many studies involving the purification and characterization of the enzymes that add the poly(A) tail (Mary Edmonds and Winters 1976). It was found that the poly(A) polymerase enzyme alone did not have any substrate specificity (Winters and Edmonds 1973; Tsiapalis, Dorson, and Bollum 1975). However, *in vivo* pulse labelling studies showed that the 3' end of an mRNA is formed by cleavage and polyadenylation at specific sites (Nevins and Darnell 1978). *In vitro* experiments using a HeLa cell lysate then demonstrated the sequence specificity of mRNA polyadenylation (J L Manley 1983), but this study did not find any cleavage activity. Moore and Sharp were the first to demonstrate site and sequence specificity in cleavage coupled with polyadenylation (C L Moore and Sharp 1984). They attributed the specificity of this polyadenylation to the endogenously produced RNA. The exogenously added RNA did not have specific polyadenylation, thus showing that transcription was somehow connected with cleavage and that transcription was not a prerequisite for polyadenylation. Yet, the factors contributing to the specificity of cleavage and polyadenylation remained elusive.

The development of improved HeLa nuclear extracts for reconstituting the splicing

reaction furthered the biochemical advancements in the field of 3'-end processing. Using non hydrolysable ATP analogs, Claire Moore showed that cleavage still occurs in the absence of polyadenylation and that cleavage and polyadenylation can occur independent of each other *in vitro* (Claire L Moore and Sharp 1985). Subsequent experiments demonstrated that the cleavage reaction was endonucleolytic in nature (C L Moore, Skolnik-David, and Sharp 1986). Studies by Butler and Pratt in 1988 showed that the mature 3'-end of yeast mRNAs were formed by endonucleolytic cleavage followed by polyadenylation and not due to transcription termination (Butler and Platt 1988).

The contribution of the RNA elements to the cleavage and polyadenylation reaction was also extensively researched. Proudfoot and Brownlee found a six-nucleotide AAUAAA sequence in all of the six mRNAs they studied (N. J. Proudfoot and Brownlee 1976). Later this sequence was shown to be important for proper formation of mRNA 3' ends by using genetic means (Fitzgerald and Shenk 1981). Higgs and colleagues showed that a single point mutation in the AAUAAA site in either human α2-globin gene or the human Fl-globin gene led to defects in mRNA 3'-end processing, and provided the molecular basis for α-thalassemia and in β-thalassemia (Higgs et al. 1983). This study was one of the first to highlight the medical importance of 3'-end processing. The progress made in our understanding of these *cis*- elements aided in the discovery of the 3'-end processing protein complex as discussed below.

Using an RNAse H protection experiment, Zarkower and Wickens found that some protein factors in HeLa nuclear extract form a complex with the AAUAAA region of the substrate RNA and that this complex formation is necessary for cleavage and polyadenylation (Zarkower and Wickens 1987). Around the same time, work from Walter Keller's lab also demonstrated the existence of an mRNA 3' end processing complex using native gel shift assays (Humphrey et al. 1987). Attempts were made to fractionate the nuclear extracts to identify the components that make up these complexes. The obvious candidates were a poly(A) polymerase and an endonuclease. The lack of substrate specificity in polyadenylation by the polymerase alone suggested that the 3' end processing reaction was made up of other unknown proteins that could confer specificity. The breakthrough came in 1988 when the labs of James Manley, Walter Keller and Joseph Nevins isolated separate complexes from HeLa nuclear

extracts (James L. Manley 1988; Christofori and Keller 1988; Mcdevitt et al. 1988). Some of these individual complexes could perform the cleavage reaction in a site-specific manner. None of these separate fractions had specific polyadenylation activity, but when combined together could restore the specificity that is found in nuclear extracts. This is highly reminiscent of RNA polymerases from eukaryotes that function efficiently only upon formation of ternary complexes with initiation and elongation factors. Figure 1.2 visually highlights some of these major milestones in the mRNA 3' end-processing field.

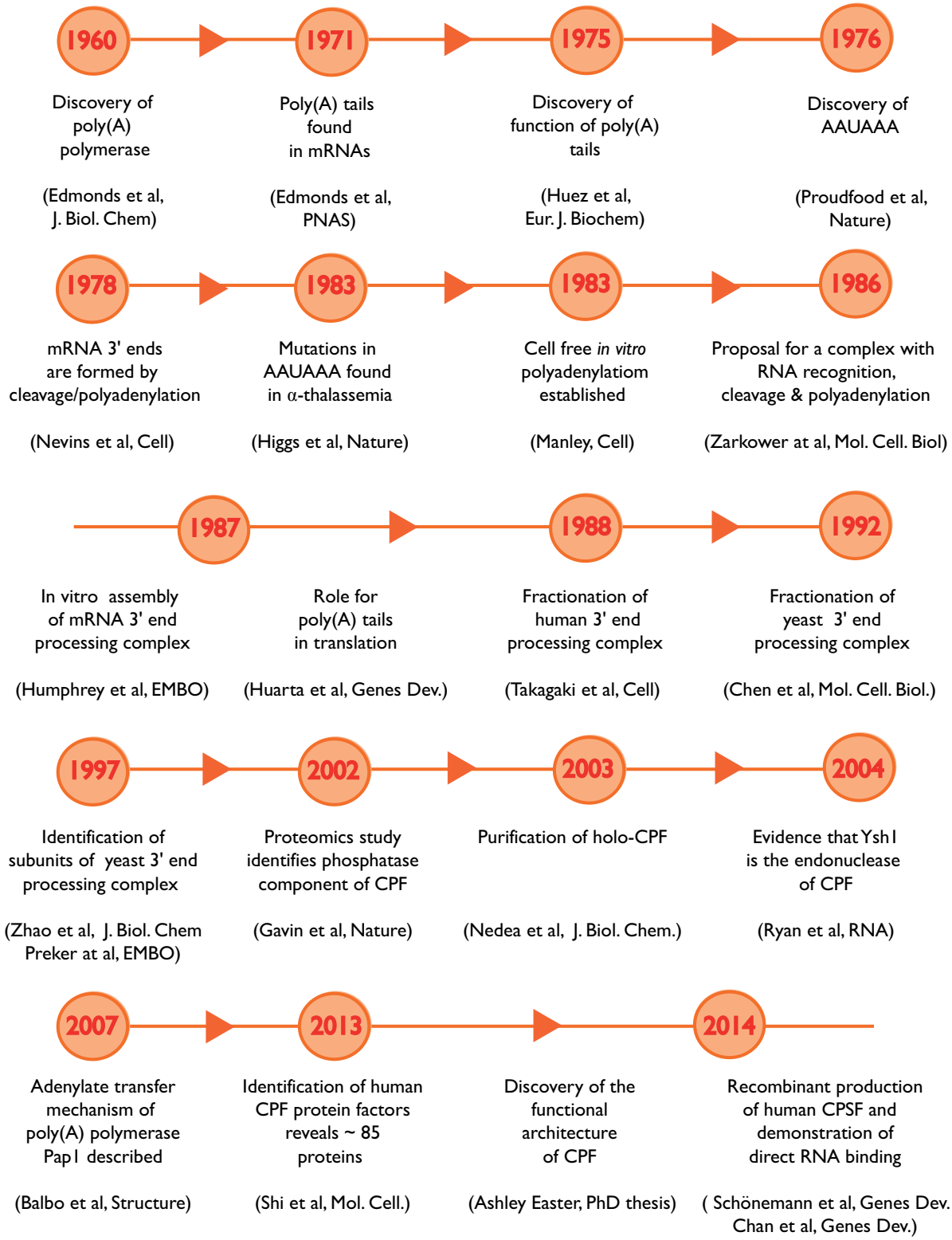


Figure 1.2: A historical timeline of some of the major discoveries in the mRNA 3' end-processing field.

1.2.2 Understanding the mRNA 3' end processing machinery

Using anion exchange chromatography of a yeast nuclear extract, Chen and Moore isolated four different fractions that when combined could reconstitute the cleavage and polyadenylation of substrate RNAs (J. Chen and Moore 1992). They named these four fractions cleavage factor (CF) I and II, polyadenylation factor (PF) I and polyA polymerase (PAP). They found that CFI and CFII when combined together could cleave a substrate RNA, whereas PFI and PAP could not cleave the RNA. However, PFI and PAP when combined with CFI and CFII could specifically polyadenylate the cleaved RNA. At this stage, the idea of a multi-protein complex being responsible for the specific cleavage and polyadenylation of mRNA 3'-ends began to emerge, but the composition of such a complex was still unknown. These above studies set the stage for detailed biochemical characterization of the mRNA 3' end-processing complex described below.

Purification of these individual factors from endogenous sources revealed the numbers and identities of their components (Preker et al. 1997; Zhao, Kessler, and Moore 1997). Endogenous purifications from the lab of Walter Keller showed that the *S. cerevisiae* Cleavage and Polyadenylation Factor (CPF) contains the proteins Cft1, Cft2, Ysh1, Pta1, Pap1, Fip1, Yth1, Pfs2 and other unidentified proteins. Proteomic studies using affinity purification of CPF coupled with mass spectrometry identified 12 known interaction partners of CPF and 7 additional components (Gavin et al. 2002). Subsequent pull downs from *S. cerevisiae* using a TAP tag on the Ref2 protein, further characterized the holo-CPF that contained 15 subunits. This new holo-CPF consisted of all the known components as well as the newly discovered components Syc1, Pti1, Swd2, and two protein phosphatases Glc7, Ssu72 (Nedea et al. 2003). Thus our current understanding of the yeast mRNA 3' end processing machinery is that it is made up of the CPF and CFI complexes (Figure 1.3).

CPF has three enzymatic activities: two protein phosphatases (Glc7 and Swd2), an endonuclease (Ysh1) and a polymerase (Pap1). Importantly, the accessory CFI has two components: CF IA and CF IB (Figure 1.3). CF IA is made up of four proteins - Rna14, Rna15 and two unknown proteins (Kessler, Zhao, and Moore 1996), which were subsequently identified as Clp1 and Pcf11 (Minvielle-Sebastia et al. 1997; Gross and Moore 2001). CF IB is made up of a single polypeptide (Hrp1) that is

critical for cleavage site selection. CF IA and CF IB act together with intact CPF to perform accurate, site-specific cleavage and polyadenylation of pre-mRNAs (Kessler et al. 1997; Kessler, Zhao, and Moore 1996; Minvielle-Sebastia, Preker, and Keller 1994).

A crude purification of the mRNA 3'-end processing machinery from HeLa nuclear extracts revealed the identities of the ~85 proteins that forms the human complex (Shi et al. 2009). Those 85 proteins not only included several 3'-end processing components but also factors associated with transcription, splicing, DNA-damage repair and translation. Bonafied components of the human 3'-end processing machinery includes the poly(A) polymerase (PAP) and four different multi-subunit protein complexes: cleavage and polyadenylation specificity factor (CPSF), cleavage factor I (CFI), cleavage factor II (CFII) and cleavage stimulatory factor (CstF). In addition to the abovementioned factors, the nuclear poly(A) binding protein PABN1, the scaffolding protein Symplekin and the CTD of pol II are additional components of the mammalian 3'-end processing machinery (Xiang et al. 2010).

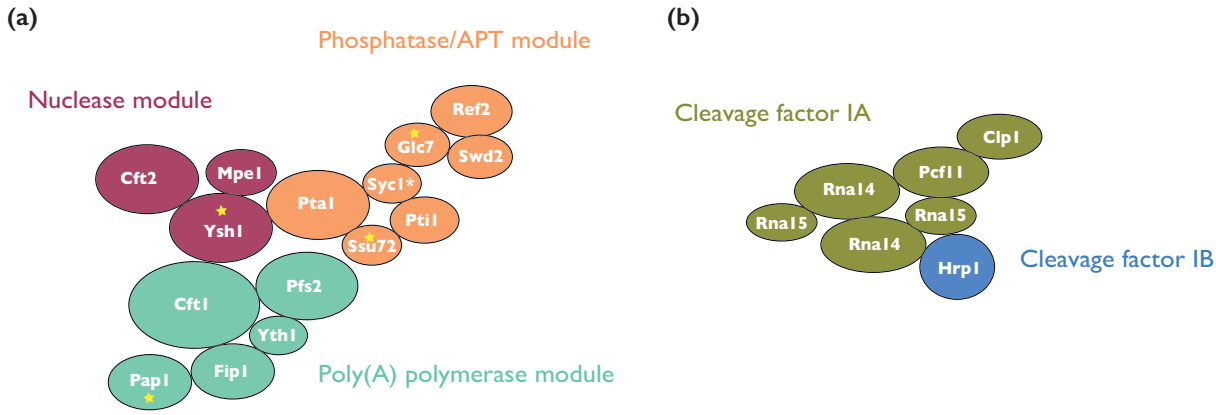


Figure 1.3: (a) The composition of *S.cerevisiae* CPF and its protein-protein interaction network from nanoESI-MS data and *in vitro* pull downs. Protein symbols are scaled down to have an area proportionate to their molecular weights. Interaction between proteins is represented by contact between the symbols. Yellow stars denote enzymes. **(b)** The composition of *S.cerevisiae* CF I and its protein-protein interaction network from *in vitro* pull downs and previous literature.

Recent work from our lab has resulted in a clear understanding of how the three-enzymatic activities of CPF are segregated in the organisation of the complex in *S. cerevisiae* (Easter 2014). Native mass spectrometry analysis of several different preparations of CPF found that CPF consists of three enzymatic modules: a six subunit phosphatase module that includes subunits Glc7 and Ssu72, a three subunit nuclease module that includes the endonuclease Ysh1, and a five subunit polymerase module that includes the poly(A) polymerase Pap1 (Figure 1.3a). This work provided biophysical evidence of the stoichiometry and composition of CPF. Thus, the previous nomenclature of the *S. cerevisiae* 3' end processing machinery that includes it's segregation into CFII, PFI and PAP has been rendered obsolete with this newfound architecture of CPF (Figure 1.3a).

1.2.2.1 Cleavage and polyadenylation factor

Outlined below is the description of proteins that make up the three-enzymatic modules of CPF.

1.2.2.1.1 The polymerase module

The polymerase module in yeast is made up of five proteins: Cft1, Pfs2, Yth1, Fip1 and Pap1 (Figure 1.4). A very similar polymerase complex was purified from human cells and was sufficient to reconstitute specific polyadenylation (Schönemann et al. 2014).

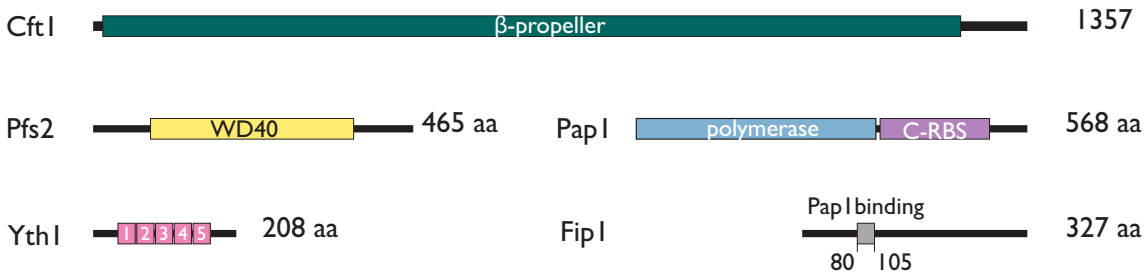


Figure 1.4: Schematic representation of yeast polymerase module subunits. Cft1 is predicted to contain three β-propeller repeats (Scrima et al. 2008). The predicted WD40 domain of Pfs2 is highlighted in yellow. Pap1 contains an N-terminal polymerase domain (in blue) and a C-terminal RNA binding surface (in purple) (Bard et al. 2000). The five zinc fingers of Yth1 are colored in pink. Fip1 is an unstructured protein with residues 80 to 105 involved in binding Pap1 (in grey).

Cft1 is rich in β-strands and is the largest of all the CPF proteins (154 kDa in size). It was identified in yeast owing to its sequence similarity to the mammalian counterpart that was known to bind AAUAAA (Stumpf and Domdey 1996). Using purified Cft1, it had been shown that the central region in the sequence of Cft1 can mediate binding to RNA (Dichtl et al. 2002). The human Cft1 (or CPSF-160) interacts with the nuclease Ysh1 (or CPSF-73), the scaffolding protein Pta1 (or

Symplekin) and the pseudo-nuclease Cft2 (or CPSF-100) to form the core of the human CPSF complex (Sullivan, Steiniger, and Marzluff 2009). Similar to its yeast counterpart, CPSF-160 is rich in β-strands and binds RNA *in vitro* (Keller et al. 1991; Dichtl et al. 2002).

Pfs2 was identified as a component of CPF during the purification of a multi-subunit complex that had specific polyadenylation activity (Preker et al. 1997). Pfs2 is a WD40 repeat-containing protein that has been proposed to mediate the interaction of CPF with the cleavage factors (Ohnacker et al. 2000). The human polymerase complex directly binds the polyadenylation signal via Pfs2 (or WDR33) and Yth1 (or CPSF 30) (Chan, Huppertz, Yao, Weng, Moresco, Yates Iii, et al. 2014). The human homologue of Pfs2 (WDR33) has an N-terminal WD40 domain and it is not uncommon for WD40 domains to mediate nucleic acid binding (Jin et al. 2016). Unlike Pfs2 that only contains the WD40 domain, WDR33 also contains a central collagen like domain and a C-terminal GPR rich sequence. The roles of the central and the C-terminal domains remain unclear.

Yth1 is another RNA binding component of the polymerase module that contains five CCCH zinc finger (ZF) domains (Tacahashi, Helmling, and Moore 2003; Barabino et al. 1997). Yth1 and its human counterpart CPSF30 are thought to recognise the RNA cleavage sites (Barabino, Ohnacker, and Keller 2000b; Chan, Huppertz, Yao, Weng, Moresco, Yates, et al. 2014). CPSF-30 interacts with the AAUAAA sequence via zinc fingers two and three (Chan, Huppertz, Yao, Weng, Moresco, Yates Iii, et al. 2014). CPSF-30 also contains a zinc knuckle at its C-terminal end that is not present in Yth1. Similarly, zinc fingers two and three of Yth1 specifically mediate the RNA binding, whereas zinc finger four and five are thought to interact with the unstructured protein Fip1 (Tacahashi, Helmling, and Moore 2003).

Fip1 was discovered as an interaction partner of Pap1 (Preker et al. 1995). Fip1 has an acidic stretch at its N-terminus, and a proline-rich C-terminal region. Human Fip1 (hFIP1) is almost two times bigger in size compared to the *S. cerevisiae* Fip1 due to the additional R-rich and RD- rich domains in the C-terminal region. The N-terminal residues 80-105 of Fip1 interact with Pap1, whereas the C- terminal residues 206-220 interact with Yth1. A crystal structure of the Fip1 peptide 80-105 bound to Pap1 reveals molecular details of this interaction (Meinke et al. 2008). Fip1 is thought to tether Pap1 to the rest of CPF via a central flexible linker and also provide a platform

for protein-protein interaction for other CPF components (Helmling, Zhelkovsky, and Moore 2001; Ezeokonkwo et al. 2011). Deletion of the last 107 residues of Fip1 (that lies after the Yth1 binding site) had an insignificant effect on yeast viability and on polyadenylation by CPF. hFIP1 binds U-rich containing RNA via its C- terminal arginine rich domain (Kaufmann et al. 2004).

Pap1 is the enzyme that synthesizes the poly(A) tail at the mRNA 3' ends (Lingner et al. 1991) and is one of the most extensively characterised subunits of CPF. Pap1 belongs to the DNA polymerase β family of enzymes. Although Pap1 loses its substrate specificity when in isolation, it has nucleotide specificity (M Edmonds 1990). The human homologue of Pap1 (PAP) acquires substrate specificity for polyadenylation upon associating with the CPSF complex (Schönemann et al. 2014). The N-terminal region of the yeast and mammalian PAPs share substantial sequence identities, whereas the C-terminal domains do not share this similarity. A deletion of the first 18 amino acids of Pap1 affects the specific polyadenylation of pre-mRNA 3' ends by CPF as this region is thought to contact other CPF subunits (Zhelkovsky, Kessler, and Moore 1995; Ezeokonkwo et al. 2012). The C-terminal end of Pap1 also has an RNA binding domain, the exact role of which is still unknown (Zhelkovsky, Kessler, and Moore 1995). Unlike yeast that contains only two poly(A) polymerase (Pap1 and Trf4), mammals have different nuclear PAPs - canonical PAP, neo-PAP, star-PAP etc (Vaňáčová et al. 2005; Laishram 2014). The canonical PAP is the most similar one to the yeast Pap1. Although the molecular details of nucleotidyl transfer are conserved between PAP and Pap1, the extended C-terminus of PAP is subjected to several post- transcriptional modifications that ultimately regulate its function. The detailed mechanism of how Pap1 adds multiple adenosine tails to the 3' end of the RNA is discussed in section 1.2.4. It is to be noted that there are no experimentally determined three-dimensional structures for any of the polymerase module components except for Pap1 and a Fip1 peptide that interacts with Pap1 (Bard et al. 2000).

1.2.2.1.2 The nuclease module

The nuclease module in yeast is made up of three proteins: Cft2, Ysh1 and Mpe1. **Cft2** is an inactive pseudo-nuclease that belongs to the β-CASP family. The exact role of Cft2 in cleavage and polyadenylation remains obscure but Cft2 has been shown to be essential for cell viability, and recognition of 3' end processing mRNA cleavage site *in vivo* (Kyburz et al. 2003). UV cross-linking experiments proposed a role for Cft2 in RNA binding (Zhao, Kessler, and Moore 1997), and pull down experiments with Pol II also suggest that Cft2 has a role in coupling 3' end processing and transcription (Kyburz et al. 2003). The crystal structure of Cft2 revealed an N-terminal metallo-β-lactamase domain and a C- terminal β-CASP domain (Mandel et al. 2006). Cft2 and Ysh1 not only share an overall sequence identify of 19.5 %, but the structural architecture of Cft2 is very similar to that of Ysh1 or CPSF-73. Interestingly, the zinc coordinating residues found in Ysh1 are not conserved in Cft2 and hence it does not bind zinc. This explains why Cft2 does not possess any nuclease activity.

Ysh1 is the endonuclease that cleaves the mRNA 3' end (Mandel et al. 2006; Zhelkovsky et al. 2006). Interestingly, strains harbouring mutations in Ysh1 have been found to be defective in splicing (Garas, Dichtl, and Keller 2008). Mutating the conserved metal coordinating residues in the metallo-β-lactamase and β-CASP domains of Ysh1 results in lethality, showing that Ysh1 requires divalent ion binding for its nuclease activity (Ryan, Calvo, and Manley 2004).

Mpe1 is a zinc knuckle containing protein that is known to be important for specificity of mRNA cleavage activity (Vo et al. 2001). Mpe1 was found to contain an ubiquitin-like (UBL) domain, a RING finger domain in addition to the zinc knuckle (Lee and Moore 2014). The zinc knuckle and RING finger have been shown to play an important role in specific RNA binding. The functional role of the UBL domain of Mpe1 remains to be explored.

1.2.2.1.3 The phosphatase module

The phosphatase module in yeast is made up of six proteins: Pta1, Ref2, Pti1, Swd2, Glc7 and Ssu72. The identities of Pti1, Swd2, Glc7 and Ssu72 in CPF purifications remained unknown and were the last of the yeast CPF proteins to be described (Dichtl et al. 2002; He et al. 2003; Gavin et al. 2002; Nedea et al. 2003). It is to be noted that some of the earlier purifications of CPF had some additional unidentified bands that could have been some of the phosphatase subunits. For example, it is plausible that the bands labelled as p 35 and p 36 in pull downs of the polyadenylation complex performed in Walter Keller's lab could be Swd2 and Glc7 (Preker et al. 1997). But interestingly none of these early purifications had any obvious bands on SDS-gels that could likely correspond to Ref-2 (Ohnacker et al. 2000). The large-scale proteomic study that involved co-immunoprecipitation of the polyadenylation complex using Pti1 as bait identified Ref2, Pti1, Glc7 and Ssu72 (Gavin et al. 2002). In absence of detailed experimental conditions (such as the number of washes between each steps) of these and earlier pull downs, it is difficult to ascertain why the phosphatase subunits were not identified earlier.

Pta1 was originally discovered as a gene that functions in pre-tRNA processing (O'Connor and Peebles 1992). Extracts prepared from a Pta1 mutant strain were found to be defective in both steps of mRNA 3' end processing (Preker et al. 1997; Zhao et al. 1999). The human homologue of Pta1, known as symplekin, is thought to act as a scaffold for protein-protein interaction within CPF (Xiang, Manley, and Tong 2012). The N-terminus of Pta1 harbours HEAT repeats similar to its metazoan counterpart and further supports the idea that Pta1 could act as a structural scaffold for CPF (Kennedy et al. 2009). Previous work from our lab has shown that Pta1 mediates the interaction between the phosphatase module and the rest of CPF via its association with the nuclease module (Figure 1.3a). This role of pta1 in bridging CPF components is further supported by previous studies where symplekin forms a tight complex with the nuclease module components Ysh1 (or CPSF-73) and Cft2 (or CPSF-100) (Sullivan, Steiniger, and Marzluff 2009). It is thought to influence the nuclease activity of CPSF-73 by bringing other regulatory proteins to the complex.

Ref2 was discovered as an RNA binding protein that plays a key role in poly(A) site selection (Russnak, Nehrke, and Platt 1995). A role for Ref2 in snoRNA formation has also been reported (Dheur et al. 2003). Interestingly, Ref2

is the only component of CPF that is not essential for the viability of the yeast cells. **Swd2** is a WD40 repeat containing protein that is involved in Pol II termination and is a shared subunit of the histone H3-K4 methylation complex, COMPASS (Roguev et al. 2001; Cheng, He, and Moore 2004). Overexpression of Ref2 can rescue the growth defect phenotype of a Swd2 deletion in yeast, highlighting a functional interaction between the two proteins (Cheng, He, and Moore 2004). The phosphatase enzymes Ssu72 and Glc7 are the newest addition to the family of CPF subunits (Edmund P. Walsh et al. 2002).

Ssu72 is the phosphatase that acts on the Ser-5 and Ser-7 of the pol II CTD and helps the initiation-elongation cycle of pol II (Krishnamurthy et al. 2004; Reyes-Reyes and Hampsey 2007; Rosado-Lugo and Hampsey 2014; Xiang, Manley, and Tong 2012). The three-dimensional structure of Ssu72 bound to its substrate (Ser5 phosphorylated pol II CTD) and in complex with the N-terminus of the scaffolding protein Pta1 (Symplekin) has been determined (Xiang et al. 2010). This structure revealed details about how Ssu72 specifically recognises the Ser5 phosphorylated CTD and how Pta1 could stimulate the phosphatase activity of Ssu72. In the structure, the CTD is bound in the active site of Ssu72 with the Ser5-Pro6 peptide bond in an unusual *cis* configuration. This unusual geometry could explain the specificity of Ssu72 for Ser5 phosphorylated CTD peptide.

Glc7 is a PP1 family phosphatase that acts on the Tyr-1 of the pol II CTD and coordinates transcription termination (Schreieck et al. 2014). The CTD of the largest subunit of Pol II consists of 26-tandem heptapeptide repeats in yeast, and 52- tandem heptapeptide repeats in humans. Almost all of the 26-heptapeptide repeats in yeast follow the consensus sequence Tyr1-Ser2-Pro3-Thr4-Ser5-Pro6-Ser7 (Eick and Geyer 2013). However, in mammals the repeats vary from the consensus sequence towards the distal end of the CTD. The CTD repeats are important for coupling transcription termination to cleavage and polyadenylation of pre- mRNAs (Hsin and Manley 2012). The CTD undergoes several post-translational modifications with phosphorylation being the most important for mRNA 3' end processing (McCracken et al. 1997; Hirose and Manley 1998). The CTD repeat is intrinsically disordered and can adopt diverse structural architecture upon interactions with different 3' end processing proteins such as Pcf11, Ssu72 and Glc7. These interactions are not only regulated by PTMs such as phosphorylation but also by other factors such

as *cis-trans* isomerisation of the peptide bonds (Xiang et al. 2010; Werner-Allen et al. 2011). The CTD modification changes throughout the different stages of transcription and the 3' end processing factors are recruited just before termination. The exact molecular details of CTD mediated transcription termination and pre-RNA cleavage is still elusive.

1.2.2.2 Cleavage Factors

CF IA and CF IB are the two cleavage factors in yeast. CF IA is a hexamer that consists of a hetero- tetramer of Rna14 and Rna15, and a single copy of Pcf11 and Clp1 (Gordon et al. 2011) (Figure 1.3b). In humans, the cleavage factors are more complicated than their yeast counterparts. There exists a similar complex known as the cleavage stimulatory factor or CstF. The three proteins that make up CstF are CstF-77 (or Rna14), CstF-64 (or Rna15) and CstF-50 (no yeast homologs). In addition to CstF, there also exists hCFI and hCFII. hCFI is made up of proteins that contain no known yeast homologs (Rüegsegger, Blank, and Keller 1998). hCFII is made up of hPcf11 and hClp1, the human homologs to yeast Pcf11 and Clp1.

1.2.2.2.1 Cleavage Factor IA

Rna14 consists of a large HAT domain in the N-terminus and a monkey tail region in its C-terminus (Paulson and Tong 2012). The HAT domain of Rna14 is made up of 12 HAT motifs and is thought to act as a protein-protein interaction platform for bringing together CPF and CFI (Ohnacker et al. 2000). Similar to Rna14, CstF-77 is a HAT domain-containing protein that dimerizes via its HAT-domain (Preker and Keller 1998; Paulson and Tong 2012; Bai et al. 2007). CstF-77 (or Rna14) interacts with CstF-64 (or Rna15) via a C-terminal proline rich region (Hockert, Yeh, and MacDonald 2010). Experiments show that extracts prepared from Rna15 or Rna14 mutant cells cannot restore mRNA 3’ end processing when mixed together. This shows that both Rna14 and Rna15 are essential for 3’ end processing and that either subunit cannot substitute for the other.

Rna15 is an RRM-domain containing protein that contributes to the RNA binding property of CF IA (Minvielle-Sebastia et al. 1991), and has been shown to bind A-rich

signals in the mRNA 3' UTR (Gross and Moore 2001). Cstf-64 binds RNA via an N-terminal RRM domain (Takagaki et al. 1992). A later study also demonstrated that Rna15 prefers binding U or G/U rich tracts of a substrate RNA (Pancevac et al. 2010). The NMR structure of the RRM domain of CstF-64 reveals the molecular details of its interaction with U or G/U rich RNA (Perez Canadillas and Varani 2003). It has been hypothesized that the proximity of the downstream U-rich and GU-rich sequences in human UTRs might enable binding of two copies of CstF-64 next to each other, connected by the CstF-77 HAT domains (Xiang, Tong, and Manley 2014). Cstf-64 contains a central proline-glycine rich region followed by an alpha helical MEARA/G repeat which is not present in its yeast homologue, Rna15. The C-terminal end of Cstf-64 and Rna15, however, is highly conserved in eukaryotes and is required for polyadenylation (Qu et al. 2007). The C-terminal conserved region of Rna15 interacts with Pcf11 and this region is necessary for efficient mRNA 3' end processing. The central hinge domain of Rna15 and the monkey tail domain of Rna14 are the minimal regions required for a stable complex formation between the two (Moreno-Morcillo et al. 2011). CstF-50 (no known yeast homologue) contains a dimerization domain in its N-terminal end and a WD40 domain in its C-terminus. Dimerization of CstF-50 along with the hetero-tetramer CstF-77/CstF-64 results in the formation of a hexameric CstF complex.

Pcf11 was identified as a component of CF IA in a two hybrid screen and a strain harbouring a Pcf11 mutant demonstrated shortening of poly(A) tails (Amrani, Minet, Wyers, et al. 1997). *In vitro* binding studies have shown that recombinant Pcf11 can interact with the CTD of pol II (Barilla, Lee, and Proudfoot 2001), but a later study from Keller and colleagues showed that the CTD binding of Pcf11 and its participation in mRNA 3' end processing are not coupled (Sadowski et al. 2003). The N-terminus of Pcf11 contains a pol II CTD binding region, the central region harbours the Rna14/Rna15 binding site and the C-terminus contains a Clp1 binding site followed by a CCHC zinc finger domain (Guéguéniat et al. 2017). **Clp1** was identified as one of the co-purification partners of CF IA that interacts tightly with Pcf11 but not with Rna14 or Rna15 (Minvielle-Sebastia et al. 1997; Gordon et al. 2011). CF IA is thought to activate pre-mRNA cleavage by CPF by mechanism that remains to be understood.

1.2.2.2.2 Cleavage Factor IB

CF IB or **Hrp1** was discovered as a suppressor of a temperature sensitive allele of the gene *NPL3* involved in mRNA export (Henry et al. 1996) UV crosslinking experiments demonstrated sequence specific binding of Hrp1 to RNA (Kessler et al. 1997; Chen and Hyman 1998), and overexpression of Hrp1 influences the cleavage site selection in yeast (Bucheli et al. 2007). Hrp1 contains a central RRM domain that mediates its interaction with Rna14. Mutants that disrupt this interaction failed to cleave and polyadenylate a substrate RNA (Barnwal et al. 2012). The same study also showed that Hrp1 binding to RNA via its RRM domains enhances its interaction with Rna14/Rna15, therefore suggesting that RNA binding may trigger the assembly of the 3' end-processing complex.

1.2.2.3 Cis-regulatory elements

The mRNA signals in budding yeast and humans that recruits and directs the 3' end processing machinery for cleavage/polyadenylation have similarities and differences. The sequences of the yeast signals and the order in which they are distributed in the 3' UTR are somewhat different when compared to human signals. *In vivo* studies on *ACT1*, *ADH1*, *CYC1*, and *YPT1* mRNAs found that the signals that direct polyadenylation site selection are dispersed across several hundred nucleotide stretches and it was unlikely that there existed a single polyadenylation signal sequence in yeast (Heidmann et al. 1992). However many human mRNA 3' UTRs contain a conserved six nucleotide AAUAAA sequence (or a close variant such as AUUAAA) known as the polyadenylation signal and is present 10 - 30 nucleotides upstream of the cleavage site (N. J. Proudfoot and Brownlee 1976; Beaudoing et al. 2000; B. Tian et al. 2005; Derti et al. 2012).

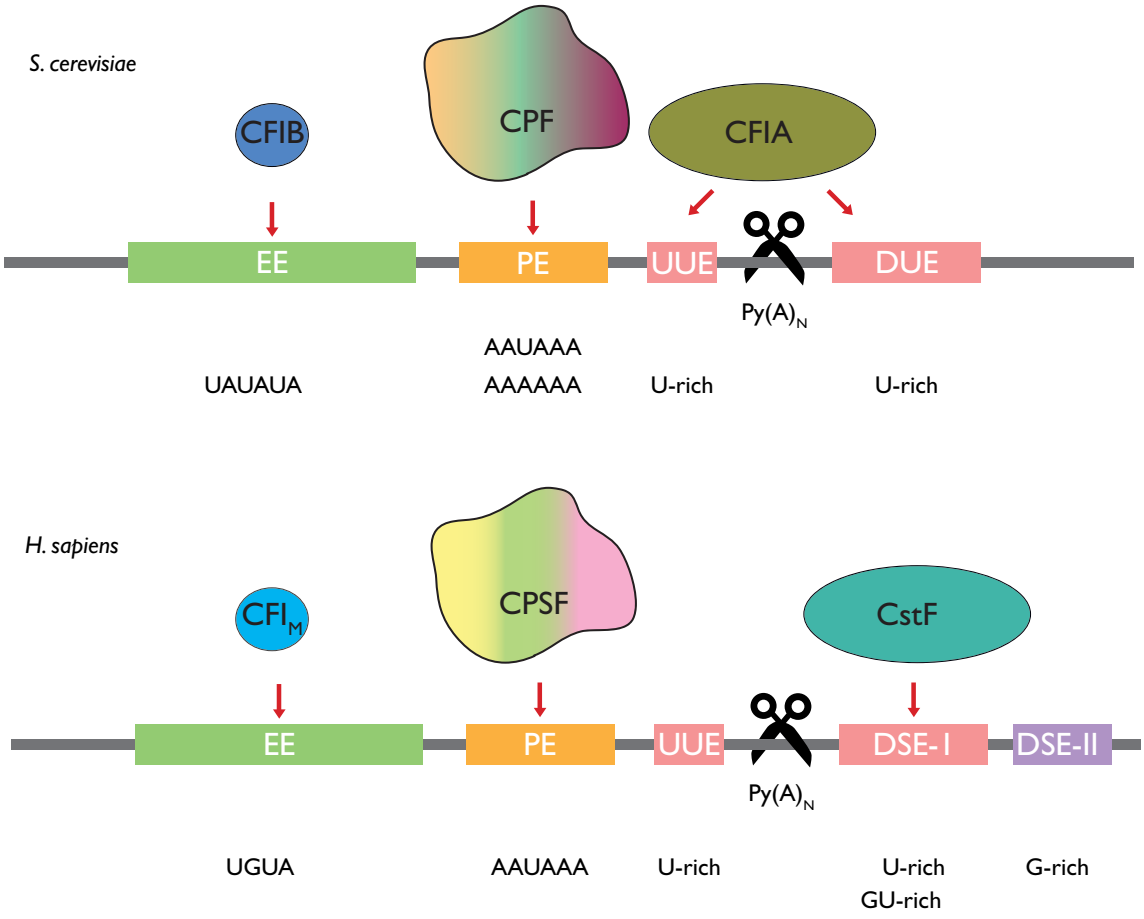


Figure 1.5: The *cis*-elements of yeast and human 3' UTR**.** EE: Efficiency element, PE: Positioning element, UUE: upstream U-rich element, DSE: Downstream U-rich element. Red arrows indicate the protein binding sites. $Py(A)_N$ is the site of cleavage. Py stands for pyrimidine.

There are five predominant motifs in yeast that constitute the *cis*-regulatory elements: the UA rich efficiency element, the A rich positioning element, the U-rich upstream element, the $Py(A)_N$ cleavage site and the downstream U rich element (Figure 1.5). In humans, biochemical studies coupled with bioinformatics have shown that the sequences downstream of the human cleavage sites are rich in U/GU sequences (Salisbury, Hutchison, and Graber 2006). The sequences upstream of AAUAAA are rich in multiple UGUA elements, two of which are specifically and simultaneously recognised by hCFI (Hu et al. 2005; Yang, Gilmartin, and Doublié 2010). Auxiliary sequences elements including the upstream U-rich sequences and the downstream G-rich regions provide protein-binding surfaces and

contribute to efficient mRNA 3'- end processing (Hu et al. 2005; Yang and Doublié 2011) (Figure 1.5).

The efficiency element consists of alternating UA dinucleotides in some genes or a mixture of U-rich and UA dinucleotide stretches in others. A consensus efficiency element was found to be a hexanucleotide UAYRUA and more specifically a UAUAUA. The Us in position one and five of this sequence are the most essential residues for mRNA 3' end formation (Irniger and Braus 1994). It is interesting to note that the efficiency elements are somewhat degenerate in different mRNAs in spite of the proposed UAUAUA consensus sequence. *GCN4* and *ADH1* use UUUUAUA as their efficiency elements (Egli, Springer, and Braus 1995). *GCN4* represent a class of mRNAs where the polyadenylation signals are strictly uni directional whereas *ADH1* polyadenylation sites are bi directional in nature (Irniger, Egli, and Braus 1991). Most of the yeast genes including *GAL7* use UAUAUA as the efficiency element. However genes such as *CYC1* and *GCN4* use different but related sequences like UAUUA, UUUAUA, UAUGUU, and UAUUUA (Egli, Springer, and Braus 1995; Guo et al. 1995). The yeast positioning elements are the ones functionally similar to the human AAUAAA. They are usually A rich stretches that dictate the position of the endonucleolytic cleavage and are located around 10 to 30 nucleotides upstream of the exact cleavage site (Heidmann et al. 1994). Removal of the positioning element from the RNA results in shift of the cleavage site (P. Russo et al. 1991). Some of the most commonly found positioning elements are UUAAGAAC, AAUAAA and AAAAAA. Deletion of the positioning element of *ADH2* and *GAL7* resulted in a decrease in the use of the strong cleavage sites (Hyman et al. 1991; Abe, Hiraoka, and Fukasawa 1990). Analysis of 1352 unique pre-mRNA 3′-end-processing sites have identified U-rich sequences both upstream and downstream of the cleavage site (Graber et al. 1999). Importantly, the human polyadenylation signals also contain similar U-rich elements near the cleavage site. Mutating either the upstream or the downstream U-rich elements to Gs did not have any effect on the cleavage activity. However, when U → G mutations were introduced upstream and downstream of the cleavage site, cleavage was strongly affected (Dichtl and Keller 2001). The interaction of these U-rich elements with the CPF or CFI subunits is yet to be studied in detail. Yeast possesses a cluster of cleavage sites downstream of the positioning element. The most common cleavage site is a pyrimidine (C or U) followed by one or two As (PyAAA), (Heidmann et al. 1992; Bennetzen and Hall 1982). Tian and Graber

provide a comprehensive review comparing the poly(A) signals in different organisms (Bin Tian and Graber 2012).

A recent study combining in cell structure probing coupled with high-throughput functional assays and bioinformatics analysis, have found that folded RNA structures near the mRNA 3′-ends facilitate cleavage and polyadenylation (Wu and Bartel 2017). This adds a new layer of regulation associated with mRNA 3′ end processing, and suggest that the contributions from RNA elements are more complex than previously thought.

1.2.3 Ysh1 and endonuclease mechanism

Ysh1 (CPSF-73 in humans) follows a general two-metal ion dependent mechanism for the endonuclease cleavage reaction (Steitz and Steitz 1993). The crystal structure of CPSF-73 in a closed state revealed an N-terminal metallo-β-lactamase domain and a C-terminal β-CASP domain (Mandel et al. 2006). The metallo-β-lactamase domain is not only found in β-lactamases but also prevalent in different protein families such as thiolesterases, glyoxalase II family and nucleases. The crystal structure reveals two zinc ions bound in the active site between the two domains. The ions have an octahedral coordination with a hydroxyl ion as one of the bridging molecules. Although this structure does not have a substrate RNA bound in the active site, conclusions can be made about the possible mechanism via which cleavage can occur. In the metallo-β-lactamase domain of CPSF-73, the bridging molecule of the two zinc ions acts as a nucleophile to hydrolyse the phosphodiester linkage (Mandel et al. 2006). The bridging hydroxyl molecule is right beneath the sulfate group that the authors argue is the chemical mimic of the phosphate group of the RNA. In a similar study of the RnaseJ enzyme, the position occupied by a sulfate ion in the *apo* structure of RnaseJ was found to be occupied by the phosphate group of an UMP in the structure bound to UMP (Sierra-Gallay et al. 2008). So the position where they have modelled the sulfate group (or the phosphate group) could be the receiver of the nucleophilic attack leading to nuclease reaction (Figure 1.6). Direct proof of the postulated two metal ion endonuclease mechanism will require a structure of the enzyme in an open state, with an RNA substrate bound in the active site, and two divalent ions coordinated octahedrally by the phosphate group and the side chains of Asp or His from Ysh1.

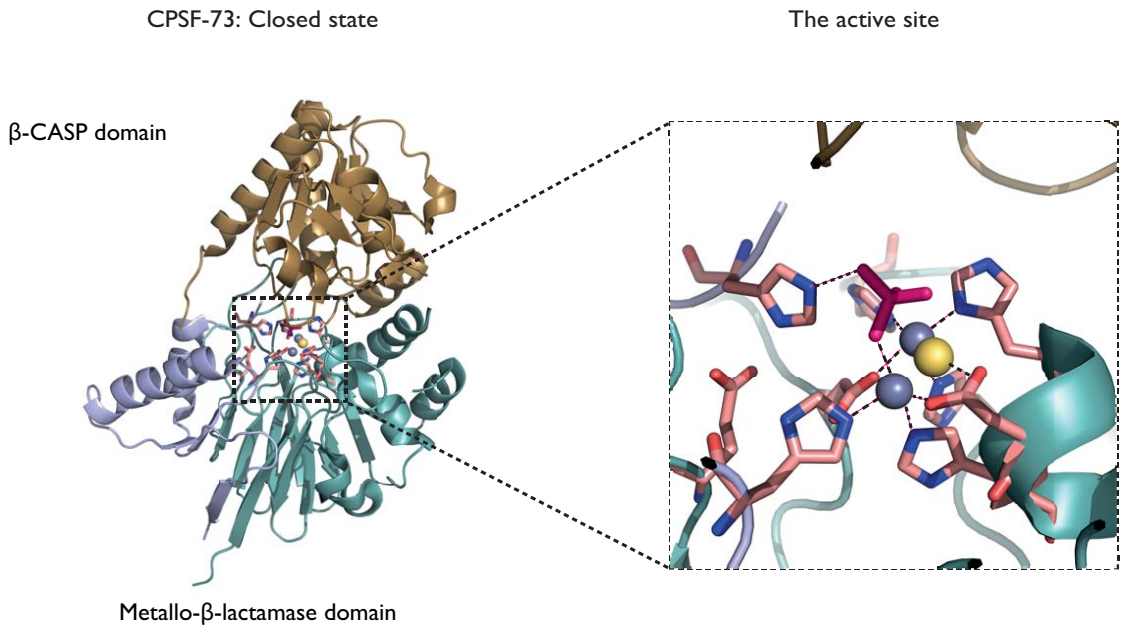


Figure 1.6: The structure of CPSF-73 (or Ysh1) in a closed state [PDB ID: 2I7V]. CPSF-73 contains an N-terminal metallo-β-lactamase domain shown in light cyan, the middle region forms a β-CASP domain shown in sand yellow and the C-terminal end also forms a part of the metallo-β-lactamases domain shown in light blue. The two bound zinc ions are shown in grey, a sulfate group (that mimics the phosphate of the substrate) is shown in pink and the bridging hydroxyl ion is shown in yellow.

1.2.4 Pap1 and polyadenylation mechanism

Pap1 is the yeast poly(A) polymerase that adds a chain of adenosines to the 3′ hydroxyl of a substrate RNA via a two-metal ion dependent nucleotidyl transferase mechanism. The crystal structure of Pap1 revealed three globular domains that surround a central substrate binding cleft of 20 Å by 30 Å by 45 Å (Figure 1.7a) (Bard et al. 2000). The three domains of Pap1 consist of a palm, finger and thumb domains. Furthermore, the arrangement of these respective domains in Pap1 is different from that seen in other template dependent polymerases (Figure 1.7a). The N-terminus of Pap1 is structurally similar to the palm domains, and the middle domain of Pap1 is functionally similar to the finger domains of pol β nucleotidyl transferases. The active site in the palm domain of Pap1 is structurally similar to the other polymerases in this family of two metal ion dependent nucleotidyl transferases. The Aspartates 100, 102 and 154 are required to coordinate the divalent cations (Figure 1.7b). The active site also shows that there are extensive contacts made by Pap1 to the mononucleotide primer.

The mechanism of poly(A) polymerase activity was determined when the structure of a catalytically inactive Pap1 mutant (D154A) was determined in complex with MgATP and a poly(A) RNA (Balbo and Bohm 2007). This 1.8 Å resolution crystal structure of the ternary complex has clear electron density for MgATP and the RNA (Figure 1.7b). However, there is no electron density accounting for a second magnesium ion, likely due to the catalytically inactive mutant used in the study.

When compared to the *apo* open structure of Pap1, this active structure is more closed with the N- terminal domain being rotated by ~23°. In this closed state, the N-terminal and C-terminal domains interact with the substrate mainly mediating contacts across the domains. Interestingly, the single stranded poly(A) chain does not form base stacking interactions. The poly(A) substrate remained solvated in the ternary structure with 24 of the protein-RNA interactions occurring through an ordered network of water molecules. Some of the water molecules that contact the base of ATP could impart substrate specificity.

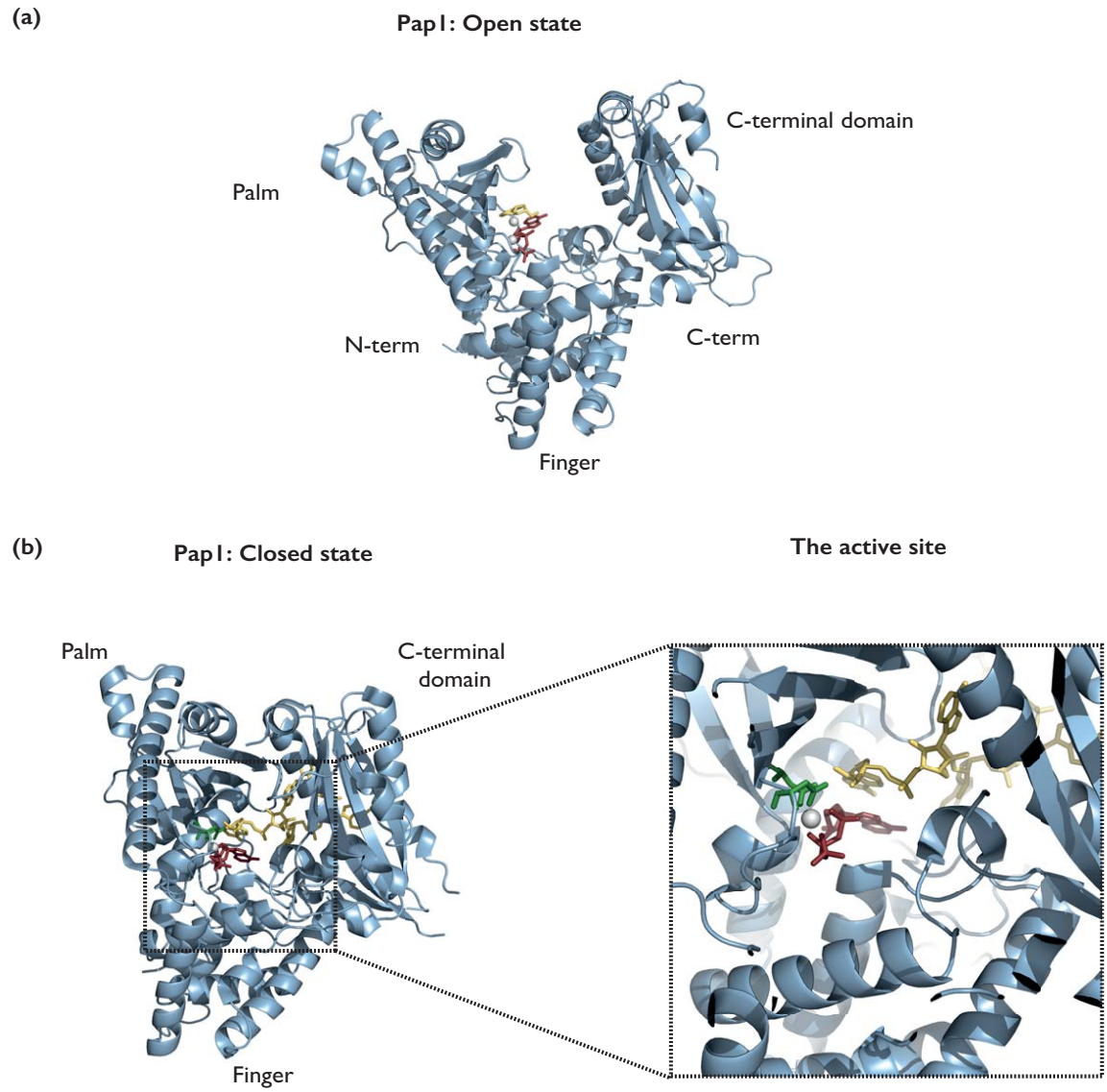


Figure 1.7: (a) The structure of Pap1 with MgATP in the open conformation [PDB ID:1FA0]. The two divalent cations are shown in grey, the phosphates of the incoming 3'-dATP is shown in ruby and the mononucleotide primer is shown in yellow. The active site cleft is ~ 20 Å by 30 Å by 45 Å. **(b)** The active site of Pap1 showing the Mg coordination and the incoming ATP in ruby poised for a nucleophilic attack on the 3' hydroxyl of the substrate RNA in yellow [PDB ID: 2Q66].

Pap1 interacts with the 2' hydroxyl groups of the adenine bases at the 3' end of the poly(A) RNA, one and two positions upstream of the 3' end. This explains why Pap1 does not prefer DNA substrates as templates for polyadenylation. The adenines at these positions are well anchored and buried completely within the active site. However, the adenine residue three positions before the 3' end remains surface exposed on the C-terminal domain of Pap1. This adenine makes base specific contacts with Pap1 residues Lys392, Glu487, Leu388 and Leu 491. The three adenosines coming after this also makes base specific contacts with highly conserved residues on Pap1.

Resembling a substrate ready for a nucleophilic attack, the 3' hydroxyl end of the poly(A) RNA is located 3.2 Å from the α phosphorous of ATP. The ternary structure provides a structural basis for nucleotide base recognition in spite of the absence of any observed direct hydrogen bonding between the base and Pap1. Base stacking interactions between the 3' end of the poly(A) RNA substrate and the incoming ATP can be observed in the structure. From a thermodynamic point of view, A-A stacking is more favourable than A-C or A-U stacking. This could explain why Pap1 is less efficient to add a poly(A) tail to an RNA 3' end that ends with C or U. The discrimination between A and G is largely due to electrostatic reasons and steric hindrances due to differences in shape (Balbo and Bohm 2007).

Detailed kinetic studies show that Pap1 switches between the open and closed conformation during nucleotidyl transfer. Formation of the closed state results in the active site assembly, and subsequently correct substrate recognition and catalysis. Interestingly, the nucleotide specificity is manifested in the V_{max} term rather than the K_m. ATP binding promotes the reaction velocity via a ground-state destabilisation mechanism. Substrate binding results in domain closure and adenylate transfer occurs in the closed conformational state. The mechanism of ground state destabilisation and catalysis in the closed state remains unanswered in the field.

1.2.5 Regulation of poly(A) tail length

It is thought that a newly synthesised RNA has a poly(A) tail length of ~60 As in *S. cerevisiae* (M Edmonds, Vaughan, and Nakazato 1971; Mclaughlin et al. 1973). Why is the length of a poly(A) tail an important cellular property that deserves attention? Depending on their lengths, poly(A) tails can either help in stabilizing the mRNA or aiding in 3' → 5' exosome mediated RNA degradation. The rate of removal of a poly(A) tail is the major determinant of mRNA degradation and subsequently it's half life. The newly added poly(A) tail of different mRNAs have similar lengths, whereas the rate of removal of the poly(A) tail in the cytoplasm is different for different mRNAs. Specific RNAs are targeted for deadenylation over others based in part on sequences in their 3' UTRs. This allows the cell to control the half-life of an mRNA as a function of the intrinsic property of the mRNA only. In addition, mRNA poly(A) tail length also has an influence on translation efficiency (Preiss, Muckenthaler, and Hentze 1998; Beilharz and Preiss 2007). Therefore the length of the poly(A) tail controls the fate of the mRNA and hence gene expression.

As CPF adds a poly(A) tail to the cleaved 3' end of a pre-mRNA, poly(A) binding proteins in the nucleus are thought to bind the poly(A) tail and somehow restrict their length. The zinc finger protein, Nab2, is the major poly(A) binding protein in the nucleus, although there are reports suggesting the presence of the cytoplasmic poly(A) binding protein Pab1 in the nucleus (Brune et al. 2005). Nab2 was discovered as a protein that associates with nuclear polyadenylated RNA *in vivo* (Anderson et al. 1993). Nab2 is also required for the efficient export of mRNAs into the cytoplasm and in addition to its role in the control of poly(A) tail length *in vivo* (Green et al. 2002; Hector et al. 2002). Although Pab1 may have a role in the regulation of poly(A) tail length in yeast (Amrani, Minet, Le Gouar, et al. 1997), Nab2 is thought to be the functionally relevant protein in poly(A) tail length control owing to its nuclear localization. However the exact amounts of Nab2 or Pab1 inside the nucleus is unknown. It was found that over-expression of Pab1 with a nuclear localisation signal cannot rescue a hyper-polyadenylation phenotype of Nab2 deletion (Hector et al. 2002).

Nab2 binds specifically to poly(A) sequences with high affinity (Kelly et al. 2007). Nab2 is made up of three functionally distinct domains. The N-terminal domain of

Nab2 is involved in poly(A) RNA export, a central domain mediates Nab2's nuclear import and the C-terminal domain harbours seven tandem zinc finger repeats, of which ZF5-ZF7 are necessary for specific poly(A) RNA binding (Soucek, Corbett, and Fasken 2012). Recognition of poly(A) RNA by Nab2 is essential for proper control of mRNA poly(A) tail length (Kelly et al. 2010), and rapid removal of Nab2 from the nucleus of yeast cells results in disappearance of polyadenylated transcripts due to enhanced mRNA decay (Schmid et al. 2015). In accordance with the *in vivo* data, addition of recombinant Nab2 to *in vitro* polyadenylation reactions of CPF results in controlled poly(A) tail addition, but the exact molecular mechanism behind how Nab2 can impart length control to the polyadenylation machinery of *S. cerevisiae* is not understood (Viphakone, Voisinet-Hakil, and Minvielle-Sebastia 2008).

In humans however, there is a model for the control of mRNA poly(A) tail length (Wahle 1995). It is based on the fully distributive property of the human poly(A) polymerase, PAP, which means it has a low affinity for RNA. The nuclear poly(A) binding protein PABN1 forms a stable ternary complex with the PAP, the human polyadenylation machinery CPSF and the substrate RNA. This transforms the polyadenylation activity of PAP from a distributive to a processive manner. After adding a poly(A) tail of ~250 As, for reasons that are currently unknown, PABN1 dislodges from this stable ternary complex resulting in the termination of this processive poly(A) tail synthesis. In cells, the mRNA polyadenylation complex consists of additional factors such as CstF, CFI and CFII that were not included in the abovementioned study. It remains to be determined whether they influence this process.

1.3 Overview of 3' end processing: links to other pathways

Maturation of pre-mRNAs transcribed by pol II requires capping at the 5' end, splicing at intronic sites, 3' end cleavage / polyadenylation and export into the cytoplasm. Although these diverse nuclear pathways function via separate protein and nucleic acid complexes, there exists a great deal of cross- talk between these processes (Tudek, Lloret-Llinares, and Jensen 2018; Giammartino and Manley 2014).

The CTD of pol II can recruit a variety of protein factors that are involved in pre-mRNA processing. The dynamic phosphorylation states of the CTD regulate this process. When transcription begins, the CTD is unphosphorylated, and becomes differentially phosphorylated throughout the different stages of transcription. We know, for example, that Ser5 phosphorylation recruits the 5' capping complex and that Ser2 phosphorylation results in assembly of the 3' end processing machinery. The Pcf11 subunit of CF IA, contains a CTD interaction domain that specifically recognises Ser2 phosphorylated CTD. It is plausible that CF IA is first recruited to the CTD through its interaction with Ser2 phosphorylation, and this brings CPF to the CTD. The Glc7 phosphatase of CPF can then dephosphorylate the Tyr1 of CTD and triggers transcription termination (Schreieck et al. 2014). As the pre-mRNA emerges out of the pol II, the various *cis*-acting elements in the RNA start to attract the binding of different proteins, eventually assembling a competent 3' end processing machinery. The CPF and CFI protein components interact both with the pol II CTD and the newly synthesized pre-mRNA. The RRM domains of Rna15 (CF IA) can recognise GU rich elements that are found downstream of the cleavage site (Pancevac et al. 2010) as well as the AAUAA of the positioning element, whereas the RRM domains of Hrp1 (CF IB) recognises AU rich efficiency element upstream of the cleavage site (Leeper et al. 2010). An open question that remains in the field is whether CF IA and CF IB bind upstream or downstream of the cleavage sites (Gross and Moore 2001; Dichtl et al. 2002; Pancevac et al. 2010; Baejen et al. 2014). CPF likely binds to the AU rich positioning element usually present ~ 20 nucleotides upstream of the cleavage site (Chan, Huppertz, Yao, Weng, Moresco, Yates, et al. 2014). After the assembly of a competent 3' end processing complex involving CF IA, CF IB and CPF, the endonuclease Ysh1 is activated and performs cleavage of the pre-mRNA at a specific site Surprisingly, some of the spliceosomal proteins can exert influence on pre-mRNA 3' end cleavage and

polyadenylation. Binding of the splicing factor U2AF 65 to the pyrimidine tract of the intron 3' site that lies upstream of the cleavage site enhances the 3' end cleavage activity (Millevoi et al. 2009). Similarly, PAP interacts directly with U2AF 65 and increases splicing efficiency (Vagner, Vagner, and Mattaj 2000). A direct interaction of the U1A snRNP with PAP has been reported to inhibit polyadenylation (Gunderson et al. 1994). However, *in vitro* assays show that U1A snRNP can stimulate polyadenylation by interacting directly with CPSF-160 (Lutz and Alwine 1994). The exact role of U1A snRNP in polyadenylation remains to be clarified.

Once the activated Ysh1 in the 3' end processing machinery carries out the pre-mRNA cleavage, it is uncertain whether there could be any conformational rearrangement of the machinery. The poly(A) polymerase Pap1 can now access the free 3' hydroxyl of the end of the pre-mRNA and start adding a tail of multiple adenosines. After reaching a tail length of ~60As, the polyadenylation reaction is terminated by a mechanism involving the poly(A) binding protein Nab2. The Nab2-bound mature (spliced and polyadenylated) mRNA is now competent for export into the cytoplasm. The Nab2-bound poly(A) tail can recruit the mRNA export adaptor complex Mex67/Mtr2 (Iglesias et al. 2010). Mex67 and Nab2 also interact directly with the nuclear pore complex and aid in the efficient export of the mature mRNA (Fasken, Stewart, and Corbett 2008). Those mRNAs that are not exported into the cytoplasm are targeted for degradation by the nuclear exosome. Studies in humans have identified a complex named PAXT that can selectively bind polyadenylated mature transcripts via PABN1 and perform exosome-mediated decay. This provides a pathway through which polyadenylated RNAs are targeted for degradation in the nucleus. Those RNAs that do not have a mature poly(A) tail, are targeted by the TRAMP complex that adds a short tail of As and subsequently triggers their exosome-mediated decay (Schmidt and Butler 2013) .

In summary, CPF (along with the cleavage factors) cleaves and adds a poly(A) tail at the 3' end of a pre-mRNA. This is a multistep process involving a large set of protein and RNA factors, and is regulated by cross talk between various other mRNA processing complexes including Pol II, spliceosome and nuclear export factors. The poly(A) tails can either stabilize the mature mRNA, aid in their export or alternatively result in nuclear exosome mediated decay or deadenylation. Research in the past two decades has aided in our understanding of the composition and functions of the

individual components of this machinery. However, very little is known about how they work together as a complex to coordinate specific and efficient mRNA 3' end processing.

1.4 Questions addressed in this dissertation

Prior work from our lab has advanced our knowledge of the architecture of CPF; the three-enzymatic activities are segregated into three different modules (Figure 1.3). Isolated Pap1 enzyme can add a multiple adenosine tail to a substrate RNA in a nonspecific manner. It has been shown in humans that the non-enzymatic subunits that associate with PAP confer specificity to the reaction. Furthermore, they transform the polyadenylation reaction from distributive to processive in nature. In chapter 2 of this dissertation, I aim to clarify the function of the non-enzymatic polymerase module subunits in yeast and provide a better understanding of how they are structurally associated with each other including the substrate RNA. I have taken an *in vitro* biochemistry approach coupled with integrative structural biology techniques.

The cleavage factors CF IA and CF IB are known to be important for proper cleavage site selection and influence the efficiency of cleavage (Minvielle-Sebastia, Preker, and Keller 1994; Kessler, Zhao, and Moore 1996; Kessler et al. 1997). Although there are prior reports concerning the role of the cleavage factors in polyadenylation, the lack of a highly pure and homogeneous *in vitro* system to study polyadenylation has hampered progress. I have used purified protein complexes to study the polyadenylation of a pre-cleaved substrate RNA *in vitro* in order to systematically test the role of specific components during the reaction. In addition, I have employed complimentary biophysical techniques to characterize the interaction between the cleavage factors and CPF. The main findings from this study are in chapter 3.

In *S. cerevisiae*, newly synthesized pre-mRNAs have a poly(A) tail length of ~ 60 As. The exact molecular mechanism behind how CPF terminates polyadenylation after 60 As remains unknown. In humans, the poly(A) binding protein PABN1 has been shown to play a role in regulating the pre-mRNA poly(A) tail lengths. *S. cerevisiae* does not have any PABN1 homologue. The tandem zinc finger-containing protein, Nab2, is thought to impart poly(A) tail length control on pre-mRNAs. By using a fourteen-subunit CPF purified from both native and recombinant sources, I have performed cleavage and polyadenylation reaction of a substrate RNA to

investigate pre-mRNA poly(A) tail length control. These results are presented in chapter 4.

2 Architecture of the Polymerase Module of CPF

The identities and functional properties of the subunits that make up the polymerase module are relatively well understood. The polymerase module in yeast is strikingly similar to a four-subunit mammalian complex that has been shown to be necessary and sufficient for specific *in vitro* polyadenylation (Schönemann et al. 2014). However, it is unknown how the five proteins constituting the module function together to facilitate polyadenylation of a cleaved substrate RNA. A comprehensive understanding of subunit interactions within the complex has remained elusive due to the complexity involved in producing the polymerase module with high yields and purity for biochemical or biophysical experiments. Furthermore, there is a lack of insight into how the polymerase module subunits are arranged in 3D space and how this gives rise to the function of the module. In this chapter, I describe a method to stringently purify high quantities of the polymerase module. I then study a purified, homogenous preparation using cryo-EM, and determine the 3D structure of the Cft1-Pfs2-Yth1 subunits of the polymerase module. Using cross-linking mass spectrometry, *in vitro* pull downs and nuclear magnetic resonance (NMR) spectroscopy, I further elucidate the molecular architecture of the polymerase module, showing how the polymerase Pap1 is flexibly tethered to the rest of the complex.

2.1 Purification of recombinant polymerase module from *Sf9* cells

In order to purify the polymerase module in the correct stoichiometry, polymerase module subunits were co-expressed and co-purified. Five genes encoding subunits of the polymerase module were cloned into one single vector using the MultiBac expression system (Figure 2.1a). Each gene had its own polyhedrin promoter and SV40 late terminator. The genes coding for Cft1, Pfs2, and Yth1 were cloned into the acceptor plasmid pACEBac1, whereas the genes coding for Pap1 and Fip1 were cloned into the donor plasmid pIDC. By means of Cre-Lox recombination, a final vector containing all polymerase module subunits was generated (section 6.2.4). The subunit Pfs2 contains an N-terminal twin StrepII-tag and the subunit Yth1 contains an N-terminal His tag. Polymerase module subunits were over-expressed in *Sf9* cells by means of baculovirus mediated over-expression system (section 6.3.3).

2.1.1 Purification of isoforms of polymerase module that vary in Pap1 stoichiometry

The polymerase module was purified by a three-step protocol (Figure 2.1b). The polymerase module was pulled down from *Sf9* lysates via the Strep-II tag on the N-terminus of Pfs2 using streptavidin resin, and eluted with desthiobiotin (Figure 2.1c). Analysis of the pull-downs by SDS-PAGE revealed a complex containing all five polymerase module subunits and additional bands whose identities were unknown (Marked by asterisks in Figure 2.1c).

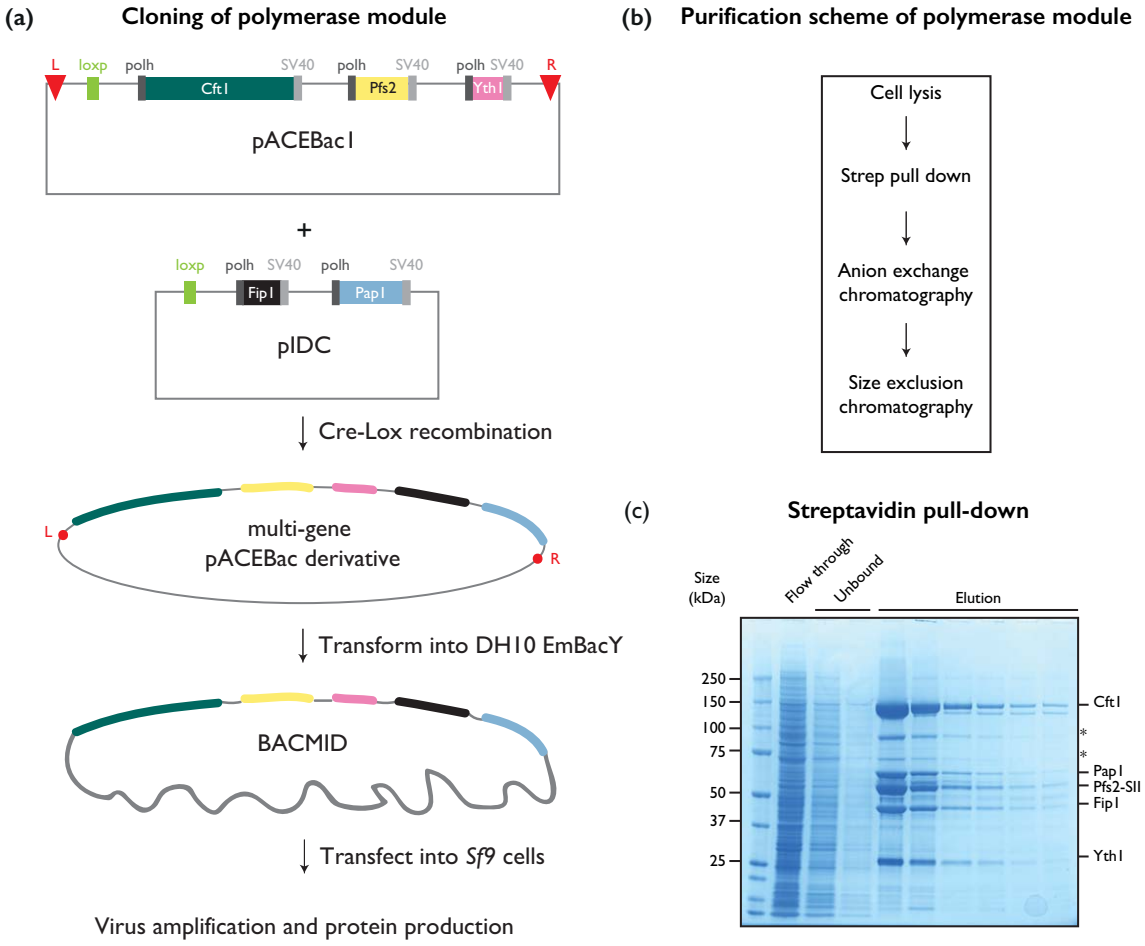


Figure 2.1: Purification of recombinant polymerase module. **(a)** Overview of steps involved in cloning of polymerase module subunits and baculovirus mediated insect cell over-expression **(b)** Purification scheme of polymerase module **(c)** Streptavidin pull down of the Sf9 cell lysate was analyzed by SDS-PAGE.

Furthermore, visual inspection of the bands revealed that the intensity of Pap1 was fainter compared to that of Pfs2. This suggested that Pap1 was either present in sub-stoichiometric amounts, or that the elution fractions from affinity purification contain multiple protein complex species. In order to further clarify this, the elution fractions were pooled and loaded onto a pre-equilibrated anion exchange column. Bound protein complexes were eluted using a very shallow gradient of increasing salt concentration. The shallow gradient allowed the resolution of three peaks at ~300 mM, ~350 mM, and ~380 mM NaCl (Figure 2.2a). The peaks are referred to as "shoulder", "peak 1" and "peak 2" respectively. SDS-PAGE analysis of the three peaks showed highly pure polymerase module complex being eluted across the fractions. Further analysis reveals that the stoichiometry of Pap1 subunit varies across the elution fractions (highlighted in a red box in Figure 2.2b). The fractions at ~300 mM NaCl seem to contain excess Pap1, and as the salt concentration is increased, complexes with stoichiometric amounts of Pap1 are eluted (~350 mM NaCl), followed by complexes containing no Pap1 (~380 mM NaCl) (Figure 2.2b).

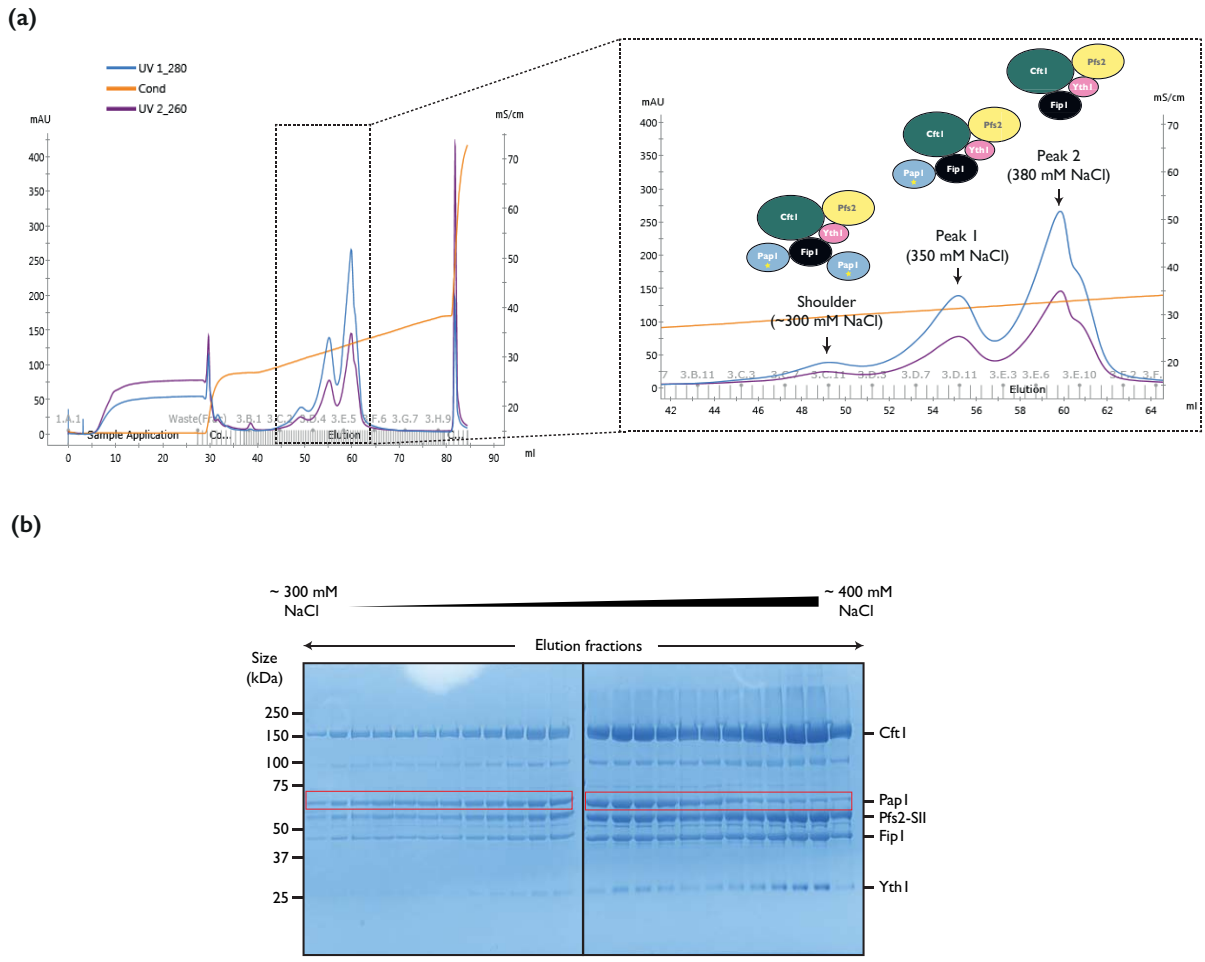


Figure 2.2: Anion exchange chromatography separates polymerase module with multiple, one or no copy of Pap1 **(a)** Anion exchange chromatography using a monoQ column. Proteins are eluted over a shallow salt gradient starting from 150 mM NaCl until 1000 mM NaCl over a 100 ml volume. **(b)** SDS-PAGE analysis of the eluted fractions reveals polymerase module containing multiple Pap1, single Pap1 or no Pap1. Varying amounts of Pap1 is highlighted by a red box.

The presence of Pap1 likely alters the ionic property of the complex therefore leading to different elution profiles for the complexes with Pap1 and without Pap1. The observation of polymerase modules with varied stoichiometry of Pap1 is highly reminiscent of earlier observations during native CPF purifications from *S. cerevisiae*. CPF species containing multiple copies of Pap1 were observed when CPF from *S. cerevisiae* is subjected to anion exchange chromatography and a very shallow salt gradient elution (Easter 2014). It is interesting to note that CPF without any polymerase was eluted first at ~280 mM salt concentration, followed by CPF containing

one copy of Pap1 and CPF with two copies of polymerase as the salt concentration is increased (Easter 2014). However, in the case of the polymerase module, the complex containing excess Pap1 subunit was eluted first, followed by the other species (Figure 2.2b). It is possible that the absence of the nuclease and phosphatase modules alter the surface chemistry of the polymerase module in comparison with full CPF, thus explaining the different elution profiles of CPF compared to polymerase module on an anion exchange column.

The three different polymerase module species from the anion exchange step were further purified using size exclusion chromatography. The size exclusion chromatograms show a well-dispersed and homogenous peak for all three species (Figure 2.3a).

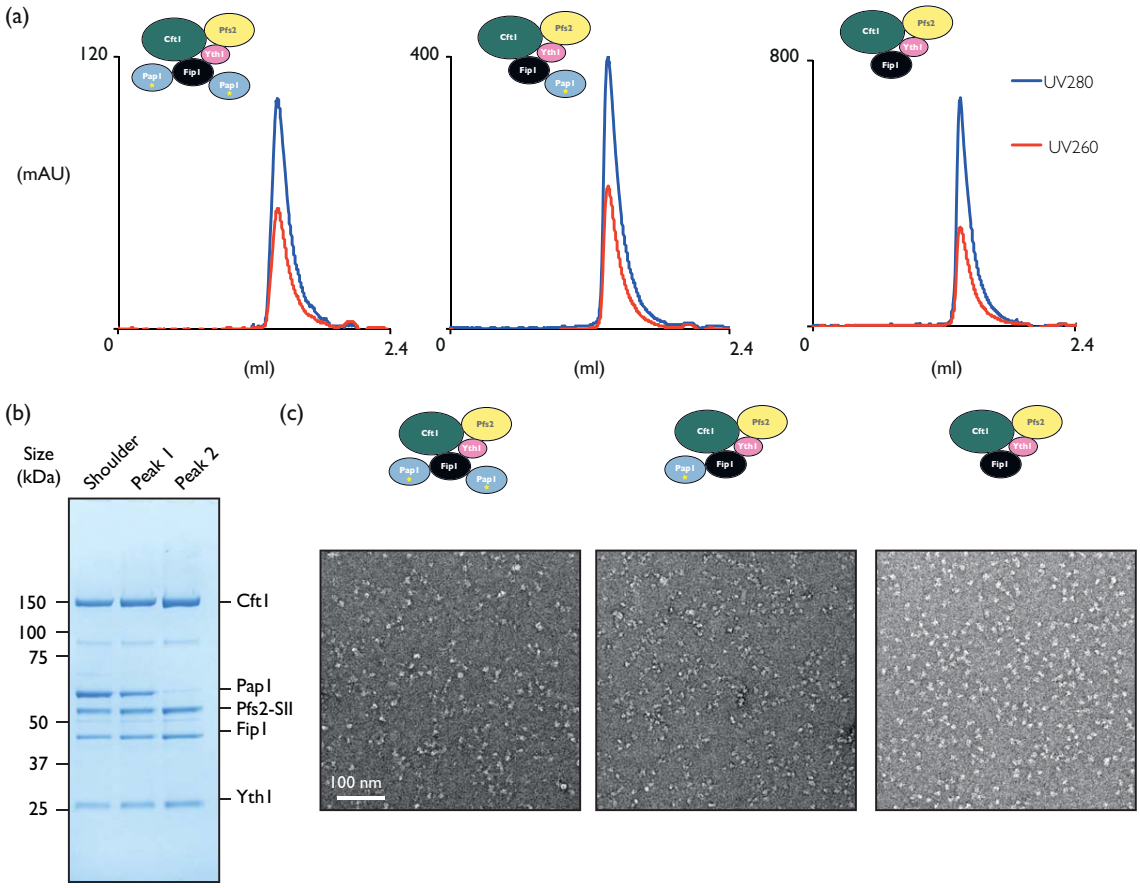


Figure 2.3: Size exclusion chromatography of the three polymerase module variants **(a)** Size exclusion chromatograms of polymerase module variants. **(b)** SDS-PAGE analysis of the polymerase module variants after size exclusion chromatography. **(c)** Representative negative stain micrographs of the three polymerase module variants.

The ratio of UV absorption at 260 nm and 280 nm suggests that there are no nucleic acid contaminants. SDS-PAGE analysis of the peak fractions further reveals highly pure polymerase module complexes (Figure 2.3b). Transmission electron microscopy images of negatively stained polymerase module further confirm sample homogeneity (Figure 2.3c). The micrographs reveal well distributed ~14 nm particles. However, micrographs of samples containing Pap1 showed smaller particles in addition to the ~14 nm polymerase module. One likely explanation is that the polymerase Pap1 dissociates from the complex during preparation of negatively stained sample. The absence of such additional small particles in the micrographs of polymerase module without Pap1 (peak 2) further supports this hypothesis.

In summary, I have established a protocol for the recombinant expression and purification of the *S. cerevisiae* polymerase module.

2.1.2 Purification of the poly(A) polymerase Pap1

In order to compare the activity of the polymerase module and the poly(A) polymerase Pap1 alone, I sought to recombinantly express and purify *S. cerevisiae* Pap1, a well-folded globular protein which has been previously purified and crystallized (Bard et al. 2000). The gene coding for *S. cerevisiae* Pap1 was cloned with a N-terminal His tag into pET-28a vector. Protein overexpression was carried out in BL21 Star (DE3) cells by IPTG induction (Figure 2.4a). The overall purification scheme is shown in Figure 2.4b Pap1 was isolated from *E. coli* lysate by affinity chromatography via the His-tag of Pap1. The eluate was subjected to ion exchange chromatography using a heparin column. The eluate from the heparin column was further purified using size exclusion chromatography (Figure 2.4c). The size exclusion chromatography revealed a void peak and a homogenous protein peak. SDS-PAGE analysis of the fractions showed the presence of highly pure Pap1 in the elution fractions. The peak fraction was further concentrated, flash frozen and stored in -80 C for use in biochemical experiments.

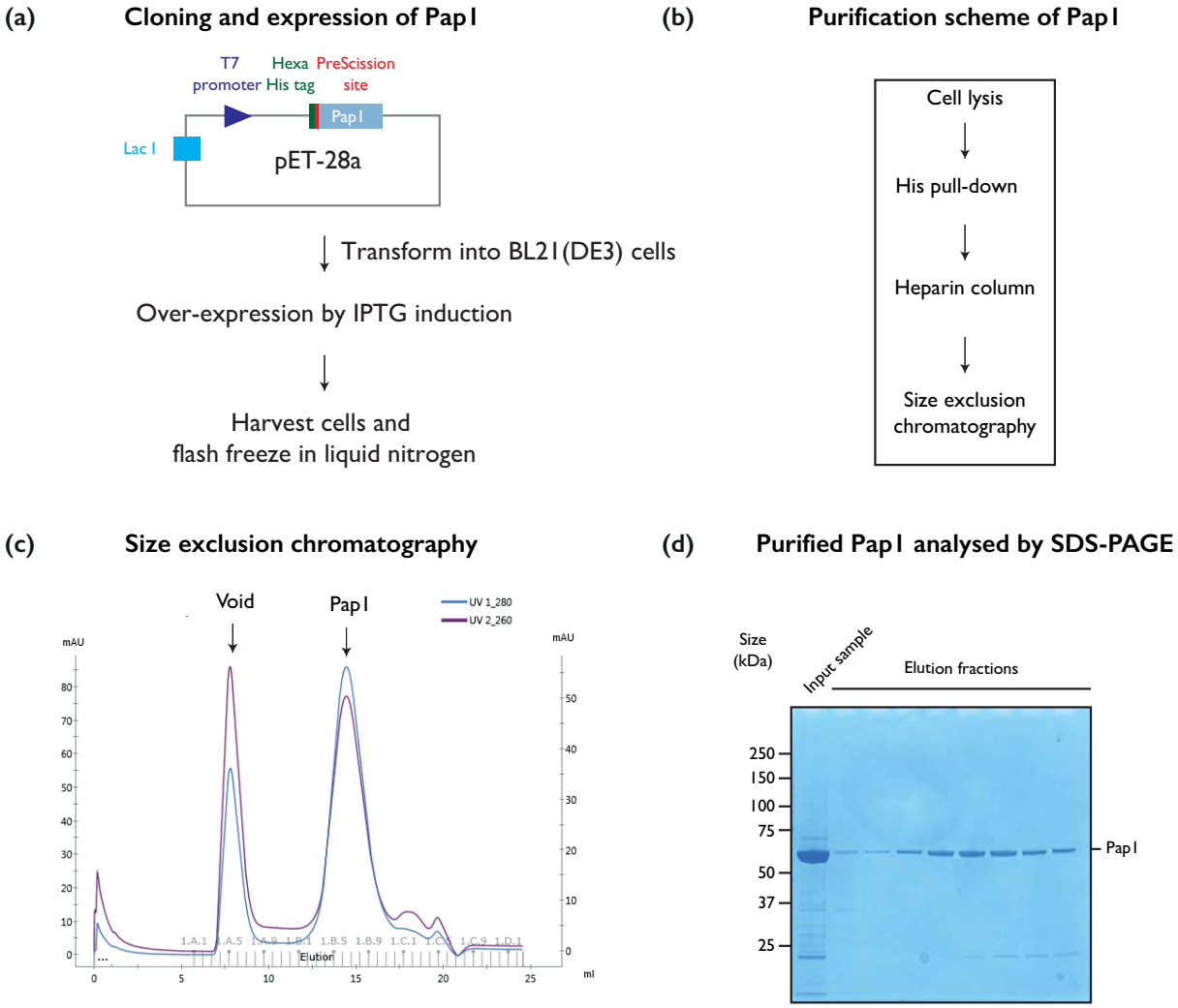


Figure 2.4: Purification of recombinant poly(A) polymerase Pap1. **(a)** Overview of steps invovled in cloning and IPTG induced bacterial over-expression of Pap1 **(b)** Purification scheme of Pap1 **(c)** Size exclusion chromatography using a S200 10/300 GL column. **(d)** SDS-PAGE of fractions from size exclusion chromatography.

2.2 Reconstitution of *in vitro* poly(A) tail addition

The recombinant production of large amounts of highly pure polymerase module has enabled us to characterize the polyadenylation activity of the complex using *in vitro* assays. Here, I describe *in vitro* polyadenylation assays of a fluorescently labeled substrate RNA.

2.2.1 Comparison of the polyadenylation activity of polymerase module and Pap1

Before setting up polyadenylation assays of Pap1 and the polymerase module complexes, it was necessary to normalize protein samples for the amount of the polymerase Pap1. This was carried out using Coomassie staining and densitometry analysis of the protein band corresponding to Pap1 (Figure 2.5a). Similarly, the polymerase module containing excess Pap1 was loaded onto an SDS-gel alongside twice the amount of isolated Pap1 (Figure 2.5b). The amount of Pap1 in the polymerase module containing excess Pap1 was found to be similar to twice the amount of Pap1 in the isolated sample, further confirming that this preparation contains two copies of Pap1 (Figure 2.5b). These samples were used in subsequent polyadenylation assays.

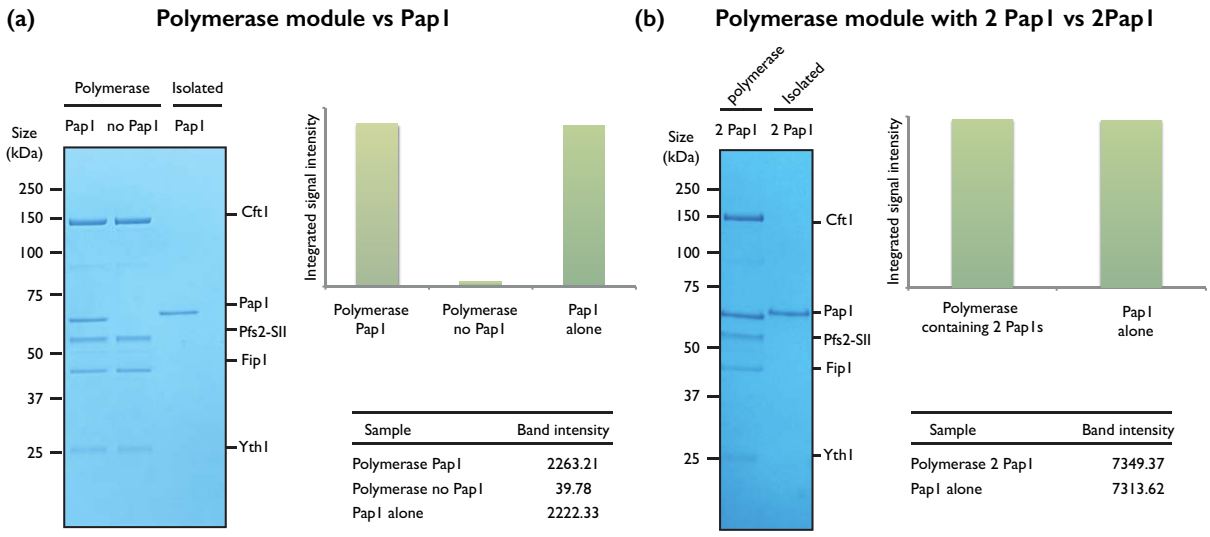


Figure 2.5: Quantification of Pap1 amounts by densitometry. **(a)** SDS-PAGE of purified polymerase module and Pap1. The gel was visualized after staining by Coomassie blue. The intensity of the band corresponding to Pap1 was analyzed by Image J (Schneider, Rasband, and Eliceiri 2012). After normalizing for the intensity of Pap1, the proteins were further used in polyadenylation assay described in figure 2.6. **(b)** SDS-PAGE of purified polymerase module variant containing 2 Pap1 and Pap1 alone. The intensity of the band corresponding to Pap1 was analyzed by Image J. After normalizing for the intensity of Pap1, the proteins were further used in polyadenylation assay described in figure 2.7

In vitro polyadenylation assays were carried out using purified polymerase module containing one copy of Pap1 and its polyadenylation activity was compared to isolated Pap1. 3' end of the yeast *CYC1* mRNA has been historically used as a model substrate to study cleavage and polyadenylation (Butler and Platt 1988). *CYC1* mRNA contains several *cis*-elements known to be important for efficient 3' end processing (Dichtl and Keller 2001). A synthetic 42-mer *CYC1* 3' UTR with a 5' 6-FAM label (hereafter referred to as 5'-FAM-*CYC1*-pc) was used as the substrate in the assays (Schmid et al. 2012). Products of polyadenylation assays were analyzed by denaturing urea-PAGE (Figure 2.6). Pap1 appears to add poly(A) tail faster than the polymerase module. All substrate RNA is polyadenylated after 8 minutes, as seen by the disappearance of the band corresponding to 5'-FAM-*CYC1*-pc. (Figure 2.6a). The differences in speed between Pap1 and polymerase module can be appreciated at early time points. Similar to the reaction containing polymerase module, all substrate RNA is polyadenylated after time 8 minutes by Pap1 (Figure 2.6b). The final length of the poly(A) tail in both cases seem similar (~260 As).

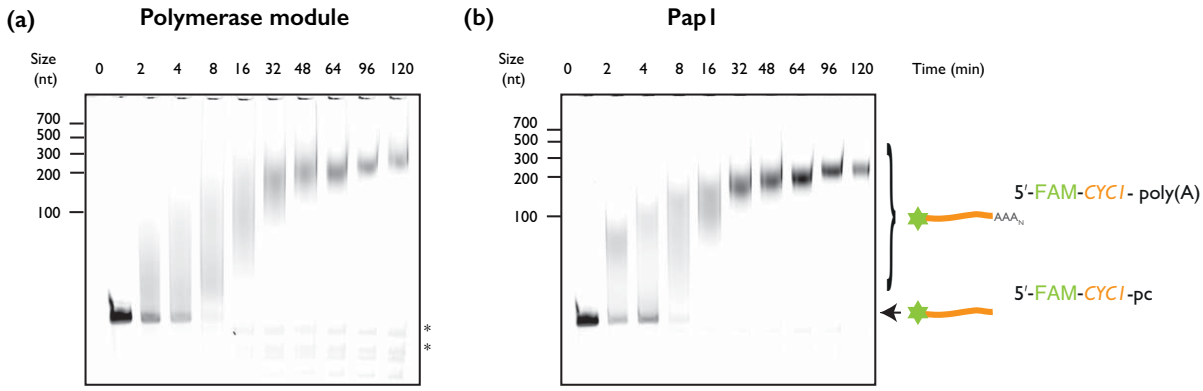


Figure 2.6: Recombinantly purified polymerase module and Pap1 can add a poly(A) tail to a pre-cleaved CYC1 RNA. Polyadenylation assay of **(a)** purified polymerase module and **(b)** Pap1 analyzed by denaturing urea-PAGE. A 5′ fluorescently labeled 42 mer CYC1 RNA that has been pre-cleaved is the substrate for polyadenylation. The assays were carried out with a final concentrations of 1 μM substrate RNA and 50 nM purified proteins. Reactions were initiated by the addition of 2 mM ATP. * denotes degraded RNA.

Next, the activity of the polymerase module containing two Pap1s was compared to that of isolated Pap1 (present in twofold excess). Isolated Pap1 was found to be faster than the polymerase module containing two Pap1s (Figure 2.7a). Moreover, polyadenylation by polymerase module containing two copies of Pap1 occurs more rapidly compared to that containing one Pap1 subunit. In all cases, polyadenylation appears to be distributive, as all substrate RNA is rapidly polyadenylated.

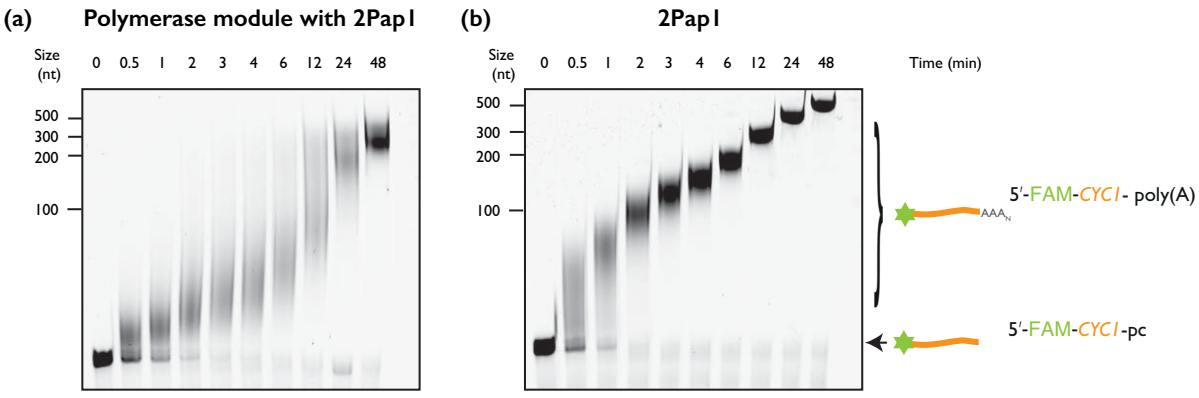


Figure 2.7: Comparing the polyadenylation of polymerase module variant and Pap1. Polyadenylation assay of purified **(a)** polymerase module with two copies of Pap1, and **(b)** Pap1 alone present in two times molar excess amounts was analyzed by denaturing urea-PAGE. A 5′ fluorescently labeled 42 mer CYC1 RNA that has been pre-cleaved is the substrate for polyadenylation. The assays were carried out with a final concentration of 400 nM substrate RNA and either 50 nM purified polymerase module or 100 nM Pap1. Reactions were initiated by the addition of 2 mM ATP.

Interestingly, the results obtained here are contradictory to the assays performed with human complex where isolated PAP (or Pap1) was less active compared with the mammalian polyadenylation specificity factor (similar to yeast polymerase module) (Schönemann et al. 2014). It is well known that isolated polymerase does not bind RNA tightly and hence adds a poly(A) tail to the substrate RNA in a distributive manner (Wahle 1991; Bienroth, Keller, and Wahle 1993; Wahle 1995). In contrast, the binding affinity of the human complex for an AAUAAA-containing RNA was reported to be ~2 nM (Schönemann et al. 2014; Chan, Huppertz, Yao, Weng, Moresco, Yates, et al.

2014), potentially explaining the increased processivity and increased activity of the human complex relative to PAP.

2.2.2 RNA binding gel shift assays of polymerase module

While the human complex binds to RNA with higher affinity than PAP alone, it is unknown if the yeast complex shows similar relative binding behaviour. To assess the RNA binding of yeast polymerase module, electrophoretic mobility shift assays (EMSAs) were carried out with a variety of fluorescently labeled substrate RNAs (Figure 2.8a). At 500 nM protein concentrations, almost all RNA was bound to the polymerase module, as observed by the disappearance of the band corresponding to free RNA. (Figure 2.8b). To assess the relative contributions of various *cis*-elements to RNA binding, point mutations were introduced in the positioning element and/or the upstream U-rich elements of the 5'-FAM-*CYC1*-pc RNA. The polymerase module appears to bind these mutant RNAs with slightly reduced affinity. Therefore, it is likely that there are additional binding sites in the RNA for the polymerase module. Furthermore, a 42-mer wild type adenovirus L3 RNA (containing the AAUAAA hexamer) with a 5' fluorescein-label was also tested for its ability to bind to the polymerase module. Interestingly, the viral RNA did not bind polymerase module as tightly as the yeast sequence (Figure 2.8c). Therefore, the yeast polymerase module displays binding specificity towards the yeast sequence.

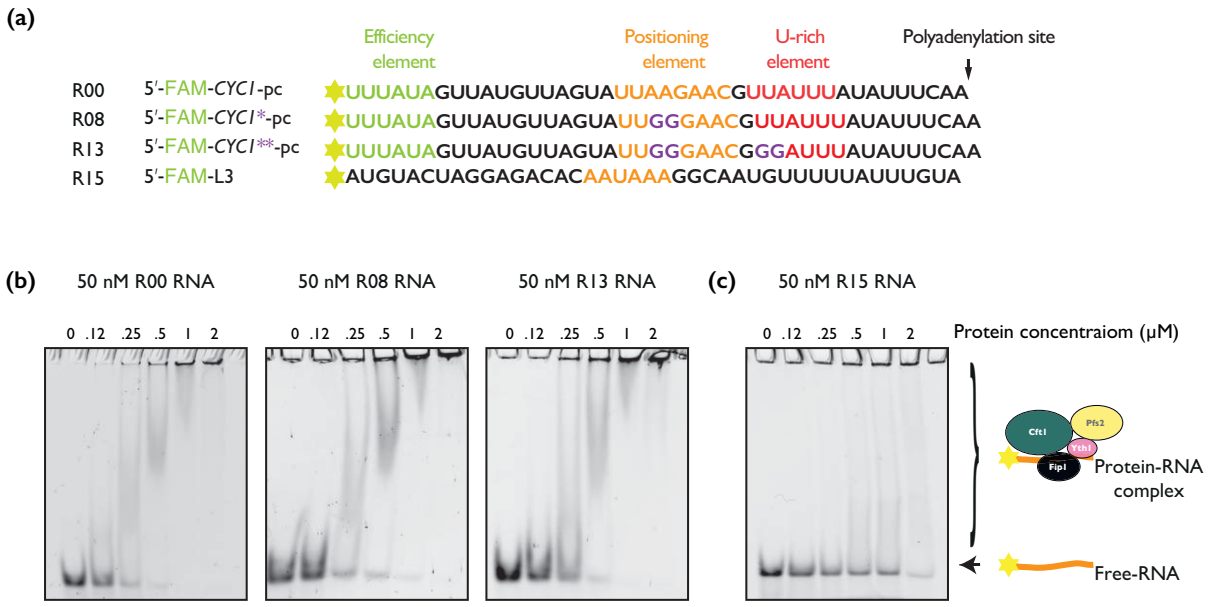


Figure 2.8: Electrophoretic mobility band shift assay (EMSA) of polymerase module with various substrate RNA **(a)** RNA sequences used in the study. **(b)** Recombinantly purified polymerase module (without Pap1) was mixed with 50 nM of a fluorescently labeled 42 nt long pre-cleaved CYC1 3′ UTR and the resulting complex was analyzed by 6% native-PAGE. **(c)** The yeast polymerase module does not bind the adenovirus L3 RNA UTR.

Nonetheless, the binding affinity of the yeast polymerase module is in the range of several hundred nanomolar lower than that shown by the human polymerase module. These results suggest that the yeast polymerase module may act less processively on the CYC1 RNA, providing a potential explanation for why the polymerase module is less active than the polymerase alone.

2.3 Structure of Cft1-Pfs2-Yth1 subunits of the polymerase module

While the non-catalytic subunits do not appear to be important for polymerase module activity *in vitro*, they are highly conserved and functionally significant *in vivo*. In order to better understand and characterize the roles of these components of the polymerase module, I sought to study the architecture and structural composition of the module using electron microscopy.

2.3.1 Negative stain electron microscopy of the polymerase module

Electron microscopy of a negatively stained protein or protein complex is a good way to assess the quality of the sample. It can provide us with valuable information such as the size (or molecular dimension) of the sample, the homogeneity or purity of the protein preparation and a low resolution initial 3D model of the specimen being studied (Kiselev, Sherman, and Tsuprun 1990). In negative stain EM, the protein of interest is mixed with a heavy metal stain such as uranyl acetate or uranyl formate. The contrast produced by a biomolecule upon interaction with a low dose of electrons is weak. However, in case of negatively stained particles, the electrons interact with the coating of the negative stain around the protein and hence produce images with very good contrast. Furthermore, negative staining also preserves the structure of the proteins in the vacuum chamber of the microscope. Given the relative ease of preparing negative stain grids, it is a recommended first step in any new structural biology project.

Previous characterization of the polymerase module (with and without Pap1) by negative stain electron microscopy showed individual ~14-15 nm particles with uniform distribution (Figure 2.3c). 2D classification of particles revealed class averages containing well-aligned polymerase module complexes (Figure 2.9a). The classes have a wide range of views and suggest that the particles are present in a variety of orientations on the grid. Interestingly, the size and shape of 2D class averages of the polymerase module look strikingly similar to those of the human 3' end processing machinery as reported in an earlier study (Shi et al. 2009).

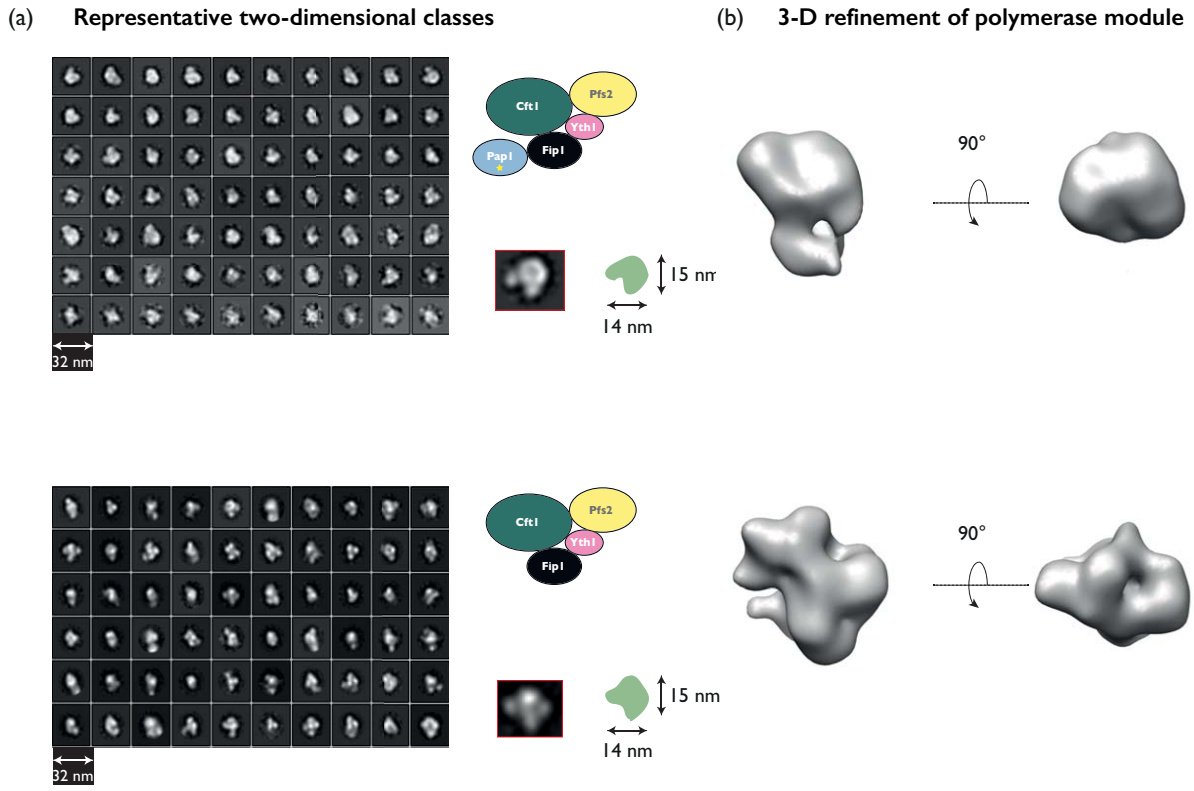


Figure 2.9: Negative stain electron microscopy of polymerase module variants **(a)** Selected 2-D class averages of aligned particles of polymerase module with one copy (top) and no copy of Pap1 (bottom). **(b)** A 3-D reconstruction of the negative stain data of the polymerase module variants.

Compared to 2D class averages of the Pap1-containing polymerase module, those lacking Pap1 appear to show more detail and display less background noise. This is consistent with the observation where micrographs of the polymerase module without Pap1 look more homogeneous and contain uniformly sized and uniformly distributed particles (Figure 2.3c). Further 3D classification of selected 2D classes showed that the polymerase module particle adopts an overall P-shape (Figure 2.9b). Similar to the 2D class averages, the complex containing Pap1 showed less structural detail compared to that lacking Pap1 (Figure 2.9b). Moreover, the sample containing Pap1 did not contain any significant additional density that could account for Pap1.

In summary, negative stain EM revealed the size and shape of the polymerase module. The polymerase module (without Pap1) appeared promising in comparison with the sample with Pap1. To further understand the molecular basis of subunit association

within the protein complex, I used cryo-EM. The structure determination of the polymerase module was done in collaboration with Ana Casañal.

2.3.2 Preliminary cryo-EM

The four-subunit polymerase module was vitrified in a layer of un-supported ice on UltrAufoil gold grids (C. J. Russo and Passmore 2014). Cryo-EM micrographs revealed well-separated particles of uniform size and distribution (Figure 2.10a). Given the particles were well-distributed and appeared intact, I proceeded to collect micrographs on a 300 kV FEI Titan Krios electron microscope equipped with a Falcon II detector. From these, I obtained 2D classes containing high-resolution features (Figure 2.10b). The overall shape of the 2D classes was consistent with those obtained from negative stain electron microscopy.

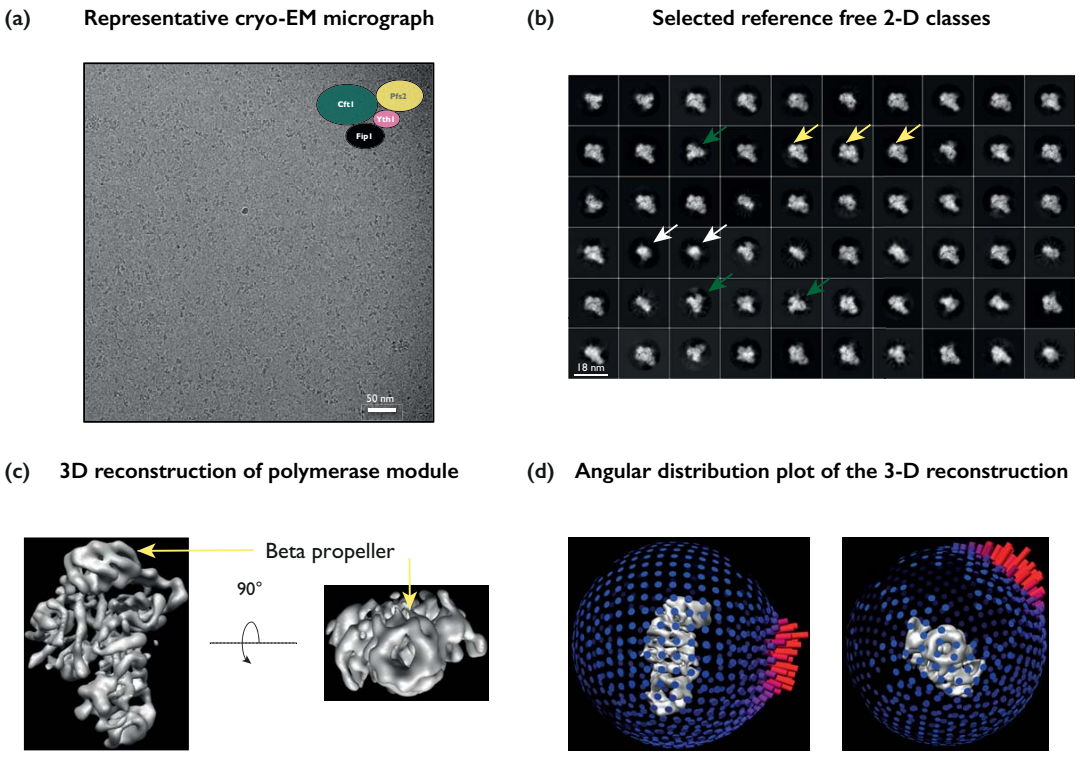


Figure 2.10: Preliminary structural characterization of polymerase module **(a)** Representative micrographs of the polymerase module without Pap1 **(b)** Selected 2D class averages of aligned particles of polymerase module without Pap1. Green arrows indicate the "Y" shaped 2-D classes. Yellow arrows indicate the additional density found on top of the "Y" shaped classes. White arrows indicate 2D classes containing a possible top view **(c)** A 3D reconstruction of the polymerase module. Yellow arrow indicates a putative beta propeller domain **(d)** Angular distribution plot of the 3D reconstruction shows preferred orientation of the aligned particles.

Upon closer inspection, the sample shows some inherent heterogeneity (Figure 2.10b). For example, some 2D classes adopt a Y-shape (dark green), whereas others contain additional density on top of the Y-shaped particle (yellow). Selected 2D classes containing extra density were subjected to 3D refinement using the 3D map of full native-CPF as an initial model (Casañal et al. 2017b). The 3D refinement, comprising 56,341 particles, yielded a map of the polymerase module at ~ 7 Å resolution (Figure 2.10c). At this resolution, clear density corresponding to a beta propeller was observed (yellow arrow). The angular distribution plot of the 3D reconstruction revealed that the side-on view was predominant, as evidenced in the 2D class averages, hindering further improvements in resolution (Figure 2.10d). The rare top views that are poorly represented in the 3D map are highlighted in the 2D classes using a white arrow. Thus, the preferred orientation problem might prevent the reconstruction of a high-resolution 3D map, in spite of obtaining 2D classes with secondary structural details. Furthermore, at this preliminary resolution, it is difficult to assign the subunits to the observed density.

More micrographs of the polymerase module were collected in order to overcome the preferred orientation problem by obtaining additional views and enriching existing rare views. The new data was combined with the previous dataset; Figure 2.11 shows a flowchart of how the combined data were processed, again revealing sample heterogeneity. From further classification, I observed the characteristic Y-shaped 2D classes (Class 6) as well as classes that contain or lack side density (pink arrow). A final refinement of the selected particles resulted in a 3D map of ~ 7 Å resolution and with enriched rare views that were not present in earlier reconstructions. It is possible that the additional data were collected at a range of ice thicknesses, resulting in different sample orientation and thus views of the protein complex. Nonetheless, despite collecting more data, there was little to no improvement in the final map resolution (Figure 2.10c, Figure 2.11), and the final resolution following post-processing was reported to be ~7 Å. Data collection and processing details are described in section 6.7.4.

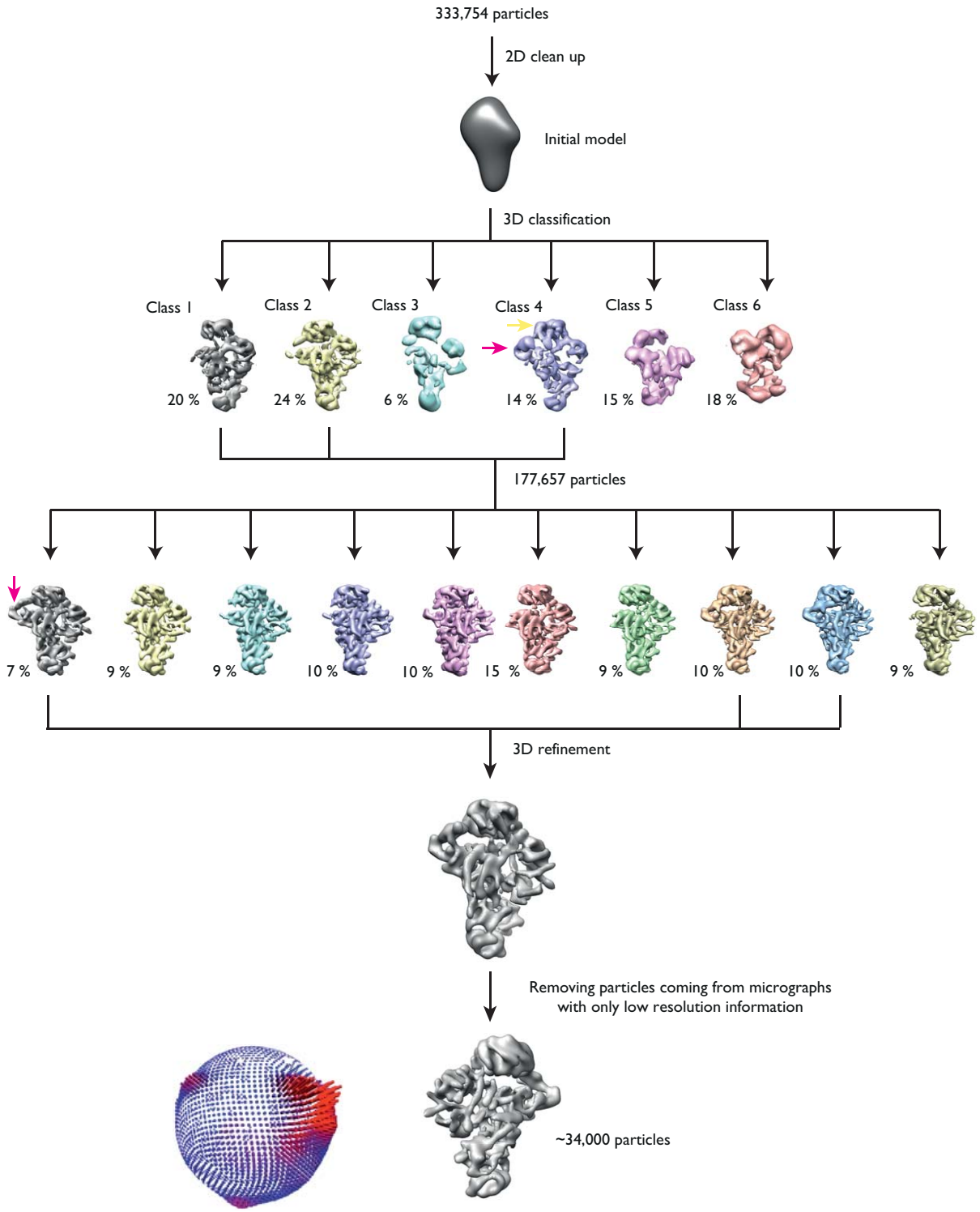


Figure 2.11: Summary of cryo-EM data processing of the polymerase module. Two data sets were collected in a Titan Krios operating at 300 kv and equipped with a Falcon-II direct electron detector. Yellow and pink arrows indicate densities that were missing in some 3D classes. 3D classification without alignments were employed to further remove incomplete density maps. The final 3D map and its particle distribution are shown.

2.3.3 Overcoming the resolution barrier

In order to improve the resolution of the polymerase module, I sought to investigate the data collection parameters that could be limiting the resolution of the map. One such parameter is the detective quantum efficiency (DQE) of the detector, which represents the signal:noise ratio contributed by the detector. An ideal detector with a DQE value of 1 will not contribute to any noise in the image, and will detect all incident electrons. All cryo-EM data were thus far collected on a Falcon-II detector in integration mode at a magnification of 47000X (1.77 Å/pixel). A comparative analysis of the DQE of three different commercially available detectors show that the K2 detector operating in counting mode has higher DQE compared to Falcon II or DE-20 at low spatial frequency (McMullan et al. 2014) (appendix 8.2.) The DQE of K2 detector increases almost linearly with decreasing spatial frequency (appendix 8.2.). This means that K2 detector will have better signal-noise at higher magnifications. Furthermore, a higher magnification allows the pixel size to be reduced, increasing the maximum attainable resolution. A higher DQE will also allow a desired resolution to be reached with reduced amounts of particles and time.

Taking these parameters into consideration, I decided to image the four-subunit polymerase module in a Titan Krios microscope operating at 300 keV and using a K2 summit detector in super-resolution mode (81000X, 1.4 Å/pixel). A Gatan Imaging Filter (GIF) with a slit width of 20 eV was used to enhance contrast. The cryo-EM micrograph of the polymerase module revealed well-separated particles of uniform size and distribution (Figure 2.12a), consistent with earlier observations. 2D classes of selected particles reveal very high-resolution secondary structural information as well as a variety of views of the complex (Figure 2.12b).

(a) Representative cryo-EM micrograph (b) Selected reference free 2-D classes

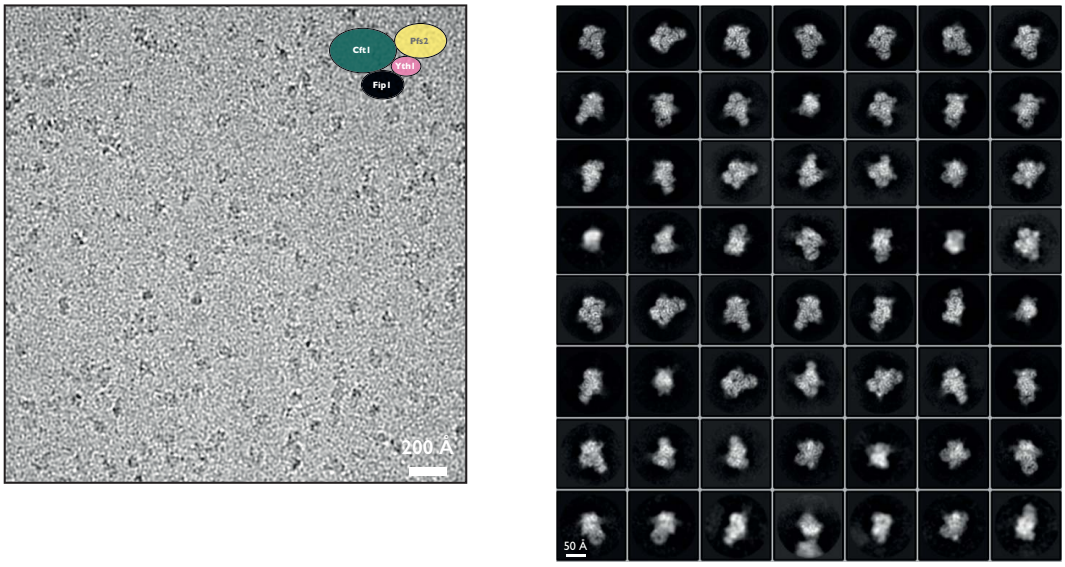


Figure 2.12: Cryo-EM analysis of the polymerase module using a K2 direct electron detector. **(a)** Representative micrograph of the polymerase module showing well-distributed and homogenous particles. **(b)** Selected 2D class averages showing high-resolution secondary structural details and a variety of views.

Two separately collected datasets were merged using the procedure illustrated in appendix 8.3. The data were independently processed and two final 3D maps were obtained separately from the two data sets. As the absolute pixel sizes on each microscope were undetermined, the micrographs needed to be rescaled relative to each other. By cross-correlating the 3D maps, a scaling factor was calculated. Once the scaling factor was calculated, all the micrographs from the first dataset were rescaled accordingly. The rescaled micrographs were subjected to a previously established data processing pipeline (section 6.7.4) and all particles were merged. Ana Casañal performed the data merging. A total of 460,167 particles from the two data sets were subjected to initial 2D clean up and 3D classification. 3D refinement following particle polishing led to a 3.6 Å resolution map of the polymerase module (Figure 2.13). Further 3D classification without alignments separated maps that either contained or lacked the additional density highlighted in pink. A final map of the polymerase module was obtained from 77,917 particles and at an overall resolution of 3.5 Å based on the gold standard Fourier shell correlation (FSC) at 0.143, which in turn is derived from comparisons between reconstructions from two independently refined half-sets.

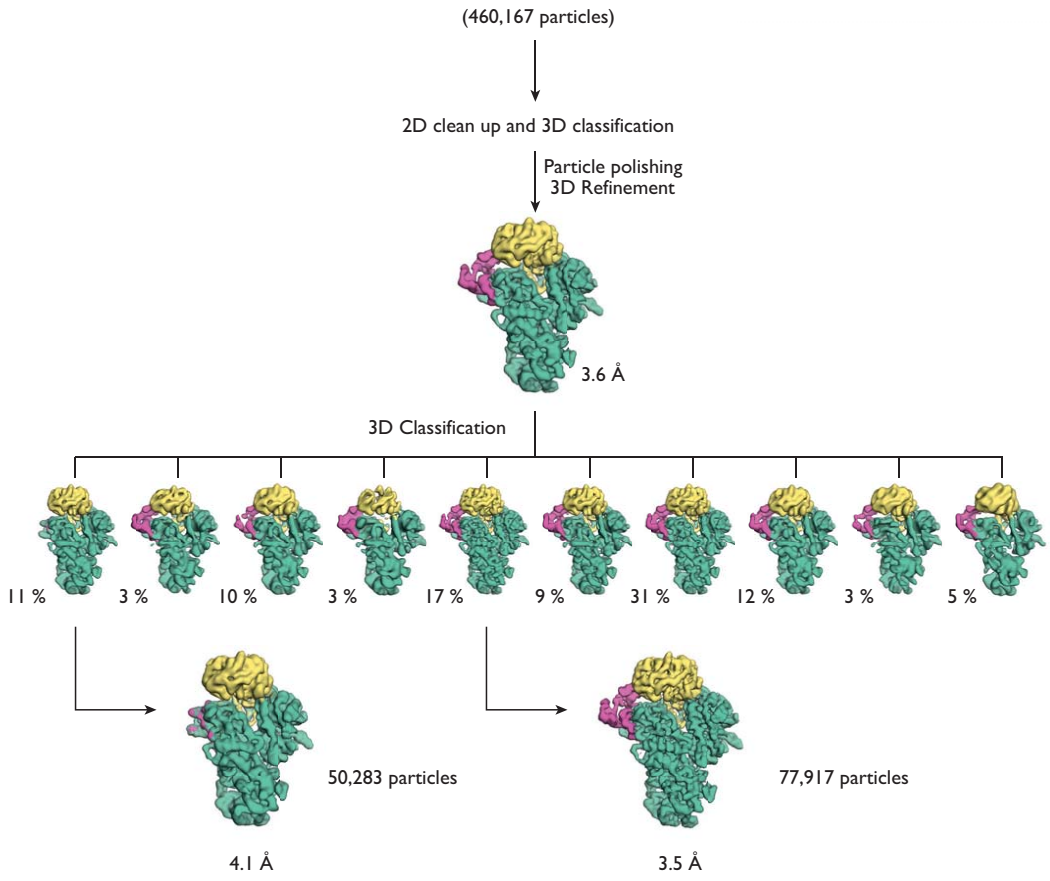


Figure 2.13: Steps in the cryo-EM data processing of the polymerase module. Particle picking followed by initial 2D and 3D classifications, resulted in 460,167 particles. They were then polished and subjected to 3D refinement, resulting in a 3.6 Å density map. Further 3D classification without alignments resulted in separation of maps that contained or lacked the density corresponding to Yth1 (shown in pink). The map containing density for Yth1 refined to 3.5 Å, whereas the map without any density for Yth1 refined to 4.1 Å.

A local resolution map of the polymerase module shows that the resolution at the core of the map is ~ 3.4 Å, which drops to ~ 4 Å towards the outer regions (Figure 2.14b). Furthermore, the orientation distribution profile of the final map showed a variety of views, contrasting with previously collected data (Figure 2.14c). I speculate that the improved data collection parameters and use of the K2 detector in super-resolution mode and the additional views obtained from collecting in different ice thickness enabled the significant improvement in the resolution of the map of the polymerase module.

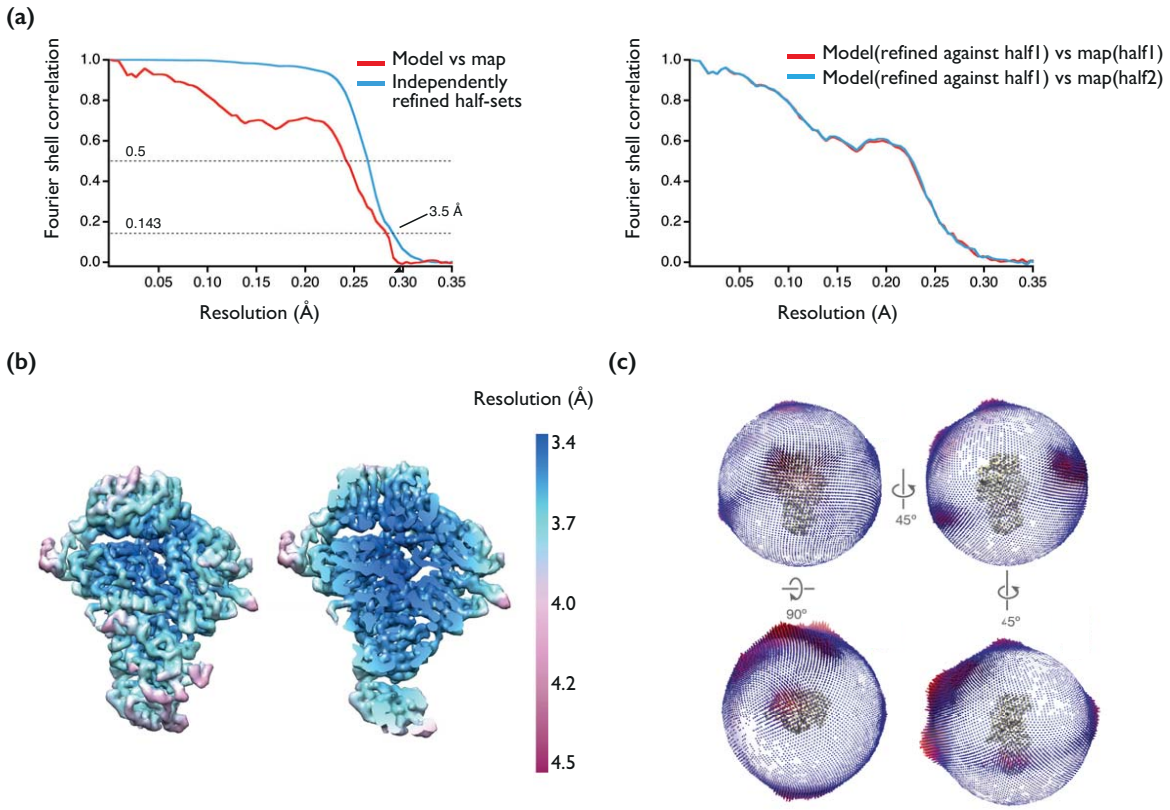


Figure 2.14: Resolution assessment of the polymerase module and orientation distribution of the particles in the final map. **(a)** Fourier shell correlation plots for gold standard refinements and model vs. map. Fourier shell correlation plots for model refined against each half map. The final map of the polymerase module at 3.5 Å resoluton. **(b)** The local resolution map shows that the resolution varies from ~ 3.4 Å at the core to ~ 4 Å at the surface. **(c)** Orientation distribution of 77,917 particles used in the final model, calculated using RELION (Scheres 2012) and visualized with *Chimera* (Pettersen et al. 2004).

2.4 Analysis of the cryo-EM structure of Cft1-Pfs2-Yth1

2.4.1 Atomic model of the polymerase module

The improved resolution of the map allowed regions of electron density to be interpreted, as side chains from alpha helices (Figure 2.15a) and beta strands (Figure 2.15b) could be clearly identified, particularly in the core of the complex. Homology models of Cft1 and Pfs2 from *Phyre* (Kelley et al. 2015), along with the crystal structures of DDB1 and DDB2, were used as templates to build the full, *de novo*, atomic model of the polymerase module. After modeling the Cft1 and Pfs2 subunits into the density map, the shape of the unmodeled density (shown in pink in Figure 2.15c) resembled that of a zinc finger protein. It is known that the Yth1 subunit consists of five zinc fingers (Barabino et al. 1997; Tacahashi, Helmling, and Moore 2003) and the crystal structure of the zinc finger domains two and three of CPSF30 has been determined (Das et al. 2008). Molecular replacement (performed by Alan Brown from the Ramakrishnan group) with the CPSF30 structure enabled the unambiguous identification of the unassigned density as that of Yth1. Appendix 8.4 details all the data collection and refinement statistics.

This model represents the first high-resolution structure available for the eukaryotic 3′-end processing complex. The cryo-EM map colored according to the subunit assignment and the corresponding model are shown in Figure 2.15c and Figure 2.15d respectively. The overall architecture of the polymerase module involves four beta propellers with two zinc fingers anchored on the side (Figure 2.15c). In the cryo-EM map, residues 1–1356 of Cft1, 27–414 of Pfs2 and 1–97 of Yth1 were well ordered and modeled. However, zinc fingers three, four, and five, and the C-terminal end of Yth1, along with the whole of Fip1, could not be modeled. Also not visible were several loops in Cft1. This is likely because those regions are disordered, as confirmed by bioinformatic predictions (appendix 8.5). The predicted disordered regions in the yeast proteins are predicted to be disordered in humans as well. For example, the C-terminus of Pfs2 (residues 412–465) that is not visible in the cryo-EM map is predicted to be disordered. Similarly, the human orthologue WDR33 contains a poorly conserved and disordered C-terminal region. Interestingly, this C-terminal disordered region in Pfs2 is not required for viability of yeast (Ohnacker et al. 2000).

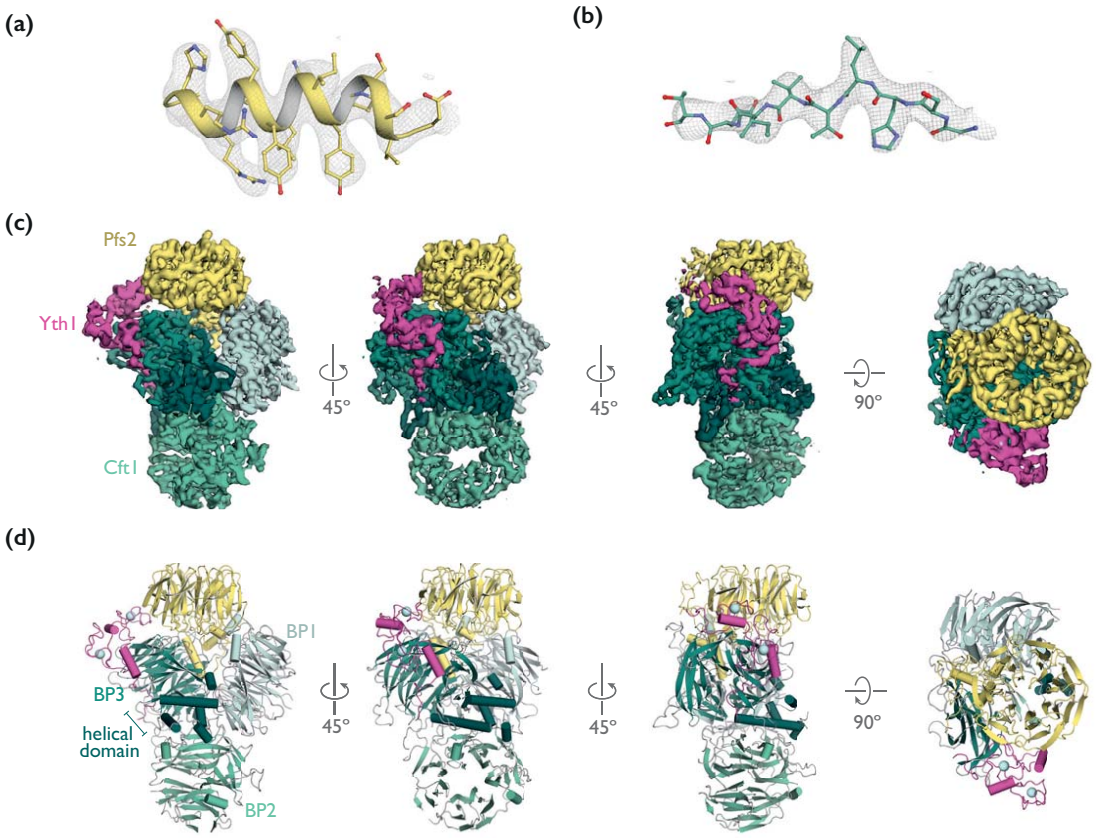


Figure 2.15: Atomic model of the Cft1-Pfs2-Yth1 subunits of the polymerase module. **(a)** Example density for an alpha helix in Pfs2. **(b)** Example density for a beta-strand in Cft1. **(c)** Cryo-EM map and **(d)** cartoon representation of the atomic model of the Cft1–Pfs2–Yth1 complex of the polymerase module. Yth1 (magenta), Pfs2 (yellow), Cft1 (green), and zinc ions (pale cyan) are depicted. The three β-propeller domains of Cft1 (BP1, BP2, and BP3) are colored in different shades.

Cft1 is a multi-domain protein that consists of three seven-bladed beta propellers followed by a C-terminal helical domain (Figure 2.15c and Figure 2.16a), forming the core of this three-subunit polymerase complex. Beta propeller (BP) 1 and BP2 are each formed of contiguous sequences. Notably, the density for BP2 was less well defined compared to BP1 and BP3, and it appeared to be more flexible (Figure 2.16a). BP3 is made up of residues mainly from the C-terminal region, but also contains one beta strand from the N-terminus, and three beta strands from the middle of Cft1, creating a rigid structural core (Figure 2.15d). A helical domain at the C-terminus of Cft1 is

located at the nexus of the three beta propellers, further securing the fold (Figure 2.15d). BP1 and BP3 interact with each other at a ~60° angle, forming a deep cavity.

Pfs2 consists of an N-terminal alpha helix followed by a central WD40 domain (Figure 2.15d). This WD40 is made up of a seven-bladed beta propeller (Figure 2.16a). The C-terminal end in the model of Pfs2 consists of a loop protruding out of the last blade of the WD40 and lies adjacent to a loop in the N-terminal end of the model (Figure 2.16.A).

Yth1 is an RNA binding protein containing five zinc fingers. An N-terminal loop and the first two zinc fingers of Yth1 are found to lie next to the BP2 of Cft1 and the WD40 domain of Pfs2 (Figure 2.15d).

2.4.2 Interaction of Pfs2 with Cft1

Previous nanoESI-MS data from our lab had reported a strong direct interaction between Cft1 and Pfs2 (Easter 2014; Casañal et al. 2017b). In agreement with this, the surface area of the interaction interface between Pfs2 and Cft1 is >4,200 Å^2 (Figure 2.16b). Almost 50 amino acids in the N-terminal region of Pfs2 are inserted into a deep cavity between BP1 and BP3 of Cft1. The alpha helix in this region of Pfs2 is likely stabilized by its interaction with Cft1 (Figure 2.16b). The WD40 domain of Pfs2 sits on top of Cft1, with the loops extending from BP1 and BP3 cradling Pfs2 to further stabilize the interaction. Many of the interacting residues between Pfs2 and Cft1 are conserved in the human orthologs CPSF160 and WDR33 (appendix 8.6).

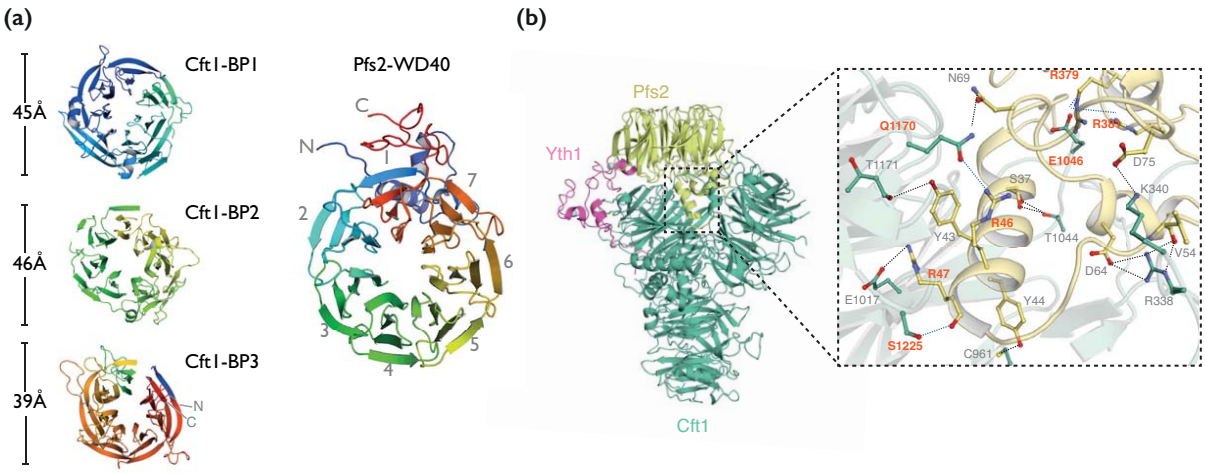


Figure 2.16: Interaction between the Cft1 and Pfs2 subunits of the polymerase module. **(a)** The three beta propellers from Cft1 are shown in ribbon diagram where the entire Cft1 chain is colored in rainbow from N- to C-terminus. Pfs2 is shown as ribbon diagram, colored in rainbow from N- to C-terminus. **(b)** Cartoon representation of the atomic model of the Cft1–Pfs2–Yth1 complex of the polymerase module. The N-terminal α-helix of Pfs2 inserts into the cavity formed between the β-propeller BP1 and BP3 of Cft1. Inset: details of the interaction between Cft1 and Pfs2. The residues involved in the interaction that are also conserved in humans are highlighted in orange. Yth1 (magenta), Pfs2 (yellow), Cft1 (green), and zinc ions (pale cyan) are depicted.

2.4.3 Interaction of Yth1 with Cft1 and Pfs2

Yth1 latches onto the Cft1-Pfs2 complex by inserting its N-terminal extended region into the central cavity of Cft1s BP3, and continues across a hydrophobic external face (Figure 2.17a). A surface representation of the structure reveals a hydrophobic patch in the interaction interface in Cft1 as well as a corresponding hydrophobic Yth1 surface, showing that the interaction is driven by hydrophobic effects (Figure 2.17b and Figure 2.17c). In agreement with the importance of the N-terminus for interaction with CPF, deletion of the first 25 residues in Yth1 results in a slow growth and temperature sensitive phenotype in yeast (Barabino, Ohnacker, and Keller 2000b).

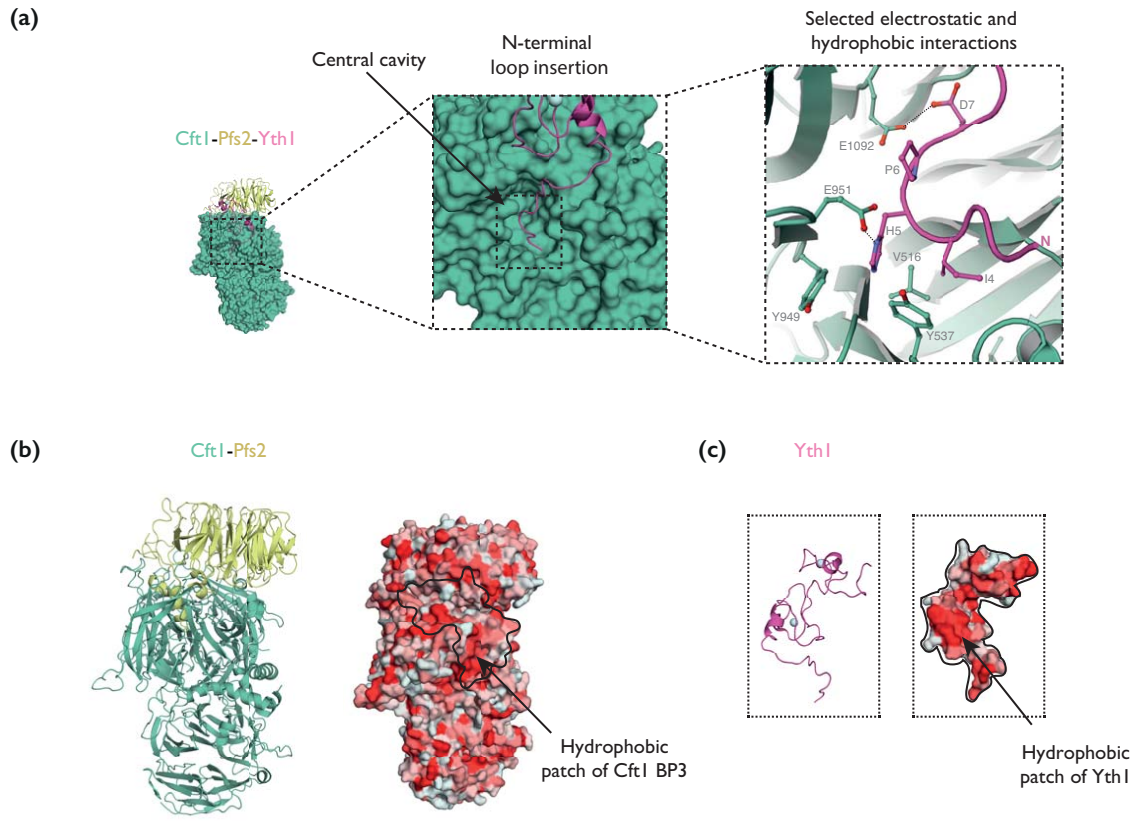


Figure 2.17: Interaction between the BP3 of Cft1 and N-terminal loop of Yth1 subunits of the polymerase module. **(a)** Cartoon representation of the atomic model of the Cft1–Pfs2–Yth1 complex of the polymerase module. An N-terminal loop in Yth1 inserts into the β-propeller BP3 of Cft1. Selected electrostatic and hydrophobic interactions between the N-terminal loop and BP3 are depicted in the inset. **(b)** Cartoon representation and the corresponding hydrophobic surface representation of Cft1-Pfs2 subunits showing a hydrophobic surface at the Yth1 interacting surface. **(c)** The hydrophobic surface rendering of Yth1 reveals a lining of hydrophobic residues that interacts with the BP3 of Cft1. Yth1 (magenta), Pfs2 (yellow), Cft1 (green), and zinc ions (pale cyan) are depicted.

The first two of the five Cys-Cys-Cys-His (CCCH) zinc fingers pack into the interface between Cft1 and Pfs2 (Figure 2.18). Zinc finger 1 (ZnF1) lies adjacent to Cft1-BP3, and second zinc finger (ZnF2) contacts both Cft1 and the side of the Pfs2 beta propeller. Furthermore, it has been shown that ZnF2 can directly bind RNA (Barabino, Ohnacker, and Keller 2000b; Tacahashi, Helmling, and Moore 2003; Chan, Huppertz, Yao, Weng, Moresco, Yates, et al. 2014). A previous study showed that mutation of Yth1-W70 to alanine causes Yth1 to dissociate from CPF (Barabino, Ohnacker, and Keller 2000a). In the structure, it can be seen that W70 forms pi-stacking interactions with the zinc-coordinating H85 in ZnF2 (Figure 2.18). Disruption of this interaction by mutating the tryptophan to alanine may destabilize zinc binding, and hence the interaction of Yth1 with the other subunits.

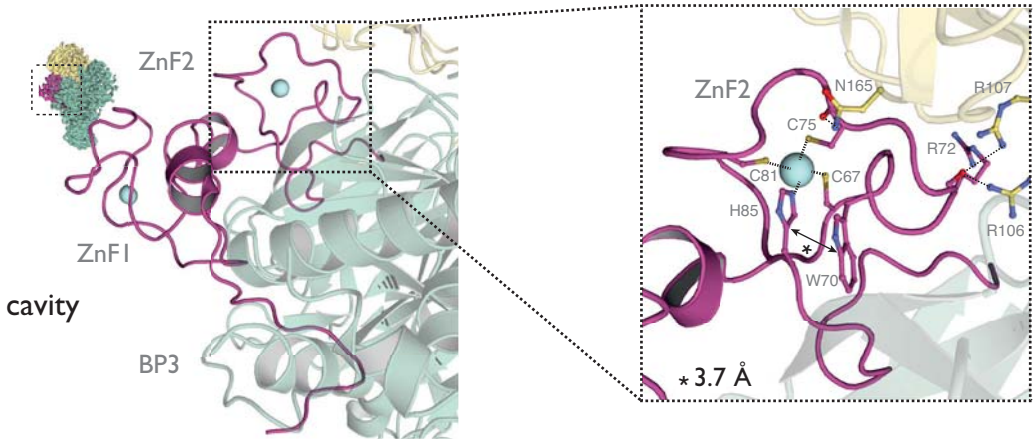


Figure 2.18: Yth1 harbors a CCCH zinc finger repeat. The first two of the five zinc fingers of Yth1 are displayed. The zinc fingers belong to the CCCH family and pack into the interface between Cft1 and Pfs2. The residue -W70 of Yth1 ZnF2 pi stacks with Yth1-H85 and stabilizes the zinc finger fold. Also shown are the hydrogen-bonding interactions between side chains of Pfs2 and backbone atoms of Yth1. Yth1 (magenta), Pfs2 (yellow), Cft1 (green), and zinc ions (pale cyan) are depicted.

2.4.4 Polymerase module has a similar architecture to DDB1-DDB2 / SF3b

The characteristic four-beta propeller arrangement of the polymerase module is very similar to other nucleic acid binding protein complexes such as DDB1-DDB2 and SF3b (Figure 2.19a) (Scrima et al. 2008; Cretu et al. 2016). DDB1-DDB2 is a complex that recognizes UV-damaged DNA, and acts as an adapter for a cullin-RING E3 ubiquitin ligase to trigger nucleotide excision repair (Scrima et al. 2008; Li et al. 2006). DDB1 has three beta propellers in a tripartite arrangement reminiscent of the Cft1 (Figure 2.19). Superposition of Cft1 onto the structure of DDB1 revealed an RMSD of 7.53 Å (Figure 2.19b). Despite not displaying significant sequence homology, DDB2 is strikingly similar to Pfs2 (RMSD = 3.3 Å) with an N-terminal alpha helix that inserts into the cavity between BP1 and BP3 of DDB1, followed by a WD40 domain.

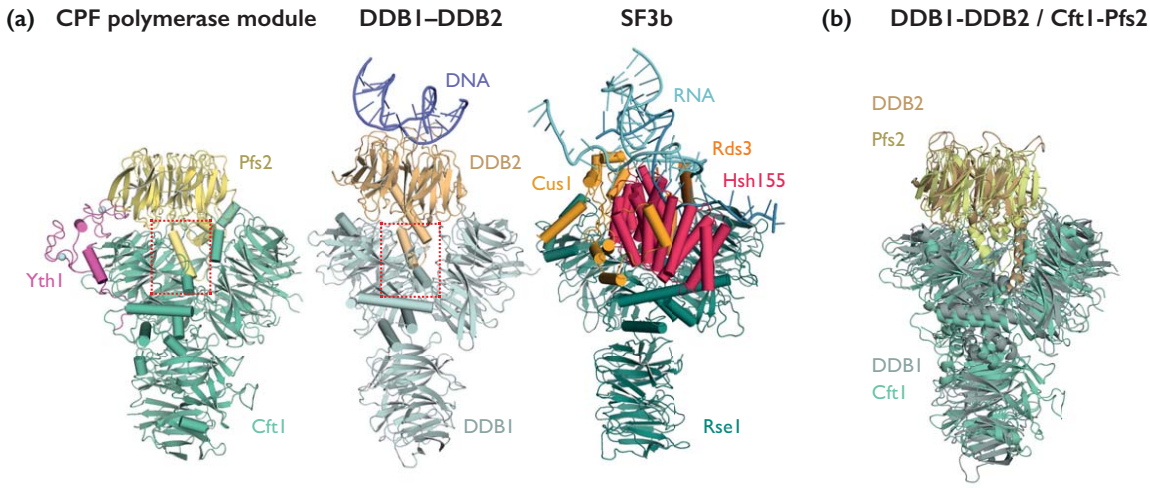


Figure 2.19: Structural comparison of Cft1-Pfs2-Yth1, DDB1–DDB2 DNA repair and SF3b splicing complexes. **(a)** Cft1, DDB1 and Rse1 contain three beta-propellers followed by a C-terminal alpha helical domain. DDB2 interacts with DDB1 by inserting an N-terminal alpha-helix into the cavity between two beta-propellers of DDB1, very similar to how Pfs2 interacts with Cft1 (highlighted by a square with dotted red lines). In SF3b, helical regions of Hsh155 and Rds3 inserts between the cavity between the beta-propellers of Rse1. **(b)** Structural superimposition of the structures of DDB1-DDB2 and Cft1-Pfs2. [PDB of DDB1-DDB2: 3ei3, PDB of Sf3b: 5gm6]

SF3b is part of the U2snRNP complex essential for pre-mRNA splicing and plays a key role in branch site recognition. The scaffolding protein Rse1/SF3b130 forms the core of the SF3B complex (Yan et al. 2016; Plaschka, Lin, and Nagai 2017). Similar to the fold of Cft1, Rse1 also contains three beta-propellers followed by a C-terminal alpha helical domain. In SF3b, two predominantly helical subunits (Hsh155 and Rds3) insert into the cavity formed between the two beta propellers of Rse1 and additionally bind another protein Cus1 (Figure 2.19a).

In summary, DDB1-DDB2, SF3b, and CPF all use a tripartite scaffold protein (DDB1, Rse1 or Cft1) to assemble a rigid and structurally stable complex. Although the core scaffolding proteins DDB1, Rse1 and Cft1 share low sequence homology (~15% sequence identity), they are highly structurally similar. Furthermore, their interaction partners (DDB2, Hsh155/Rds3 or Pfs2) inserts an N-terminal alpha helix into a cavity between BP1 and BP3, but the exact interaction mechanism is not conserved. These structural similarities may suggest that these complexes are evolutionarily related and adapted for different nucleic acid binding and processing complexes. The complexes also contact other binding partners through their beta propellers (e.g. DDB1-BP2 binds cullin-RING proteins). Similarly, Cft1 may contact other CPF subunits through the beta propeller domains.

2.4.5 A potential RNA binding surface on Pfs2

Using iCLIP experiments, previous studies have shown that WDR33 can directly bind RNA. Furthermore, numerous biochemical experiments support the RNA binding property of ZnF2 of Yth1. Interestingly, the aforementioned complexes that are similar to the polymerase module also directly bind nucleic acids. In the SF3b complex, subunits Hsh155 and Cus1 are known to contact RNA. On the other hand, the DDB2 substrate receptor of DDB1-DDB2 complex directly binds UV-damaged DNA lesions. A hairpin on the top surface of the beta propeller of DDB2 (rich in basic amino acid residues) extends into the minor groove of DNA, pushing the damaged base into a binding pocket. Similarly, the equivalent surface on Pfs2 contains a cluster of conserved lysine, arginine and aromatic residues that could form a putative RNA binding surface (Figure 2.20a). The electrostatic surface potential of the polymerase module reveals a patch of positively charged residues on top of Pfs2 and on the ZnF2 of Yth1 that lies adjacent to Pfs2 (Figure 2.20b). Together, the Pfs2 top surface and the ZnF2 of Yth1 could provide a composite RNA binding platform. A previous study implicated the central region (residues 586 - 749) of Cft1 in contacting RNA directly (Dichtl et al. 2002). This region lies within the BP2 of Cft1 in the cryo-EM structure. Visualization of the electrostatic surface potential of BP2 revealed a positively charged patch enriched in lysine, phenylalanine, and arginine (Figure 2.20c).

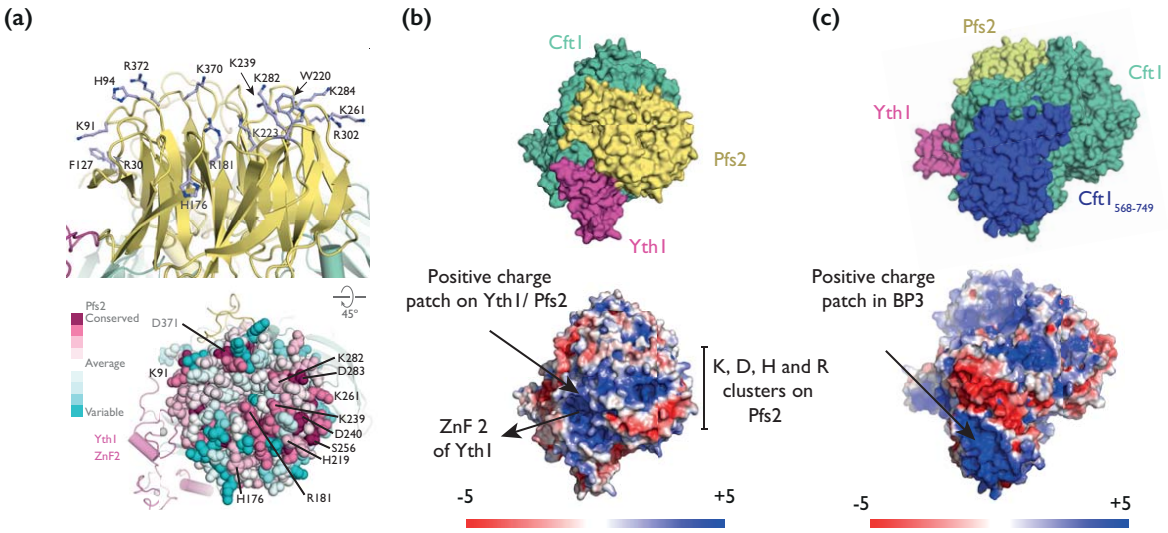


Figure 2.20: Potential RNA binding regions in Cft1–Pfs2–Yth1 complex of the polymerase module. **(a)** The top region of Pfs2 has several surface exposed lysines, arginines, histidines and phenylalanines. The putative RNA binding residues in Pfs2 are also conserved across 35 Pfs2 orthologs. **(b)** Electrostatic surface potential of the Cft1–Pfs2–Yth1 reveal a patch of positively charged residues on top of Pfs2 and around zinc finger 2 of Yth1. **(c)** The region in BP3 of Cft1 previously shown to bind RNA (Dichtl and Keller 2001) is coloured in blue in the surface representation. Interestingly, electrostatic surface potential reveal a stretch of positively charged residues in BP3 that might potentially contact RNA. Electrostatic surface potential is plotted onto the solvent-accessible surface (blue is positive, red is negative, in the range ± 5 kT/e).

2.4.6 Comparing the structures of the yeast and human polymerase module subunits

The cryo-EM structure of the human equivalent of the polymerase module bound to an AAUAAA-containing RNA was recently determined (Sun et al. 2018; Clerici et al. 2018, 2017). The structure revealed four beta propellers in a tripartite arrangement similar to Cft1-Pfs2. Structural alignment of the human polymerase components with the yeast subunits revealed a high degree of similarity (Figure 2.21b). In contrast to the structure of the yeast polymerase module, the RNA-bound structure also contained the third zinc finger of CPSF30 in addition to ZnF1 and ZnF2 (Figure 2.21a). It is likely that the presence of RNA makes ZnF3 less flexible, allowing the zinc finger to be visualized. The structures also reveal that the ribophosphate backbone of the RNA substrate adopts an S-shaped structure. The residues A1, A4, and A5 in AAUAAA make base specific contacts with residues in CPSF30 and WDR33 (Figure 2.21c). Interestingly, these amino acid residues are also conserved in Yth1. Of note, U3 and A6 form a non-canonical Hoogsteen base pair, and are sandwiched by π- π stacking between F153 and F43 of WDR33 (Figure 2.21d). Nonetheless, the residue F43 is not conserved in Pfs2, providing a potential explanation for why yeast-positioning elements differ from the human AAUAAA.

In summary, the human structure bound to RNA provides insights into the molecular mechanism behind how AAUAAA is specifically recognized by the human polymerase module. Moreover, it also suggests how the yeast and human machinery, despite having a similar overall architecture, might have distinct structural features that are involved in RNA recognition.

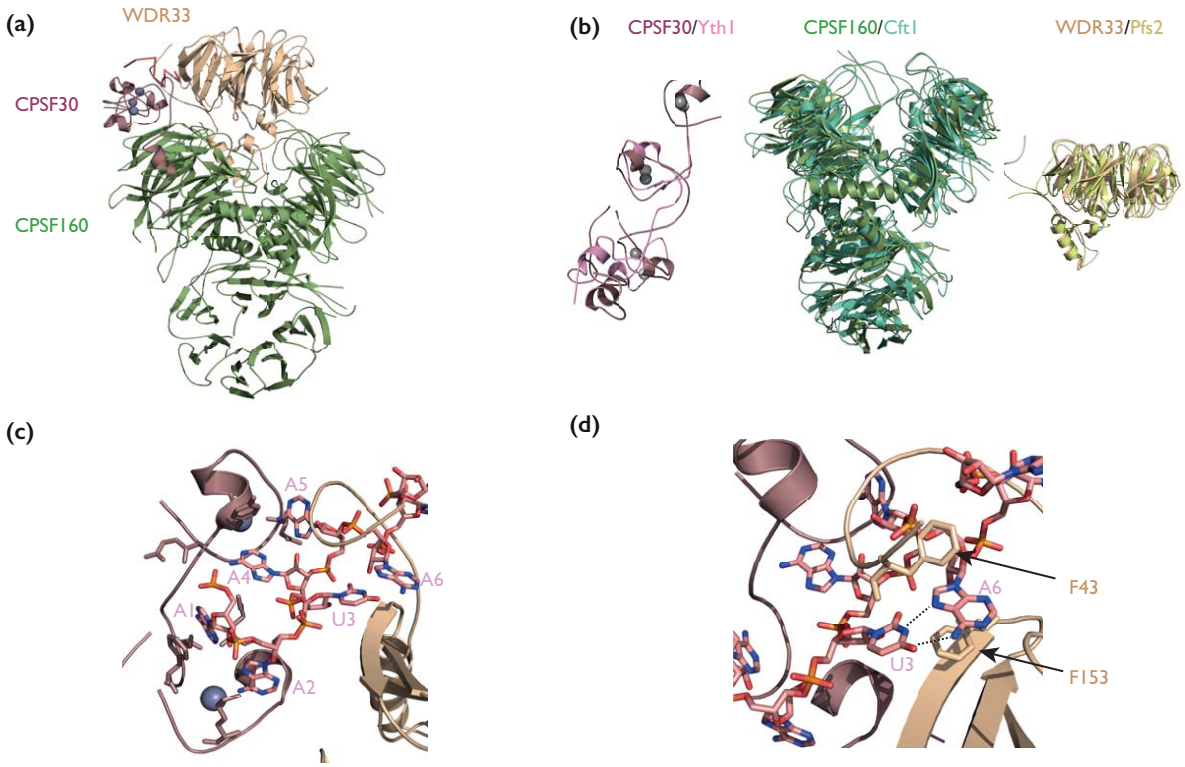


Figure 2.21: Structure of the human polymerase module. **(a)** Overall structure of CPSF160 - WDR33- CPSF30 bound to an AAUAAA containing RNA **(b)** Structural alignment of the polymerase module subunits from humans and yeast **(c)** AAUAAA recognition by CPSF30 and WDR33 **(d)** U3 and A6 forms a Hoogsteen base pair and is sandwiched between F153 and F43 of WDR33. CPSF160 in smudge, WDR33 in wheat, CPSF30 in dirty violet, RNA in pink and zinc ions in grey.

2.4.7 A mechanism for disruption of mRNA 3' end processing by influenza protein NS1A

It is known that the influenza viral protein NS1A binds to host CPSF30 and inhibits the production of antiviral mRNAs. Shutting down normal host 3' end processing is also beneficial to the viral life cycle. A crystal structure of the C-terminal domain of the NS1A protein bound to human CPSF30 reveals the molecular details of this interaction, including a conserved CPSF30 binding pocket in NSIA (Das et al. 2008). Although it is known that NS1A blocks RNA binding by CPSF30, the exact molecular mechanism behind this remains unclear. The high sequence and structural similarities between the human and yeast polymerase module subunits enabled us to hypothesize how NS1A hijacks the 3' end processing machinery. Superposing the crystal structure of CPSF30-NS1A CTD on the Cft1-Pfs2-Yth1 cryo-EM structure revealed that NS1A may dislodge Yth1/CPSF30 from Pfs2/WDR33, as binding of Yth1/CPSF30 to NS1A is mutually exclusive with binding to Pfs2/WDR33. Furthermore, NS1A binding could obstruct the proposed composite RNA binding site on the polymerase module (Figure 2.22). Interestingly, the N-terminal RNA binding domain (NRD) is connected to the C-terminal domain via a flexible linker. It is plausible that, upon binding the polymerase module, NRD of NS1A could sequester the host mRNA and hence result in inhibition of mRNA processing.

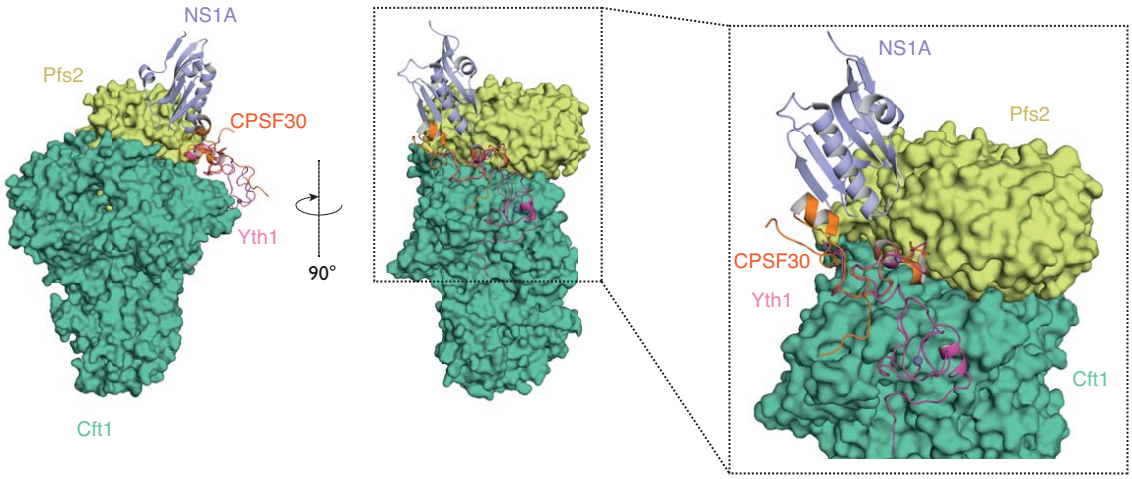


Figure 2.22: NS1A hijacks the 3'-end processing machinery. Docking the crystal structure of a complex of the influenza protein NS1A with the zinc fingers 2 and 3 of human CPSF30 into the polymerase module structure. The zinc finger 2 of CPSF30 and Yth1 superpose well on top of each other (yeast and human structures superimposed in magenta and orange, respectively). Yth1 (magenta), CPSF30 (orange), NS1A (light purple), Pfs2 (yellow), Cft1 (green), and zinc ions (dark grey) are depicted.

2.5 Pap1 is flexibly tethered to the polymerase module

The cryo-EM structure of the Cft1-Pfs2-Yth1 complex reveals for the first time how the RNA binding subunits of the polymerase module are arranged in 3D space. However, we don't yet have a detailed understanding of how Pap1 associates with the polymerase module and how the interaction influences its function. In order to characterize the molecular topology of polymerase module and its interconnection with Pap, I used an integrative structural biology approach as described below.

2.5.1 Cryo-EM demonstrates the flexible association of Pap1

In order to understand how Pap1 interacts with the polymerase module, I studied the "peak 1" fraction containing polymerase module (with Pap1) by cryo-EM (Section 3.1.1). This fraction was predicted to have one stably-associated Pap1 subunit (Sections 3.1.1 and 3.2.1). Cryo-EM micrograph of polymerase module containing Pap1 shows well-separated, homogeneous particles (Figure 2.23a). 2D classification revealed polymerase module class averages with high resolution-information; in particular, the central Cft1-Pfs2-Yth1 core of the polymerase module is clearly visible (Figure 2.23b). Interestingly, some 2D classes (red) contain an additional horseshoe shaped density, characteristic of the structure of Pap1. Such a characteristic density is seen in the 2D projection of the 3D crystal structure of Pap1 (blue) (Figure 2.23c). Interestingly, Pap1 appears to be flexible with respect to the central Cft1-Pfs2-Yth1 density, as it is not found at a consistent location relative to the central core. Comparison of the 2D class averages of the polymerase module with and without Pap1 further supports the flexible nature of Pap1 (Figure 2.23d).

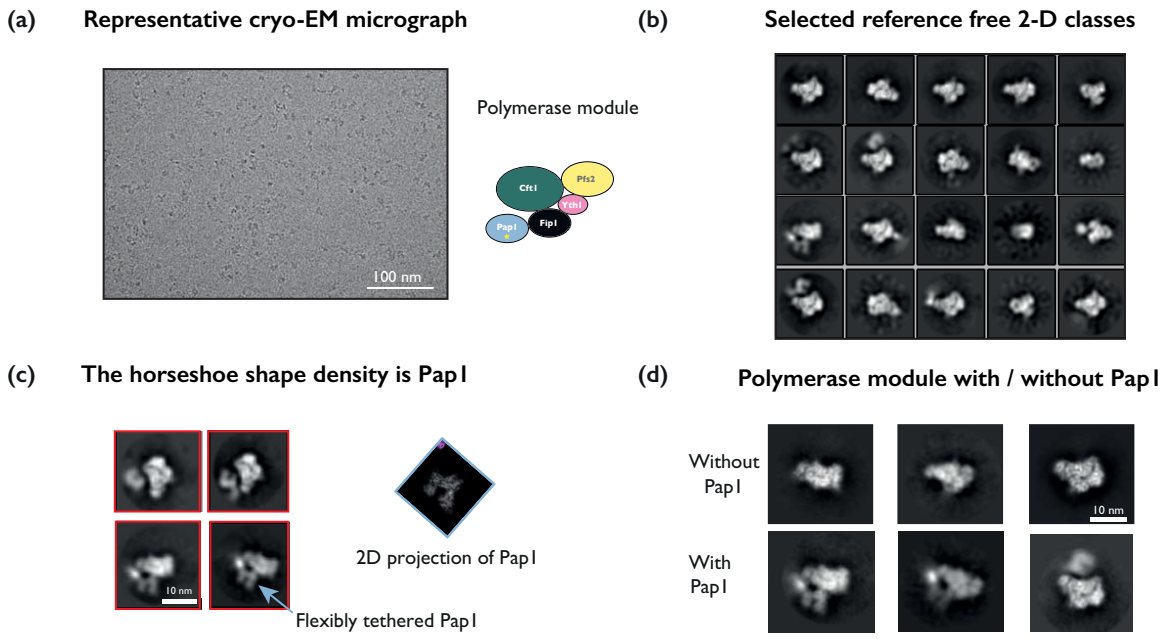


Figure 2.23: Cryo-EM analysis of the polymerase module containing Pap1. **(a)** Representative micrograph of the polymerase module containing Pap1. **(b)** Selected 2D class averages showing high-resolution secondary structural details and a variety of views. **(c)** A characteristic horseshoe shaped Pap1 density is seen near the polymerase module. The 2D classes reveal that Pap1 is flexibly tethered to the polymerase module (shown by blue arrow). **(d)** Comparing 2D classes with similar views of polymerase module without Pap1 and with Pap1.

Unfortunately, only ~ 0.5 % of all particles contained additional density that corresponds to Pap1. It is possible that Pap1 is present in other particles but is obscured by the core of the polymerase module. Alternatively, the flexibility of Pap1 could mean that most particles containing Pap1 do not average together to yield a 2D class with clear Pap1 density. Another possible explanation could be that Pap1 interaction is disrupted by the sample preparation process. In order to overcome the problem of Pap1 flexibility, I crosslinked the polymerase module containing two Pap1 subunits using BS3. The polymerase module containing two Pap1 subunits was used as the concentration of Pap1 would be higher in the sample, increasing the probability of obtaining Pap1-associated polymerase module. The cross-linked sample was further purified on a size exclusion column (Figure 2.24a) and eluted fractions were analyzed by tris-acetate PAGE (Figure 2.24b). The chromatogram revealed that the un-

crosslinked and crosslinked samples were eluted at the same volume, suggesting that crosslinking did not result in the aggregation or formation of oligomers of the polymerase module.

I studied the crosslinked sample by cryo-EM. Cryo-EM micrographs of the cross-linked polymerase module (with two Pap1s) showed significantly bigger particles (Figure 2.24c). 2D classification of selected particles resulted in class averages containing high-resolution information. Upon further analysis, the 2D class averages appear to consist of a dimer of the polymerase module (red arrows in Figure 2.24d). The 2D classes also contain density that was not well resolved (blurred density) at different positions along the polymerase module (green arrows in Figure 2.24d). Similar to the un-crosslinked sample, it is possible that Fip1 and Pap1 are present at different positions with respect to the core of the polymerase module, resulting in their densities being averaged out during 2D classification.

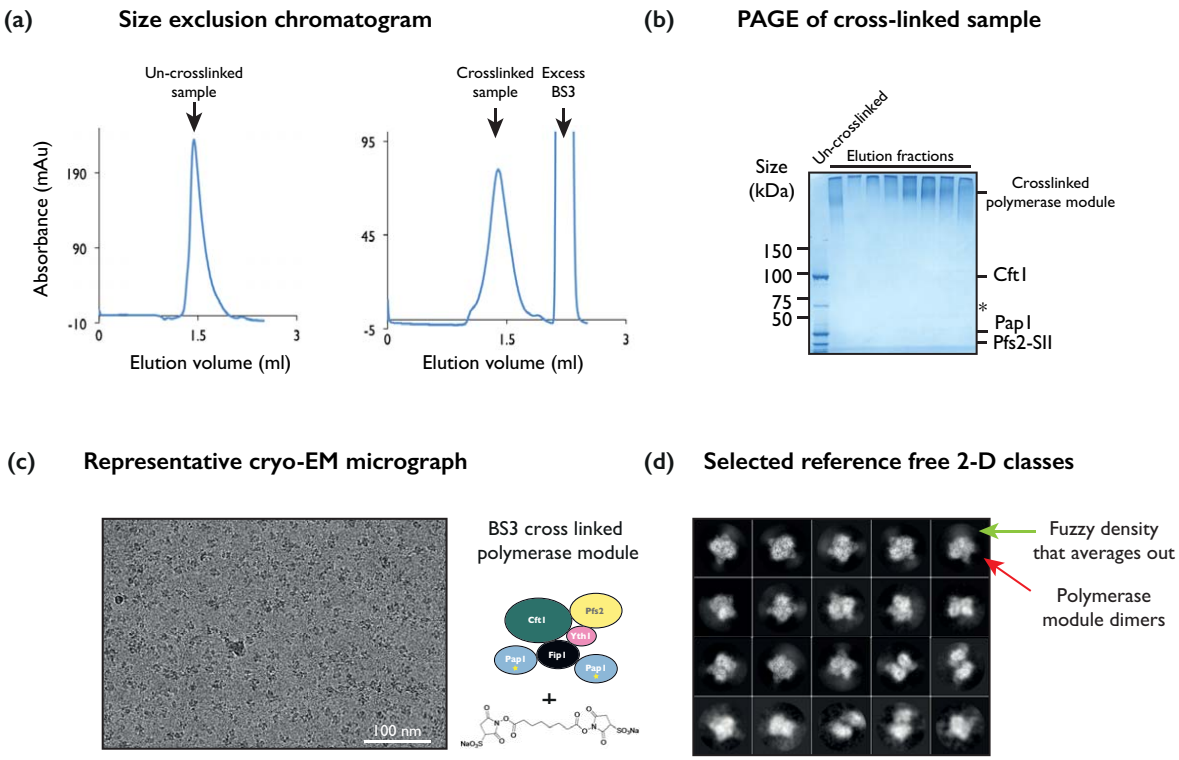


Figure 2.24: Cryo-EM analysis of cross-linked polymerase module containing two Pap1s. **(a)** Size exclusion chromatography of the cross-linked polymerase module with two Pap1s. The sample runs at similar elution volume compared to the un-crosslinked complex. **(b)** Tris-acetate-PAGE of cross-linked polymerase module containing two Pap1s. The high molecular band at ~ 330 kDa contains cross-linked polymerase module. * represents a possible Cft1 degradation band. **(c)** Representative micrograph of a cross-linked sample of the polymerase module containing Pap1. **(d)** Selected 2D class averages showing high-resolution secondary structural details and a variety of views. Red arrow indicates possible polymerase module dimers. Green arrow indicates an extra region of fuzzy density.

Thus, preliminary cryo-EM analysis suggests that Fip1 and Pap1 are flexible relative to the stable polymerase module core. Further biochemical optimization is likely required before further analysis of the Pap1-bound polymerase module by cryo-EM.

2.5.2 Cross-linking mass spectrometry reveals extensive inter-subunit interaction

In order to interrogate the architecture of a Pap1-bound polymerase module, the complex (with and without Pap1) was crosslinked using BS3, which covalently links lysines, serines, threonines or tyrosines that are within 27 Å of each other. The observed inter- and intra-molecular crosslinks were in agreement with the atomic models for Cft1, Yth1 and Pfs2 (Figure 2.25a, appendix 8.7). The intra-molecular cross-links found within Pap1 are also consistent with the crystal structure of Pap1 (Bard et al. 2000) (Figure 2.25a).

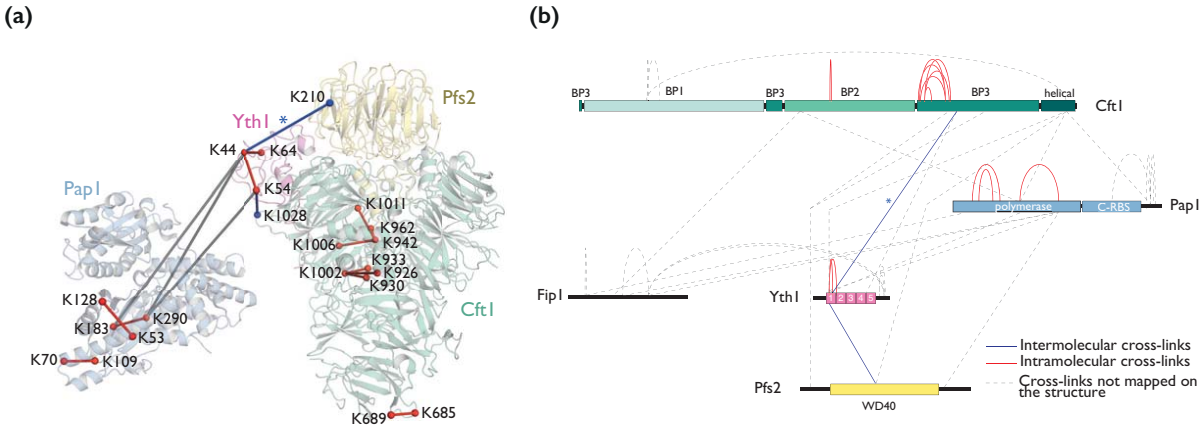


Figure 2.25: Cross-linking mass spectrometry of the polymerase module. **(a)** Lysine-lysine intra- (red) and inter- (blue) protein cross-links were mapped onto the cryo-EM structure of Cft1–Pfs2–Yth1 and the crystal structure of Pap1 (PDB:2Q66). The cross-link marked with an asterisk is 33 Å, but has a similar xQuest (Leitner, Walzthoeni, and Aebersold 2014) score to other validated cross-links and appears reasonable upon inspection of the structure. Lysine-lysine cross-links between Pap1 and polymerase module are shown in grey. **(b)** Linkage map showing all identified lysine-lysine crosslinks. Intermolecular cross-links are shown in blue; intramolecular in red and the cross-links that could not be mapped on any available 3D structures are shown by a dotted line.

Inspection of the crosslinking data revealed that Fip1 K219 crosslinks to Yth1 K182, K191 and K196 (Figure 2.25b), in agreement with previous work where Fip1 residues 206–220 were shown to bind to the C-terminal half of Yth1 (Barabino, Ohnacker, and Keller 2000b; Helmling, Zhelkovsky, and Moore 2001; Tacahashi, Helmling, and Moore 2003). Fip1 also crosslinks with the polymerase domain of Pap1 (Figure 2.25b). A previous crystallographic study has identified the C-terminal domain of Pap1 as Fip1 binding (Meinke et al. 2008). Interestingly, Pap1 crosslinked with all other polymerase module subunits, including Fip1, the C-terminal helical domain of Cft1, zinc finger 1 of Yth1, and the C-terminal region of Pfs2. This agrees with previous data suggesting a close association of Pap1 with Cft1 and Fip1 (Easter 2014), and with previous literature (Murthy and Manley 1992; Meinke et al. 2008). It is likely that the observed crosslinks of Pap1 with all subunits of polymerase module reflects the flexibility of Pap1 relative to the central core, and that chemical cross-linking captures the transient interactions of Pap1. Thus suggesting that the Pap1 interface with the complex is more extensive.

Thus, the data obtained by crosslinking mass spectrometry support both our model of the polymerase module core, as well as previous models of how Fip1 and Pap1 are associated with the complex (Ezeokonkwo et al. 2011; Helmling, Zhelkovsky, and Moore 2001; Tacahashi, Helmling, and Moore 2003; Barabino, Ohnacker, and Keller 2000a; Barabino et al. 1997). The promiscuous crosslinking of Pap1 also supports my observation that Pap1 appears to be flexible relative to the polymerase module core.

2.5.3 Pull downs reveal the molecular topology of the polymerase module

To clarify the nature of the interaction of Pap1 with the polymerase module core, and to dissect the contributions of Fip1 and Yth1 to this interaction, I carried out pull-downs of the polymerase module containing truncated versions of Fip1 and Yth1. The truncations include deletion of the C-terminal 91 residues in Yth1 (referred to as Yth1 ΔZnF45C), residues 145 - 170 in Fip1 (referred to as Fip1Δ145-170), and the potential Yth1 binding region in Fip1 (referred to as Fip1 Δ180-220) (Figure 2.26a).

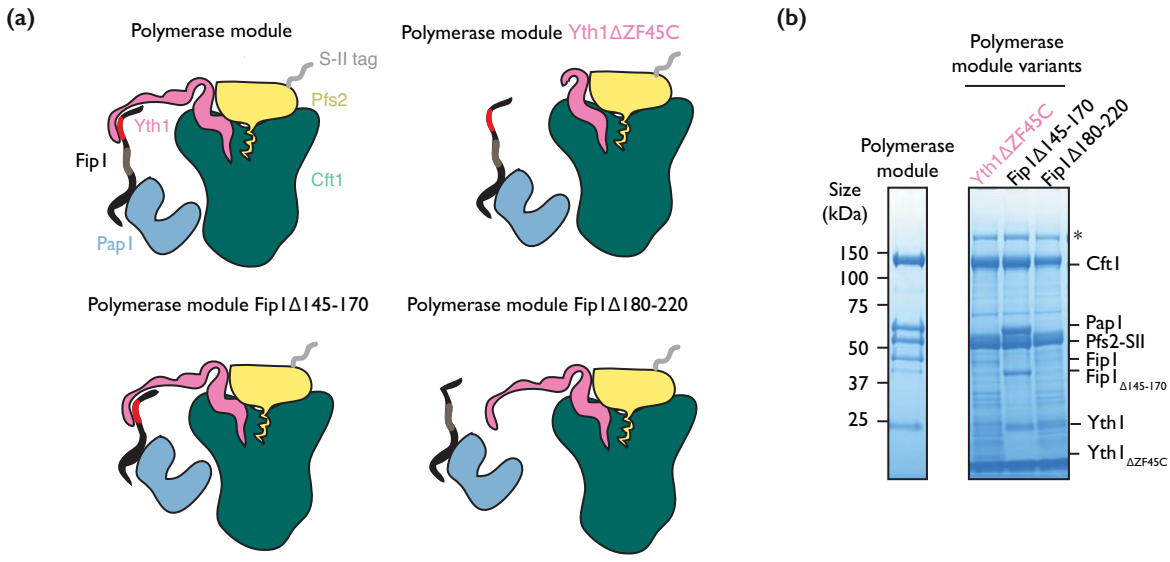


Figure 2.26: Physical interactions linking Pap1 to the polymerase module. **(a)** A cartoon representation of the various mutants that were employed in the pull downs. The red region in Fip1 highlights the potential interaction interface with Yth1. **(b)** Pull down assay of polymerase module constructs harboring truncations in Fip1 and Yth1. The polymerase module pull downs were carried via the S-II tag on Pfs2 subunit using streptavidin resin and the elution was analyzed by SDS-PAGE.

All polymerase module constructs were expressed in *Sf9* insect cells using baculovirus-mediated insect cell over-expression. Streptavidin pull downs were performed using the Strep-II tag in the Pfs2 subunit and the isolated complex from *Sf9* cell lysate was analyzed on SDS-PAGE (Figure 2.26b). As expected, the elution fraction from the wild type complex contains all five polymerase subunits. However, deletion of the C-

terminal 91 residues in Yth1 yields a polymerase module complex without any Fip1 or Pap1 (Figure 2.26b). Similarly, deletion of Fip1 residues 180-220 did not pull-down Fip1 or Pap1 (Figure 2.26b). Interestingly, deletion of Fip1 residues 145-170 did not affect association of Fip1 or Pap1 with the complex (Figure 2.26b). A recent study on the human polymerase complex also identified a conserved domain in FIP1 that associates with the polymerase module via its interaction with ZnF45 of CPSF30 (Clerici et al. 2018). In agreement with this, the residues 180-220 of yeast Fip1 lie within this conserved domain in FIP1 whereas the residues 145-170 lie outside this domain.

In summary, pull-down experiments have revealed the molecular requirements of Yth1, Fip1, and Pap1 association with the polymerase module core. Specifically, the C-terminal 91 residues of Yth1 and a central region of Fip1 (180-220) appear to be necessary for the stable association of Fip1 and Pap1 to the polymerase module core.

2.5.4 Fip1 directly contacts Zinc finger 4 of Yth1

Previous nanoESI-MS studies from the lab discovered that endogenous CPF purified from *S. cerevisiae* contains multiple copies of Pap1 (Casañal et al. 2017b). Similarly, the recombinantly expressed and purified polymerase module consisted of three variants containing two, one, or no copies of Pap1 (Figure 2.2b and Figure 2.5). While the Pap1 binding site in Fip1 has been well characterized and lies within residues 90-105 (Helmling, Zhelkovsky, and Moore 2001), it remains unknown where the additional copy of Pap1 binds in the complex containing two copies. Furthermore, nanoESI-MS data also revealed that CPF containing two copies of Pap1 always contained two copies of Fip1 (Casañal et al. 2017b). It is therefore plausible that Yth1 harbors two binding sites for Fip1. In order to investigate the molecular details of the Fip1-Yth1 interaction, I used NMR spectroscopy in collaboration with Conny Yu.

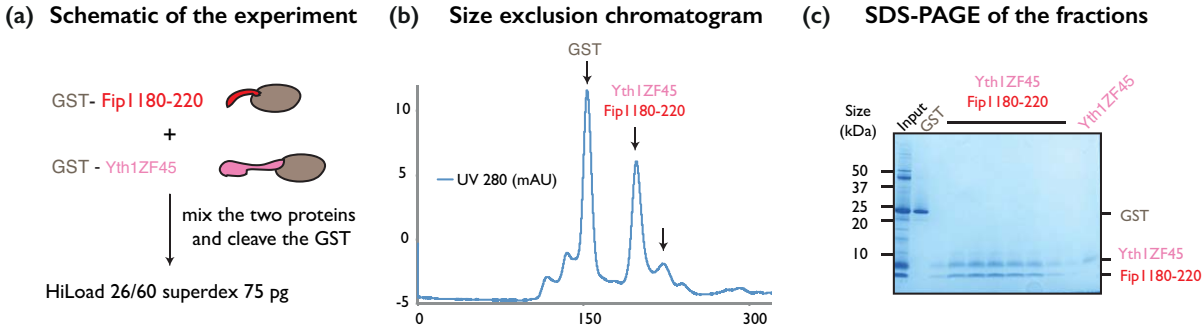


Figure 2.27: Characterization of a minimal interaction surface between Yth1 and Fip1. **(a)** Schematic for the formation of a minimal Yth1-Fip1 complex. A construct of Yth1 ZF45 with a N-terminal GST tag and a construct of the residues 180-220 of Fip1 with an N-terminal GST tag are mixed, the GST tags are cleaved by the use of precission protease, and the mixture is subjected to size exclusion chromatography. **(b)** Size exclusion chromatogram (HiLoad 26/600 Superdex 200 pg) of a complex of ZF 4 and 5 of Yth1, with residues 180 - 220 of Fip1. **(c)** SDS -PAGE of the peak fraction reveals a stoichiometric complex formation.

Fip1 residues 180-220, which were shown to be important in tethering Fip1 to the polymerase module core (Section 3.5.3), were fused to an N-terminal GST tag. The ZnF45 of Yth1 were separately fused to an N-terminal GST tag. These two protein

constructs were overexpressed in *E. coli* and purified by affinity chromatography. Purified GST-Fip1 180-220 was then mixed with purified GST-Yth1 ZnF45, and the GST tags on the two proteins were cleaved (Figure 2.27a). The resulting sample was subjected to size-exclusion chromatography to resolve the cleaved GST and the remaining proteins (Figure 2.27b). Analysis of the three resulting peaks by SDS-PAGE revealed that the first peak contained GST alone, the second contained a complex of Fip1 180-220 and Yth1 ZnF45, and the third contains excess Yth1 ZnF45 (Figure 2.27c). SEC-MALS analysis revealed that this was a 1:1 complex with one copy of Yth1 ZnF 45 bound to one copy of Fip1 180-220 (appendix 8.8).

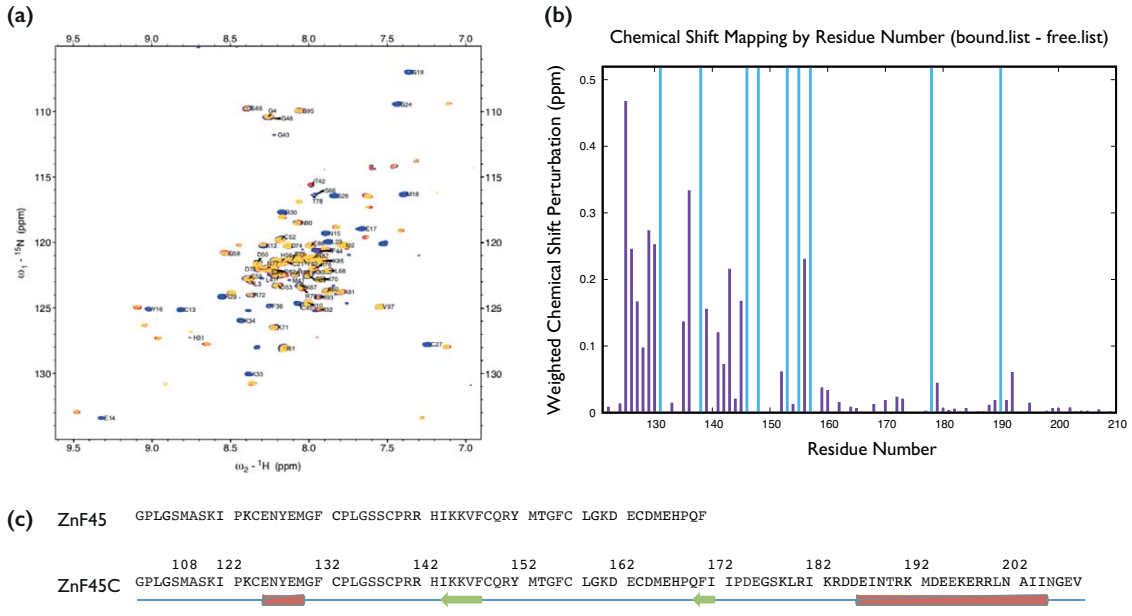


Figure 2.28: Yth1 ZnF 45C shows major chemical shift perturbation upon Fip1 binding. **(a)** Overlay of 1H-15N 2D BEST-TROSY spectra of Yth1 (118-208) in its free from (blue) and upon adding Fip1 (180-220) addition. Fip1 was added in equal molar ratio (red) and in 2:1 excess (yellow). The labelled cross-peaks showed the backbone assignment of free Yth1. **(b)** Weighted chemical shift perturbation in (a) was mapped against residue number of Yth1. The identity of cross-peaks in Yth1/Fip1 spectrum was determined by mapping and partial assignment. Cross-peaks with line broadening upon Fip1 addition was indicated with a blue bar. **(c)** Sequence of Yth1 (118-170) and Yth1 (118-208) and secondary structure of free Yth1 (118-208) calculated by TALOS (Shen et al. 2009) based on chemical shifts of C_α, C_β and C_O.

Crosslinking mass spectrometry of the polymerase module had identified crosslinks between Fip1 and the C-terminus of Yth1 (after ZnF45) (Figure 2.25.B, appendix 8.7). Thus, I hypothesized that the C-terminal ~50 residues after ZnF5 could also associate with Fip1, providing a second binding site. NMR spectroscopy was used to map the residues in Yth1 involved in Fip1 binding.

Backbone assignment of Yth1 ZnF45C (residues 188-208) was carried out using a ^{13}C, ^{15}N labeled sample at 278K to achieve a complete assignment. The assignment was then transferred to 298K for binding studies at this temperature. Major chemical shift perturbations were observed in Yth1 ZnF domain 4 (residues 122-160) upon the addition of equal molar Fip1 180 - 220, with weighted chemical shifts larger than 0.1 ppm. The largest chemical shift differences occur in 124-131, which coincides a helical region identified from secondary structure calculations based on chemical shifts. It is speculated that upon Fip1 binding, there is a significant change in the chemical environment around Yth1 residues 124-131, highlighting that these residues may be essential for Yth1/Fip1 interactions.

To address the stoichiometry of Yth1/Fip1 complex, excess Fip1 180 - 220 was titrated into Yth1 ZnF45C and no further perturbation of chemical shifts were observed. This concludes that there is no second binding site for Fip1 180 - 220 on Yth1 ZnF45C, and it agrees with SEC-MALS data that two proteins bind in 1:1 stoichiometry.

2.6 Discussion

In order to biochemically and structurally characterize the polymerase module, I established protocols for the recombinant expression and purification for the protein complex. *In vitro* reconstitution of the polyadenylation reaction with pure components revealed that the polymerase module surprisingly had similar activity compared to the individual polymerase (Figure 2.6 and 3.7). Moreover, Pap1 appeared to add longer poly(A) tails compared to the polymerase module. As previously mentioned, this is in contrast to previous findings in humans where the polymerase module is efficient and specific in polyadenylation in comparison to PAP (Schönemann et al. 2014), likely due to an increased RNA binding affinity (~2 nM) for an AAUUAA-containing RNA (Clerici et al. 2018; Sun et al. 2018) relative to PAP. From the EMSAs performed in this chapter, the apparent affinity of the yeast polymerase module for *CYC1* RNA is between 500 nM to 1 μM (Figure 2.8). Therefore, the relatively low affinity of the polymerase module for RNA provides an explanation for why its activity is not higher than that of Pap1 alone. Compared to an *in vivo* scenario, it is likely that the yeast complex is missing additional RNA binding surfaces from other CPF subunits or cleavage factors. The association of the polymerase module with additional RNA binding proteins may stimulate polyadenylation activity, in comparison with Pap1 alone. Further experiments addressing this question are described in Chapter 3. Interestingly, the yeast polymerase module bound the adenovirus L3 RNA with lower affinity in EMSAs (Figure 2.8). Therefore, it is seen that the RNA binding subunits within the polymerase module can impart a degree of specificity to the complex. The recombinant production of high amounts of polymerase module described in this chapter will now allow us to perform biophysical experiments to characterize the kinetics of RNA binding and the implication on polyadenylation, relative to the isolated polymerase. Furthermore, other yeast RNAs with less idealized efficiency, positioning, and U-rich elements should be tested in polyadenylation and binding assays. This will allow us to observe whether the low affinity, distributive activity is conserved among other RNA species.

The cryo-EM structure of Cft1-Pfs2-Yth1 subunits revealed, for the first time, the overall architecture of the polymerase module and the interactions between the three proteins. Cft1 and Pfs2 share an extensive interaction surface; most notably, an N-terminal alpha helix of Pfs2 insets into a cavity formed between two beta propellers in

Cft1 (Figure 2.16). An N-terminal loop of Yth1 inserts into the cavity of one of the beta propellers of Cft1, anchored by several ionic as well as hydrophobic interactions (Figure 2.17). The first two zinc fingers in Yth1 also associate at the interface of Cft1 and Pfs2. Furthermore, the structure validates previous biochemical and genetic data highlighting the importance of the amino acid regions, which we show to be important in inter-protein interactions (Ohnacker et al. 2000; Barabino, Ohnacker, and Keller 2000b; Tacahashi, Helmling, and Moore 2003). The structure of the polymerase module has also revealed that the complex forms a structural platform for other polymerase module subunits, bringing together the flexibly tethered Fip1 and Pap1 as well as, potentially, RNA, via a composite binding surface formed by ZnF23 and the top of the Pfs2 beta propeller domain (Figure 2.29b).

During the course of this study, the structure of the human CPSF160-WDR33-CPSF30 was reported (Sun et al. 2018; Clerici et al. 2018, 2017). The yeast and the human complexes share a high degree of structural and sequence similarity (Figure 2.21). The structure of the human complex was determined in complex with RNA, thus providing molecular insight into how the mammalian poly(A) site AAUAAA is recognized by CPSF. The structure also provides a potential explanation for why the yeast RNA positioning element does not contain AAUAAA and is more degenerate compared to mammalian elements. The structure of the yeast polymerase module bound to RNA will provide further insights into RNA recognition and enable better understanding of the difference between the yeast and human RNA *cis*-elements.

Intriguingly, the structure of Cft1-Pfs2 is highly similar to other nucleic acid binding complexes such as DDB1-DDB2 and SF3b. In particular, the mode of interaction by helix insertion into a cavity lined by beta propellers (seen in the Pfs2-Cft1 and DDB2-DDB1 interactions) appears to be conserved. This suggests a shared evolutionary origin for these complexes.

Cryo-EM of the polymerase module containing Pap1 provides evidence for the flexible tethering of Pap1 with the polymerase module (Figure 2.23). Furthermore, crosslinking mass spectrometry of the complex shows that Pap1 crosslinks with all subunits of polymerase module, highlighting the dynamic nature of Pap1 (Figure 2.25). Docking the crystal structure of Pap1 (PDB ID: 1FAO) onto the cryo-EM structure of Cft1 using HADDOCK (van Zundert et al. 2016; Cyril Dominguez, Rolf Boelens, and Bonvin

2003), and using cross-linking mass spectrometry data as restraints, shows that Pap1 binds to the BP2 and BP3 of Cft1 (Figure 2.29a). In the context of our cryo-EM and crosslinking mass spectrometry data, it is more likely that Pap1 does not interact strongly with Cft1, but rather lies adjacent to it in 3D space.

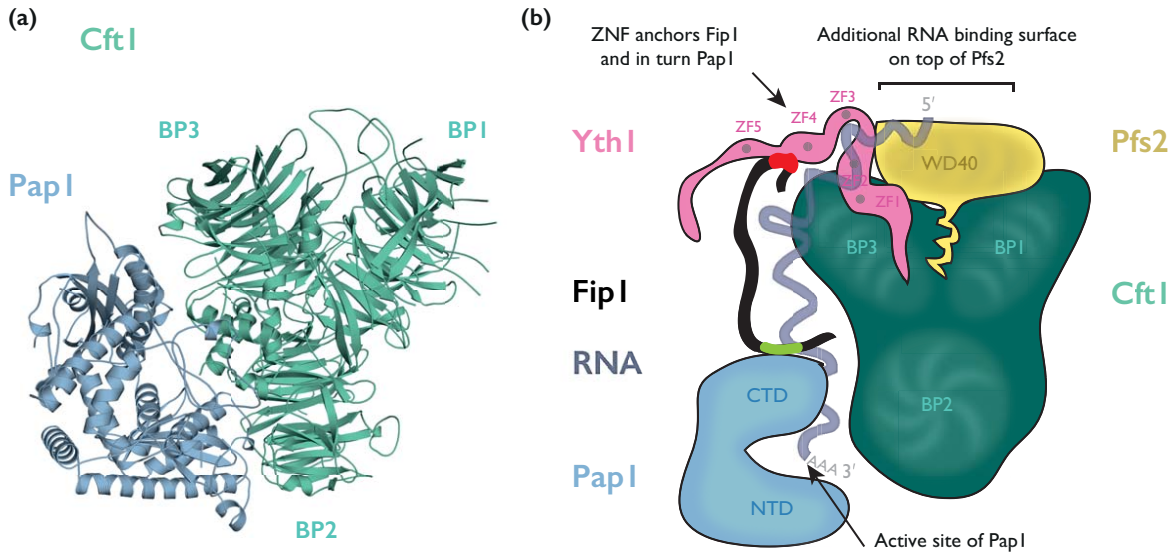


Figure 2.29: A model for the architecture of the polymerase module. **(a)** A model of the crystal structure of Pap1 docked onto the cryo-EM structure of Cft1 using HADDOCK. Data from cross-linking mass spectrometry experiment were given as primary interaction restraints while running HADDOCK. **(b)** A cartoon representation of the topology and architecture of the polymerase module showing the inter-protein interactions. The individual domains within the proteins that are involved in the interactions are highlighted. Fip1 (black), Pap1 (blue), Yth1 (magenta), Pfs2 (yellow), Cft1 (dark green), and zinc ions (dark grey) are depicted. The Yth1 binding site in Fip1 is depicted in red, the Pap1 binding site in Fip1 is shown in green. The RNA is depicted in transparent neon.

Pull-down experiments of the polymerase module using truncated constructs of Fip1 and Yth1 reveal that Pap1 is anchored to the polymerase module via Fip1 (Figure 2.26). Disruption of the interaction between Fip1 and Yth1 causes Pap1 to disassociate from the complex. An earlier study showed that replacing the unstructured flexible linker in Fip1 with that of human replication protein A resulted in reduction in the efficiency of

polyadenylation *in vitro* (Ezeokonkwo et al. 2011), highlighting the importance of the flexible linker of Fip1 in Pap1 activity. It should be noted that this region is known to mediate interactions with other components of the 3' end-processing machinery, including CF IA (Helmling, Zhelkovsky, and Moore 2001; Ezeokonkwo et al. 2011). Therefore, it is imperative to study the effects of the flexible linker of Fip1 on polyadenylation in the context of the complete 3′-end processing machinery.

Previous studies from our lab discovered the presence of multiple copies of the polymerase Pap1 in endogenously-purified CPF (Easter 2014). The molecular basis of the association of multiple polymerases to CPF, as well as the *in vivo* relevance of this complex, is unknown. Further analysis by mass spectrometry revealed that two copies of Pap1 are always accompanied by two copies of Fip1 in the complex (Easter 2014; Casañal et al. 2017b). Therefore, it is possible that there are two Fip1 binding sites in CPF. While Fip1 is anchored to the polymerase module via Yth1 (Figure 2.26), NMR experiments revealed that only ZnF4 of Yth1 binds Fip1 residues 180-220. Surprisingly, Fip1 binding did not cause any major chemical shift perturbation in the ZnF5 or the C-terminal end of Yth1, therefore suggesting that there is no additional binding site for Fip1 (Figure 2.28). Nonetheless, it is plausible that Fip1 sequences downstream of residue 220 might have interactions with additional surfaces in Yth1. Further experiments with various truncations of Fip1 and Yth1 are needed to characterize this interaction.

In this chapter, the structural architecture of the polymerase module was investigated using *in vitro* biochemistry, cryo-EM, mass spectrometry and NMR (Figure 2.29b). Given that WD40 domains are frequently found as protein-protein interaction hubs (van der Voorn and Ploegh 1992), it is plausible that other components of the 3' end-processing machinery could interact with the beta-propellers of Pfs2/Cft1 or the unstructured regions in Fip1. The association of other CPF subunits or the cleavage factors with the polymerase module could also provide additional RNA binding surfaces, thereby increasing the binding affinity for RNA. The role of additional proteins in polyadenylation will be further discussed in Chapter 3.

3 CF IA stimulates polyadenylation by tethering CPF to RNA

Isolated Pap1 can add a poly(A) tail to an mRNA substrate *in vitro* in a manner similar to the polymerase module. However, in cells, CPF functions together with the cleavage factors IA and IB for specific and efficient polyadenylation (Gross and Moore 2001). Both CF IA and CF IB are essential for specific cleavage *in vitro* (Kessler, Zhao, and Moore 1996). Specifically, it has been shown that CF IB is important for mRNA cleavage site selection (Minvielle-Sebastia et al. 1998; Kim Guisbert, Li, and Guthrie 2006). It is thought that CF IB competes with CPF for RNA binding and hence restricts CPFs endonuclease activity to the poly(A) site (Dichtl and Keller 2001). NMR experiments have revealed that CF IB binds AU-rich efficiency elements, and the Rna15 subunit of CF IA binds to the neighbouring A-rich positioning elements (Leeper et al. 2010).

On the other hand, CF IA alone is sufficient to activate cleavage, but the exact mechanism behind how it activates cleavage remains unknown (Dichtl and Keller 2001). Specific RNA binding by the cleavage factors occurs as a result of multiple weak protein-RNA and protein-protein interactions (Noble et al. 2004; Leeper et al. 2010; Barnwal et al. 2012). Mutants defective in RNA binding of cleavage factors result in imprecise cleavage, and are lethal *in vivo* (Kessler et al. 1997; Gross and Moore 2001; Pérez-Cañadillas 2006). In mammals, both CstF (homologs of Rna14 and Rna15) and CF II_m (containing homologs of Pcf11 and Clp1) are interestingly required only for cleavage. In the absence of any cleavage factors, human CPSF is sufficient to reconstitute specific and efficient polyadenylation reaction *in vitro.*

Previous studies using purified individual CF IA subunits have highlighted the importance of CF IA in polyadenylation (Gross and Moore 2001). Using fractionated yeast extracts and purified subunits of CF IA and CF IB; the study showed that all five-cleavage factor subunits are required for reconstitution of *in vitro* polyadenylation activity. However, characterizing the exact nature of this interaction and its significance on polyadenylation without highly pure preparation of CPF is challenging. To investigate the role of yeast CF IA in polyadenylation, I carried out *in vitro* assays in the presence of cleavage factor sub-complexes. Next, I identified a direct interaction

between CPF and CF IA, mediated by Rna14/Rna15 and polymerase module subunits. Using a variety of techniques such as cross-linking mass spectrometry, hydrogen-deuterium exchange mass spectrometry, and cryo-EM, I characterize the dynamic interaction between Rna14/Rna15 and the polymerase module. Together, I propose a model for the stimulation of polyadenylation by CF IA and thus clarify the function of CF IA in mRNA polyadenylation.

3.1 Cleavage Factor CF IA Stimulates Polyadenylation by CPF

In order to reconstitute polyadenylation *in vitro* using components of the 3′-end processing machinery, I purified CPF and CF IA from endogenous and recombinant sources respectively. Native CPF was purified from a yeast strain harboring a C-terminal TAPS-tag (tandem-affinity purification using protein A and streptactin) in the *REF2* gene (Easter 2014). The TAPS tag consists of an N-terminal 3-C protease site followed by a Strep-II tag (SII), two TEV protease sites and two repeats of *S. aureus* protein A IgG-binding domain. The purification included two tandem-affinity steps. First, lysate was incubated with IgG beads to purify the complex by its protein A tag; the SII tag was then affinity purified using StrepTactin resin. Eluate from the StrepTactin pull-down was finally subjected to anion exchange chromatography to separate the APT complex from the fourteen-subunit CPF complex. The purification is described in detail in Ashley, 2014. The identities of the purified proteins were verified by tandem mass spectrometry. Cleavage factor IA subunits were recombinantly expressed in *E.coli* and purified using a protocol adapted from a previous study (Gordon et al. 2011) and described in section 6.4.3.

An *in vitro* polyadenylation assay of substrate RNA 5'-FAM-*CYC1*-pc was carried out using purified CPF, in the presence and absence of CF IA (Figure 3.1b). The reaction products were analysed using denaturing urea-PAGE. CPF adds poly(A) tail to a pre-cleaved RNA substrate in a distributive manner, similar to the assays performed using the polymerase module (Figure 3.1a and Figure 2.6a). Interestingly, the addition of recombinant CF IA to the reaction results in the stimulation of polyadenylation by CPF (Figure 2.6b). The stimulatory effect is particularly clear at earlier time points (t=2, 4, 8 min), where the band corresponding to RNA rapidly increases in size. Densitometry analysis of the urea-polyacrylamide gels revealed that the final length of the poly(A) tails is longer in the presence of CF IA (Figure 3.1). The two observations suggest that CF IA stimulates polyadenylation activity of CPF, potentially via a direct protein-protein interaction. Previous studies suggest that the WD-40 containing Pfs2 could interact with the HAT-domain containing Rna14 subunit of CF IA, linking the cleavage factors with CPF (Ohnacker et al. 2000). *In vitro* pull-downs have also shown a direct interaction between Fip1 and Cft1 of CPF, and Rna14 subunit of CF IA (Murthy and Manley 1992; Preker et al. 1995; Ezeokonkwo et al. 2011). The assays performed

above provide the first insight towards the functional relevance of the interactions between CF IA and CPF.

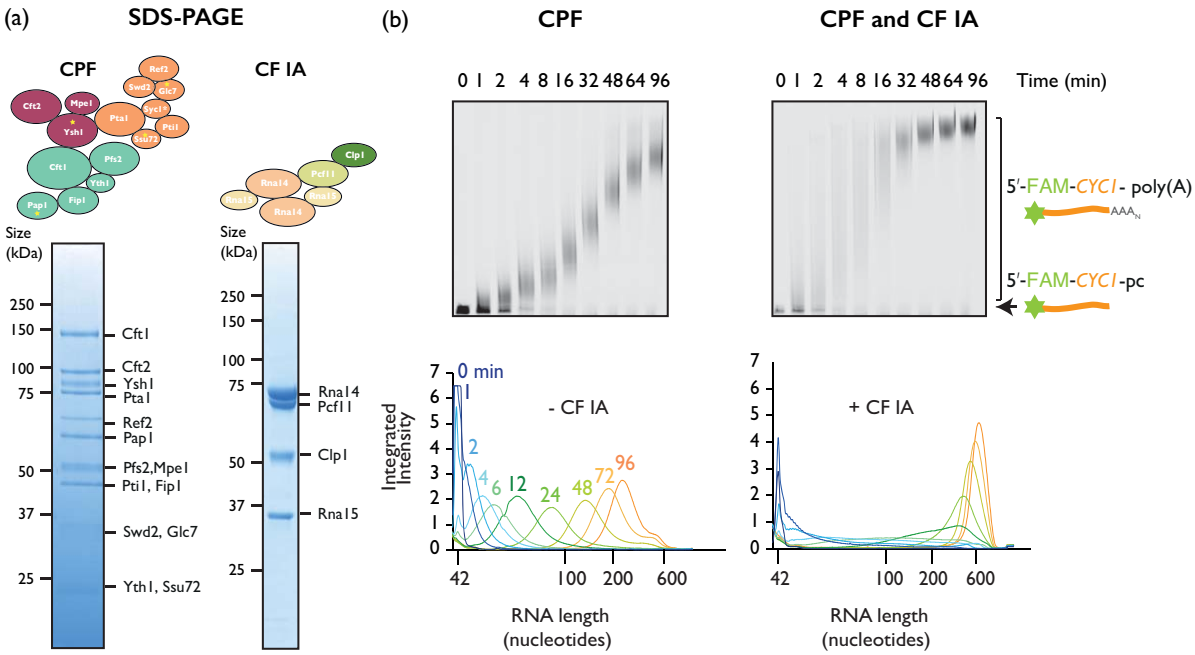


Figure 3.1: Cleavage factor IA stimulates polyadenylation by CPF. **(a)** SDS-PAGE of CPF and CF IA used in the assays **(b)** Polyadenylation assay of purified CPF analyzed by denaturing urea-PAGE. A 5′ fluorescently labeled 42 mer *CYC1* RNA that has been pre-cleaved is the substrate for polyadenylation. The assays were carried out with a final concentration of 400 nM substrate RNA, 50 nM purified CPF, 400 nM CF IA and 2 mM ATP. Densitometry analyses of the gels are plotted below; RNA length is marked based on an RNA ladder.

Notably, three out of the five-polymerase module subunits i.e. Fip1, Pfs2 and Cft1 have been shown to bind to CF IA in the past. I hypothesized that CF IA would be able to stimulate polyadenylation activity in a minimal reconstituted system containing the five-subunit polymerase module. To test this, I carried out *in vitro* polyadenylation assays of 5'-FAM-*CYC1*-pc using the polymerase module in the presence and absence of CF IA (Figure 3.2a and Figure 3.2b). In agreement with the hypothesis, CF IA can stimulate the activity of the polymerase module (Figure 3.2b). Densitometry analysis revealed that the poly(A) tails are longer and were added faster in the presence of CF IA (Figure 3.2b). CF IA consists of four subunits – a heterotetramer of Rna14/Rna15 and a

heterodimer of Pcf11–Clp1 (Gordon et al. 2011); among these, the Rna14/Rna15 heterotetramer mediates RNA binding by CF IA (Noble et al. 2004).

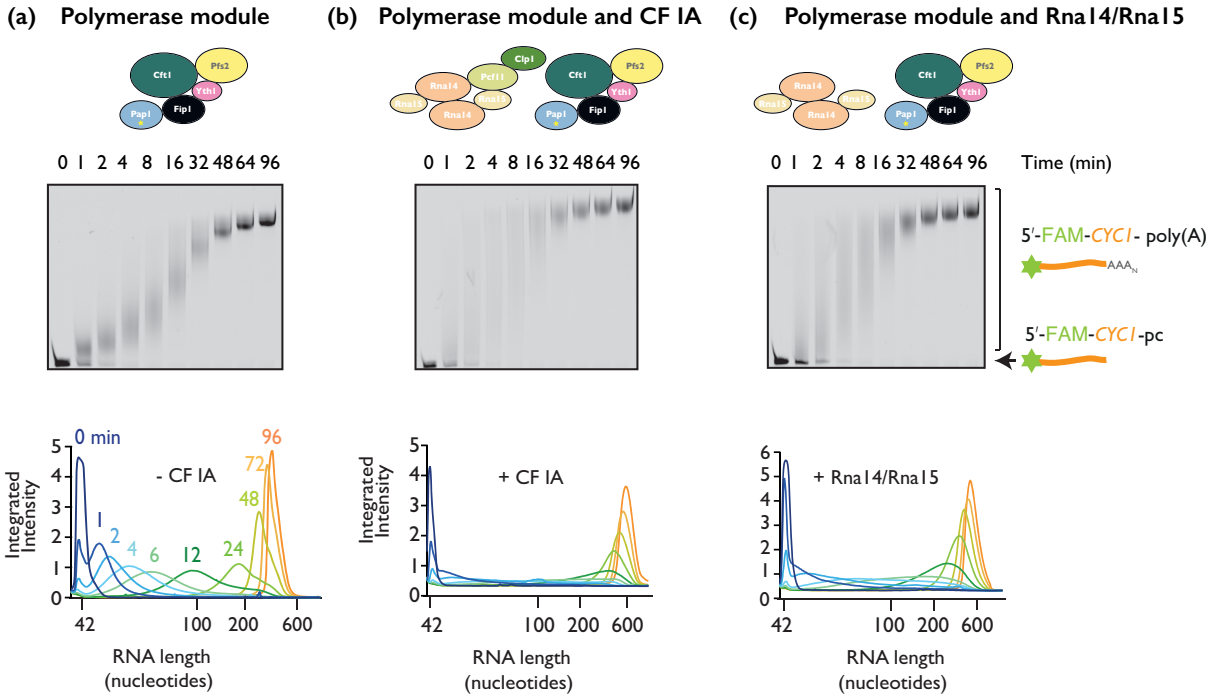


Figure 3.2: Cleavage factor IA stimulates polyadenylation by the polymerase module. Polyadenylation assay of purified polymerase module analyzed by denaturing urea-PAGE. A 5′ fluorescently labeled 42 mer *CYC1* RNA that has been pre-cleaved is the substrate for polyadenylation. **(a)** Assay carried out with only polymerase module and RNA. **(b)** Assay carried out with polymerase module and CF IA. **(c)** Assay carried out with polymerase module and Rna14/Rna15. The assays contained a final concentrations of 400 nM substrate RNA, 50 nM purified polymerase module, 400 nM CF IA or Rna14/Rna15 and 2 mM ATP. Densitometry analyses of the gels are plotted below; RNA length is marked based on an RNA ladder.

To dissect the contribution of subcomplexes of CF IA to polyadenylation, I performed *in vitro* assays of 5'-FAM-*CYC1*-pc with the polymerase module in the presence of Rna14/Rna15. Surprisingly, Rna14/Rna15 is sufficient to stimulate the polyadenylation activity of polymerase module (Figure 3.2c). Moreover, the polyadenylation activity of the polymerase module in reactions containing CF IA or Rna14/Rna15 is highly similar (Figure 3.2b-c), suggesting that Rna14/Rna15 can account for the full stimulatory activity of CPF by CF IA.

Furthermore, CF IA (and Rna14/Rna15) can stimulate the polyadenylation activity of a polymerase module containing two copies of Pap1 (Figure 3.3), demonstrating that the stimulatory effect is independent of Pap1 stoichiometry.

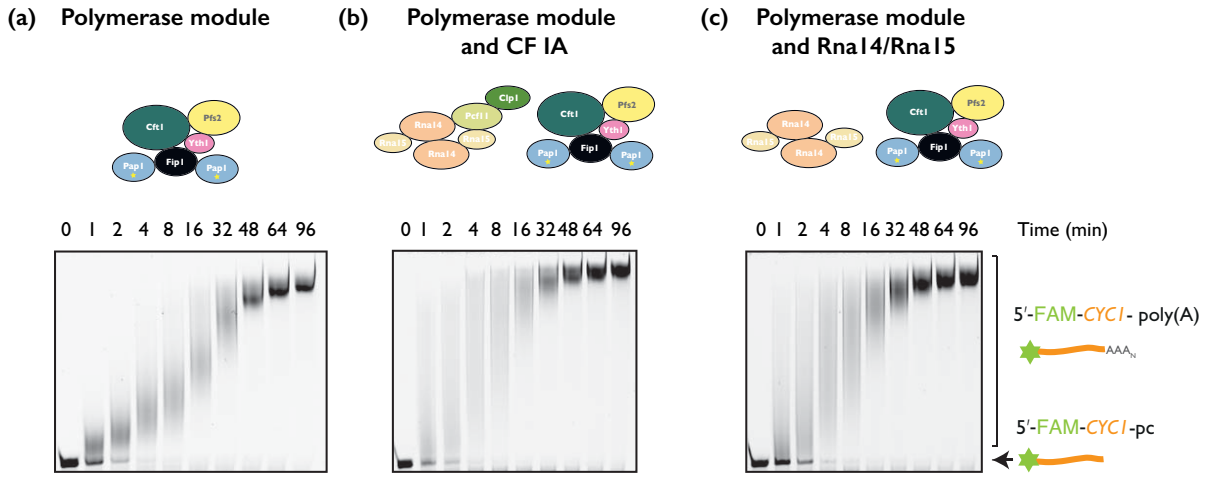


Figure 3.3: Cleavage factor IA stimulates polyadenylation by polymerase module with two copies of Pap1. Polyadenylation assay of purified polymerase module analyzed by denaturing urea-PAGE. A 5′ fluorescently labeled 42 mer *CYC1* RNA that has been pre-cleaved is the substrate for polyadenylation. **(a)** Assay carried out with only polymerase module and RNA. **(b)** Assay carried out with polymerase module and CF IA. **(c)** Assay carried out with polymerase module and Rna14/Rna15. The assays contained a final concentrations of 400 nM substrate RNA, 50 nM purified polymerase module, 400 nM CF IA or Rna14/Rna15 and 2 mM ATP.

In addition to CF IA, CF IB also harbours an RNA-binding RRM domain and can bind tightly to the AU rich efficiency elements (Pérez-Cañadillas 2006). To probe the role of CF IB in polyadenylation, I performed *in vitro* polyadenylation assays of 5'-FAM-*CYC1*-pc using polymerase module in the presence and absence of CF IB. The addition of CF IB does not stimulate the activity of the polymerase module as the rate of polyadenylation is similar whether or not CF IB is present (Figure 3.4). Therefore, CF IB does not affect polyadenylation and likely does not interact with the polymerase module.

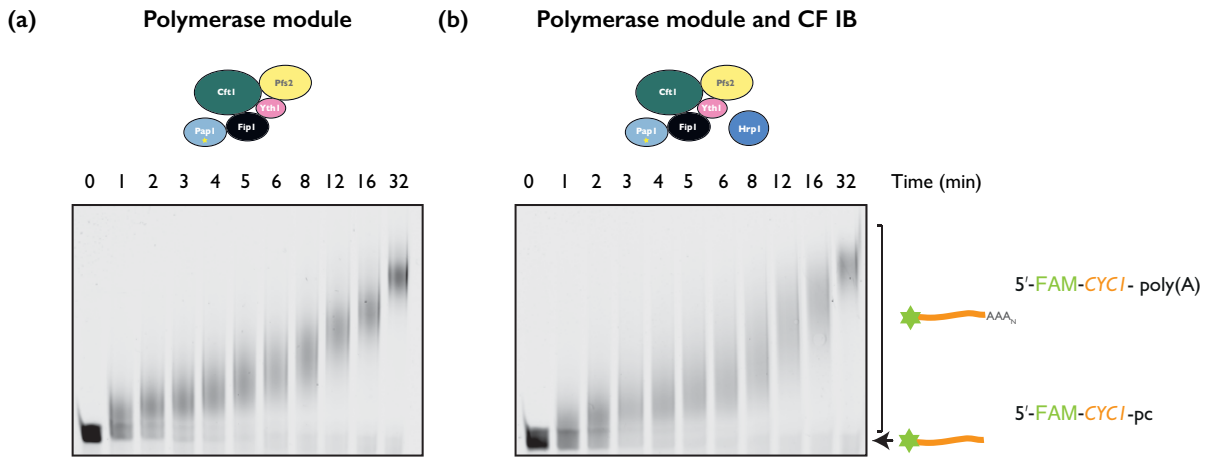


Figure 3.4: CF IB does not stimulate polyadenylation. Polyadenylation assay by purified polymerase module analyzed by denaturing urea-PAGE. A 5′ fluorescently labeled 42 mer *CYC1* RNA that has been pre-cleaved is the substrate for polyadenylation. **(a)** Assay carried out with only polymerase module. **(b)** Assay carried out with polymerase module and CF IB. The assays contained a final concentrations of 400 nM substrate RNA, 50 nM purified polymerase module, 400 nM CF IB and 2 mM ATP.

In agreement with an earlier study (Gordon et al. 2011), my experiments using a minimal reconstituted system suggest that the Rna14/Rna15 subunits of CF IA are sufficient to stimulate polyadenylation by the five-subunit polymerase module and CPF. I sought to further investigate the individual subunit contributions and the unexplored mechanistic basis of the observed stimulatory effect.

3.2 CF IA does not affect polyadenylation by Pap1

Earlier experiments demonstrated that the *in vitro* polyadenylation activity of isolated Pap1 enzyme is similar to that of a five-subunit polymerase module (Figure 2.6). This is surprisingly different from the human homolog. In humans, the isolated PAP enzyme is distributive, but its incorporation within CPSF makes the polyadenylation reaction processive due to the additional RNA binding surfaces provided by CPSF (Schönemann et al. 2014; Chan, Huppertz, Yao, Weng, Moresco, Yates Iii, et al. 2014). The biochemical studies performed so far do not explain the presence of accessory yeast polymerase module subunits, and whether these non-catalytic subunits affect the *in vitro* polyadenylation activity of the polymerase module.

In order to better understand the contribution of the polymerase module subunits, I performed *in vitro* assays of 5'-FAM-*CYC1*-pc using the isolated Pap1 enzyme, in the presence and absence of CF IA. CF IA does not stimulate the polyadenylation activity of Pap1 (Figure 3.4). In fact, Pap1 activity is slightly reduced in the presence of CF IA (compare early time points, e.g. t = 4, 8, 16 min in Figure 3.5.a and Figure 3.5b). It is possible that, in the absence of CF IA, Pap1 can freely access substrate RNA to add poly(A) tails in a distributive manner as described earlier. However, the presence of CF IA could obstruct Pap1 from accessing the substrate RNA, reducing Pap1 activity (Figure 3.4.b). The results here show that CF IA cannot stimulate the activity of isolated Pap1 and that additional polymerase module subunits are required for this simulation (Figure 3.2). Such an observation underscores the functional importance of other polymerase module subunits in the full polyadenylation activity of CPF and the polymerase module in the presence of CF IA.

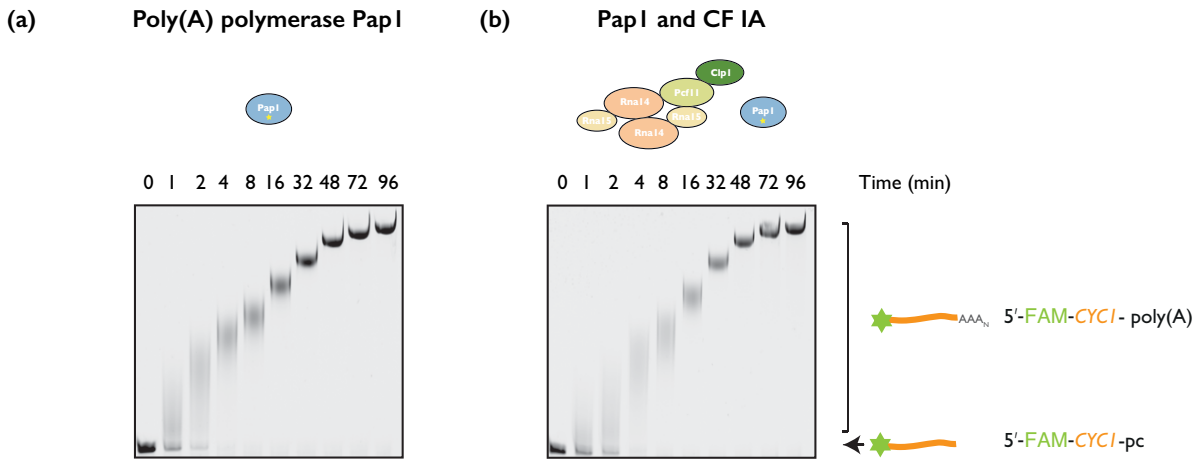


Figure 3.5: Activity of isolated Pap1 is not stimulated by CF IA. Polyadenylation assay by purified Pap1 analyzed by denaturing urea-PAGE. A 5′ fluorescently labeled 42 mer *CYC1* RNA that has been pre-cleaved is the substrate for polyadenylation. (a) Assay carried out with only Pap1. (b) Assay carried out with Pap1 and CF IA. The assays contained a final concentrations of 400 nM substrate RNA, 50 nM purified Pap1, 400 nM CF IA and 2 mM ATP.

3.3 CF IA interacts with the polymerase module of CPF

The observation that CF IA (Rna14/Rna15) stimulates the polyadenylation activity of CPF and the polymerase module led me to hypothesize that stimulation occurs via a direct protein-protein interaction between CF IA and CPF subunits. To investigate the interactions between CF IA and CPF subunits, I carried out *in vitro* pull-down experiments on immobilized CPF subunits.

3.3.1 *In vitro* pull-downs

To study if CPF directly interacts with CF IA, I carried out pull-downs using SII-tagged CPF sub-complexes as the bait proteins, and CF IA and CF IB proteins as the prey.

The following CPF sub-complexes were used as bait: a recently described eight-subunit core-CPF (Core-CPF), the polymerase module with Cft2 (Polymerase + Cft2), the five-subunit polymerase module (Polymerase); the polymerase module without Pap1 (Polymerase – Pap1); Cft2; isolated Pap1; and a hetero-dimer of Ysh1/Mpe1 (Figure 3.6a). Core-CPF consists of all five polymerase module subunits and three nuclease module subunits and has been shown to be sufficient for reconstituting pre-mRNA cleavage (Hill et al. 2019). Control reactions for the pull downs were performed with Strep beads that did not contain any bait proteins.

Pull-down reactions were first carried out using the full CF IA complex as the prey protein complex. Elution fractions from the pull-downs were analysed by SDS-PAGE (Figure 3.6b-c). CF IA binds to core-CPF and polymerase module (both with and without Pap1), as demonstrated by the presence of bands corresponding to CF IA subunits in the pull-down (Figure 3.6b). On the other hand, no interaction was observed between CF IA and Cft2, Pap1 or Ysh1/Mpe1 (Figure 3.6b). These results show that the core four polymerase module subunits (i.e. Cft1, Pfs2, Yth1 and Fip1) mediate the interaction between CPF and CF IA. This provides a potential explanation for why the activity of isolated Pap1 is not stimulated by the addition of CF IA, as Pap1 alone cannot interact with CF IA, whereas the polymerase module binds to the cleavage factor.

Next, to understand which subunits of CF IA contact CPF; pull-downs were carried out with the Rna14/Rna15 dimer of dimers (Figure 3.6c) or the Pcf11/Clp1 subcomplex (Figure 3.6d) as prey. Rna14/Rna15 can interact with core-CPF and polymerase module (both with and without Pap1). This is consistent with previous polyadenylation assays where Rna14/Rna15 stimulates polyadenylation to a similar degree compared to CF IA (Section 4.1, Figures 4.2b-c). In contrast, Pcf11/Clp1 did not stably interact with CPF. Interestingly, CF IB does not interact strongly with CPF sub-complexes (Figure 3.6e), supporting earlier observations where addition of CF IB to *in vitro* assays did not have any stimulatory effect on polyadenylation (Section 4.1, Figure 3.4b).

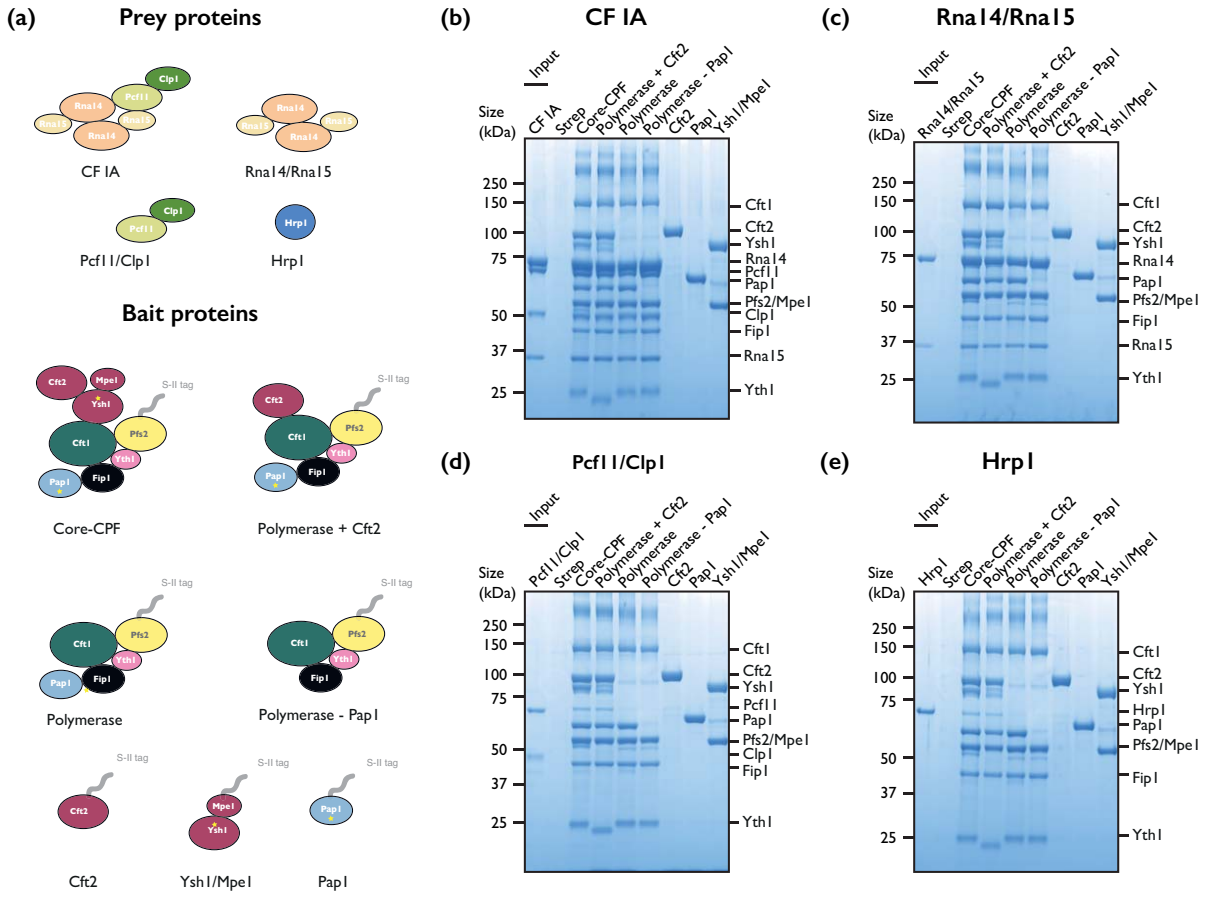


Figure 3.6: CF IA is recruited to CPF via direct interactions between Rna14/Rna15 and the polymerase module. **(a)** Schematic of the protein complexes used in the pull-downs. CPF proteins and sub-complexes were immobilized on StrepTactin resin (bait). **(b)** Untagged CF IA, **(c)** Rna14–Rna15, **(d)** Pcf11–Clp1 and **(e)** Hrp1 (CF IB) were used as prey. Coomassie-stained SDS-PAGE of pull-down experiments. The input lane contains the purified prey proteins that were added. As a negative control, prey proteins alone were added to StrepTactin resin.

To further characterize the interaction between CPF and CF IA, a minimal complex containing the polymerase module (without Pap1) and the Rna14/Rna15 subunits, present in three-fold molar excess, were mixed *in vitro* and separated using a size exclusion column (Figure 3.7a). Polymerase module alone elutes at a volume of ~1.4 ml (in pink), Rna14/Rna15 complex elutes at ~1.65 ml (in green) and the complex of polymerase module with Rna14/Rna15 elutes at ~1.25 ml (in yellow) (Figure 3.7a). The earlier elution of the Rna14/Rna15-polymerase module complex relative to each individual complex suggests that the apparent molecular weight is increased. SDS-PAGE analysis of the eluted fractions reveals the presence of all CF IA and polymerase module subunits (Figure 3.7b). Thus, Rna14/Rna15 and the polymerase module associate stably and the resulting complex can be separated by size exclusion chromatography. Therefore, a protocol to produce a complex containing polymerase module and Rna14/Rna15 subunits has been established. This allows further characterization of the protein-protein interactions using biophysical techniques.

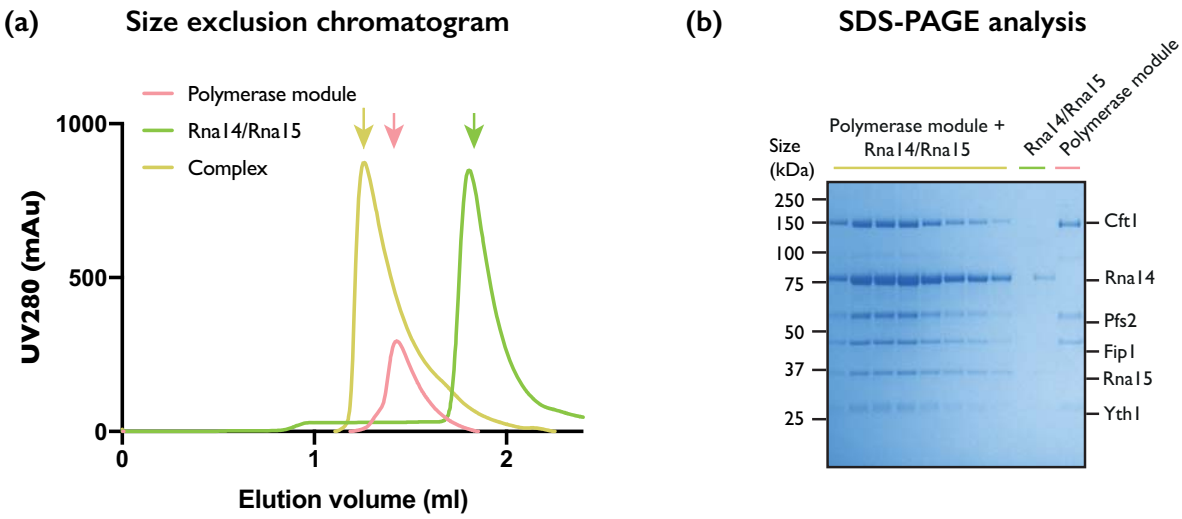


Figure 3.7: Rna14/Rna15 co-elutes with the polymerase module on a size exclusion column. **(a)** Purified polymerase module and Rna14/Rna15 complexes were mixed together in 1:3 molar ratio, subjected to size exclusion chromatography on a Superose 6 Increase 3.2/300 column (in yellow). As a control, samples containing polymerase module alone (in pink) and Rna14/Rna15 (in green) alone were run on the same column. **(b)** SDS-PAGE analysis of fractions from the size exclusion column.

3.3.2 Hydrogen Deuterium Exchange of Polymerase Module-Rna14-Rna15

Hydrogen deuterium exchange coupled with mass spectrometry (HDX) is a powerful biophysical technique that is used to study protein-protein and protein-ligand interactions (Shukla et al. 2014; Seckler, Barkley, and Wintrode 2011; Masson, Jenkins, and Burke 2017). Furthermore, it is also widely employed to study protein conformational dynamics (Wales and Engen 2006). Hydrogen atoms bound to the backbone amide group in proteins readily exchanges with protons in the aqueous solution. In HDX, deuterated water (D_2O) replaces the aqueous environment surrounding the protein of interest. Thus, solvent-exposed amide hydrogens will be exchanged for deuterium. On the other hand, amide hydrogens that are not exposed to solvent will exchange much less readily. The hydrogen deuterium exchange process can be monitored in a mass spectrometer owing to the difference in mass between the two elements. By comparing different conditions, information on protein-protein interactions or conformational changes can be obtained, as long as they affect the extent to which amide hydrogens are shielded or exposed to the deuterated solvent, as well as the timescale of hydrogen-deuterium exchange.

HDX experiments were performed for the following three samples: five-subunit polymerase module with Pap1, four-subunit polymerase module without Pap1, and four-subunit polymerase module in complex with Rna14/Rna15. The interaction of Pap1 with the polymerase module subunits has been well characterized (Section 3.5). Thus, comparison of HDX data from the polymerase module (with and without Pap1) should enable us to validate the previously identified Pap1 interaction surface on Fip1.

Cft1 had ~66% coverage across its sequence, Pfs2 ~80%, Yth1 ~68%, and Fip1 ~80%. Overall, no difference in deuterium exchange rates were observed in Cft1, Pfs2 and Yth1 across the three samples analyzed. Thus, the interaction of Pap1 or Rna14/Rna15 with the polymerase module did not significantly induce any conformational change in Cft1, Pfs2, or Yth1. However, residues 86-101 of Fip1 were protected from deuterium exchange in the presence of Pap1 (Figure 3.8a). Thus, this normally solvent-exposed region of Fip1 likely becomes buried upon the interaction of Pap1. Interestingly, this region of Fip1 contains the peptide sequence (residues 85 - 105) that had been previously shown to interact with Pap1, and a crystal structure describing the interaction

is available (Meinke et al. 2008), thereby validating HDX as a bona fide technique to study protein-protein interactions for CPF and associated complexes.

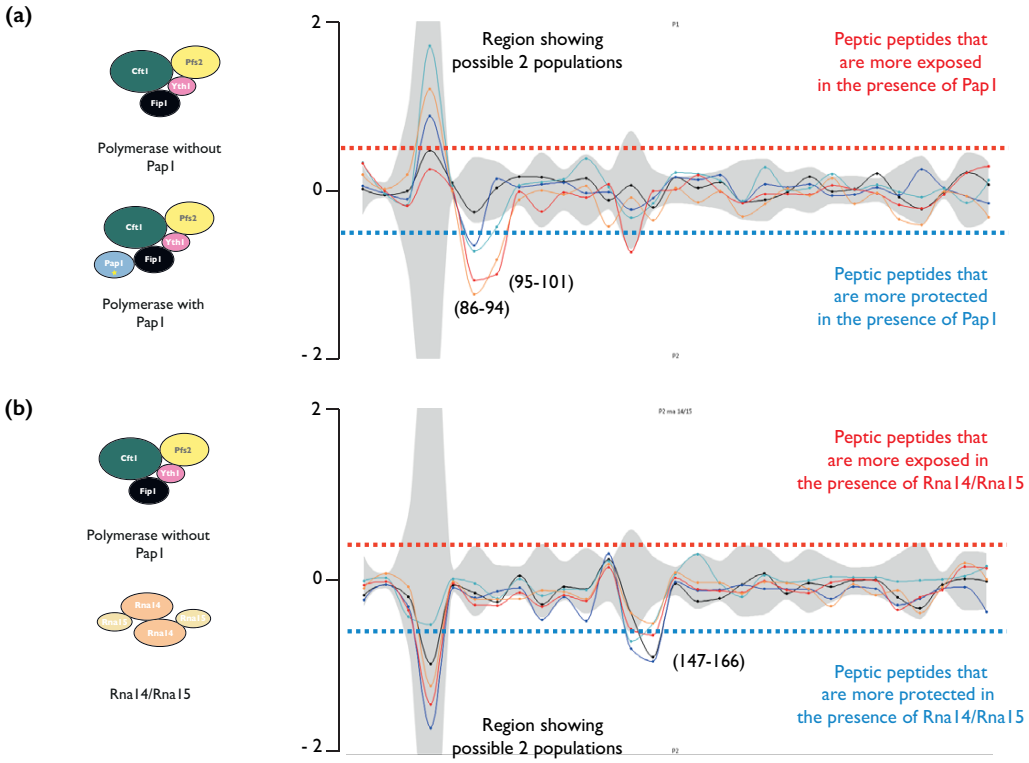


Figure 3.8: Rna14/Rna15 interacts with residues 147 to 166 of Fip1. Hydrogen-deuterium exchange mass spectrometry difference plot showing peptides of Fip1 that are exposed (positive) or protected (negative) by **(a)** Pap1 and **(b)** Rna14/Rna15. Triplicate data from four time points 0.3 sec (orange), 3 sec (red), 30 sec (green) and 300 sec (blue) are show. Dotted line in grey represents the significance threshold. Grey shading represents the standard deviation of all charged states and replicates per peptide.

Moreover, amino acids 147-166 of Fip1 was protected from deuterium exchange in the presence of Rna14/Rna15 (Figure 3.8.B). This region of Fip1 lies within the "conserved domain" or "CD" of Fip1 (Clerici et al. 2017). Thus, it is likely that Rna14/Rna15 interacts with Fip1 via its CD. In addition, previous *in vitro* pull-downs have identified the N-terminal region of human Fip1 (containing the conserved domain) as an interaction partner of Cstf-77 (Kaufmann et al. 2004). Therefore, it is likely that

this interaction between Fip1 and Rna14/Rna15 is conserved across different species, underscoring its functional importance.

3.3.3 Rna14/Rna15 can stimulate the activity of polymerase module with a Fip1 truncation

The role of Fip1 in tethering Pap1 to the polymerase module is well established (Section 3.5). Previous studies have highlighted the importance of this flexible conserved domain of Fip1 in polyadenylation (Ezeokonkwo et al. 2011). Therefore, in addition to its role in linking Pap1 to the remainder of the polymerase complex, Fip1 may also be involved in the allosteric regulation of activity of the polymerase module. I hypothesized that Rna14/Rna15 could interact with this flexible region of Fip1 and allosterically induce a conformational change that results in stimulation of the enzymatic activity of Pap1 within the polymerase module.

In order to validate the findings from HDX that region 147-166 of Fip1 interacts with Rna14/Rna15, and to test the aforementioned hypothesis, I made truncations within the CD of Fip1. I cloned a five-subunit polymerase module containing Fip1 where the amino acids 145 to 170 were deleted (Fip1Δ145-170). The truncated polymerase module containing Fip1Δ145-170 was purified using a similar protocol to that of the wild type polymerase module (Section 6.4.1). The purified complex was analysed by SDS-PAGE alongside wild type polymerase module (Figure 3.9a). The truncated polymerase module was functionally active and added poly(A) tails to a substrate RNA 5'-FAM-*CYC1*-pc in a manner very similar to the wild type polymerase module (Figure 3.9b). Interestingly, Rna14/Rna15 can still stimulate the activity of this truncated polymerase module containing Fip1 Δ145-170 (Figure 3.9b). Thus, the deletion of Rna14/Rna15 interaction surface on Fip1 does not affect the ability of Rna14/Rna15 to influence polyadenylation activity.

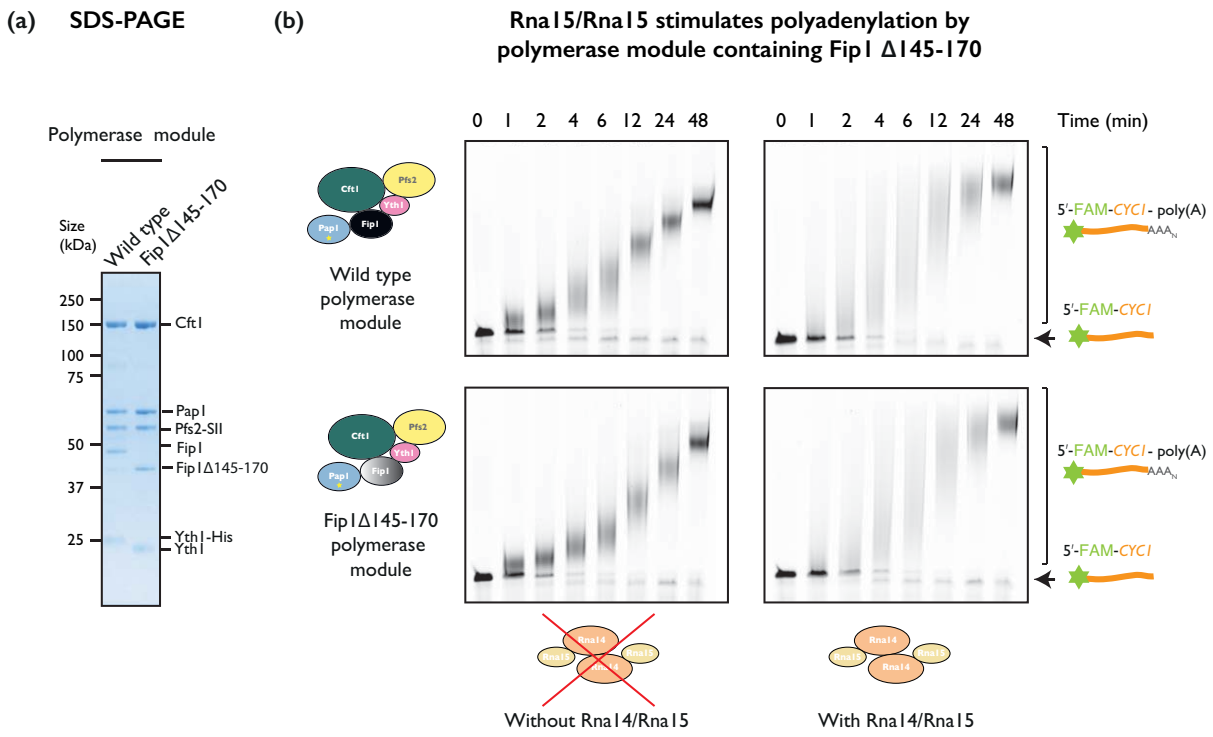


Figure 3.9: Deletion of the RNA14/Rna15 interaction surface in Fip1 does not affect polyadenylation stimulation. **(a)** SDS-PAGE analysis of the purified polymerase module. The deletion of residues 145 to 170 in Fip1 does not affect the assembly and purification of intact polymerase module. **(b)** Polyadenylation assays analyzed by denaturing urea-PAGE. A 5′ fluorescently labeled 42 mer *CYC1* RNA that has been pre-cleaved is the substrate for polyadenylation. Rna14/Rna15 stimulated the polyadenylation activity of both the wild type and Fip1 Δ145-170 containing polymerase module. The assays were carried out with a final concentration of 400 nM substrate RNA, 50 nM purified polymerase module (WT or Fip1 Δ145-170), 400 nM Rna14/Rna15 and 2 mM ATP**.**

3.4 Rna14 HAT domains interact with the polymerase module

As Rna14/Rna15 can stimulate the activity of the polymerase module containing Fip1Δ145-170, it is likely that the CF IA sub-complex can interact with other subunits of the polymerase module. A limitation of the HDX experiment is its modest sequence coverage of polymerase module proteins, and thus the interaction surface between Fip1 and polymerase module subunits could lie outside the covered sequences. For example, the beta-propellers in Pfs2 or Cft1 can act as a protein-protein interaction hub for the structurally rigid HAT domains in Rna14 (Ohnacker et al. 2000). Similarly, in the structure of the human SF3b core complex, the HEAT repeats containing SF3155 is found to interact with the beta-propellers of SF3b130 (Cretu et al. 2016).

Thus, to fully elucidate the mechanistic basis of CF IA interaction with the polymerase module, there was the need to use orthogonal biochemical or biophysical methods. I chose to further investigate the interaction using cross-linking mass spectrometry and cryo-EM.

3.4.1 Cross-linking mass spectrometry of polymerase module - Rna14/Rna15 complex

The complex containing polymerase module and Rna14/Rna15 subunits was cross-linked with BuUrBu (Disuccinimidyl Dibutyric Urea) (Müller et al. 2010). Similar to BS3, BuUrBu contains N-hydroxysuccinimide (NHS) ester functional groups and can covalently link two lysine, serine, threonine or tyrosine molecules that are within 27 Å of each other.

A total of 82 inter- and intra-subunit crosslinks were found. Several of the crosslinks within the polymerase module subunits were consistent with the BS3 crosslinking data described earlier (Section 3.5.2). Residues K47, K44 and S38 in the N-terminal region of Fip1 cross-linked to the C-terminal HAT domain of Rna14, consistent with previous *in vitro* pull downs where the N-terminus of Fip1 was found to interact with Rna14 (Kaufmann et al. 2004) (Figure 3.10). In addition to the previously characterized Fip1-Rna14 interaction, K60 in the N-terminus of Yth1 also crosslinked with residues K170, K219 and K28 in the N-terminal HAT domain of Rna14. This region of Rna14 also cross-linked with the C-terminus of Pfs2. These results hint at the possibility that the HAT domain of Rna14 lies in close proximity to Yth1 ZnF2 and the WD40 of Pfs2 as these domains are structurally proximal (Section 3.15). Moreover, the HAT domain of Rna14 crosslinked with other polymerase module subunits, including various regions of Cft1 (Figure 3.10), which is consistent with a previous study, demonstrating a close association between CPSF160 and CstF-77 (Murthy and Manley 1992). It is thus possible that Rna14 harbours multiple interaction sites on the polymerase module. Interestingly, no crosslinks were found between Rna15 and polymerase module subunits. One explanation could be that the region of interaction between Rna15 and the polymerase module subunits do not contain any lysines, serines or tyrosines. Alternatively, there may not be any direct or proximal contacts between Rna15 and the polymerase module.

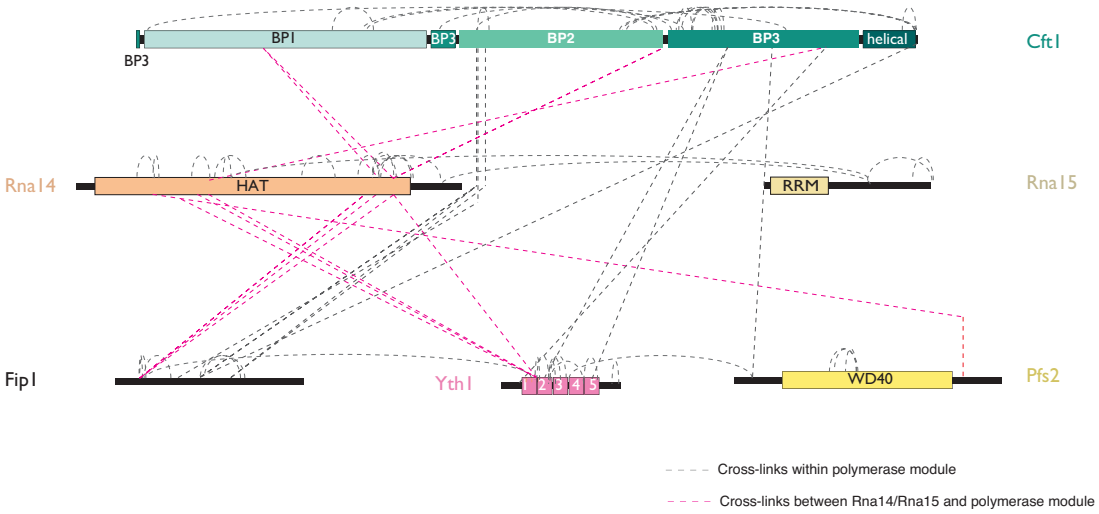


Figure 3.10: Cross-linking mass spectrometry of the polymerase module - Rna14/Rna15 complex. Linkage map showing all identified lysine-lysine, lysine-serine and lysine-tyrosine crosslinks. Cross-links between the polymerase module and Rna14/Rna15 are shown in pink dotted lines.

3.4.2 Cryo-EM of polymerase module - Rna14/Rna15 complex

To better understand how Rna14/Rna15 interacts with the polymerase module subunits, and to gain insights into how the two complexes bind the substrate RNA, a ternary complex consisting of the four-subunit polymerase module, Rna14/Rna15 subunits and *CYC1*-pc RNA was assembled on a size exclusion column (Figure 3.11a). The peak fraction containing the entire complex eluted from the column at ~1.25 ml elution volume, similar to the previous complex lacking RNA (Section 4.3.1). Comparison of the absorbance at 260 nm to that at 280 nm confirmed the presence of nucleic acid in the peak fractions (Figure 3.11a). SDS-PAGE analysis of the peak fractions revealed the presence of all six-protein subunits in the preparation (Figure 3.11b). Cryo-EM micrographs of the ternary complex revealed ~15 nm individual particles that were (Figure 3.11c). Further inspection revealed mono-disperse but heterogeneous particles. 2-D class averages of selected aligned particles show the characteristic polymerase module 2-D classes (shown by green arrowhead in Figure 3.11d, and Figure 2.10b). Also seen were 2-D classes containing a ~15 nm boomerang shaped density. This density resembles projections of the crystal structure of the Rna14 HAT domain from *Kluyveromyces lactis*, which revealed a ~15 nm elongated dimer of the HAT domains (Paulson and Tong 2012). Similar boomerang shaped 2-D classes were obtained in a previous electron microscopy study of *S. cerevisiae* Rna14-Rna15 (Noble et al. 2004). Thus, the ~15 nm elongated, kinked density (shown by cream arrowhead in Figure 3.11d) likely corresponds to the dimer of HAT domains formed by Rna14. It is interesting to note that in 2-D classes highlighted in red boxes, the HAT domain appears to associate with the bottom region of the polymerase module (BP2). In agreement with this observation, crosslinks were found between Y914 and T915 of Cft1 and K551 in the HAT domain of Rna14.

Overall, the 2-D classes shown here highlight the heterogeneity of the sample under investigation. In some of the representative 2-D classes, the particles are aligned on the Rna14 HAT domain density, resulting in a blurred density corresponding to the polymerase module (highlighted by grey arrows in Figure 3.11d). Conversely, certain 2-D classes were aligned on the polymerase module, and contain additional blurry boomerang-like density surrounding the central module (classes 2, 6). This reflects a scenario where Rna14/Rna15 interacts with polymerase module at many sites. As a result, the particles of the complex are unable to align as one fixed conformation.

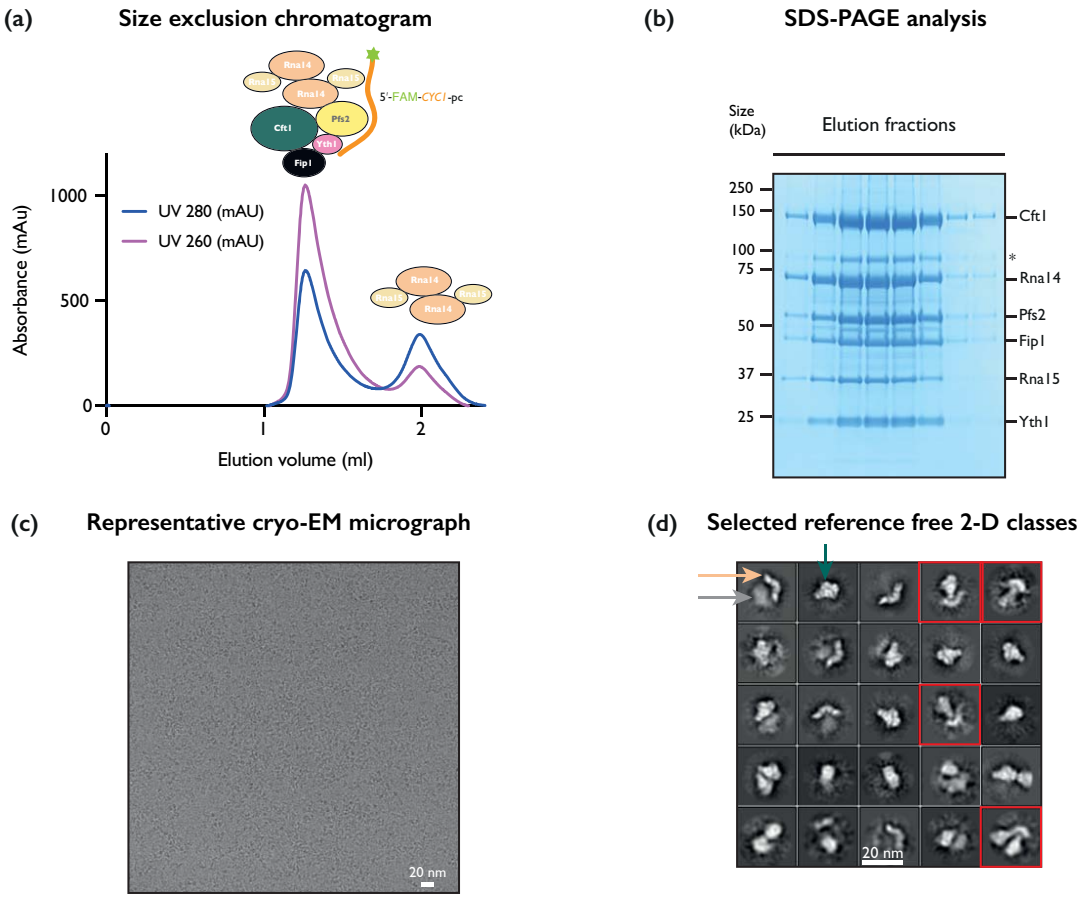


Figure 3.11: Cryo-EM analysis of the polymerase module - Rna14/Rna15 - *CYC1* RNA complex. **(a)** Size exclusion chromatography of the ternary complex **(b)** SDS-PAGE of the peak showing all the polymerase module subunits and Rna14/Rna15. **(c)** Representative cryo-EM micrograph of polymerase module - Rna14/Rna15-CYC1 complex taken at 59k X magnification. **(d)** Selected 2D class averages of aligned particles reveals sample heterogeneity. The HAT-domain of Rna14 is highlighted by cream arrowhead. The isolated polymerase module is highlighted by green arrowhead. Grey arrowhead shows the fuzzy density found near the HAT-domain in many of the classes. Classes highlighted by red squares show the HAT-domains of Rna14 positioned below the beta-propeller of Cft1.

Further stabilization of the sample is therefore likely required to proceed with cryo-EM analysis. One possible remedy is to use a mild cross-linker to fix the complex in one particular conformation.

Despite the presence of RNA in the sample, no density for RNA could be observed in the 2-D classes. A likely explanation is that the RNA used in the study (42 nucleotides long) is too small to be observed in 2-D classes at the current magnifications used in imaging. A high-resolution 3D reconstruction would be necessary to validate the presence of RNA.

In summary, cryo-EM studies have shown that the HAT domain of Rna14 subunit interacts with the polymerase module. It is likely that Rna14 interacts with polymerase module at many sites. This is further supported by cross-linking data where the HAT domain cross-links across all the subunits of polymerase module. It is possible that the conformation where Rna14-Rna15 is bound to BP2 is preferred, as they constituted 2-D classes, which were best resolved in preliminary cryo-EM analysis. These data also suggest that Rna14 harbours the main binding site of CF IA to the polymerase module. However, in the absence of any cross-linking and cryo-EM data on the Rna15 subunit, one cannot rule out the possibility that Rna15 could harbour a polymerase module interaction surface.

3.5 An RNA binding mutant of Rna14/Rna15 fails to stimulate polyadenylation

While CF IA stimulates the polyadenylation activity of CPF and CPF interacts with CF IA via Rna14/Rna15 and polymerase module subunits, the exact mechanism underlying the stimulation remains unknown. There are at least two possible mechanisms of stimulation. Firstly, because HDX experiments found an Rna14/Rna15 interaction surface in the conserved flexible domain of Fip1, it was hypothesized that the Rna14/Rna15 creates a conformational change in this region of Fip1, thereby modulating Pap1 activity. On the other hand, *in vitro* assays showed that the activity of the polymerase module could be stimulated by Rna14/Rna15 even in the absence of this central region of Fip1. Thus, another possible mechanism by which Rna14/Rna15 stimulates polyadenylation is by providing additional RNA binding surface for the polymerase module. Together with the surface-exposed positively charged residues on top of Pfs2 and the ZnF2/3 of Yth1, the RRM domain of Rna15 can form a composite RNA binding surface. Such a cooperative binding would increase the overall affinity for RNA and result in stimulation of polyadenylation.

In order to test this hypothesis, I generated an RNA binding mutant of Rna14/Rna15 (Figure 3.12a). It is thought that Rna14/Rna15 binds RNA mainly through the RRM domain of Rna15, whose structure has been determined in complex with a GU dinucleotide (Pancevac et al. 2010). In the structure, guanine is stacked against the aromatic ring of Y21 (Figure 3.12a). Furthermore, heteronuclear NMR experiments suggests that the aromatic rings of Y61 and F63 could be a part of a nucleobase binding site (Pancevac et al. 2010). I cloned, expressed and purified a mutant Rna14/Rna15 that carried a Y21A, Y61A, F63A triple mutant in the RRM domain of Rna15 (Section 6.4.5). EMSA experiments show that the mutant Rna14/Rna15 has a significant reduction in its ability to bind RNA (Figure 3.12b). In contrast with the wild-type Rna14/Rna15, *in vitro* assays show that the mutant Rna14/Rna15 does not stimulate the polyadenylation activity of polymerase module (Figure 3.12c). Interestingly, the polymerase module has reduced activity in the presence of the mutant Rna14/Rna15 in comparison to an assay performed without any Rna14/Rna15 (Figure 3.12c). The RNA binding mutant of Rna14/Rna15 can likely still interact with polymerase module and

therefore sequesters the polymerase module from binding RNA. This may explain why there is a reduced polyadenylation activity in the presence of mutant Rna14/Rna15.

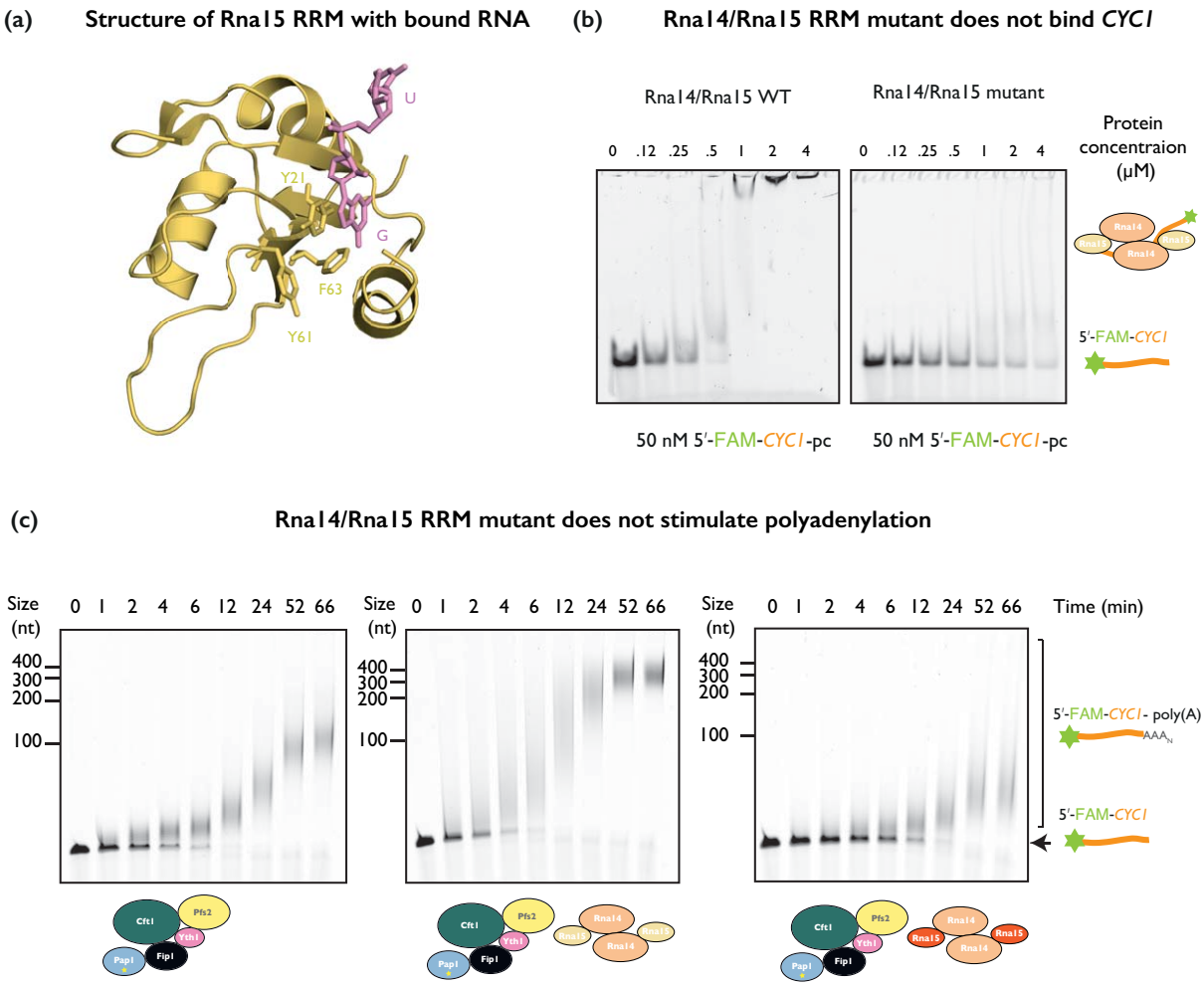


Figure 3.12: Rna14/Rna15 stimulates polyadenylation by tethering the polymerase module to the RNA. **(a)** X-ray crystal structure of the RRM domain of Rna15 bound to GU (PDB ID: 2X1F) showing residues involved in base recognition. **(b)** An RRM domain triple mutant Y21A, Y61A, F63A in Rna15 loses its ability to bind RNA as shown by the EMSA experiment. **(c)** A mutant of Rna14/Rna15 that does not bind RNA, no longer stimulates polyadenylation of the polymerase module.

The inability of the mutant Rna14/Rna15 to stimulate polyadenylation suggests that CF IA (and Rna14/Rna15) enhances the activity of CPF (and polymerase module) by providing additional RNA binding surfaces. However the allosteric model for polyadenylation stimulation cannot be completely ruled out in the absence of complete structural information about the CF IA - polymerase module interaction.

3.6 Model for stimulation of polyadenylation by CF IA

Taken together, these results enable us to hypothesize a model of polyadenylation by the polymerase module. In the absence of any CF IA (and Rna14/Rna15), the polymerase interacts with RNA at ~ 500 nM affinity (figure 2.8). In this scenario, the polymerase module is weakly processive and only adds ~10 to 15 As before the polymerase module falls off the product RNA (Figure 3.2a). The polymerase module is then competent to bind a new substrate RNA to further continue addition of poly(A) tails. This model of polyadenylation is illustrated in Figure 3.13a.

On the other hand, the presence of CF IA provides the polymerase module with additional RNA binding surfaces via the RRM domain of Rna15 subunit. In this scenario, the complex of polymerase module and CF IA (or Rna14/Rna15) binds more tightly to RNA and the resulting complex has a longer half-life. The polymerase module then adds poly(A) tail to the substrate RNA without the complex falling off the RNA. This tethering of the polymerase module to the substrate RNA by CF IA results in the longer poly(A) tails being rapidly added. This agrees with a previous study showing that physical tethering of Pap1 to RNA results in an increase in its polyadenylation activity (Ezeokonkwo et al. 2011). Such a model of processive polyadenylation is illustrated in Figure 3.13b.

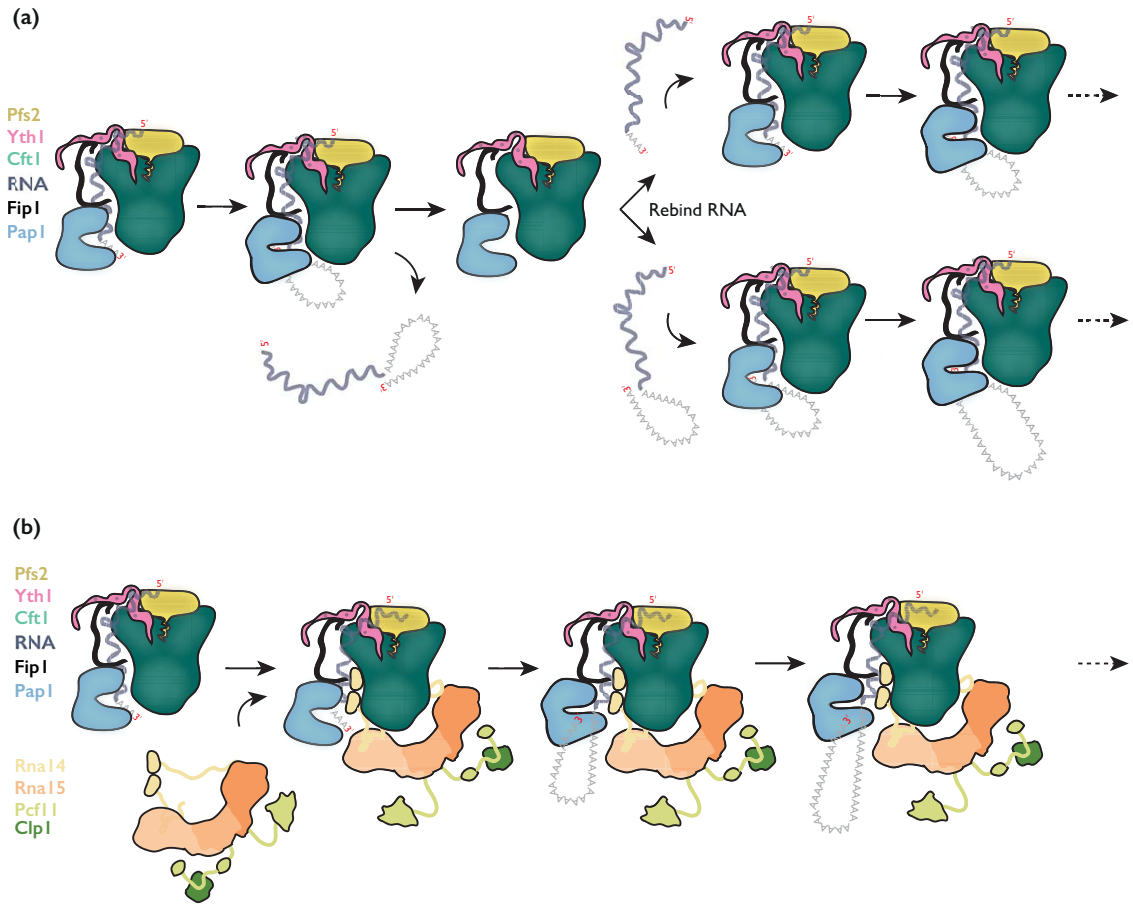


Figure 3.13: A model for polyadenylation stimulation by cleavage factor IA. Model obtained by combining data from cryo-EM, cross-linking–mass spectrometry, biochemical assays and in vitro pull-downs. **(a)** In the absence of the cleavage factor IA, polymerase module binds the substrate RNA, adds ~ 10 - 15 As and falls off the product RNA. Then the free polymerase module binds another substrate or product RNA, continues adding poly(A) tails in this manner. **(b)** In the presence of CF IA, the polymerase module remains tethered to the substrate RNA. Rna14/Rna15 interacts with the polymerase module, and provides additional RNA binding surface. This results in stimulation of polyadenylation.

4 Progress Towards Understanding Poly(A) Tail Length Control

As pre-mRNA emerges from Pol II, specific *cis*-acting elements within its 3′ UTR are bound by CPF and CF I. This results in the assembly of the entire 3' end processing complex poised to carry out an endonucleolytic cleavage at the poly(A) site, followed by the addition of a poly(A) tail to the free 3' hydroxyl of the terminal nucleotide.

In mammals, the poly(A) tails of mRNAs are synthesized to ~ 200 As. *In vitro* biochemical studies have highlighted the role of the nuclear poly(A) binding protein PABN1 in poly(A) tail length control (Wahle 1995). PABN1 bound to a short poly(A) tail promotes the interaction between CPSF and PAP. This interaction results in the cooperative stimulation of poly(A) tail addition by PAP. PAP then adds ~ 200 to 250 As in a single processive synthesis step (Kühn et al. 2009). PABN1 can measure the length of the poly(A) tail and terminates the processive polyadenylation by disrupting the interaction between CPSF and PAP.

In *S. cerevisiae*, on the other hand, the poly(A) tail length of newly made pre-mRNAs is ~ 60 A's (Mclaughlin et al. 1973; Groner, Hynes, and Phillips 1974). Despite the absence of a functionally similar protein for PABN1 in *S. cerevisiae*, previous studies have highlighted the role of Pab1 and Nab2 in the regulation of mRNA poly(A) tail length control (Amrani, Minet, Le Gouar, et al. 1997; Minvielle-Sebastia et al. 1997; Schmid et al. 2012, 2015). It has been shown in yeast that the cytoplasmic poly(A) binding protein Pab1 physically associates with CF IA and is involved in the control of poly(A) tail length *in vitro* (Amrani, Minet, Le Gouar, et al. 1997; Minvielle-Sebastia et al. 1997). It was hypothesized that poly(A)-associated Pab1 could recruit the deadenylase complex Pan2/Pan3, which in turn trims the hyperadenylated tails to appropriate lengths. However, a later study revealed that *in vitro* poly(A) tail lengths were regulated in the absence of Pab1 dependent poly(A) nuclease activity, thus suggesting that poly(A) tail lengths were controlled via a different mechanism not involving Pan2/Pan3 (Dheur et al. 2005). The nuclear poly(A) binding protein Nab2 has also been implicated in control of poly(A) tail length both *in vivo* and *in vitro* (Hector et al. 2002). Nab2 binds poly(A) RNA with high affinity (K_d = 30 nM) and this interaction is mediated by the tandem CCCH family of zinc finger repeats in its C-

terminal end (Kelly et al. 2007; Aibara et al. 2017). Cells expressing a mutant allele of Nab2 that can no longer strongly bind poly(A) RNA, results in defects in poly(A) lengths (Kelly et al. 2010).

Previous biochemical studies have clarified the role of Nab2 in poly(A) tail length control and have provided valuable clues towards a mechanism for the same (Hector et al. 2002; Viphakone, Voisinet-Hakil, and Minvielle-Sebastia 2008). In the current model for poly(A) tail length control, Nab2 interacts with CPF components or likely alters the poly(A) conformation resulting in restriction of polyadenylation by Pap1. This leads to poly(A) tails of proper lengths being sythesized. However, several questions including why the poly(A) tail addition stops exactly after ~ 60As and what are the molecular players that trigger the termination of polyadenylation after ~ 60As remain unanswered. The lack of a pure and homogenous preparation of the 3' end processing complexes has hindered the systematic dissection of length control mechanism *in vitro*. Furthermore, we do not yet understand the contributions of CPF towards poly(A) tail length control, and how does Nab2 influence this. To address this major obstacle in the field, I describe the construction, recombinant expression and purification of the fourteen-subunit CPF complex. Moreover, I describe a robust and reproducible protocol for the *in vitro* reconstitution of cleavage and polyadenylation of a substrate mRNA 3' UTR using CPF from recombinant or endogenous expression. The experiments were designed taking into consideration the exact molarity of the reaction components. Remarkably, the experiments presented herein reveal the intrinsic capacity of CPF to restrict poly(A) tail length in the absence of Nab2. Further, establishing a recombinant system to produce CPF has now provided a means to make mutant complexes that was hitherto impractical. This will act as a future tool for dissecting the molecular mechanism of poly(A) tail length control by CPF.

4.1 CPF restricts poly(A) tail length

4.1.1 CPF purified from *S. cerevisiae* has intrinsic poly(A) tail length control

Native CPF was purified as described in section 3.1. A flow chart describing the purification protocol is illustrated in Figure 4.1a. The identities of the purified proteins were assessed by tandem mass spectrometry.

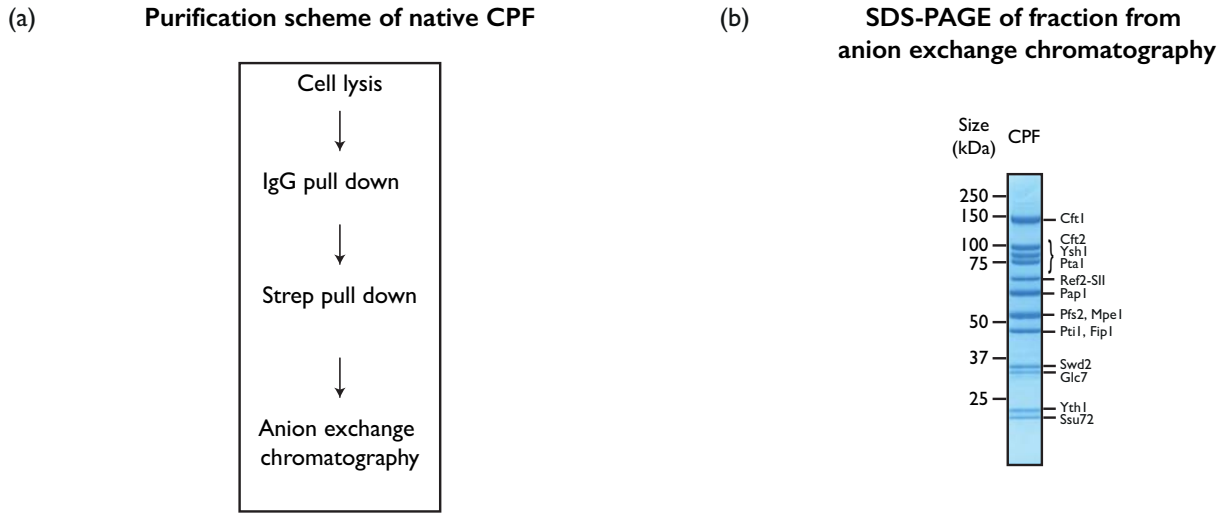


Figure 4.1: Purification of native CPF from *Saccharomyces cerevisiae* **(a)** Flow chart describing the different steps involved in the purification of native CPF. **(b)** The peak from the MonoQ anion exchange chromatography was analyzed using SDS-PAGE. The preparation contains all fourteen-subunits of CPF from *S. cerevisiae*.

Previous work from our lab has shown that addition of Nab2 to an *in vitro* polyadenylation assay results in RNAs with a poly(A) tail length mimicking *in vivo* lengths of ~ 60 As. *In vitro* cleavage and polyadenylation assays carried out in the absence of a poly(A) binding protein, only led to hyper-polyadenylation of the substrate RNA (Easter 2014).

I performed coupled cleavage and polyadenylation assays of an *in vitro* transcribed 3′ UTR of the *CYC1* mRNA. The assays consisted of CPF purified from yeast (Figure 4.1a) as well as CF IA and CFI B purified from *E. coli* (section 6.4.3, 6.4.4). The

reactions contained 100 nM substrate RNA, 50 nM CPF, and the cleavage factors were added in threefold excess (300 nM). It is found that polyadenylation of the cleaved RNA was surprisingly stalled after addition of around ~150 As (Figure 4.2a). Quantification of the reaction products by densitometry show a distribution of poly(A) tail lengths ranging from ~80 As to ~200 As (Figure 4.2b). It is to be noted that the length of the observed poly(A) tails is longer than what is observed *in vivo* in the presence of Nab2. Furthermore, not all the cleaved RNA 5′ fragments are polyadenylated. A band corresponding to the cleaved RNA (5'-*CYC1*-pc) remains throughout the time-scale of the experiment (Figure 4.2a). This observation is in contrast with previous assays (Casañal et al. 2017a).

The main difference between the current assay and the one performed previously lies in the concentrations of RNA and protein factors. In the earlier study, 87.5 nM RNA, 18 nM CPF and 48 nM CF IA/CF IB were used (Easter 2014). In the current study, 100 nM RNA, 50 nM CPF and 300 nM CF IA/CF IB are used (hereafter referred to as length control conditions). Under the current conditions, the cleavage factors are present in threefold excess over the substrate RNA. This could result in the occupancy of cleavage factors on RNAs being higher for a given RNA concentration. Furthermore, it has been shown earlier that CPF directly interacts with CF IA via the polymerase module (Figure 3.6). Therefore, CPF, CFs and the product RNA could form a stable ternary complex after addition of ~ 80 to ~ 200As, explaining why a fraction of the cleaved RNA (5'-*CYC1*-pc) remains unprocessed without a poly(A) tail. To test this hypothesis, assays carried out with increasing amounts of the CPF or CFs or reducing the RNA concentration should result in the disappearance of the 5'-*CYC1*-pc band. Alternatively, it is also plausible that the newly added poly(A) tail is bound by the excess cleavage factors in the absence of any poly(A) binding proteins.

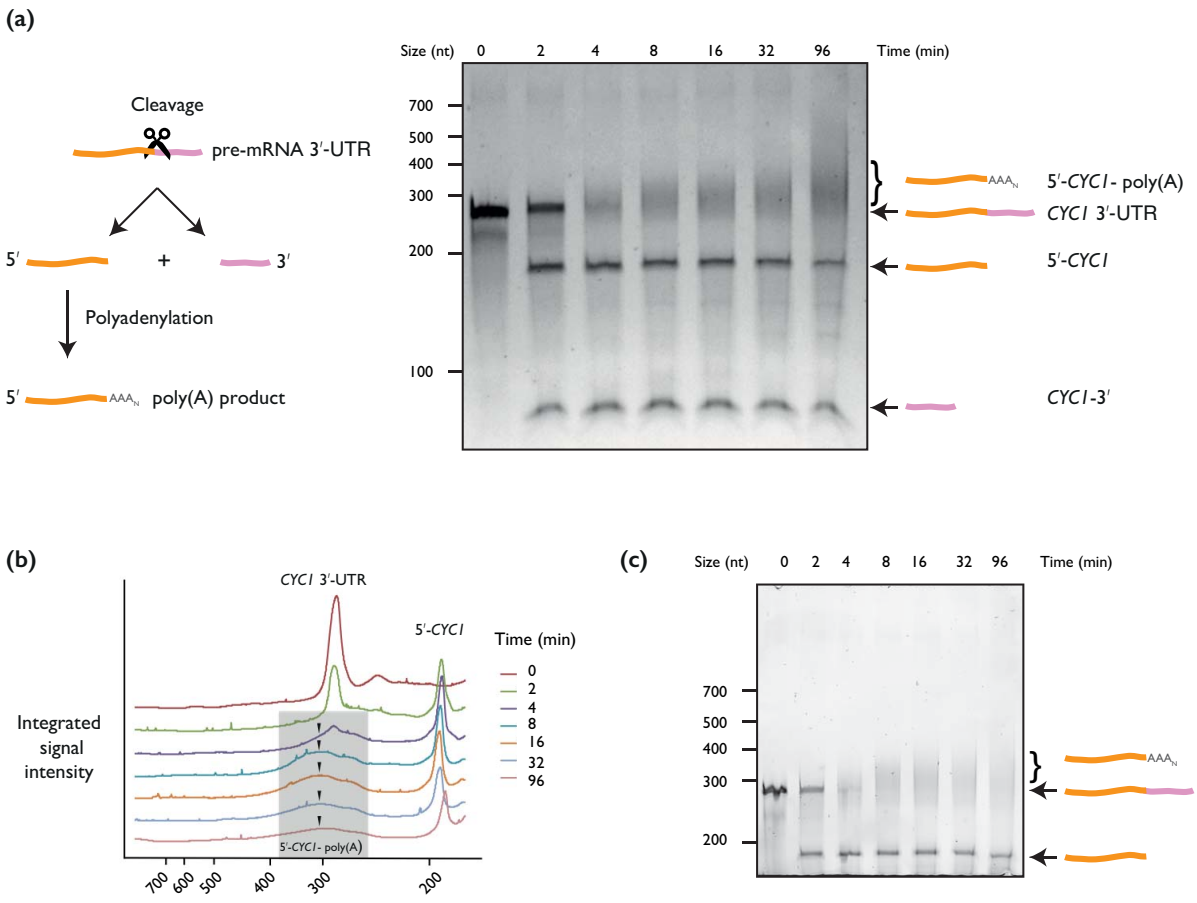


Figure 4.2: CPF has intrinsic capacity to restrict the poly(A) tail length to ~ 150As. **(a)** Schematic of cleavage and polyadenylation of a substrate RNA 3′ UTR (in left). The 5′ cleavage product shown in orange gets polyadenylated after cleavage. Coupled cleavage and polyadenylation assay of purified CPF analyzed by denaturing urea-PAGE (in right). *CYC1* is the substrate RNA. Cleavage products are 5′-CYC1 (in orange) and *CYC1*-3′ (in pink). **(b)** Quantification of the band intensities of the cleavage and polyadenylation assay to estimate the length distribution of the poly(A) tails. Highlighted in a grey box are the band intensities corresponding to polyadenylated RNAs. Arrows point the peaks of the poly(A) tail intensities. **(c)** Denaturing urea-PAGE of the assay performed for band quantification. The gel is run longer compared to the gel in (a), resulting in disappearance of the downstream cleavage product from the view.

When assays were performed with lower amounts of cleavage factors (150 nM each), it is seen that the 5'-*CYC1*-pc disappears eventually (Compare time points 96 and 128 in Figure 4.3a and 4.3b). Furthermore, under such conditions, CPF no longer restricts the length of the poly(A) tails at longer time points. This observation supports my hypothesis that, at concentrations lower than the possible K_d for RNA binding, CFs (and in turn CPF) disassociates from the poly(A) containing product RNA. The tight binding between CPF/CF and the product RNA with poly(A) tail is reduced. The CFs are now available for acting on the 5'-*CYC1*-pc.

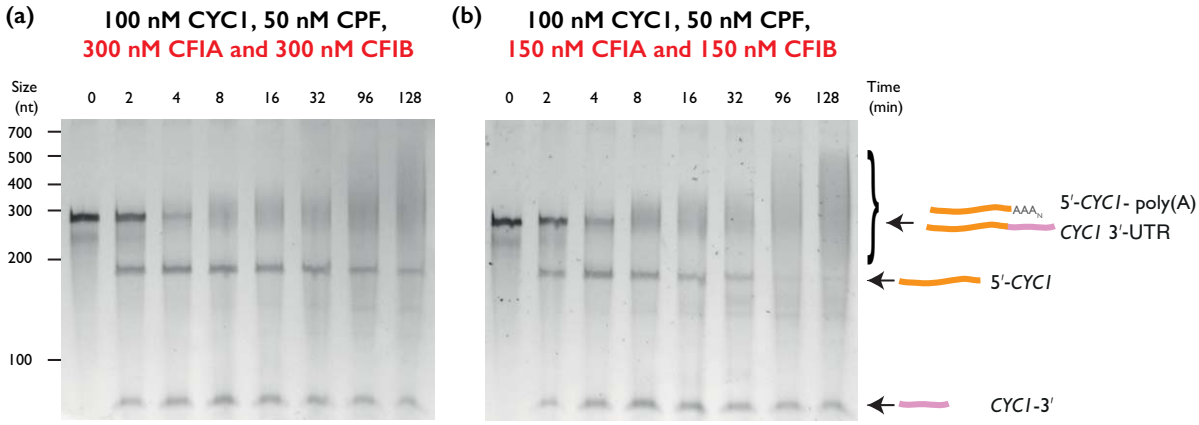


Figure 4.3: Influence of the concentration of the cleavage factors on poly(A) tail length control. Coupled cleavage and polyadenylation assay of purified CPF analyzed by denaturing urea-PAGE. *CYC1* is the substrate RNA. Cleavage products are 5′-*CYC1* (in orange) and *CYC1*-3′ (in pink). **(a)** Regulated poly(A) tail addition by CPF is observed when CF IA and CF IB are present at 300 nM. **(b)** Unregulated poly(A) tail addition by CPF when cleavage factors are present at 150 nM. (Compare the 96 and 128 min time points).

Thus, it has been found that CPF and cleavage factors have an intrinsic capacity to restrict the length of the poly(A) tails.

4.1.2 Intrinsic length control is not salt dependent

It has been observed that CF IA has a tendency to precipitate in buffers containing salts lesser than ~ 100 mM in concentration. The aforementioned intrinsic length control assays are carried out in a well-defined polyadenylation buffer that contains 75mM potassium acetate (section 6.6.1.1), raising the possibility that the observed length control exhibited by CPF could be an artifact caused by precipitating proteins. To test whether the observed intrinsic poly(A) tail length control of CPF is sensitive to salt concentrations, I performed cleavage and polyadenylation assays in buffers containing 150 mM and 225 mM potassium acetate. The results show that the intrinsic ability of CPF to restrict poly(A) tail length is not dependent on the salt concentration. Under both the salt conditions, a distribution of poly(A) tails from ~80 to 200 As were observed (Figure 4.4). However, at later time points in the assay carried our at 225 mM salt, the distribution of poly(A) containing RNAs is wider. Interestingly, the rate of cleavage at 225 mM salt was slower than when compared to that at 150 mM salt (Figure 4.4b). Such changes in cleavage and polyadenylation at higher salt conditions is likely due to a reduction in protein-RNA binding, suggesting that the interactions are possibly ionic in nature.

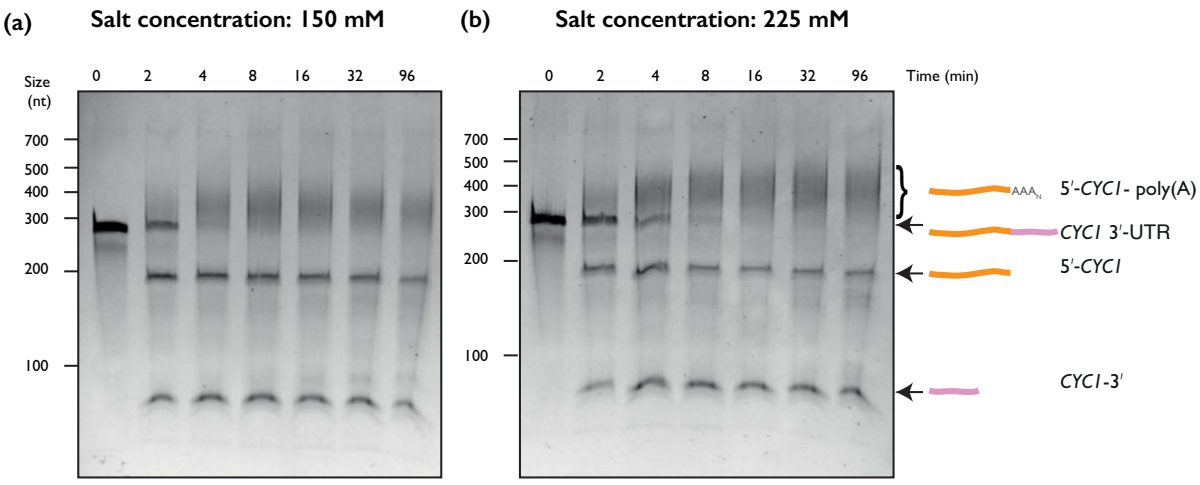


Figure 4.4: The intrinsic capacity of CPF to restrict the poly(A) tail length to ~150As is not disrupted at higher salt concentration. Coupled cleavage and polyadenylation assay of purified CPF analyzed by denaturing urea-PAGE. *CYC1* is the substrate RNA. Cleavage products are 5′-*CYC1*(in orange) and *CYC1*-3′ (in pink). CPF has intrinsic capacity to restrict the poly(A) tail length to ~ 150 As. **(a)** The reaction carried out in a buffer containing 150 mM salt **(b)** The reaction carried out in buffer containing 225 mM salt. The rate of cleavage is reduced at higher salt conditions likely reflecting the disruption of ionic interactions between protein-RNA.

4.1.3 Cleavage is not coupled to intrinsic poly(A) tail length control

Previous work has shown that cleavage and polyadenylation operate as independent reactions (Claire L Moore and Sharp 1985). Furthermore, there is no significant difference in the rates of cleavage or polyadenylation when the reactions were performed either in a coupled or uncoupled manner (Casañal et al. 2017a). However, the functional connection between cleavage and the newly observed intrinsic poly(A) tail length control by CPF is not known. To test this potential coupling, polyadenylation only assays were carried out using a pre-cleaved *CYC1* RNA (5'-*CYC1*-pc) and under length control conditions (100 nM RNA, 50 nM CPF and 300 nM CF IA, 300 nM CF IB). The data show that cleavage activity of CPF is not coupled to its intrinsic poly(A) tail length restriction property (Figure 4.5). CPF has the intrinsic capacity to restrict the poly(A) tail length of a pre-cleaved RNA to ~ 80 - ~ 200 As independent of the initial cleavage step.

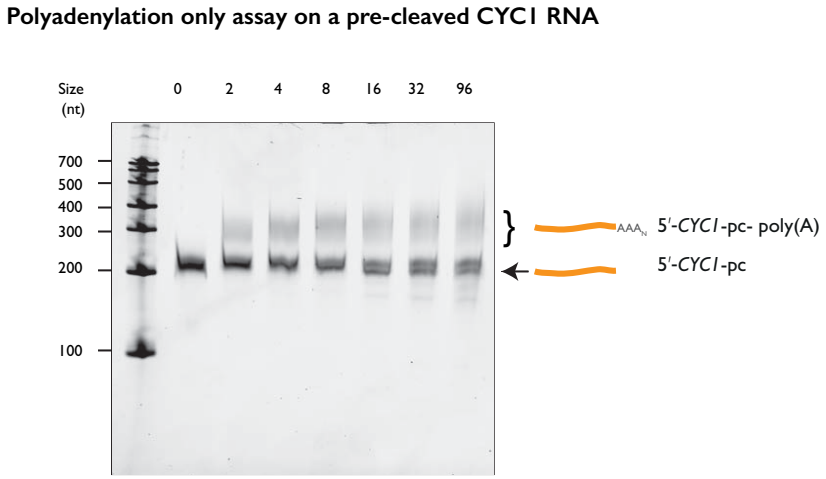


Figure 4.5: Cleavage activity of CPF is not coupled to it's intrinsic poly(A) tail length restriction property. Polyadenylation assay of purified CPF analyzed by denaturing urea-PAGE. *CYC1*-pc is the substrate RNA. CPF has intrinsic capacity to restrict the poly(A) tail length of a pre-cleaved RNA to ~ 150 As .

4.1.4 The entire 3' end processing machinery is required for length control

In order to dissect the contributions of the individual components of the 3' end processing machinery towards poly(A) tail length control, I carried out dropout experiments (Figure 6). CPF could restrict the lengths of newly added poly(A) tails only when the assay contained both CF IA and CF IB (Figure 4.6). In the absence of CF IA, the cleavage activity of CPF was highly impaired and only non-specific polyadenylation of the un-cleaved *CYC1* RNA was observed (Figure 4.6b). In contrast, in the absence of CF IB, cleavage could still occur (Figure 4.6c). However, polyadenylation by CPF was not restricted as the polyadenylated RNAs were seen as a smear in the gel and have a wide length distribution. Interestingly, unlike the reactions that contained both CF IA and CF IB, the 5'-*CYC1*-pc product disappeared in this assay. The absence of CF IB might disrupt the hypothesized co-operative binding of CPF and CF IA to RNA, allowing the protein complexes to act on any remaining substrate RNAs. Taken together CF IA and CF IB are required and have unique roles in imparting intrinsic length control on the growing polyA tail by CPF.

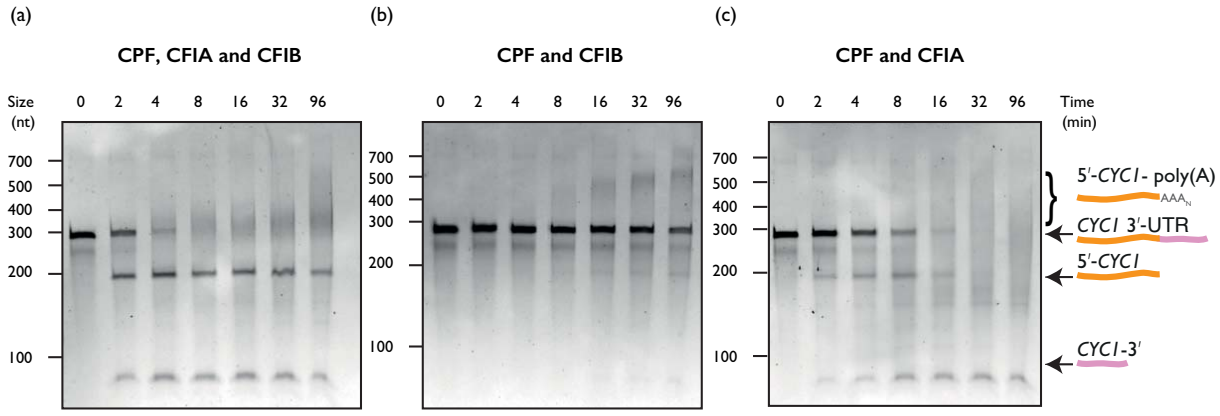


Figure 4.6: Cleavage factors are required for regulated poly(A) tail by CPF. Coupled cleavage and polyadenylation assay of purified CPF analyzed by denaturing urea-PAGE. *CYC1* is the substrate RNA. Cleavage products are 5′-*CYC1* (in orange) and *CYC1*-3′ (in pink). **(a)** Reactions carried out using CPF, CF IA and CF IB under length control conditions. **(b)** Reactions carried out in the absence of CF IA did not result in efficient cleavage and led to longer poly(A) tails. **(c)** In the reactions carried out in the absence of CF IB, all the 5′-CYC1cleaved product are consumed. However regulated poly(A) tail addition is not observed.

4.1.4.1 The absence of CF IA or CF IB results in hyper-polyadenylation

In cleavage and polyadenylation assays of *CYC1* RNA carried out in the absence of CF IA, cleavage was highly impaired (Figure 4.6b). Hence, the role of CF IA in regulated poly(A) tail addition could not be examined. To clarify CF IA function, polyadenylation only assays of a pre-cleaved *CYC1* RNA were carried out in the presence and absence of CF IA/CF IB. As mentioned in section 4.1.4, the newly made poly(A) RNAs had a tail length of ~ 80 to ~ 200 As in reactions with CPF, CF IA and CF IB (Figure 4.7a). Removing either CF IA or CF IB from the reaction results in RNAs with longer poly(A) tails (Figure 4.7b and 4.7c). This hyper polyadenylation is also observed in polyadenylation assays carried out with only CPF and substrate RNA (Figure 4.7d).

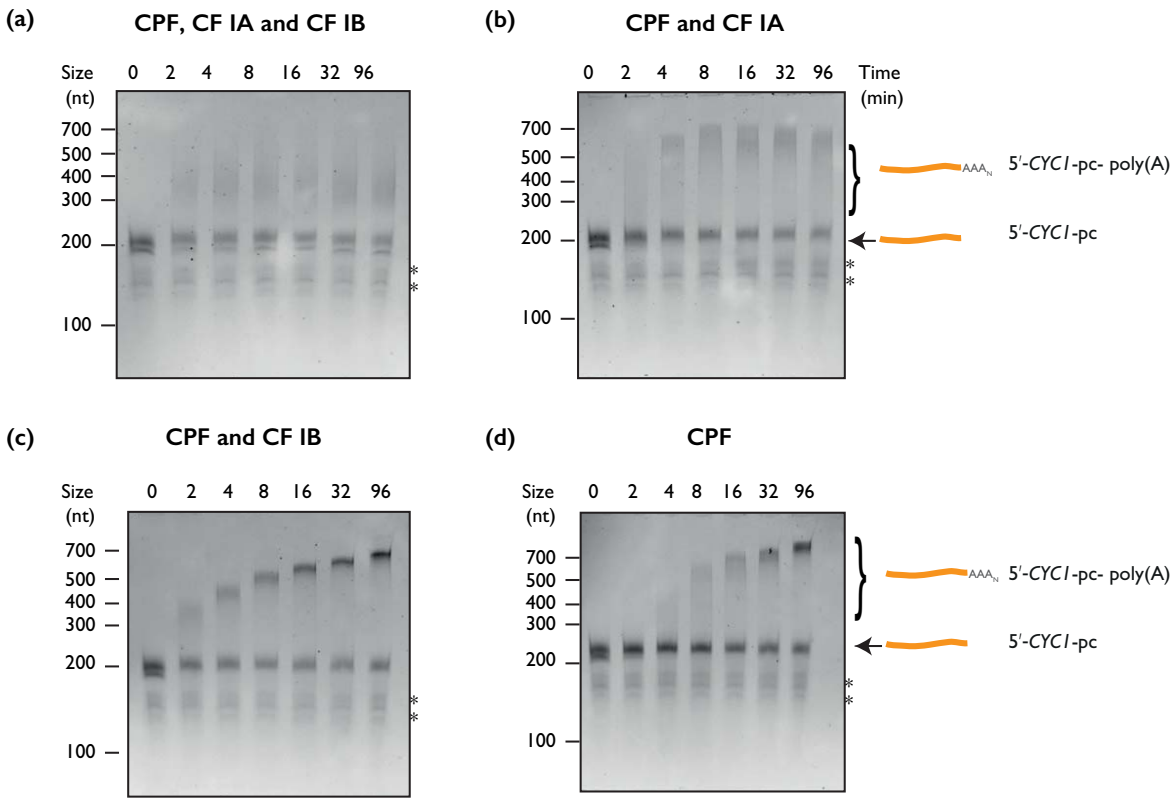


Figure 4.7: Cleavage factors are required for regulated poly(A) tail addition by CPF. Polyadenylation assay of purified CPF analyzed by denaturing urea-PAGE. *CYC1*-pc is the substrate RNA. **(a)** CPF has intrinsic capacity to restrict the poly(A) tail length of a pre-cleaved RNA to ~ 150 As. Reactions carried out in the absence of **(b)** CF IB or **(c)** CF IA or **(d)** both led to hyperadenylation. Marked in asterisk (*) are contaminant RNA bands.

In summary, in the absence of any poly(A) binding proteins, CPF functions together with CF IA and CF IB in order to restrict the poly(A) tail length of the substrate RNA to ~80 A's - ~200 A's. From our preliminary observations, I speculate that CPF and the cleavage factors form a stable complex with the polyadenylated RNA, eventually resulting in the stalling or slowdown of polyadenylation. When either CF IA or CF IB is dropped out from the assays, CPF no longer forms a stable complex with the polyadenylated RNA. Free CPF is available to add poly(A) tails to the remaining substrate RNA, eventually leading to hyper-polyadenylation.

4.1.4.2 Core-CPF does not have intrinsic poly(A) tail length control

A minimal 8-subunit core-CPF is sufficient for cleavage and polyadenylation of a substrate RNA *in vitro* (Hill et al. 2019). To understand whether this minimal CPF would be sufficient to restrict the poly(A) tails of a substrate RNA, cleavage and polyadenylation assays of *CYC1* RNA was carried out with core-CPF and the cleavage factors under length control conditions. Surprisingly, core-CPF does not have the intrinsic ability to restrict the lengths of poly(A) tails in the assays (Figure 4.8). Thus, the phosphatase module of CPF is important in poly(A) tail length restriction. This result further supports the requirement of the entire 3' end processing machinery in regulated poly(A) tail addition.

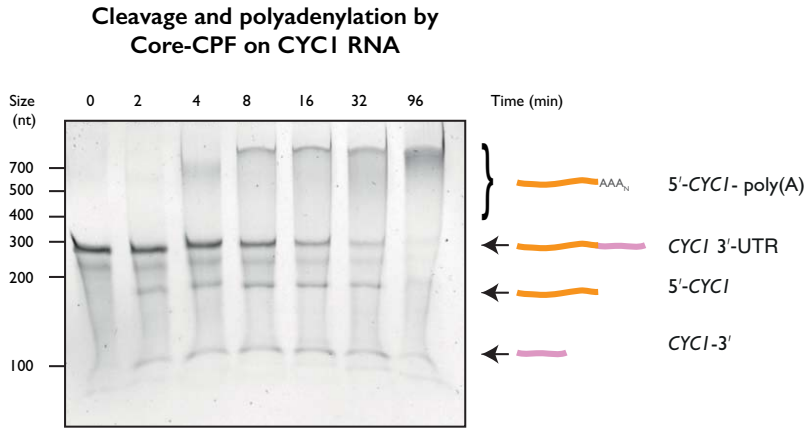


Figure 4.8: Core-CPF does not have an intrinsic capacity to restrict the poly(A) tail length. Coupled cleavage and polyadenylation assay of purified core-CPF analyzed by denaturing urea-PAGE. *CYC1* is the substrate RNA. Cleavage products are 5′-*CYC1* (in orange) and *CYC1*-3′ (in pink). The cleaved 5′-*CYC1* gets hyper-polyadenylated.

The above three sections highlight the contributions from the protein components of the 3' end processing machinery to restrict poly(A) tail length. However, contributions from *cis*-elements in the substrate RNA remain to be explored.

4.1.5 Cleaved RNA remains bound to CPF and Cleavage factors

In my current model, under non-length control conditions CPF adds an initial ~ 80 to ~ 200 As in the initial burst of polyadenylation and then falls off the substrate RNA. It can then re-bind and repositions itself on the polyadenylated RNA to continue adding more As. In the assays carried under length control conditions, CPF and the CFs remain bound to the polyadenylated RNA product. Under these conditions, CPF does not dissociate from the poly(A) RNA product to reposition itself for further polyadenylation. Such a tight binding between the 3' end processing complex and the polyadenylated RNA could prevent CPF from extending poly(A) tails beyond ~ 200 As.

In order to test this hypothesis, I carried out *in vitro* pull down experiments of CPF from the cleavage reaction mixture (Figure 4.9a). I started by performing cleavage only experiments to optimize wash conditions for the pull-downs. Once conditions for the pull-downs are established, I intend to carry out the pull-down experiments for cleavage coupled with polyadenylation assays. CPF was first immobilized onto StrepTavidin beads that were pre-equilibrated with the poly(A) reaction buffer. Excess unbound CPF was washed off. A cleavage reaction master mix containing CF IA, CF IB and the substrate RNA *CYC1* is then added to the CPF loaded Streptavidin beads. On beads cleavage assay was carried out for 30 minutes. The slurry was gently centrifuged to separate the beads from the solution. The beads were then washed three times with wash buffer containing 0.01% Tween-20. CPF was then eluted from the beads using a denaturing buffer. The elution fraction was then analyzed by denaturing urea-PAGE (Figure 4.9b). In parallel, a control reaction was carried out using a similar protocol but without any CPF on the Streptavidin beads. This would inform us whether the RNA binds to the Streptavidin beads in a nonspecific manner. Lane 1 shows the un-cleaved substrate RNA used in the reaction (Figure 4.9b). Lane 2 shows the cleavage reaction products from slurry containing CPF, CF IA and CF IB (Figure 49B). Lane 3 shows the cleavage reaction products in the slurry without any CPF (Figure 4.9b). There is no cleavage reaction in the absence of CPF. Lanes 4 and 5 show the unbound excess RNA that is separated from the Streptavidin beads (Figure 4.9b). Lane 6 shows the cleaved RNA products that were eluted from the Streptavidin beads containing CPF and the cleavage factors (Figure 4.9b). It can be seen that CPF and the cleavage factors remain bound to both the upstream and downstream RNA

products after cleavage. Lane 7 shows the elution fraction from the assay carried out in the absence of any CPF (Figure 4.9b). Compared to lane 6, only a trace amount of non-specific RNA binding can be seen in lane 7 (Figure 4.9b). These results show that CPF and the cleavage factors remain bound to both the fragments of the cleaved RNA products.

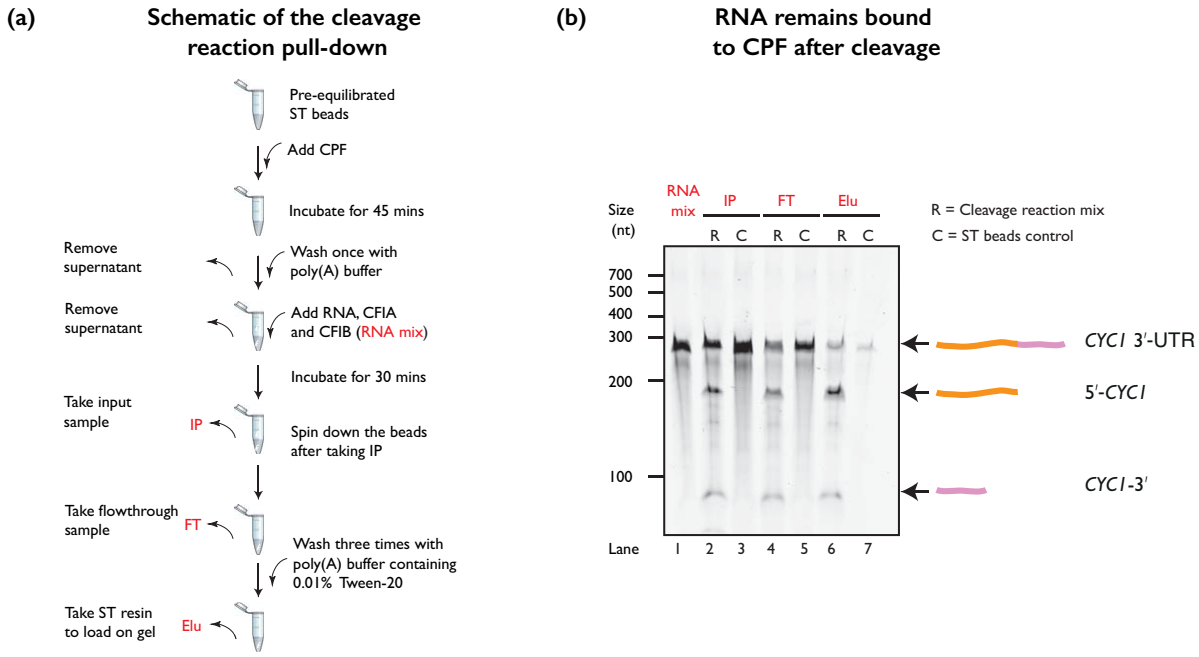


Figure 4.9: The cleaved RNA remains bound to CPF. **(a)** Schematic of the pull-down carried out on cleavage assay of CPF. CPF was immobilized on to StrepTactin (ST) beads, which was then incubated with a cleavage factor-RNA mixture. The on-bead cleavage reaction was performed for 30 minutes. The complex bound to ST beads were washed three times with buffer containing 0.01% Tween-20. Samples were taken for analysis (as indicated in red in the schematic). **(b)** The ST beads pull downs of cleavage assays with CPF analyzed by denaturing urea-PAGE. *CYC1* is the substrate RNA. Cleavage products are 5′-*CYC1* (in orange) and *CYC1*-3′ (in pink). IP = Input protein-RNA mix, FT = Flow through or unbound protein-RNA mix, Elu = Elution of the bound complex.

It remains to be tested whether CPF and the cleavage factors also remain bound to polyadenylated RNA products. Repeating the aforementioned pull-downs in the presence of ATP would clarify this. Furthermore, it would be interesting to repeat the same pull down experiment under different protein concentration conditions to observe whether complex formation would be disrupted.

4.2 Length control: Molecular Ruler or Kinetic Effect?

Despite our investigations, the exact molecular mechanism behind how CPF restricts poly(A) tail synthesis is unknown. More specifically, it is unknown if CPF possesses a molecular ruler that can "measure" the number of As being added, or if the observed poly(A) tail length restriction results from kinetic effects associated with cleavage factor and RNA binding. In order to address the aforementioned hypotheses, I produced a cleaved *CYC1* RNA containing ~30 As at the 3′ end. The molecular ruler hypothesis would suggest that polyadenylation is terminated when the poly(A) tail reaches ~ 200 As. On the other hand, a kinetic effect hypothesis would result in RNA products with poly(A) tails longer than ~ 200As.

4.2.1 Purification of mature polyadenylated RNA

A pre-cleaved *CYC1* 3' UTR (5'-*CYC1*-pc) harbouring 30 As at the 3′ end was synthesized as follows. In brief, the RNA was generated by PCR of a *CYC1* RNA 3′ UTR, with a forward primer containing the T7 polymerase start site and a reverse primer ending in poly $(T)_{30}$. The PCR product was used as a template for *in vitro* transcription. The resulting RNA was analyzed by denaturing PAGE. The desired band from the gel was excised and purified by electro-elution. A detailed protocol is described in sections 6.5 and 6.2.10.

4.2.2 CPF cleaves a mature polyadenylated RNA

A cleavage and polyadenylation assay was carried out using the *CYC1*-A30 RNA as a substrate. Surprisingly, rather than continuing adding poly(A) tails to the mature RNA, CPF cleaved the 30 As and added adenosines to the 182 nt 5'-*CYC1*-pc product (Figure 4.10). This is surprising, as the RNA contains only adenosines downstream of the cleavage site. It is well known that a U-rich downstream element after the pre-mRNA cleavage sites in UTRs plays an important role in enhancing the efficiency of cleavage (Dichtl and Keller 2001).

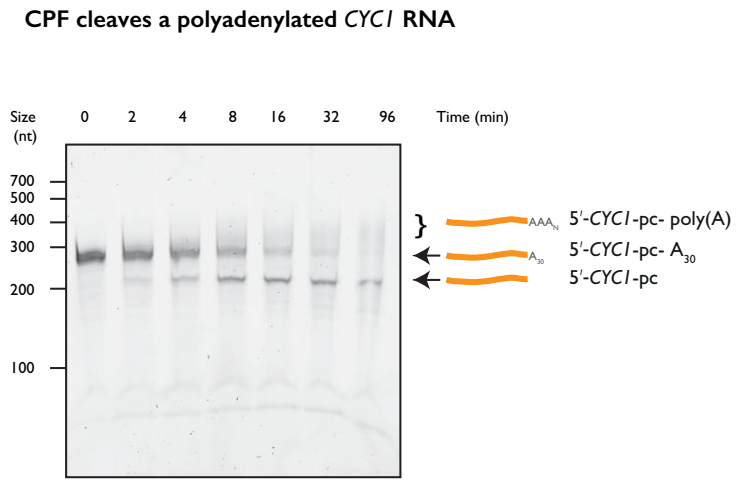


Figure 4.10: CPF cleaves a polyadenylated *CYC1* RNA. Polyadenylation assay of purified CPF analyzed by denaturing urea-PAGE. A polyadenylated *CYC1*-pc with 30 As is used as the substrate. CPF cleaves the 30As and polyadenylates the cleaved *CYC1*-pc product.

A similar observation has been previously reported (Viphakone, Voisinet-Hakil, and Minvielle-Sebastia 2008), showing that CPF could cleave a mature polyadenylated RNA in the absence of Nab2 or Pab1, and that poly(A) binding proteins protect the mature RNAs from being re-cleaved. The assays carried out here do not contain any poly(A) binding proteins. Nonetheless, previous work from our lab has shown that the cleavage factors CFI A and CFI B do not bind to poly(A) stretches, whereas the 8-subunit core-CPF can weakly bind poly(A) RNA (Hill et al. 2019). This suggests that CPF could putatively bind the downstream As in a manner similar to its binding to downstream Us, thereby providing this unexpected cleavage. This also poses the question of whether CPF doesn't re-cleave a poly(A) tail *in vivo*. It is likely that the poly(A) tail is immediately bound by poly(A) binding proteins that protects it from being re-cleaved.

4.2.3 The eight subunit core-CPF does not re-cleave a polyadenylated RNA

To assess if core-CPF can recapitulate the cleavage of *CYC1*-A30 by CPF, I repeated cleavage and polyadenylation assays of *CYC1*-A30 with core-CPF and the cleavage factors (Figure 4.11). Surprisingly, core-CPF did not cleave a mature polyadenylated RNA. Instead, it continued to add poly(A) tails to the mature transcript. Consistent with our previous observations, assays with core-CPF resulted in hyper-polyadenylated transcripts (Figure 4.8, Figure 4.11). As the only difference between core-CPF and CPF is the phosphatase module, it is conceivable that the phosphatase module is involved in re-cleavage and poly(A) tail length restriction, either by providing additional RNA binding surfaces or by providing an allosteric effect in cleavage and polyadenylation activity. The above result could act as a starting point towards further research aimed at understanding how phosphatase module of CPF contributes to cleavage.

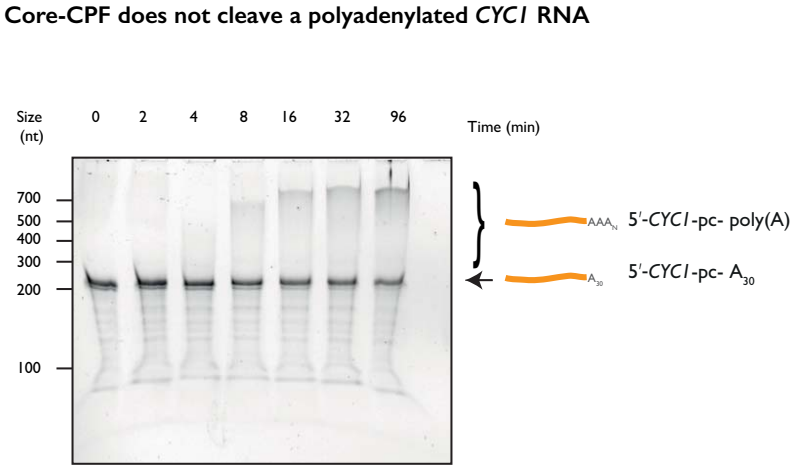


Figure 4.11: Core-CPF does not cleave a polyadenylated *CYC1* RNA.
Polyadenylation assay of purified core-CPF analyzed by denaturing urea-PAGE. A mature polyadenylated *CYC1*-pc with 30 As is the substrate RNA. Core-CPF continues to polyadenylate the mature RNA substrate. There is not poly(A) tail length control.

The findings described so far reveal for the first time that CPF/CFs have intrinsic length control of polyadenylation and that this property is dependent on the ternary CPF-CF I assembly on a substrate RNA. Despite these findings, it remains unclear how CPF is able to restrict poly(A) tails to certain lengths. The observed re-cleavage of previously polyadenylated RNAs hinders experiments to test the hypotheses of poly(A) tail length control mechanism. There are two ways by which the re-cleavage of RNA could be circumvented. Firstly, the length control polyadenylation assays could be carried out in the presence of a small molecule cleavage inhibitor, preventing re-cleavage. Nonetheless, there is no known inhibitor of *S. cerevisiae* Ysh1, complicating this approach. The alternative would involve establishing a recombinant system to produce fourteen-subunit CPF. Establishing a recombinant system to produce CPF will provide a means to make mutant complexes that was not previously possible. For example, creating a Ysh1 catalytic mutant within CPF will allow us to carry out polyadenylation assay of *CYC1*-A30, enabling us to understand the molecular mechanism of poly(A) tail length control by CPF.

4.3 Production and Characterization of a 14-subunit recombinant CPF

The full fourteen-subunit CPF has been traditionally purified in the lab using a TAPS-tag on the Ref2 subunit (Nedea et al. 2003; Schreieck et al. 2014). Although the preparation yields a pure protein complex (Figure 4.1b), the yield of the purification is very low. Starting from ~1 kg of *S. cerevisiae* cells typically yields ~1 mg of CPF. This process thus requires the impractical use of a fermenter to obtain sufficient CPF to carry out extensive biochemical assays. Furthermore, the entire purification process involves multiple carefully controlled steps and is difficult to reproduce consistently. The process is also laborious and time consuming. Finally, purification from endogenous sources often results in CPF, which is contaminated by trace amounts of other yeast proteins. Moving to a fully recombinant system for CPF production would lead to sufficient yields from only a few liters of insect cells. Furthermore, the larger amount of protein will enable more stringent purification of the CPF complex. Here I describe my efforts towards producing a fully recombinant fourteen-subunit CPF.

4.3.1 Producing a fully recombinant CPF (rCPF)

There are several ways to recombinantly assemble a protein complex. A cost effective method would involve using a bacterial overexpression system consisting of a bacteriophage T7 promoter (Studier and Moffatt 1986). A straightforward approach to reconstitute multi-protein complexes is to purify the individual proteins separately using such a system, mix the individual components, and carry out size exclusion chromatography to purify the larger complex. Simpler complexes such as the cleavage factor I A sub-complexes Rna14/Rna15 and Pcf11-Clp1 can be recombinantly expressed using pETDuet and pRSFDuet vectors respectively (Noble et al. 2004; Gordon et al. 2011). In the aforementioned vector systems, individual genes are cloned into different open reading frames (ORFs), and both the genes are simultaneously expressed in bacterial cells. However, eukaryotic protein machineries including CPF often require co-translational assembly in order to form a properly folded complex. In addition, eukaryotic proteins also often carry post-translational modifications (PTMs) that are unique. Using a bacterial expression system to express yeast or mammalian proteins could lead to lack of PTM patterns. The requirement for complex PTMs can be met by using mammalian expression systems such as HeLa or HEK-293 cells (Bandaranayake and Almo 2014). However, scaling up of mammalian protein production platforms for higher yields of proteins remains difficult. The usage of baculovirus-mediated insect cell over-expression allows both high yield and preservation of most PTMs.

The baculovirus-mediated insect cell over-expression system uses a late polyhedron promoter that includes a tetra nucleotide TAAG transcription start site (O'Reilly, Miller, and Luckow 1994). The baculovirus rod-shaped capsid usually encloses a circular, double stranded and supercoiled DNA of ~ 80 to 180 kbp (Harrap 1972; Chambers et al. 2018). The capsid dimensions can be expanded to make space for bigger genomes, making it suitable for use in recombinant protein production. Current protocols use a Tn7 transposition system in bacterial cells in order to generate bacmids. Bacmids are shuttle vectors containing the genes of interest (GOI) and can be propagated in both bacterial and insect cells. Bacmids are transfected into adherent insect cells to produce the initial virus. The initial virus is subjected to further rounds of amplification to increase the titer before being used to infect large scale insect cell cultures.

Some of the commonly used baculovirus systems include MultiBac and OmniBac systems (Bieniossek et al. 2012; Thimiri Govinda Raj, Vijayachandran, and Berger 2014). These systems enable the expression of multiple subunits from a single baculoviral DNA, which is advantageous because only one DNA construct is transfected into the insect cells in order to produce the protein complex, reducing heterogeneity. One of the drawbacks of the MultiBac or OmniBac systems is the time required to generate large DNA constructs by restriction enzyme based cloning. In addition, the Cre-loxP based recombination method of the MultiBac system makes it challenging to control the number of genes that can be incorporated into the final vector to be used in protein expression.

The use of a Gibson assembly reaction for the generation of baculoviral expression vectors has now enabled the user to have precise control over the number of gene copies that are introduced into the final vector (Weissmann et al. 2016). The usage of Gibson assembly as opposed to traditional techniques has enabled rapid and efficient generation of many expression vectors in parallel (Gibson et al. 2009; Weissmann et al. 2016). This new modified system is referred to as biGBac. In this system, the efficiency of the Gibson assembly is greatly enhanced due to computationally identified "optimal" linker overhang sequences, characterized by melting temperatures, propensity to form higher order structures and minimal probability to result in incorrect assembly. The first step of cloning in biGBac includes assembling the GOIs, each containing their own promoter and terminator, into the pBig1 family of vectors, each of which can be used to assemble up to five gene of interest. Multi-gene fragments from the pBig1 family of vectors can be integrated into the linearized pBig2 family of vectors in a hierarchical fashion by unique linker sequences from the pBig1 plasmids flanking the multi-gene insert. In this manner, a final pBig2 vector that can contain up to twenty-five GOI can be assembled. This method was adopted for the production of recombinant CPF. More details about the biGBac system can be found in methods sections 6.2.2 and 6.2.9.

4.3.1.1 Cloning a full 14-subunit complex into a single vector

In order to produce the entire CPF recombinantly, I cloned all fourteen subunits of CPF, including a C-terminally tagged Ref2 subunit similar to endogenous purifications, into a single vector using the biGBac system. The polymerase module subunits (Cft1, Pfs2, Yth1, Fip1 and Pap1) were assembled into pBIG1a, the nuclease module subunits (Cft2, Ysh1 and Mpe1) into pBIG1b, five of the phosphatase module subunits (Ssu72, Pti1, Glc7, Ref2-SII and Swd2) into pBIG1c and the sixth phosphatase subunit (Pta1) into pBIG1d (Figure 4.12 a). The genes of interest contained a Swa1 restriction site at both ends, whereby Swa1 digestion can be used to confirm the incorporation of the individual CPF genes into the pBIG1 family of vectors. Following incorporation into the pBIG2 plasmid, PacI restriction sites either side of the multigene fragments enable confirmation of the incorporation of fragments into pBIG2 (Figure 4.12b).

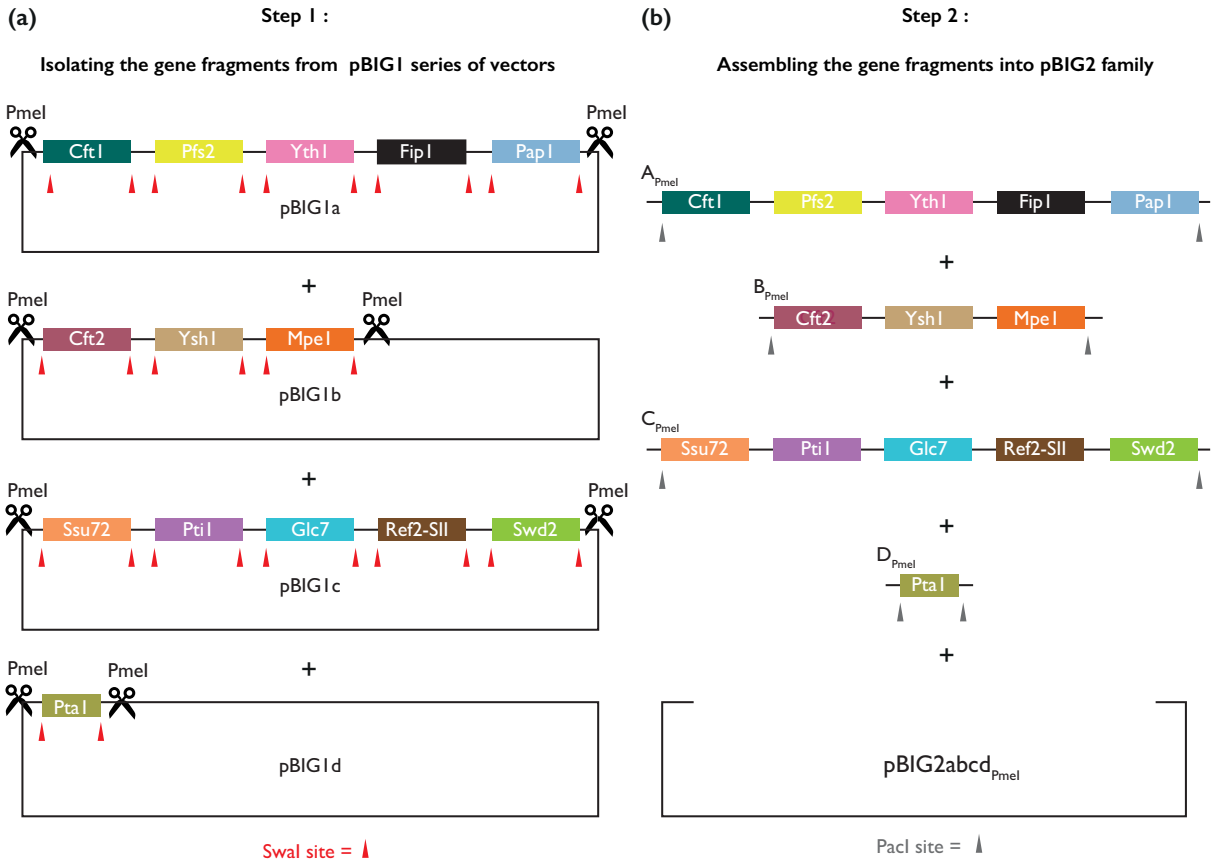


Figure 4.12: Cloning strategy to make recombinant CPF. **(a)** The individual genes of CPF are first cloned into pBIG1 family of vectors by means of Gibson assembly. The presence of individual genes within pBIG1 is verified by digesting with SwaI. Each one of the genes contain a SwaI digestion site both upstream and downstream as shown by a red arrow head. The gene fragments from each one of the pBIG1 series of vectors are isolated by digesting with PmeI enzyme. **(b)** The gene assemblies isolated by PmeI are assembled in a hierarchical order into a pBIG2 vector that has also been digested with PmeI. The incorporation of the gene assemblies into the pBIG2abcd is verified by digestion with PacI enzyme. The gene cassettes are flanked by PacI sites as shown by a grey arrowhead.

The incorporation of such gene fragments into pBIG2abcd was verified by using diagnostic restriction digestion tests and PCR using gene specific primers (Figure 4.13a, 4.13b). Digesting pBIG2abcd with Pac1 would release all the gene expression cassettes and would result in a pattern with five bands. Four of those bands would correspond to each one of the expression cassettes and one band for the backbone of pBIG2abcd (Figure 4.13a). After screening more than ~60 colonies, two colonies

referred to as CPF1 and CPF2 were found to have Pac1 digestion pattern that could likely correspond to the correct clone. The presence of all the fourteen genes within CPF1 and CPF2 clones was further confirmed by carrying out PCRs with gene-specific primers (Figure 4.13b).

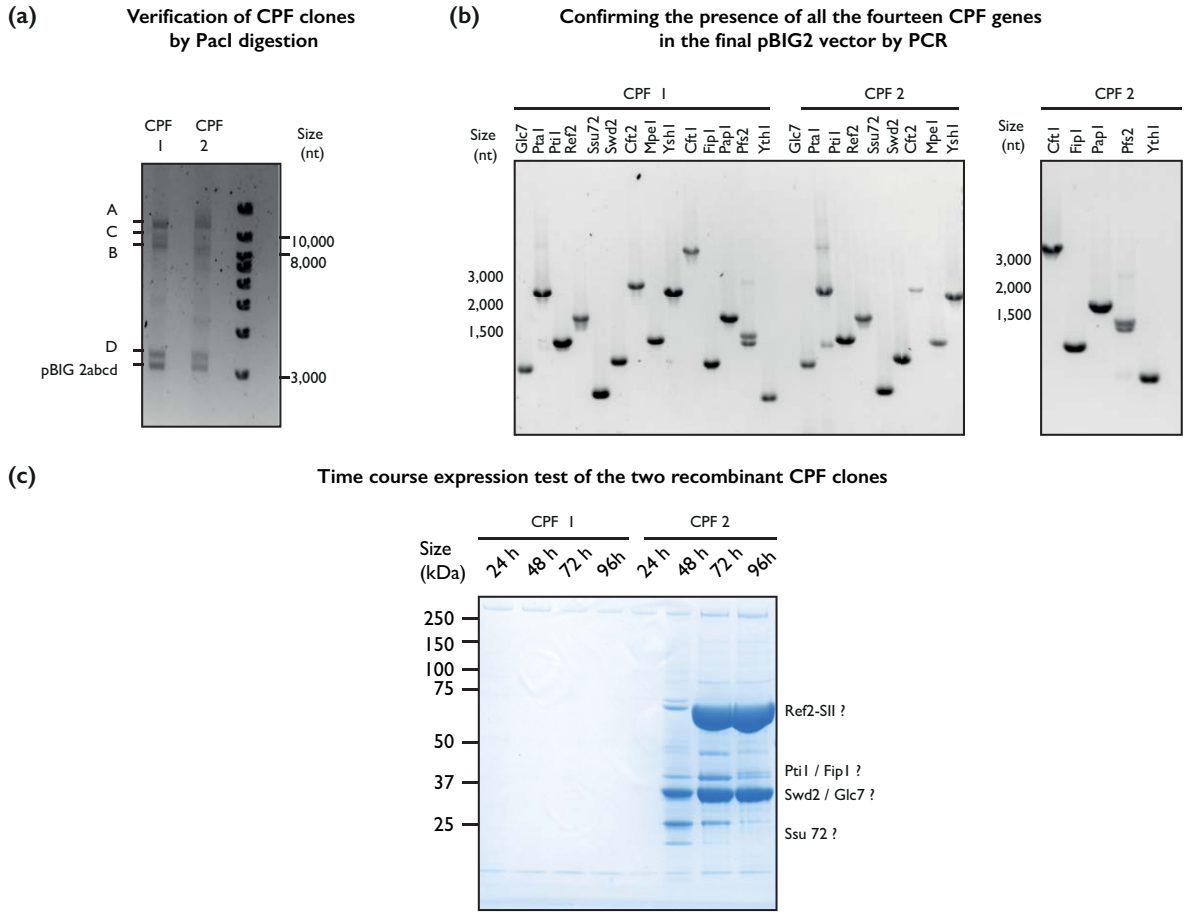


Figure 4.13: Verification of the cloning of recombinant CPF and testing the protein expression. (a) The enzyme Pac I cleaves the pBIG2 series of vectors at the cloning site. Digestion pattern would reveal the size of the gene fragments inserted into pBIG2. In case of CPF cloning, the expected size of the gene fragments are A ~12000, C ~8100, B ~8900, D ~3200 and pBIG2 ~3000 nucleotides. (b) The construct CPF 1 and CPF 2 are further verified for the presence of all the fourteen subunits by performing PCR using gene specific primers. The products were analyzed on a 1% agarose-TAE gel (c) Time course expression test of a virus carrying all fourteen CPF subunit genes. Protein complexes were pulled down with Streptavidin beads. Ref2 subunit contains two strep tags. The isolated proteins were analyzed by SDS-PAGE.

4.3.1.2 Attempts at over-expressing the 14-subunit bacmid

Once the gene composition of CPF1 and CPF2 was verified by PCR, the ~44 kb plasmids were transformed into EMBACY^{TM} *E. coli* cells. Bacmids were generated from the cells using the protocol described in section 6.3.3.1. The isolated bacmids were transfected into *Sf9* cells to produce a primary virus that is further amplified to yield higher titer viruses. The resulting virus was used to infect 50 ml cultures of *Sf9* cells to carry out protein expression tests. The pull-downs revealed that a full fourteen-subunit CPF was not produced using neither CPF1 nor CPF2 viruses (Figure 4.13c). CPF1-infected *Sf9* cells yielded no protein, likely because all CPF genes were removed from the viral genetic material. The size of the gene cassette that contains all fourteen CPF subunits is ~ 44 kb, approximately 50% of the viral genome. Under selection pressure, it is often the case that inserted gene constructs are eliminated from the genome. On the other hand, CPF2-infected *Sf9* cells yielded a sub-complex of CPF containing Ref2, Pti1 (or Fip1), Swd2 (or Glc7) and Ssu72 subunits. However, the full stoichiometric CPF complex was not pulled down. It is plausible that the nuclease and polymerase module subunits (core-CPF) were not expressed in the cells because their genes were similarly ejected out from the virus. An alternative explanation is that the genes corresponding to core-CPF were in fact expressed, but do not get pulled down via the Strep-II tag on the phosphatase module subunit Ref2. Thus, the strategy of simultaneously expressing all fourteen subunits of CPF from a single vector did not yield full CPF and an alternative strategy was required to generate recombinantly expressed full CPF.

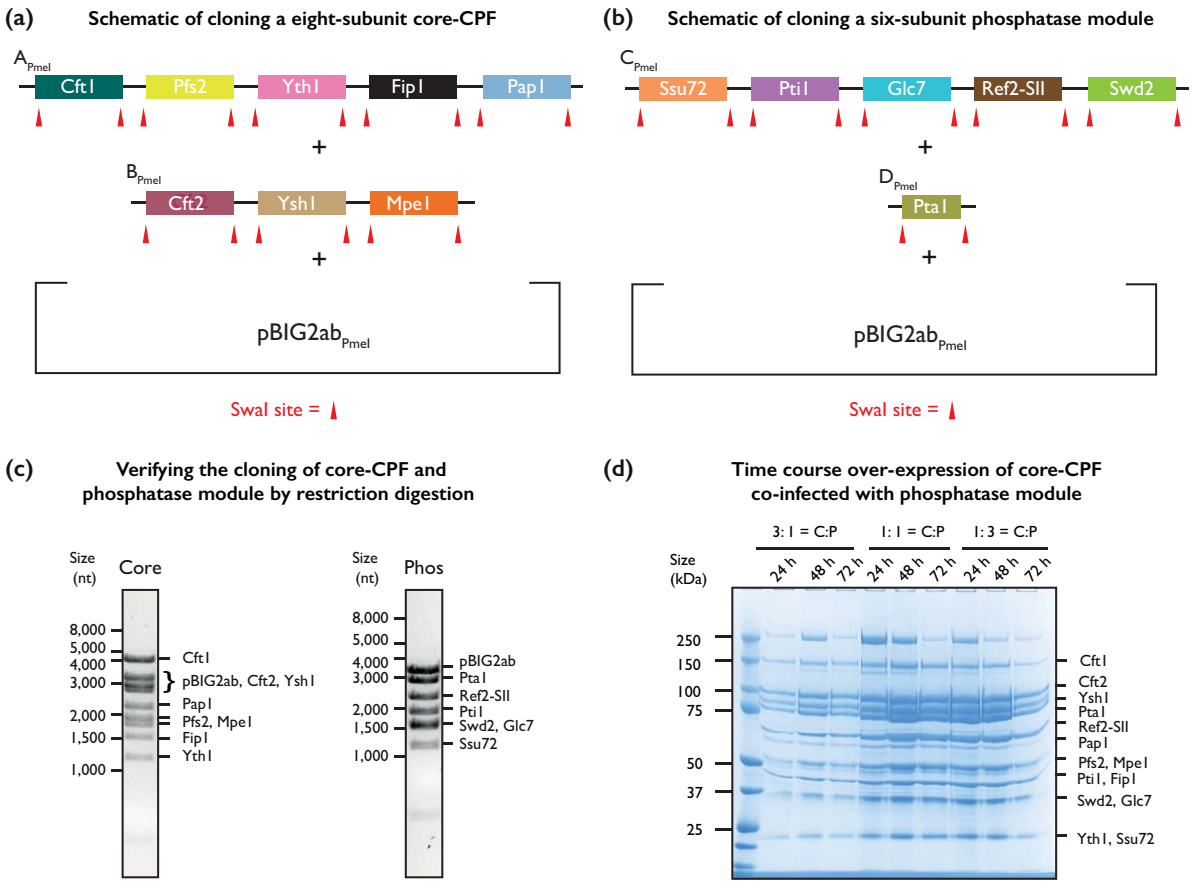


Figure 4.14: Revised strategy to make rCPF. Schematic representation of the strategy used to produce recombinant CPF. **(a)** The nuclease and polymerase module was cloned into a single vector pBIG2ab. **(b)** The six-subunit phosphatase module was cloned into a separate pBIG2ab **(c)** The constructs core and phos are verified for the presence of all core-CPF and phosphatase module subunits respectively. Swa1 digestion of the corresponding pBIG2ab vectors were analyzed on a 1% -agarose TAE gel **(d)** Time course expression test of co-infecting a virus carrying all the core-CPF genes with a virus with all the phosphatase subunits. The infections were performed in different ratios of the core and phos viruses. Protein complexes were pulled down with Streptavidin beads. Ref2 subunit contains two strep tags. The isolated proteins were analyzed by SDS-PAGE.

4.3.1.3 Co-infection of core-CPF and phosphatase module viruses results in production of recombinant full CPF

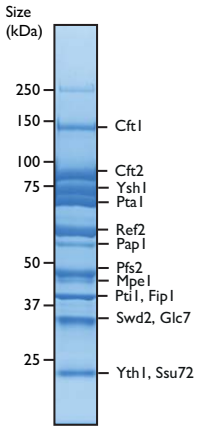
Previous work from our lab had shown that mixing *Sf9* cell pellets overexpressing core-CPF and the phosphatase module during lysis and subsequent purification from the mixture of cells did not result in the assembly of the full CPF. This highlights the importance of *in vivo* assembly in the formation of CPF, despite the modules being functionally separate.

I therefore adopted an approach where core-CPF and the phosphatase module would be co-expressed from separate vectors (viruses) in the same cells. I cloned an eight-subunit core-CPF into the pBIG2ab vector without any affinity tags and a six-subunit phosphatase module into a separate pBIG2ab vector with a Strep-II tag on the Ref2 subunit (Figure 4.14a and Figure 4.14b).

The clones (named "Core" and "Phos") were verified by performing restriction digestion using Swa1. This released individual gene fragments contained within the pBIG2ab vector (Figure 4.14c). The verified clones were then transformed separately into EMBACY^{TM} cells to generate bacmids and viruses of each multigene cassette were produced separately. High-titer viruses containing Core and Phos were produced and mixed together in different ratios for infections. I tested three different ratios (Core:Phos of 1:3, 1:1, and 3:1 by volume) for co-infection of *Sf9* cells. Pull-downs were carried out from *Sf9* cell lysate (Figure 4.14d). All virus ratios resulted in the production of a full fourteen-subunit CPF as analyzed by SDS-PAGE (Figure 4.15a). Interestingly, the amount of complex obtained from *Sf9* cells was higher when Core and Phos viruses were mixed in equal volume amounts (1:1 ratio) or when Phos was present in excess (1:3). This is likely because the affinity tag is on the phosphatase module subunit Ref2. Thus, when the phosphatase module is present in excess, the total amount of CPF being pulled down is maximized. The identities of the fourteen different proteins were further examined using tandem mass spectrometry to ensure that the preparation contained all bona-fide *S. cerevisiae* CPF proteins and that there were no contaminants or homologues from *Sf9* cells incorporated into CPF. Figure 15b lists the fourteen identified CPF subunits and describes the total number of spectra uniquely assigned to the individual proteins identified.

Here I show a reliable and robust protocol for the over-expression of full CPF.

(a) SDS-PAGE of CPF strep pull down



(b) Identification of subunits of CPF produced using Sf9 cells

Protein	Species	Exclusive spectrum count	Percentage coverage
Cft1	*S.cerevisiae*	205	70
Cft2	*S.cerevisiae*	310	76
Ysh1	*S.cerevisiae*	384	82
Pta1	*S.cerevisiae*	393	70
Ref2	*S.cerevisiae*	238	72
Pap1	*S.cerevisiae*	163	67
Pfs2	*S.cerevisiae*	65	69
Mpe1	*S.cerevisiae*	272	62
Pti1	*S.cerevisiae*	126	69
Fip1	*S.cerevisiae*	82	55
Swd2	*S.cerevisiae*	132	60
Glc7	*S.cerevisiae*	63	65
Yth1	*S.cerevisiae*	65	65
Ssu72	*S.cerevisiae*	212	66

Figure 4.15: Tandem mass spectrometry confirms the subunit identities of recombinant CPF (rCPF) **(a)** SDS-PAGE of recombinant CPF isolated from *Sf9* cells using Strep pull down **(b)** Tandem mass spectrometry confirms the presence of all fourteen CPF subunits in the recombinant preparation from Sf9 cells. The total number of spectra uniquely assigned to a protein is represented as exclusive spectrum count.

4.3.2 Purifying and characterizing recombinant CPF

4.3.2.1 Overall purification strategy

The higher yields of recombinantly expressed CPF allow more stringent purification to be carried out. I developed a purification protocol for recombinant CPF (hereafter referred to as rCPF) resulting in a highly pure and homogenous preparation.

The various steps in the purification of rCPF are illustrated in figure 4.16a. The first step involves pulling down the Ref2 subunit of CPF via its C-terminal Strep-II tag. SDS-PAGE of the protein complexes isolated after Streptavidin pull downs reveal a ~ 200 kDa contaminant along with the other CPF bands (shown by * in figure 4.16b). Furthermore, the eluted fractions did not contain stoichiometric amounts of the CPF subunits. The phosphatase and nuclease module subunits are present in higher amounts relative to the polymerase module subunits (Figure 4.16b). To remove the additional contaminating proteins and to clarify the subunit heterogeneity present within CPF, the eluted fractions from Strep pull downs were subjected to anion-exchange chromatography. The bound proteins were separated by a shallow gradient elution starting from 150 mM KCl to 1 M KCl containing buffer (Figure 4.16c).

The chromatogram of the anion-exchange step shows two protein peaks (peak 1 and peak 2) and two nucleic acid rich peaks (peak 3 and peak). SDS-PAGE analysis of peak 1 revealed a nine-subunit nuclease-phosphatase complex (Figure 4.16d). The existence of such a complex has not been previously described before. During the purification of native CPF, we usually isolate a seven-subunit APT complex along with full CPF (Lidschreiber et al. 2018). APT is a seven-subunit complex made up of all the six phosphatase module subunits along with the protein Syc1 (Nedea et al. 2003; Lidschreiber et al. 2018). In the recombinant system, however, Syc1 is not present. I hypothesize that the Syc1-interacting region in the phosphatase module interacts with Ysh1 instead as the C-termini of Ysh1 and Syc1 are homologous. In this complex, it is possible that Ysh1 can be further associated with Mpe1 and Cft2, resulting in the novel nuclease-phosphatase module. Interestingly, later experiments revealed that the nuclease-phosphatase complex is unable to cleave a substrate *CYC1* RNA (appendix 8.9), highlighting the importance of the polymerase module subunits in the nuclease activity of CPF. Peak 2 contains full CPF as revealed by SDS-PAGE

(Figure 4.16d). It is interesting to note that the stoichiometry of Pap1 seems to vary across the peak (highlighted by a pink box). This is reminiscent of the shallow gradient anion-exchange purification of native CPF carried out previously in the lab, where CPF associated with two, one, or no copies of Pap1 was observed (Easter 2014). Purification of the recombinant polymerase module described in section 2.1.1 also revealed a similar polymerase complex that contains two, one, or no copies of Pap1. The CPF with differential Pap1 stoichiometry can partially explain the non-Gaussian nature of the profile. This can be further explained by the lack of the ~200 kDa *Sf9* contaminant at higher salt elution. The fractions without this contaminant (and containing one copy of Pap1) were used in further purification, as highlighted by curly brackets in Figure 4.16d.

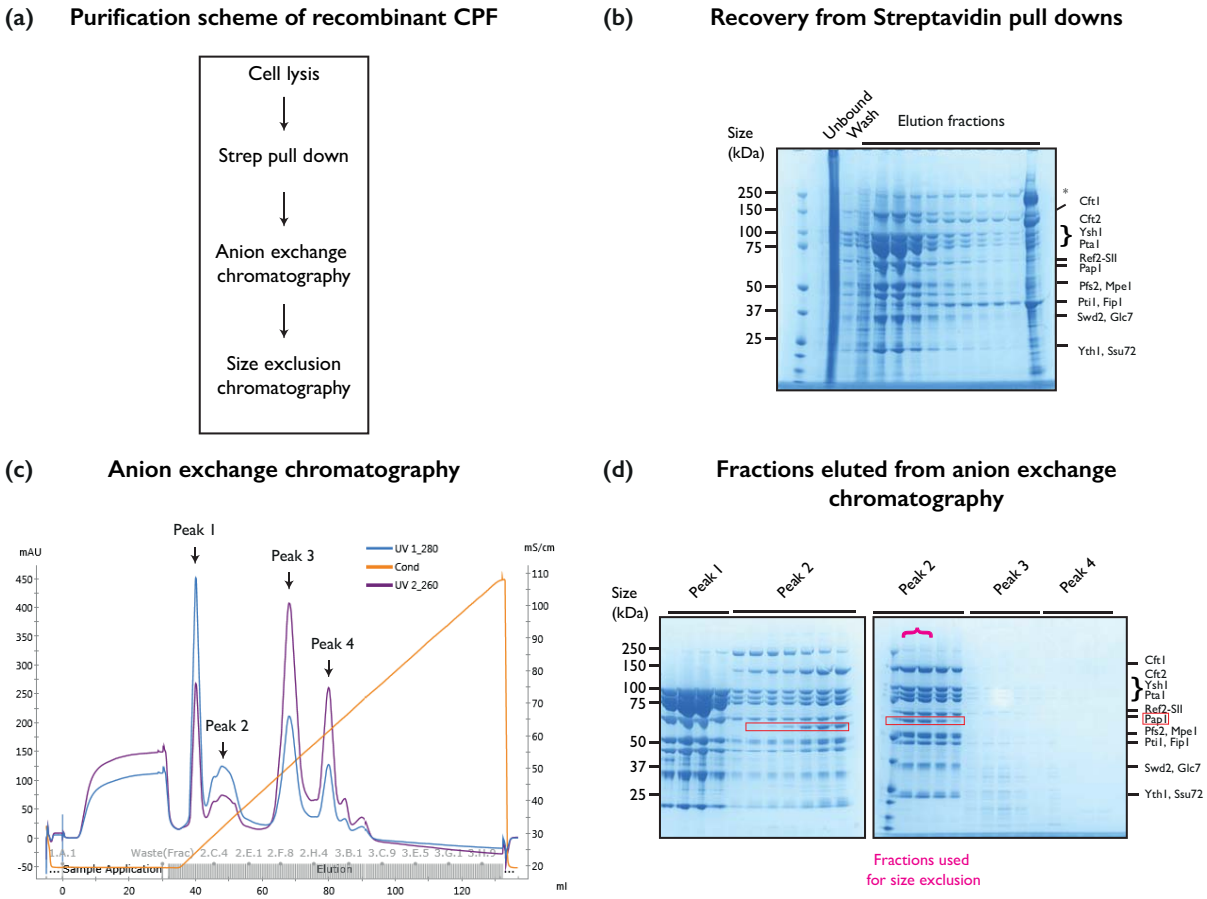


Figure 4.16: Purification of recombinant CPF (rCPF). **(a)** Flow chart describing the different steps involved in the purification of rCPF. **(b)** Streptavidin pull down of the Sf9 cell lysate was analyzed by SDS-PAGE. **(c)** Anion exchange chromatography using a Resource Q column. **(d)** The four different peaks observed in the anion exchange step were analyzed using SDS-PAGE. Peak 1 was found to contain the nuclease-phosphatase module. Peak 2 contains the full recombinant CPF. It can be seen that across peak 2, the stoichiometry of Pap1 polymerase is varying (highlighted by a red box). Peak 2 fractions highlighted by pink curly bracket were pooled together for further purification by size exclusion chromatography.

CPF fractions from anion exchange chromatography were further purified by size exclusion chromatography (Figure 4.17a). CPF eluted from the Superose 6 Increase 3.2/300 column at 1.25 ml. SDS-PAGE of this peak revealed a highly pure CPF preparation with stoichiometric amounts of each subunit (Figure 4.17b). The CPF sample was concentrated, flash frozen in liquid nitrogen, and stored at -80 °C for future biochemical experiments. The frozen sample was later thawed and run on a size exclusion column to assess its integrity. It was found that the frozen CPF did not show significant aggregation or differences in behavior on a size exclusion column (appendix 8.10).

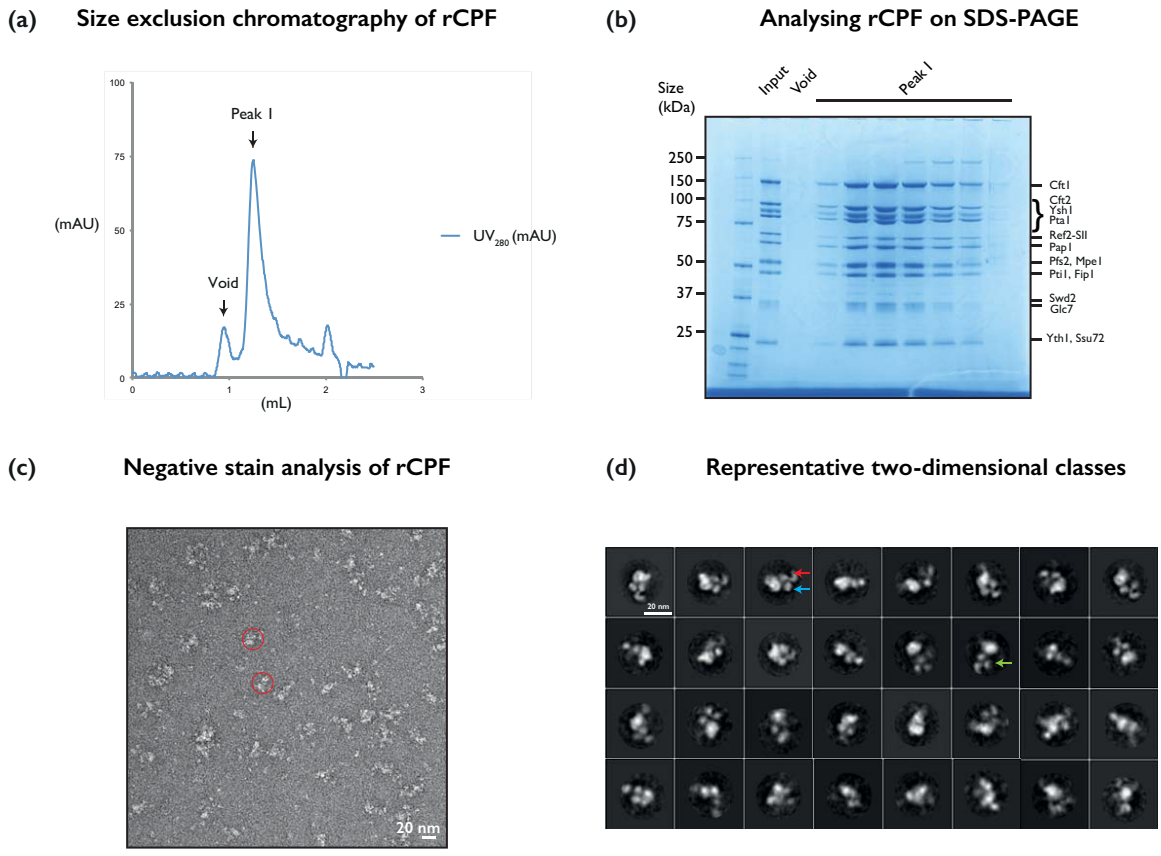


Figure 4.17: Negative stain electron microscopy of rCPF (a) Chromatogram of rCPF run on a Superose 6 Increase 3.2/300 column in SEC buffer. (b) The peak fractions were analyzed using SDS-PAGE. (c) Negative stain electron microscopy analysis of rCPF. A representative micrograph at 21K x magnification and -1.7 μM defocus. A red circle highlights a single particle of CPF. (d) Selected 2-D class averages of aligned particles highlighting sample heterogeneity. Red arrow points to the density likely to be Ysh1, blue arrow to Cft2 and green arrow to phosphatase subunits.

4.3.2.2 Negative stain electron microscopy of rCPF

To further assess sample homogeneity, I carried out negative stain electron microscopy of un-crosslinked rCPF immediately following size exclusion chromatography. The micrographs revealed well-distributed individual particles that were roughly ~20 nm in their maximum dimension (highlighted by a red circle in figure 4.17c). However, aggregates and smaller particles were also visible. Selected 2-D class averages of aligned CPF particles reveal distinct ~20 nm particles (Figure 4.17d). The best-aligned density looks very similar to the Cft1-Pfs2-Yth1 subunits of the polymerase module (Figure 2.9a). The scaffold seems to anchor several other modules that appear as globular densities. The globular density appears to be very dynamic with respect to the central scaffold, likely reflecting the inherent flexibility of CPF. The identities of the densities cannot be ascertained. However, the density shown using blue arrows in Figure 4.17d could correspond to Cft2 (Hill et al. 2019), the beta propeller containing protein known to closely associate with the polymerase module. The density shown using red arrow in Figure 4.17d could correspond to the endonuclease Ysh1 that is flexibly tethered to CPF via Mpe1 (Hill et al. 2019). The densities resembling three circular dots (highlighted using green arrows) in figure 4.17d could be the phosphatase module subunits.

In summary, the 2-D class averages show that the sample contains ~20 nm particles and that subunits within CPF are flexibly tethered to the polymerase module, which acts as the central scaffold.

4.3.3 Biochemistry of rCPF

Having established a protocol for the production of highly pure milligram quantities of recombinant CPF from practical volumes of insect cells, I then tested the biochemical activity of the complex. *In vitro* assays reveal that rCPF can cleave and polyadenylate a substrate *CYC1* RNA (appendix 8.11).

4.3.3.1 rCPF does not have inherent length control

As rCPF showed similar biochemical activity compared to endogenously purified CPF, I sought to test if rCPF could restrict the length of poly(A) tails in a similar manner. Coupled cleavage and polyadenylation assays of an *in vitro* transcribed *CYC1* RNA were performed using recombinant *S. cerevisiae* CPF purified from *Sf9* cells, as well as CF IA and CF IB recombinantly purified from *E.coli*. Surprisingly, despite the apparent identical subunit composition and reaction conditions, rCPF does not restrict the length of the poly(A) tails in assays (Figure 4.18b), unlike native CPF (Figure 4.2a). Urea-PAGE analysis of the reaction shows the polyadenylated products as a widely distributed smear ranging from ~ 100 A's to 600 A's.

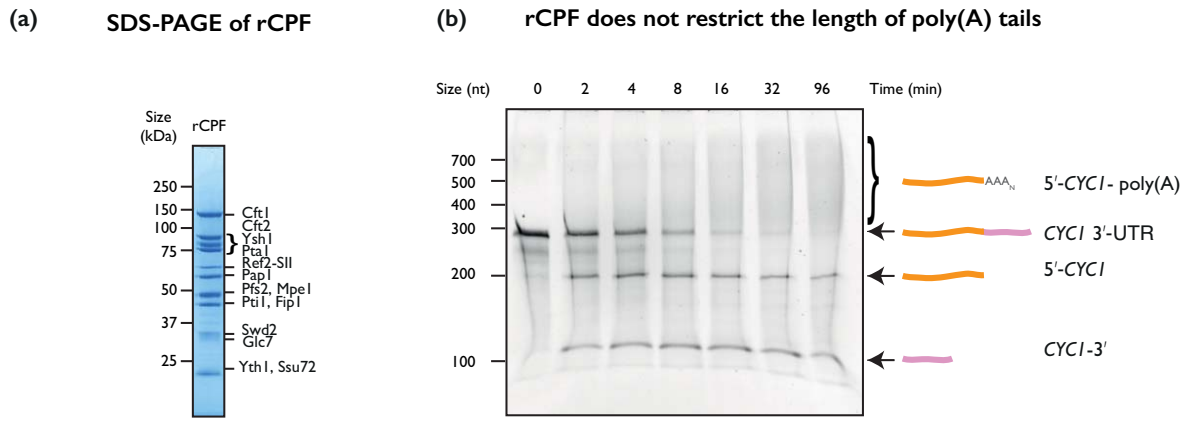


Figure 4.18: rCPF does not have the intrinsic capacity to restrict the length of poly(A) tails **(a)** SDS-PAGE showing purified rCPF after size exclusion chromatography that was used in the assay. **(b)** Coupled cleavage and polyadenylation assay of purified rCPF analyzed by denaturing urea-PAGE. *CYC1* is the substrate RNA. Cleavage products are 5′-*CYC1* (in orange) and *CYC1*-3′ (in pink).

The intrinsic length control property of CPF is sensitive to the relative concentrations of CPF and the cleavage factors to that of the substrate RNA. In endogenous preparations of CPF, the final concentration is in the range of 0.5 mg/ml, whereas for rCPF, this is around 2 mg/ml. Significant error in the estimated concentrations could potentially explain the lack of poly(A) tail length control by rCPF. To clarify this hypothesis, more accurate estimates of protein concentration may be required, and length control assays should be carried out while titrating the amount of rCPF. An alternative explanation could be that the native CPF preparation contains trace amounts of impurities or other co-factors that could influence its biochemical property, due to the less stringent protein purification process, whereas the preparation of rCPF is sufficiently stringent and contains no such impurities or co-factors.

It has been shown that phosphorylation of Pap1 affects its ability to add a poly(A) tail (Mizrahi and Moore 2000). It is not uncommon to find different levels of posttranslational modifications in proteins from higher eukaryotes when compared to the yeasts or bacterial counterparts. Therefore, a third possible explanation to why rCPF does not have an intrinsic ability to restrict the poly(A) tail length could stem from different PTMs in native and recombinant preparations.

4.3.3.2 rCPF and native CPF has different post-translational modifications

To understand whether recombinant expression of CPF influences the pattern of its posttranslational modifications, tandem mass spectrometry was used. Samples from a Streptavidin pull-down of rCPF from *Sf9* cell lysate and were analyzed by SDS-PAGE and compared to endogenously purified CPF from *S. cerevisiae* cells. The bands corresponding to each of the individual CPF subunits were excised from the gel and subjected to tandem mass spectrometry (detailed protocol in methods 6.8.2). rCPF and native CPF harbour very different phosphorylation patterns (Figure 19). Only the phosphatase subunits Pta1 and Ssu72 contain similar phosphorylation mark in both the yeast and insect cell preparations. In light of lack of any structural data on full CPF, it is unclear how these phosphorylation marks might influence the structure and function of CPF.

(a) Phosphorylation sites identified in rCPF

Protein	Residues phosphorylated
Cft2	S583
Pta1	S366, S370, T380, S500, S553
Ysh1	S338, S441
Mpe1	S221
Fip1	S15, S27, S38, S50, S163
Ssu72	S43

(b) Phosphorylation sites identified in native CPF

Protein	Residues phosphorylated
Pta1	T378, T380, S390
Ref2	S396, S438
Mpe1	S199
Pfs2	Y9
Fip1	S56
Pti1	S267, S268, T271, S272, S274, S407, S410
Ssu72	S43

Figure 4.19: Recombinant and native CPF preparations contain different phosphorylation patterns. Tandem mass spectrometry reveals different phosphorylation marks in the two CPF preparations. The total number of spectra uniquely assigned to a protein is represented as exclusive spectrum count.

4.4 Discussion

Previous assays performed in the absence of Nab2 resulted in hyper-polyadenylation of the substrate RNAs. It is thought that Nab2 can bind to the elongating poly(A) tail and restrict the lengths of the poly(A) tails to ~ 60 As. Surprisingly, the nuclear poly(A) binding protein Nab2 is not essential for *in vitro* poly(A) tail length control, as demonstrated in this chapter. I have demonstrated that CPF and cleavage factors have the intrinsic ability to restrict the lengths of newly added poly(A) tails to ~80-200 A's. Interestingly, poly(A) tail lengths observed here are longer than the *in vivo* poly(A) tail length i.e. ~ 60As. Nonetheless, the mechanism by which CPF and the cleavage factors impart poly(A) tail length control without Nab2 remains unknown.

By dissecting the requirements of different protein factors for poly(A) tail length restriction, I have found that cleavage factors CF IA and CF IB are essential for intrinsic poly(A) tail length control. When the RNA is at a concentration of 100 nM, CPF at 50 or 100 nM, and the cleavage factors are present in threefold excess, poly(A) tail length control by CPF is observed. It is hypothesized that CFs are present presumably in excess of their Kd for RNA binding, resulting in the formation of a stable complex with the product RNA. In order to clarify this, the exact affinities of CF IA and CF IB towards binding substrate (and product) RNA have to be determined. One candidate for RNA binding is the Rna15 subunit of CF IA, which contains RRM domains that are known to contact U and G/U rich sequences. The affinity of the RRM domain of Rna15 to binding a poly(U) RNA has been estimated to be 10 μM (Pancevac et al. 2010). Nonetheless, there are no reported studies of the exact RNA binding affinities of highly pure cleavage factors. However, estimation of affinities from EMSA experiments reveal a Kd of ~ 250 nM for CF IA - RNA binding (Hill et al. 2019). Understanding the RNA-binding affinities of cleavage factors could allow the estimation of binding saturation under reaction conditions, potentially providing an explanation for the stoichiometry of cleavage factors required for poly(A) tail length restriction to be observed. These experiments should be accompanied by further polyadenylation assays where the cleavage factors are titrated to observe the minimum concentrations (and stoichiometry) required to observe poly(A) tail length control, and to observe if a larger proportion of the un-polyadenylated 5′ *CYC1*-pc becomes polyadenylated.

In general, enzymes do not remain tightly bound to their substrates or products, in order to allow efficient catalysis. I hypothesized that CPF, along with the cleavage factors, remain tightly bound to the polyadenylated RNA (figure 4.20). This likely results in termination of poly(A) tail addition after ~ 80 to ~ 200 As. To this end, I have shown via *in vitro* pull down experiments that the cleaved RNA products remain bound to CPF and the cleavage factors. This pull-down has also enabled us to understand the experimental conditions at which such pull-downs can be unambiguously performed. The next step would be to carry out pull-downs of cleavage coupled with polyadenylation reactions. This will provide insight into whether the polyadenylated RNA product remains bound to CPF (and the cleavage factors). Further single molecule fluorescence experiments on CPF/CF I bound RNA will be required to confirm this hypothesis.

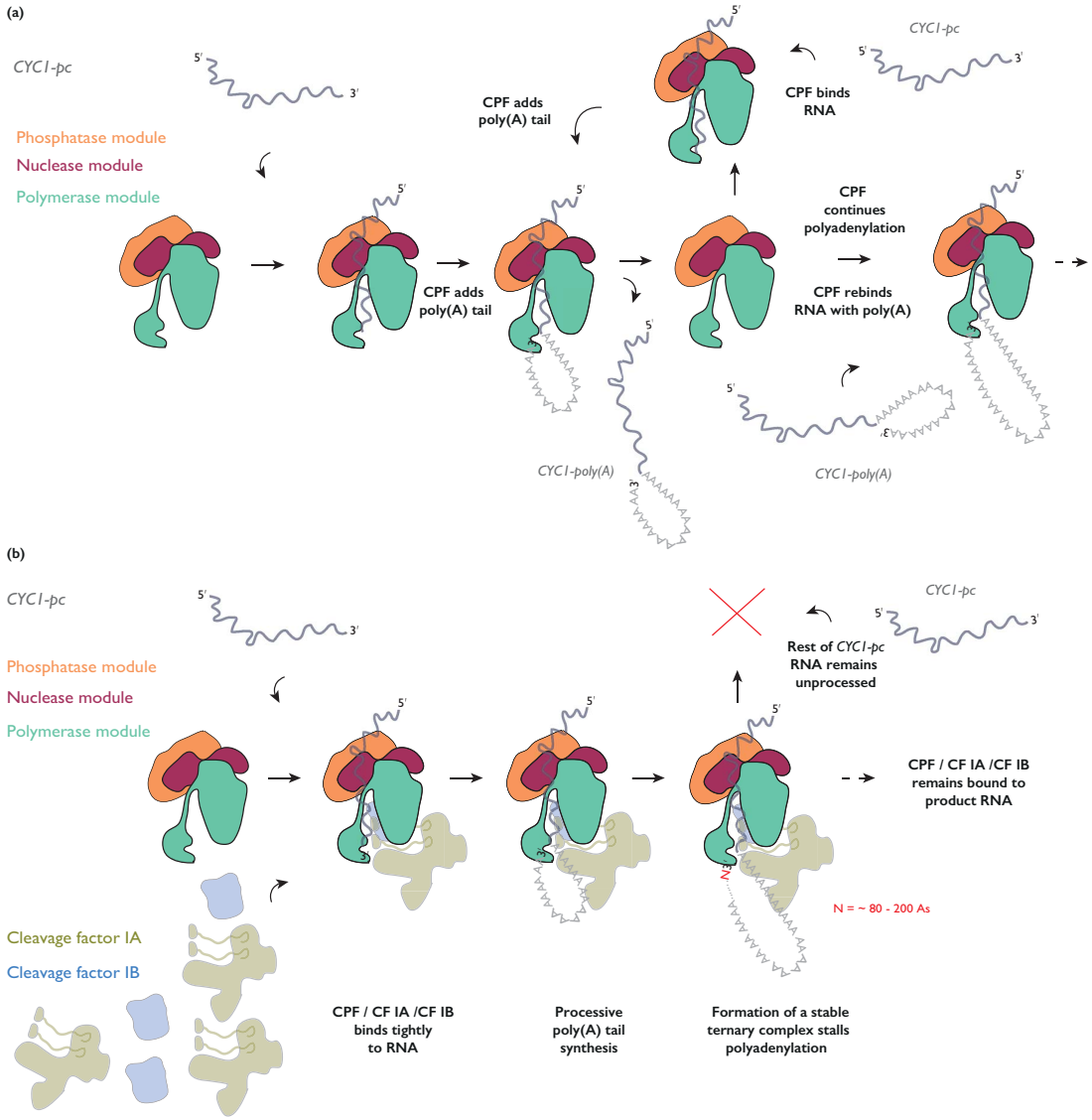


Figure 4.20: A model for poly(A) tail length control by CPF, CF IA and CF IB.

The *in vivo* poly(A) tail length of a model RNA in the absence of Nab2 is longer than ~ 60As (Schmid et al. 2015). The assays described here could recapitulate a situation in the cell, when Nab2 is unavailable to act on the newly made poly(A) containing mRNAs. This may happen during a rapid increase in transcription rate or in case of mRNA export defects when all the available Nab2 would be sequestered by poly(A) containing RNAs in the nucleus. The intrinsic poly(A) tail length control property of CPF described here would ensure that the mRNAs are not massively hyper–adenylated and that nuclear ATP is not rapidly depleted by polyadenylation. These newly made RNAs with poly(A) tail lengths of ~200 A's could either be subjected to exosome mediated decay or later be bound by Nab2, and exported into the cytoplasm. Thus, the intrinsic poly(A) tail length control of CPF discussed in this chapter could be a **quality control pathway** that prevents abnormal long polyadenylation of transcripts. It will be interesting to know how Nab2 can regulate polyadenylation alongside CPF's intrinsic capacity to control poly(A) tail lengths. Perhaps Nab2 can bind and sequester polyA RNAs from the CPF/CFI complex, and shuttle such sequestered poly(A) containing RNAs for export to the cytoplasm. This scenario can be tested using the *in vitro* pull down described in 4.1.5, as the addition of Nab2 to the pull-downs of cleavage and polyadenylation assays may sequester the polyadenylated RNA product from CPF and the cleavage factors. Furthermore, addition of Nab2 to polyadenylation assays may release the tightly bound CPF so that it can act on cleaved but non-polyadenylated transcript. Such experiments could clarify the role of Nab2 in termination of polyadenylation.

A method to recombinantly produce a fourteen-subunit rCPF has been described. The purity and homogeneity of the purified proteins were assessed by size exclusion chromatography, SDS-PAGE, and negative stain electron microscopy. rCPF could cleave a substrate RNA and add a poly(A) tail and showed comparable activity to endogenously purified CPF. Surprisingly, rCPF differed from the natively purified CPF in its ability to intrinsically impart poly(A) tail length control. At this point, the reason behind such a difference remains unclear. However, tandem mass spectrometry revealed different phosphorylation patterns in the subunits of native and recombinantly produced CPF. It will be interesting to de-phosphorylate both native and recombinant CPF using a generic phosphatase, and repeating polyadenylation assays under length control conditions. Such an experiment could elucidate the importance of post-translational modifications in the intrinsic ability of CPF to restrict

poly(A) tail lengths. Alternatively, treatment of the CPF preparations with kinases from yeast such as Ctk1 could write the missing marks and prime CPF for intrinsic poly(A) tail length control.

The lack of protocols to efficiently purify large quantities of highly pure CPF has thus far hindered our progress towards understanding the structure and function of CPF. For the first time, the recombinant system allows us to generate high amounts of sample, allowing us to perform the necessary experiments (i.e. cross-linking mass spectrometry, hydrogen–deuterium exchange coupled with mass spectrometry, fluorescence polarization studies) to thoroughly biophysically characterize CPF. These studies open new experimental possibilities to study the protein-protein interaction network within CPF, affinity measurements for substrate binding, etc. Establishing a recombinant system to produce CPF has also provided a means to make mutant complexes that was hitherto impractical. Thus, I have generated a new tool that facilitates progress towards our understanding the mechanism of cleavage and polyadenylation by CPF.

5 Conclusion and Future Directions

In this dissertation, I have studied the mRNA 3' end polyadenylation reaction by using a biochemical and structural approach. The main questions under investigation were as follows:

1. Why does Pap1 exist in the context of a complex and what is the functional significance of this?
2. What are the roles of the non-enzymatic subunits in CPF?
3. How are the CPF subunits arranged in 3D space?
4. How do accessory factors influence the function of CPF?
5. What are the mechanisms governing how poly(A) tail lengths are regulated?

In the next few sections, I summarize the important findings of this dissertation and provide future perspectives.

5.1 Structural Architecture of the Polymerase Module

By using baculovirus mediated insect cell over-expression, I expressed and purified a five-subunit polymerase module of CPF. Biochemical assays revealed that isolated Pap1 and the polymerase module add poly(A) tail to a substrate RNA in a similar manner. Cryo-EM analysis of a four-subunit polymerase module (without Pap1) resulted in a 3.5 Å resolution map. An atomic model of almost all of the Cft1, Pfs2 and Yth1 subunits were built into the map, revealing an architecture that was highly similar to other nucleic acid processing complexes involved in DNA damage repair and splicing. Around the same time, the structures of the human CPSF160-WDR33-CPSF30 in complex with an AAUAAA containing RNA were published (Sun et al. 2018; Clerici et al. 2018), revealing a high degree of similarity between the human and yeast complexes. The structures also provide insights into understanding how the yeast and human complexes might employ slightly different mechanisms to recognize substrate RNA. This is further reflected in the differences in the yeast and human *cis*-RNA elements.

In the future, a 3D structure of the yeast polymerase module in complex with a substrate RNA will enable better appreciation of the differences. Specifically, the structure will elucidate the determinants of RNA binding specificity. Furthermore, cryo-EM and cross-linking mass spectrometry have revealed that Pap1 is flexibly associated with the polymerase module. An RNA-bound polymerase module sample might aid in stabilizing Fip1 and Pap1 that are currently not visible in the cryo-EM map.

A previous study had shown that a terminal deoxyadenosine residue at the 3' end of a pre-cleaved RNA enables a stable protein -RNA complex formation (Zarkower and Wickens 1987). Performing an *in vitro* polyadenylation assay (as described in section 2.2) in the presence of cordycepin (abscence of ATP) may allow in stabilizing Pap1 and polymerase module on RNA. This may result in an improved sample for structural studies by cryo-EM. The key to the aforementioned future direction would be in identifying the right substrate RNA, as yeast mRNA sequences do not contain a conserved consensus recognition sequence, unlike AAUAAA in human mRNAs. For example, two of the most well studied substrates for cleavage and polyadenylation i.e *CYC1* and *GAL7* share a considerable amount of differences in RNA elements (Zhao et al. 1999; Dichtl and Keller 2001). One possible method to overcome this challenge

would be to characterize the RNA sequence (and structure) specificity of polymerase module. RNA Bind-n-Seq is a recently published method that allows the determination of high-affinity binding motifs (Lambert et al. 2014).

5.2 Effects of Cleavage Factors

In chapter 3 of this dissertation, I have shown that the accessory cleavage factor CF IA (but not CF IB) can stimulate poly(A) tail addition of CPF. By using a minimal *in vitro* reconstituted system, I further show that Rna14/Rna15 subunits of CF IA are sufficient to stimulate poly(A) tail addition by polymerase module. Interesting, CF IA (or Rna14/Rna15) has no effect on polyadenylation activity of Pap1. Thus, the non-enzymatic subunits in polymerase module (and hence CPF) are required for its stimulation by CF IA. *In vitro* pull downs reveal that CF IA directly contacts CPF through an interaction between Rna14/Rna15 and the non-enzymatic polymerase module subunits. By using a mutant Rna14/Rna15 that can no longer strongly bind RNA, I show that RNA binding of Rna14/Rna15 is important for its polyadenylation stimulation property. I present a model where CF IA provides accessory RNA binding surfaces for polymerase module to bind RNA, tethering polymerase module to the substrate RNA and resulting in processive polyadenylation.

This model of polyadenylation can be compared to the model for human poly(A) tail addition. In both cases, the polymerase enzyme appears to be distributive in nature. However, in humans, the incorporation of PAP within CPSF makes the polyadenylation reaction processive by providing additional RNA binding surfaces (Schönemann et al. 2014; Chan, Huppertz, Yao, Weng, Moresco, Yates, et al. 2014). In contrast, in yeast, Pap1 incorporation into the polymerase module does not markedly increase polyadenylation activity. Instead, processive polyadenylation is only achieved when Rna14/Rna15 or CF IA is present, whereby the affinity for RNA of the whole 3'-end processing complex is increased. Unlike the yeast machinery, the human complex achieves such high substrate affinities without the need for any cleavage factor subunits. Such a difference between the yeast and human machinery could stem from the subtle differences in sequences and sequence preferences for *cis*-RNA elements.

A method to assemble a complex between polymerase module and Rna14/Rna15 *in vitro* has been presented. This now allows us to study not only the RNA binding affinities of the individual components but also perform detailed kinetic analysis. I speculate that the k_{off} for RNA binding by the polymerase module will be lower when it binds RNA cooperatively with CF IA (or Rna14/Rna15), explaining the differences in activity in the presence or absence of CF IA.

Furthermore, it is imperative to test the effects of CF IA on polyadenylation using a different substrate RNAs. This will inform us whether this mechanism is conserved across different *cis*-RNA elements.

It is known that cleavage precedes polyadenylation *in vivo* and that CF IA activates cleavage by CPF. However, the exact mechanism by which CF IA activates cleavage is unknown. A recent study from our lab has shown that Rna14/Rna15 alone cannot activate cleavage by CPF and that Pcf11/Clp1 are required for its activation (Hill et al. 2019). This raises the possibility that activation of cleavage by CF IA could not simply be an effect of RNA binding. A future experiment would be to test the effect of the RNA binding mutant of CF IA (similar to the one described here) on activation of cleavage. The observed effect will raise new questions about the activation mechanism of CF IA. It will be interesting to know whether CF IA activates CPF endonuclease and polymerase activities via two different mechanisms.

By using a combination of biophysical techniques including HDX, cross-linking mass spectrometry and cryo-EM, I have characterized the interaction between Rna14/Rna15 and polymerase module. Although HDX experiments revealed that the Fip1 subunit of polymerase module potentially interacts with Rna14/Rna15, biochemical assays suggest that there might be multiple interaction surfaces between the two complexes. Cryo-EM analysis showed that the HAT-domain of Rna14 likely associates with the Cft1 subunit of polymerase module. In agreement with this, several cross-links were found between them. In the absence of a high-resolution structure of the complex, molecular details of these interactions remain unknown, impeding our understanding of the possible allosteric effects of cleavage factors on the function of polymerase module. However, the preliminary cryo-EM analysis presented in this dissertation revealed the inherent heterogeneity (or flexibility with respect to each other) of the polymerase module - Rna14/Rna15 sample. Mild cross-linking using a water-soluble cross-linker like BS3 might aid in fixing the complex in one conformation, possibly reducing the flexibility.

In the cell, CPF functions alongside with CF IA and CF IB. Therefore the model presented here on the role of CF IA in polyadenylation may reflect a scenario happening *in vivo*. Similarly, there are many other protein-nucleic acid complexes in the cell that could potentially influence the function of CPF. A previous *in vitro* study using human

proteins has show that U1A snRNP complex stimulates polyadenylation by interacting with CPSF-160 (Lutz and Alwine 1994). Further research is needed to understand the effects of other complexes on cleavage and polyadenylation, which might clarify the roles of the other non-enzymatic subunits of CPF subunits.

5.3 Polyadenylation mechanism

Using highly pure protein components, I have investigated the polyadenylation activity of a fourteen-subunit CPF in an *in vitro* biochemical system. Surprisingly, the assays revealed that CPF has an intrinsic ability to stall polyadenylation after adding ~ 80 - 200 As without the need for the nuclear poly(A) binding protein Nab2. Drop out experiments then revealed that the cleavage factors CF IA and IB and the phosphatase module of CPF were indispensible for this observed poly(A) tail length control.

While investigating the mechanism of intrinsic length control, I serendipitously discovered that CPF can cleave the poly(A) tail of a *CYC1*-A30 *in vitro*. It has been shown that the presence of the nuclear poly(A) binding protein Nab2 can prevent such re-cleavage events in cells (Viphakone, Voisinet-Hakil, and Minvielle-Sebastia 2008). This raises several interesting questions. Firstly, it is not known if CPF can re-cleave it's own product RNA with poly(A) tail in an *in vitro* assay in the absence of Nab2. Secondly, under length control conditions, significant amounts of the cleaved 5'-*CYC1*-pc RNA remains without a poly(A) tail. It is not known if these non-polyadenylated RNA products are the result of CPF re-cleavage. Thirdly, I demonstrated that a minimal eight-subunit core-CPF cannot re-cleave the poly(A) tail of *CYC1*-A30 RNA, highlighting the contribution of the phosphatase module towards the cleavage activity of CPF. We know that CF IA and CF IB do not tightly bind poly(A) RNAs, and that core-CPF can only weakly bind to poly(A) stretches (Hill et al. 2019). Nonetheless, it is unknown whether the presence of the phosphatase module in CPF enables CPF to bind tightly to poly(A) stretches and result in re-cleavage activity. Alternatively, phosphatase module subunits could impart an allosteric influence on the cleavage activity of CPF. Investigations of the role of phosphatase module in influencing the cleavage of CPF would clarify the significance of the presence of all CPF subunits in recapitulating the *in vivo* behaviour of CPF.

So far, all assays demonstrating intrinsic length control have been carried out using either a 259 nt *CYC1* full or a 182 nt pre-cleaved *CYC1* as a model substrate. It remains to be tested whether the 3' end processing machinery will have intrinsic poly(A) tail length control on a shorter *CYC1* such as the well characterized pre-cleaved *CYC1* 42-mer. Such an experiment will enable us to understand the contributions from the *cis*-elements of *CYC1* such as the efficiency or positioning

element, to the observed intrinsic length control. Additionally, in order to understand whether this could be conserved across a variety of CPF substrates, it is important to test different RNA substrates. The other most commonly studied model RNA is *GAL7*. However, *in vitro* assays of a full-length *GAL7* using purified CPF and cleavage factors resulted in no detectable RNA cleavage of the RNA (appendix 8.12). A similar observation was reported earlier where a wild-type *GAL7* RNA could not be cleaved by CPF (Dichtl and Keller 2001). I have thus identified and produced the 3' UTR of two new RNAs: *GCN4* and *MFA2* (appendix 8.13). *GCN4* consists of a complex polyadenylation signal that differs greatly from the most commonly studied *CYC1* RNA (Egli, Springer, and Braus 1995). *MFA2* is a very small mRNA where a long-range interaction between the 5' and 3' UTR has been shown *in vivo* (Doktycz et al. 1998). The choice of such RNAs (with very different properties compared to *CYC1*) would enable us to elucidate whether the observed intrinsic length control phenomenon is conserved across diverse RNA elements. Additionally, it will serve as a platform to investigate the effects of various *cis*-acting elements towards cleavage and polyadenylation.

5.4 Final conclusions

The field of mRNA 3'-end processing has greatly advanced since the discovery of poly(A) polymerase in 1960. The tremendous progress made can be attributed to the advancements made in modern molecular biology including recombinant DNA technology, sequencing and protein engineering. Although the identities of the individual components of the mRNA 3'-end processing machinery were known, there remained a lack of understanding of how the components were structurally organised into one whole complex and how this affects the function. Taking advantage of the recent progress in the fields of recombinant protein production and cryo-EM, I was able to visualise the intricate and extensive contacts between different protein subunits that constitute a major part of CPF. Together with various biophysical techniques and *in vitro* assays described in this dissertation, I provide new insights into how CPF adds poly(A) tails and how it restricts the length of poly(A) tails. Moreover, I have also described a method to recombinantly produce a full fourteen-subunit CPF. This will have a long lasting impact on our ability to probe the function and structure of CPF. We are only beginning to understand the workings of this poorly studied yet essential step in the central dogma of molecular biology.

6 Materials and Methods

6.1 Common reagents and methods

6.1.1 Generic buffers

The following are the description of some of the commonly used buffers:

- Elution Buffer (EB): 10 mM Tris pH 8.0.
- Tris Borate EDTA (TBE): 10 mM Tris-borate pH 8.0, 1 mM EDTA pH 8.0.
- Tris EDTA (TE): 10 mM Tris pH 8.0, 1 mM EDTA pH 8.0.
- Competent cell buffer 1 (CCB1): 10 mM MES, 100 mM $RbCl_2$, 50 mM $MnCl_2$, 10 mM CaCl2, pH 5.8 with acetic acid.
- Competent cell buffer 2 (CCB2): 10 mM MOPS, 100 mM $CaCl_2$, 10 mM $RbCl_2$, 15% glycerol v/v, pH 6.5 with NaOH.
- Phosphate buffered saline (PBS): 137 mM NaCl, 10 mM Phosphate, 2.7 mM KCl, pH 7.4.

The buffers used in protein purifications are described in later sections.

6.1.2 Media

6.1.2.1 Bacterial overexpression

6.1.2.1.1 Liquid media

- 2xTY: 10 g yeast extract, 16 g tryptone, 5 g NaCl, ~ 2.5 ml 1 M NaOH (to pH 7.4), adjust to 1 L with milliQ H_20 (MQ), autoclave.

- LB: 5 g yeast extract, 10 g tryptone, 10 g NaCl, 900 ml MQ, ~2.5 ml 1 M NaOH(to pH 7.0), adjust to 1 L with MQ, autoclave.

- SOC: 5 g yeast extract, 20 g tryptone, 0.6 g NaCl, 0.2 g KCl, 900 ml MQ, 3.5 ml 1M NaOH (to pH 7.2), adjust to 1 L with MQ, autoclave, add: 10 ml 1 M $MgCl_2$, 10 ml 1 M $MgSO_4$, add 10 ml 20% w/v glucose immediately before use.

6.1.2.1.2 Plates

- TYE: 5 g yeast extract, 10 g tryotone, 8 g NaCl, 15 g agar, ~2.5 ml 1 M NaOH (to pH 7.0), adjust to 1 L with MQ, autoclave.

- LB: 15 g agar, 5 g yeast extract, 10 g tryptone, 10 g NaCl, 900 mL MQ, ~2.5 mL 1M NaOH (to pH 7.5), adjust to 1 L with MQ, autoclave.

6.1.2.2 Insect cells over expression

- *Sf*9 Media: Insect-XPRESS™ protein-free insect cell medium with L-glutamine from Lonza.

6.1.2.3 E.coli minimal media for isotope labelling

- K-MOPS: 10x stock: 1.0 M MOPS, 0.1 M Tricine mix, pH to 8.0 with KOH (800 ml);10 mM $FeCl_2$ (10 ml); 1 mM $CaCl_2$ (10 ml); 1 mM $MgCl_2$ (10 ml), 20 mM H_3BO_3, 5 mM $CoCl_2$, 2 mM $CuCl_2$, 10 mM $MnCl_2$, 2 mM $ZnCl_2$, 2 mM $NaMoO_4$ mix (10 ml); 4.0 M NaCl (10 ml); dilute 10x stock to 1x with MQ and autoclave; add directly before use 20 mM NH_4Cl, 4 mM KPi, 1% w/v glucose, 10 mg thiamine, 2 mg D-biotin, 2 mg choline chloride, 2 mg folic acid, 2 mg niacinamide, 2 mg D-phantothenic acid, 2 mg pyridoxal, 2 mg riboflavin. K-MOPS minimal medium containing $15NH_4Cl$ was used for 2-D NMR experiments. For 3D experiments, media also contained [13C] glucose.

6.1.2.4 Antibiotics

Antibiotics were made at the following stock concentrations in autoclaved MQ H_2O: kanamycin 50 mg/ml in MQ H_2O, chloramphenicol 35 mg/ml in 100% EtOH and

gentamycin 10 mg/ml in MQ H_2O. The final concentrations of the antibiotics are also mentioned as follows: kanamycin 50 μg/ml, chloramphenicol 35 μg/ml and gentamycin 10 μg/ml in liquid media or 7 μg/ml in bacmid plates.

6.1.3 Cells and strains

Strain	Genotype
TOP10	F-mcrAΔ(mrr-hsdRMS-mcrBC) Φ80lacZΔM15 ΔlacX74 recA1 araD139 Δ(araleu)7697 galU galK rpsL (StrR) endA1 nupG
BL21 Star (DE3)	F-ompT hsdSB(rB–,mB–) gal dcm rne131 (DE3)
DH10EmBacY	F-mcrA Δ(mrr-hsdRMS - mcrBC) Φ80lacZΔM15 ΔlacX74 recA1 endA1 araD139 Δ(ara, leu)7697 galU galK λ-rpsL nupG (bEmBacY, pTn7helper)

Table 6.1: *E.coli* strains used in this dissertation

6.1.4 Making competent cells

Terence Tang and James Stowell prepared the competent cells. Four colonies of a desired *E.coli* strain from a freshly streaked plate is grown in 5 mL of LB media at 200 rpm (37 °C) for 2 h until an OD_{550} of 0.3. 2 mL of this pre-culture is used as a starter culture to inoculate 40 mL of pre-warmed LB and grown at 200 rpm (37 °C) to an OD_{550} of 0.45.

The next few steps were all carried out on ice. The cells grown to OD_{550} of 0.45 were decanted into 50 mL falcon tubes, incubated on ice for 15 min, pelleted at 1300 rcf (4 °C) for 15 min, and suspended in 2 mL of CCB1 by swirling the tube gently. The total

volume was brought to 16 mL with CCB1, and the suspension was incubated on ice for 15 min. Cells were gently harvested at 1300 rcf (4 °C) for 15 min. The pellet was resuspended in 1.6 mL CCB2 and the suspension was incubated on ice for 15 min. The suspension was aliquoted either as 50 or 100 µL volume into 1.5 mL tubes, slowly frozen, and stored at -80 °C.

6.1.5 Plasmid transformation

~ 10 ng of plasmid DNA was gently mixed with 50 µl of chemically competent *E. coli* in a 1.5 mL tube and incubated for 30 min on ice. The tube was swiftly transferred into a water bath at 42 ºC for 45 sec, and subsequently transferred on to ice for 2 minutes. 400 µl of SOC medium were added to the tubes and the cells were recovered at 200 rpm (37 ºC) for 1 hr.

For transformations of intact plasmids, 50 µl cell suspension were plated on TYE/antibiotic plates. For transformations of ligation reactions, cells were pelleted at 500 rcf for 1.5 min, the supernatant decanted, resuspended in 50 µl SOC medium and plated on TYE/antibiotic plates.

6.1.6 Bacmid transformation

For transformations of pBIG1/2 family of plasmids into EmBacY cells to generate bacmids, the transposition / recovery step was performed for three to four hours. Following such a long recovery, 500 µl of cell suspension were plated on TYE/antibiotic plates. Plates were incubated overnight at 37 ºC.

6.1.7 Plasmid isolation

A single colony of a desired *E.coli* strain from a freshly streaked plate is grown in 5 mL of LB media (with the appropriate antibiotic) at 200 rpm (37 °C) overnight.
The cells were pelleted at 3200 rcf, 10 min, 4 ºC and the supernatant discarded. Plasmids were isolated using the QIAprep Spin Miniprep kit(Qiagen), following the standard protocol.

6.1.8 Nucleic acids quantification

The absorbance at 260 nm on a NanoDrop ND-1000 spectrophotometer (Thermo Scientific) was measured. The extinction co-efficients reported in table 6.24 were used to calculate the nucleic acid concentration.

6.1.9 DNA Sanger sequencing

5 µL of the following samples per reaction was provided to Source Bioscience (Cambridge, UK) for Sanger sequencing: 100 ng/µL of plasmid DNA and 3.2 µM primer DNA. The samples were diluted from their stocks in milliQ water. The sequencing data was analysed by MacVector (MacVector, Inc.).

6.1.10 Nucleic acid electrophoresis

6.1.10.1 DNA agarose gel

1 g of agarose (BioGene) was dissolved in 100 mL 1x TAE by microwaving the mixture on high (900 W). After cooling to ~50 ºC, SybrSafe (Life Technologies) was added, and the mixture was poured into a running apparatus. The gel sets in ~ 45 min after which the combs were removed. The samples are loaded after TAE buffer completely covers the gel and the electrodes. Gels were run at constant 100 V until the bands were separated as required. Gels were visualized on a Gel Doc XR+(BioRad).

The following is the 6x concentration of a DNA loading dye: 30% w/v glycerol, 0.25% w/v bromphenol blue, 0.25% w/v xylene cyanol FF.

6.1.10.2 RNA polyacrylamide gel

RNA acrylamide gels were casted using the MINI-Protean electrophoresis system (Biorad). Gel recipes for 6%TBE-urea and 15% TBE-urea gels are given below (Table 6.2). In the case of a preparative gel, the solution was poured between two 20 x 18.5 cm glass plates with 2 mm spacers and was sealed with polypropylene-based sellotapes.

Once the gesl were set, the tapes were removed and the gels were pre-run in 1 x TBE buffer at 30 W for 15 min using a PowerPac HV power supply (BioRad). Then appropriate amounts of sample was loaded and the gel was run for ~ 10 to 20 min depending on the size of the RNA sample. Preparative gels were run for ~ 2 hours until the loading dye runs off the gel.

6% Urea acrylamide	**15% Urea acrylamide**
220 g Urea	189 g Urea
50 ml 10x TBE	45 ml 10x TBE
75 ml 40% Acrylamide/Bis Solution, 19:1	169 ml 40% Acrylamide/Bis Solution, 19:1
up to 500 ml Nuclease free H_20	up to 450 ml Nuclease free H_20

Table 6.2: Composition of the Urea-acrylamide solutions used to cast the gels

The following is the recipe for making a 6% - urea acrylamide mix: Dissolve the urea in ~ 300 ml of milliQ water on a stirrer. Set the temperature to 90 °C while dissolving the urea. As it is dissolving, add the required amounts of 19:1 bis acrylamide and 10x TBE to this mixture. After all the urea has dissolved, make up to 500 ml or 450 ml with MQ and cool down the mixture.

For analytical gels, 10 ml of the urea-acrylamide mixture was mixed with 10 µl of TEMED (Sigma) and 100 µl of Ammonium persulphate solution. For preparative gels, the above mixture was scaled up five times.

2x RNA loading dye contained the following: 960 µl of Formamide (Thermofisher), 10 µl of 0.08 % w/v bromophenol blue, 0.08 % w/v 10 µl of xylene cyanol and 20 µl 0.5 M EDTA.

6.1.11 SDS-PAGE

4 µl OF 4x LDS (106mM Tris HCl, 141mM Tris base, 2% LDS, 10% glycerol, 0.51mM EDTA, 0.22mM SERVA Blue G250, 0.175mM Phenol Red, pH 8.5) (Thermo Fisher)

sample buffer was added to 12 µl of protein samples up to ~ 500 nm in concentration. For higher concentrations of proteins, 2 µl OF 4x LDS was mixed with 6 µl of proteins. Samples were run on a pre-cast 15 well 4-12% Bis-Tris gradient (Thermofisher) gel at 180 volts for 60 minutes. MOPS based running buffer was used for the electrophoresis. For the analysis of cross-linked protein samples, 3-8% tris-acetate gels (Thermofisher) were used and were run at 140 volts for > 2 hours. This was to allow the cross-linked protein complexes to run into the wells. In case of the Yth1 and Fip1 constructs for NMR, 20% tris-tricine gels were used. The gels were then stained with a coomassie based InstantBlue™ (Expedion) for several hours. The stain was then removed and replaced with MQ water for de-staining. The de-stained gels were visualized by imaging on a Gel Doc™ (Biorad).

6.2 Cloning

6.2.1 Restriction based cloning

Using 'traditional cloning', a plasmid was digested with either one or two restriction enzymes. The gene of interest to be cloned into the plasmid is produced by PCR and the PCR product is digested with the same set of enzymes that were used to cut open the plasmid. The digested gene of interest is then ligated into the plasmid that is cut open.

6.2.1.1 PCR

To amplify genes from plasmid DNA, PCR reactions using High Fidelity Phusion DNA polymerase (NEB) were performed. PCR primers were obtained at the 100-nmol scale from Sigma Aldrich and suspended to a final concentration of 100 μM in milliQ. A standard 50 μL PCR reaction mixture is described in Table 6.3 and the reaction conditions are given in Table 6.4. The reactions were performed in a Veriti 96-well Thermal Cycler (Applied Biosystems).

Component name	Final concentration
Nuclease free water (DEPC)	N/A
5X Phusion HF buffer (NEB)	1X
dNTPs	200 μM
Forward primer	0.5 μM
Reverse primer	0.5 μM
Template DNA	~10 ng
DMSO	~2 to 3 %
Phusion HF enzyme (NEB)	1 unit per 50 μl reaction

Table 6.3: Setting up a cloning PCR

The annealing temperature (Tm) for the PCR reaction was calculated for a set of primers by using the NEB Tm Calculator v1.10.1. The extension time in the annealing step was calculated based on the size of the gene being clones. For example, for Cft1 gene which is ~ 4.6 kb in size, an extension time of ~ 70 sec was used

Number of cycles	Step	Temperature	Time
1	Initial denaturation	98 °C	30 sec
30	Annealing	98 °C	10 sec
		Tm °C	30 sec
		72 °C	15 sec per kb
1	Extension	72 °C	10 min
1	Hold	4 °C	Until further use

Table 6.4: PCR conditions

.

6.2.1.2 PCR purification and gel extraction

50 µl of the PCR reaction product is mixed with 10 µl of 6x DNA loading dye and is run on a 1 % agarose TAE gel. The desired bands are then excised and purified using the QIAquick Gel Extraction kit (Qiagen), following the standard protocol and eluted in 30 - 50 µl EB.

6.2.1.3 Restriction digestion

The purified PCR products and the host vector, into which the gene of interest is about to be cloned, are digested with the same set of enzymes.

Component name	Final concentration
Nuclease free water (DEPC)	N/A
10x buffer	1X
PCR product or plasmid	~1 µg
Enzyme 1	10 units
Enzyme 2	10 units

Table 6.5: Setting up a restriction digestion for cloning

The choice of the 10x buffer is dictated by the combination of the enzymes used in the cloning. This can be decided by analyzing the list of available enzyme and their activity in the respective buffers using the tool NEBcloner v1.3.9.

6.2.1.4 Ligation

The products of the restriction digestion are analysed by running them in a 1% agarose TAE gel. The desired bands are then excised and purified using the QIAquick Gel Extraction kit (Qiagen), following the standard protocol and eluted in 10 µl EB. The EB is pre-warmed at 70 °C before usage. The following protocol was used to ligate the gene of interest (PCR product) into the host vector.

Component name	Final concentration
Nuclease free water (DEPC)	N/A
T4 DNA Ligase Buffer (10x)	1X
Vector	0.02 pmol
Insert	0.06 pmol
T4 DNA Ligase	1 µl per 10 µl reaction

Table 6.6: Setting up a sticky end DNA ligation

NEBioCalculator v1.9.0 is used to calculate the mass of insert required to set up ligation reactions at different ratios of vector:insert. The assembled reaction is incubated at 16 °C for 20 minutes. All of the reaction is transformed into TOP10 competent cells and plated on TYE plates with appropriate antibiotics depending on the vector being used.

6.2.1.5 Verification of clones

Plasmids from a few different colonies (usually I screened four colonies per new construct) were extracted based on Section 6.1.7 and their sequences verified based on Section 6.1.9.

6.2.2 biGBac cloning

The biGBac system allows for the cloning of up to twenty-five individual genes into one single vector. This method involves the usage of computationally optimal DNA linker sequences, which allows efficient assembly of DNA fragments by Gibson cloning in order to produce such large synthetic genomes (Weissmann et al. 2016). The DNA linker sequences for Gibson assembly were selected in such a manner that the propensity to form incorrectly assembled products is minimized.

The first step in this system is to clone the gene of interest (GoI) using Gibson assembly or restriction digestion based cloning techniques into a vector containing an expression cassette with a polyhedrin (polh) promoter and a SV40 late terminator (SV40). In our lab, GoIs were first cloned either into pACEBac1, pIDC or pIDS.

Then, gene expression cassettes are amplified by PCR using primers that contain the alpha, beta, gamma, delta, epsilon or omega Gibson linker overhangs (Weissmann et al. 2016).

Primer overhangs	Sequence
Alpha	AACGCTCTATGGTCTAAAG
Beta	AAACGTGCAATAGTATCCAGTT
Gamma	AAACATCAGGCATCATTAGGTTT
Delta	AAACTAAGCTATGTGAACCGTT
Epsilon	AAACCAAGTCAATGTCAGTGTTT
Omega	AACCCCGATTGAGATATAGATT

Table 6.7: Sequences of optimized linker DNA involved in Gibson assembly

The first level of multi-gene assembly involves carrying a Gibson assembly reaction with up to five genes (PCR products from the earlier step) and one of the five pBIG1 series of vectors (pBIG1a - pBIG1e). A modified version of the biGBac system (Weissmann et al. 2016) was used in our lab. An equivalent version of the vectors pBIG1a,b,c,d,e were created by cloning the necessary Gibson overhangs, spacers and

Swa1 sites into pACEBac1. The resulting pBIG1 series vectors were selectable using gentamycin rather than ampicillin and spectinomycin.
It is important to note that the first GoI that will occupy the 5' end of pBIG1 vector needs to have the alpha linker sequence in the 5' end. Similarly, the final GoI occupying the 3' end of pBIG1 vector needs to carry the omega linker sequence in the 3' end. For example, the first GoI will have alpha sequence in its 5' end and beta sequence in the 3' end. The second GoI will have beta sequence in its 5' end and gamma sequence in its 3' end. If one wishes to clone only two GoIs into pBIG1 vector, then the second GoI needs to carry omega sequence in its 3' end.

The above step leads to the generation of a polygene cassette (PGC) in circular pBIG1 vector. The pBIG1 series of vectors can be propagated in *E.coli* cells. In order to verify whether all the desired GoIs have been successfully incorporated into pBIG1 vectors, I made use of the Swa1 restriction sites that flanks all the GoIs. So upon Swa1 digestion of the pBIG1 vectors containing the GoIs and analysing the resulting digestion pattern by running a TAE agarose gel, one can confirm the incorporation of all the GoIs into pBIG1 series of vector. Furthermore, PmeI sites flank the PGCs in the pBIG1 vectors. So an alternative approach to confirm the incorporation of all the GoIs is to perform a diagnostic restriction digestion with PmeI.

The digestion of pBIG1 vectors with PmeI also results in generation of a PGC that is flanked by optimized Gibson linker sequences at their ends. These linker sequences can then be used in a second of Gibson assembly to assemble up to five PGCs from the five different pBIG1 series of vectors, into one of the four pBIG2 series of vectors. An equivalent version of the vectors pBIG2ab,abc,abcd,abcde from the original biGBac system were created by cloning the necessary Gibson overhangs, spacers and Swa1 sites into pACEBac1. In order to differentiate the selection between pBIG1 and pBIG2 series of vectors, an additional chloramphenicol resistance gene was added to the pBIG2 equivalents. Mathias Girbig from the Carter lab created the pBIG equivalent versions.

The vector pBIG2ab can accommodate the PGCs from pBIG1a and 1b. Similarly pBIG2abc can accommodate PGCs from pBIG1a, 1b and 1c. And so on. By doing so, a final vector pBIG2abcde containing five PGCs and a total of twenty-five genes can be assembled.

6.2.2.1 PCR for making gene expression cassettes

The PCR protocol is same as section 6.2.1.1. The only difference is the choice of primers. The primers used for carrying out the PCR are listed as follows:

Name	Sequence
pB_pIDC_CasI_F	AACGCTCTATGGTCTAAAGATTTAAATCGACCTACTCCGGAATATTAATAGATC
pB_pIDC_CasI_R	AAACGTGCAATAGTATCCAGTTTATTTAAATGGTTATGATAGTTATTGCTCAGCG
pB_pIDC_CasII_F	AAACTGGATACTATTGCACGTTTAAATCGACCTACTCCGGAATATTAATAGATC
pB_pIDC_CasII_R	AAACATCAGGCATCATTAGGTTTATTTAAATGGTTATGATAGTTATTGCTCAGCG
pB_pIDC_CasIII_F	AAACCTAATGATGCCTGATGTTTAAATCGACCTACTCCGGAATATTAATAGATC
pB_pIDC_CasIII_R	AAACTAAGCTATGTGAACCGTTTATTTAAATGGTTATGATAGTTATTGCTCAGCG
pB_pIDC_CasIV_F	AAACGGTTCACATAGCTTAGTTTAAATCGACCTACTCCGGAATATTAATAGATC
pB_pIDC_CasIV_R	AAACCAAGTCAATGTCAGTGTTTATTTAAATGGTTATGATAGTTATTGCTCAGCG
pB_pIDC_CasV_F	AAACACTGACATTGACTTGGTTTAAATCGACCTACTCCGGAATATTAATAGATC
pB_pIDC_Cas w_R	AACCCCGATTGAGATATAGATTTATTTAAATGGTTATGATAGTTATTGCTCAGCG
pIDS_CasI_F	AACGCTCTATGGTCTAAAGATTTAAATGGCACCTAGGTATCGATACTAGTATAC
pIDS_CasI_R	AAACGTGCAATAGTATCCAGTTTATTTAAATGTACCCGTAGTGGCTATGGCAGGG
pIDS_CasII_F	AAACTGGATACTATTGCACGTTTAAATGGCACCTAGGTATCGATACTAGTATAC

	AAACATCAGGCATCATTAGGTTTATTTAAATGTACCCGTAG
pIDS_CasII_R	TGGCTATGGCAGGG
	AAACCTAATGATGCCTGATGTTTAAATGGCACCTAGGTATC
pIDS_CasIII_F	GATACTAGTATAC
pIDS_CasIII_	AAACTAAGCTATGTGAACCGTTTATTTAAATGTACCCGTAG
R	TGGCTATGGCAGGG
pIDS_CasIV_	AAACGGTTCACATAGCTTAGTTTAAATGGCACCTAGGTATC
F	GATACTAGTATAC
pIDS_CasIV_	AAACCAAGTCAATGTCAGTGTTTATTTAAATGTACCCGTAG
R	TGGCTATGGCAGGG
	AAACACTGACATTGACTTGGTTTAAATGGCACCTAGGTATC
pIDS_CasV_F	GATACTAGTATAC
pIDS_Casw_	AACCCCGATTGAGATATAGATTTATTTAAATGTACCCGTAG
R	TGGCTATGGCAGGG

Table 6.8: Sequences of cloning primers used in the PCR for making gene expression cassettes

The abovementioned primers do not contain any gene specific sequences. Before assembling the multi-gene complex, the individual subunit genes are cloned into pACEBAC1 or pIDC or pIDS. The choice of the cloning primer depends on the vector within which the GoI lies. The GoI is PCRed out of pACEBac1 or pIDC or pIDS. 1 µl Dpn1 (NEB) is added to the 50 uL PCR reaction mixture, incubated for 1 h at 37 °C. The product is purified according to protocol described in section 6.2.1.2.

6.2.2.2 Gibson assembly

Before assembling the GoI PCR products or gene expression cassettes into pBIG1 vector, it is necessary to linearize the vector. ~ 1 µg of pBIG1 vector is digested using Swa1 enzyme in a 10 µl reaction containing NEB 3.1 buffer (Table 6.12). The digestion is carried overnight at 25 °C. The linearized pBIG1 series of vectors are then purified according to protocol described in section 6.2.1.2.

Component name	Concentration (pmol)
Swa1 digested pBIG1	~ 0.02
Gene expression cassette 1	~ 0.06 to 0.1
Gene expression cassette 2	~ 0.06 to 0.1
Gene expression cassette 3	~ 0.06 to 0.1
Gene expression cassette 4	~ 0.06 to 0.1
Gene expression cassette 5	~ 0.06 to 0.1
Nuclease free water	up to 5 µl

Table 6.9: Making the DNA mix to be used in the Gibson assembly

The DNA mixture is prepared as described above. The Gene expression cassettes are added in three or five fold molar excess compared to the pBIG1 vectors. The final DNA mixture volume should not exceed 5 µl. The DNA mix is then added to 15 µl of the Gibson mix 1 (KGB 1), mixed well and incubated in a thermo cycler at 50 °C for 60 minutes.

Component name	Amount
1M Tris-HCl pH 7.4	3000 µl
1M $MgCl_2$	300 µl
10 mM dNTPs	600 µl
1 mM DTT	300 µl
PEG-8000	1.5 g
50 mM NAD	600 µl
Nuclease free water	up to 6000 µl

Table 6.10: Preparation of 5X Isothermal assembly buffer (IAB)

Component name	Volume (µl)
Taq DNA Ligase (NEB)	160
T5 exonuclease (NEB)	1.5
Phusion HF DNA (NEB)	20
1X IAB	320
Nuclease free water	700

Table 6.11: Preparation of KGB 1

After 60 min, transform all 20 µl of the Gibson reaction mix into chemically competent TOP10 cells. Plate the cells on TYE plates containing Gentamycin.

6.2.2.3 Verifying clones by Swa1 digestion

From the TYE Gentamycin plates, pick a few colonies and isolate plasmid DNA from them according to section 6.1.7. Swa1 digestion is used to verify the incorporation of the gene expression cassettes into pBIG1 series of vectors.

Component name	Final concentration
Nuclease free water (DEPC)	up to 10 µl
10x NEB 3.1 buffer (NEB)	1 µl
Plasmid	~1 µg
Swa1 (NEB)	1 unit

Table 6.12: Preparing Swa1 digestion reaction

The reaction is carried out for 4 hours at 25 °C. 10 µl of the reaction mixture is analysed by gel electrophoresis as described in section 6.1.10.1. The pBIG1 constructs showing the correct or expected Swa1 digestion pattern are sequenced as described in section 6.1.9.

6.2.2.4 Cloning into pBIG2 series of vectors

Once the identities of the pBIG1 series of vectors were verified and it was confirmed that they carry all the necessary gene expression cassettes, they were digested with Pme1 to be cloned into pBIG2 series of vectors. pBIG2 series of vectors are linearized by digestion with Pme1 using a protocol similar to table 6.12. The reaction is carried out for 2 hours at 37 °C. The linearized pBIG2 series of vectors are then purified according to protocol described in section 6.2.1.2.

The second Gibson assembly reaction is set up as follows:

Component name	**Concentration (pmol)**
Pme1 digested pBIG2abcd	~ 0.02
pBIG1a	~ 0.06 to 0.1
pBIG1b	~ 0.06 to 0.1
pBIG1c	~ 0.06 to 0.1
pBIG1d	~ 0.06 to 0.1
pBIG1e	~ 0.06 to 0.1
5X IRB	4 µl
Nuclease free water	up to 12.5 µl
Pme1	1 µl

Table 6.13: Making the DNA mix to be used in the second step of Gibson assembly

The DNA mix is incubated at 37 °C for 2 hours. Then 2.5 µl of Gibson mix 2 (KGB 2) is added to the reaction and is incubated at 50 °C for 60 min. Transform all 20 µl of the Gibson reaction mix into chemically competent TOP10 cells. Plate the cells on TYE plates containing chloramphenicol.

Component name	Volume (μl)
Taq DNA Ligase (NEB)	24
T5 exonuclease (NEB) [Diluted 1:30 in 1X IAB]	3
Phusion HF DNA (NEB)	3

Table 6.14: Preparation of KGB 2

6.2.2.5 Verifying clones by Pac1 digestion

From the TYE chloramphenicol plates, pick a few colonies and isolate plasmid DNA from them according to section 6.1.6. Pac1 digestion is used to verify the incorporation of the polygene cassettes into pBIG2 series of vectors.

Component name	Final concentration
Nuclease free water (DEPC)	up to 10 μl
10x CutSmart Buffer (NEB)	1 μl
Plasmid	~1 μg
Pac1 (NEB)	1 unit

Table 6.15: Preparing Pac1 digestion reaction

The reaction is carried out for 1 hour at 37 °C. 10 μl of the reaction mixture is analysed by gel electrophoresis as described in section 6.1.10.1. The pBIG2 constructs showing the correct or expected Pac1 digestion pattern are sequenced. 35 μl of the plasmid DNA (at ~ 40 to 65 ng/μl) are submitted for complete plasmid sequencing at MGH CCIB DNA Core, Massachusetts General Hospital.

The subsequent steps involved in biGBac cloning leading to protein expression are described in section 6.3.3. The following sections 6.2.3 until 6.2.9 provide details about the specific protein or protein complexes that were cloned for study in this dissertation.

6.2.3 Pap1

The *PAP1* gene was cloned into a pET-28a(+) vector, fusing it to an N-terminal Histag. The gene was cloned from an already existing plasmid pMK-RQ containing the *PAP1* gene. The following were the primers used for cloning:

Forward primer (FP): 5′-GGAATTC CATATG ATGAGCAGCCAGAAAGTT-3′

Reverse primer (RP): 5′- CCCA AAGCTT TTAATTAACATCAACTGCTG -3′

The PCR product and the circular pET-28a(+) vector was digested with NdeI and HindIII. The PAP1 gene expression cassette was ligated into linear pET-28(+) according to section 6.2.1.3 and 6.2.1.4.

6.2.4 Polymerase module

The polymerase module constructs were made by Gillian Dornan using the MultiBac^{Turbo} Expression System (EMBL) and the individual subunit genes were synthesised by GeneArt. Cft1, 8His-3C-Yth1 and Pfs2-3C-SII were cloned into pACEBac1, and Pap1 and Fip1 were cloned into pIDC. All the cloning was done using BamHI/XbaI sites. The polymerase module construct was assembled using cre-lox recombination (Bieniossek et al. 2012).

6.2.4.1 Polymerase module truncations

The biGBac system was used for cloning for the polymerase module variants described in section 3.5.3. Splicing by overlap extension was used to generate truncations in the Fip1 gene (Higuchi, Krummel, and Saiki 1988). Fip1 gene in pIDC was used as the template for the splicing PCR. The final gene cassettes were digested with BamhI and XbaI before introducing them into pACEBac1.

- The primers used in making Fip1 Δ190-220 are as follows:

5′-GGAATTC GGATCCCTCGAG ATGAGCAGCAGCGAAGATG-3′

5′-ACGCGGATTGTAATCCTGCTGCGGTTTTTCTTTCAGAACTTC-3′

5′-GAAGTTCTGAAAGAAAAACCGCAGCAGGATTACAATCCGCGT-3′

5′-GCCCCATCTAGAGGTACCTCATTATTATTTGCTATTCTGGTTCTGATTCTG-3′

- The primers used in making Fip1 Δ145-170 are as follows:

5′-GGAATTC GGATCCCTCGAG ATGAGCAGCAGCGAAGATG-3′

5′-GATGCCCACGCTATCAAAAATTGCGGTAACACCCTGATTTG-3′

5′-CAAATCAGGGTGTTACCGCAATTTTTGATAGCGTGGGCATC-3′

5′-GCCCCATCTAGAGGTACCTCATTATTATTTGCTATTCTGGTTCTGATTCTG-3′

- Yth1 ΔZnF45C was created using pACEBac1-Yth1 as the template and the following primers were used:

5′- GGAATTC GGATCCCTCGAG ATGAGCCTGATTCATCC-3′

5′- GCCCCA TCTAGA GGTACC TCA TTA CGGATCAATATGCAGATATTG -3′

The PCR product was digested with BamHI and XbaI before introducing them into pACEBac1. The pACEBac1 plasmids containing the truncated versions of Fip1 or Yth1 were isolated, their sequences verified as described in section 6.1.9. A polymerase module vector containing all the five-subunit genes including the truncated Fip1 or Yth1 was generated as described in section 6.2.2.1 and 6.2.2.2. Pfs2 contained a C-terminal twin strep tag.

6.2.5 CF IA/CF IB

CF IB or *HRP1* gene was cloned into a pOPINB vector (Oxford Protein Production Facility) and verified by sequencing, yielding an N-terminally His-tagged Hrp1.

The plasmids used in the over-expression of CF IA were provided by Andrew Bohm (Tufts School of Medicine, Boston). The genes coding for RNA14 and RNA15 were cloned into a pETDuet vector to enable co-expression of the proteins. Rna15 harboured an N-terminal 6xHis tag. Pcf11 and Clp1 were cloned into a pRSFDuet vector. Pcf11 harboured an N-terminal 6xHis tag.

6.2.5.1 Rna14/Rna15 RRM mutants

A QuikChange Lightning kit (Agilent) was used to generate a triple point mutant in Rna15. The residues Y21, Y61 and F63 were mutated into alanine. His tagged RNA15 gene in pACEBac1 was used as a template for the QuikChange PCR. First, a double mutant Y61A, F63A was generated in pACEBac1-His-Rna15 using the following protocol.

Component name	Amount
Nuclease free water (DEPC)	39 μl
10x QC buffer (Agilent)	5 μl
dNTPs (10 mM)	1 μl
Forward primer (10 μM)	1 μl
Reverse primer (10 μM)	1 μl
Template DNA (200 ng/μl)	0.5 μl
Quik solution (Agilent)	1.5 μl
QuikChange Lightning (Agilent)	1 μl

Table 6.16: Setting up a QuikChange Lightning PCR

The primers to make the point mutants were generated using QuikChange Primer Design program (Agilent).

Number of cycles	Step	Temperature	Time
1	Initial denaturation	95 °C	2 min
18	Annealing	95 °C	20 sec
		60 °C	10 sec
		68 °C	2 min
1	Extension	68 °C	5 min
1	Hold	4 °C	Until further use

Table 6.17: PCR setup

The PCR products were treated with 1 µl DpnI at 37 °C for 1 hour. 2 µl of the reaction mixture was transformed into XL10 Gold ultra competent cells (Agilent) according to section 6.1.5. Plasmid was isolated from several colonies and the sequence verified by sequencing (Section 6.1.7 and 6.1.9). The generated double mutant of Rna15 was then used as a template to perform a second QuikChange PCR where Y21A mutation was introduced. The final pACEBac1 plasmid contained 6xHis tagged RNA15 gene with Y21A, Y61A and F63A mutations. A pBIG1a containing both RNA14 and RNA15 genes were created as described in section 6.2.2.1 and 6.2.2.2.

6.2.6 Yth1 and Fip1 constructs for NMR

Fip1 residues 180-220, Yth1 residues 118-170 (ZnF 4 and 5) and Yth1 residues 118-208 (ZnF 4, 5 and the C-terminal end) were cloned into pGEX-6P-2 (Addgene) using the BamHI and EcoRI sites as described in section 6.2.1. pIDC-Fip1 and pACEBAC1-Yth1 were used as the template plasmid for the cloning PCR.

The following primers were used in the cloning:

- Fip1 residues 180-220

5′- ATTAA GGATCC ATG ATTGATCCGGAAGTTCTG-3′

5′- ATAAT GAATTC TTATTA CAGTTTTTCTTGACGGTGC-3′

- Yth1 residues 118-170 (ZnF 4 and 5)

5′- ATTAA GGATCC ATG GCAAGCAAAATTCCGAAA-3′

5′- ATAAT GAATTC TTATTA AAACTGCGGATGTTCCATATCA-3′

- Yth1 residues 118-208 (ZnF 4, 5 and the C-terminal end)

5′- ATTAA GGATCC ATG GCAAGCAAAATTCCGAAA -3′

5′- ATAAT GAATTC TTATTA TTAAACTTCACCGTTAATAATGGCG -3′

6.2.7 Phosphatase module

All the six individual subunit genes were synthesised by GeneArt. All the genes were first cloned into pACEBac1 from pMK vector using BamHI and XbaI sites as described in section 6.2.1. The gene cassettes were digested from pMK vector, separated by running a 1% agarose TAE gel and purified using the QIAquick Gel Extraction kit (Qiagen). The inserts were ligated into pACEBac1 as described in section 6.2.1.4. The clones were then verified by sequencing. Once the genes were cloned into pACEBac1, Pta1 was cloned into pBIG1a; Ssu72, Pti1, Glc7, Ref2-SII and Swd2 were cloned into pBIG1b as described in section 6.2.2.1 and 6.2.2.2. A pBIG2ab vector containing all the six phosphatase module subunits was then assembled using the second Gibson assembly (KGB2) as described in 6.2.2.4.

6.2.8 Core-CPF

Polymerase module genes were cloned into pBIG1a as described in section 6.2.2.1 and 6.2.2.2. There were two pBIG1a polymerase module constructs generated. One of the constructs contained Pfs2 with a twin Strep tag. This tagged polymerase module was later used in assembling core-CPF (Figure 3.6). The second polymerase module construct generated did not contain any tags (Figure 5.14a). This untagged polymerase module was later used in assembling the full fourteen-subunit CPF (Figure 5.14d).

Similarly the nuclease module subunits were cloned into pBIG1b. Core-CPF was generated by assembling the polymerase and nuclease polygene cassettes into pBIG2ab as described in 6.2.2.4.

Nuclease module subunit Cft2 was cloned into pBIG1b. Polymerase module polygene cassette from pBIG1a was then assembled together with Cft2 from pBIG1b to generate a six-subunit polymerase module containing Cft2 (Figure 3.6).

6.2.9 Full CPF

The strategy used to clone a full fourteen-subunit CPF is illustrated in figure 5.12. An untagged polymerase module was assembled into pBIG1a (referred to as construct A), untagged nuclease into pBIG1b (referred to as construct B), a twin Strep tag containing Ref2, Ssu72, Pti1, Glc7, Swd2 were assembled into pBIG1c (referred to as construct C) and a lone Pta1 was coned into pBIG1d (referred to as construct D). I tried to assemble all the CPF subunits into one single vector namely pBIG2abcd. Initial attempts at generating pBIG2abcd containing CPF using the protocol described in section 6.2.2.4 failed. This second step of Gibson assembly was trickier than previously thought. If assembled correctly into one single vector, the entire CPF polygene cassette is expected to be around 44,000 nt in length. The individual polygene cassettes are of the following size: A ~12000, B ~ 8900, C ~8100, D ~ 3200 and the pBIG2abcd backbone ~ 3000 nucleotides. Many often, the Gibson assembly reactions yielded no colonies after transformation into TOP10 cells. The transformation efficiency of such big gene assemblies is probably very low. Next, in the protocol for the second Gibson assembly, only the receiving vector is linearized prior to the reaction (Section 6.2.2.4). The pBIG1 series of plasmid containing the polygene cassette are linearized simultaneously together in 1x IRB by the addition of Pme1 (Table 6.13). This likely results in reduced efficiency of Gibson assembly. Next, I tried the Gibson assembly in a variety of different conditions in addition to the protocol described in 6.2.2.4. First, I tried to increase the total amounts of all the individual plasmids used in the assembly. Next, I resorted to a 'cleaner' second Gibson assembly step. Along with the pBIG2abcd plasmid, the pBIG1a,b,c and d plasmids were also linearized by Pme1 in separate reactions. The desired polygene cassettes (A,B,C and D) were then separated from the

backbone of pBIG1 by running a 1% TAE agarose gel and purified using the QIAquick Gel Extraction kit (Qiagen). Now, we have purified linear vector pBIG2abcd, and polygene cassettes A, B, C and D with desired Gibson overhangs. I set up a Gibson assembly reaction as described as follows:

Component name	Concentration (pmol)
Pme1 digested pBIG2abcd	0.04
Polymerase module cassette	0.09
Nuclease module cassette	0.028
Ssu72-Pti1-Glc7-Ref2SII-Swd2' cassette	0.014
Pta1	0.017
Nuclease free water	up to 5 µl

Table 6.18: Making the DNA mix to be used in the CPF Gibson assembly

The DNA mixture is prepared as described above. The DNA mix is then added to 15 µl of the Gibson mix 1 (KGB 1), well suspended and incubated in a thermo cycler at 50 °C for 60 minutes. The rest of the protocol is similar to the one described in section 6.2.2.2. Interestingly, there were only very few colonies grown on the TYE plates containing chlormaphenicol. After screening several (~ 60) colonies by PacI digestion, two colonies CPF1 and CPF2 were found to contain the correct Pac1 digestion pattern (Figure 5.13a).

6.2.10 RNA production

DNA coding for the 3′ UTR of *CYC1* mRNA cloned into pIDTSmart Kan was purchased from IDT. The construct named *CYC1* 3′-UTR contained 259 nucleotides from the 3′ UTR of *CYC1* mRNA. The construct named 5′-CYC1-pc contained 182 nucleotides from 3′ UTR of *CYC1* mRNA. The 5′-CYC1-pc RNA ends in the poly(A) site whereas *CYC1* 3′-UTR RNA contained sequences downstream of the poly(A) site as well. *CYC1* 3′-UTR was synthesized as described in Jana Wolf, 2014.

PCR primers were designed such that sequence corresponding to 5′-*CYC1*-pc (used in Figure 5.5) or 5′-*CYC1*-pc-A30 (used in Figure 5.10) were amplified from the vector plasmid pIDTSMART-Kan-EcoR1-T7-CYC1-BamH1. The vector contains the T7 RNA polymerase transcription promoter site TAATACGACTCACTATAGGG just before the sequence of our desired RNA. The PCR products would then contain the T7 promoter site and the DNA coding for our RNA of interest. This PCR product is then used as a template for *in vitro* transcription reaction (section 6.5).

The following are the primers used to amplify 5′-CYC1-pc

5′- CCCGTGTAAAACGACGGCCAGTTTA-3′
5′- TTGAAATATAAATAACGTTCTTAATACTAACATAAC-3′

The following are the primers used to amplify 5′-CYC1-pc-A30

5′- CCCGTGTAAAACGACGGCCAGTTTA-3′
5′- TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTGAAATATAAATAA-3′

Similarly the 3′ UTR of *GAL7* mRNA cloned into pIDTSmart Kan was purchased from IDT. The construct named *GAL7* 3′-UTR contained 251 nucleotides from the 3′ UTR of *GAL7* mRNA. The following are the primers used to amplify *GAL7* 3′-UTR from the plasmid pIDTSMART-Kan-EcoR1-T7-GAL7-BamH1

5′- CCCGTGTAAAACGACGG-3′
5′- TACAGATAATGATGTCATTATTATATATATATATATATATTGCTACTCC-3′

For the production of *GCN4* and *MFA2* RNAs, I ordered a gBlocks gene fragments from IDT containing the DNA sequence of the respective RNAs. The 417-nucleotide long *MFA2* gene fragment contains the full *MFA2* gene including its 5′ UTR and 3′ UTR. The 300-nucleotide long *GCN4* gene fragment contains only the 3′ UTR sequence. PCR was employed to amplify the gene fragments for use in *in vitro* transcription reaction. The forward primer used contains T7 Polymerase promoter site.

The following are the primers used to amplify 5′-UTR *MFA2* 3′-UTR

5′-
TAATACGACTCACTATAGGGAGAGACACCAGCGAGCTATCATCTTCATACA
A-3′
5′- ATAATTAAAAAAAAAATCGAATGTAATGGGTGG -3′

The following are the primers used to amplify *GCN4* 3′-UTR

5′- TAATACGACTCACTATAGGG TTTCATTTACCTTTTATTTTATATTTTTTA -
3′
5′- ACACGTTAATATGGTGGAGTCAGCTGAGA -3′

6.3 Protein expression

6.3.1 Bacterial expression

Plasmids containing the gene of interest were transformed into BL21 (DE3) star cells as described in section 6.1.5. A single colony was picked from the transformation plates and a 50 ml 2xTY starter culture with appropriate antibiotics was set up in a 250 mL Corning Erlenmeyer flask (Sigma) and incubated overnight at 37 °C, 180 rpm in an multitron standard shaker (Infors-ht). The next day, 1L of 2xTY media containing the appropriate antibiotic is inoculated with 10 ml of the starter culture in a 2L glass Erlenmeyer flask. The cells are grown until a $0D_{600}$ of 0.6 and protein expression is induced by the addition of 1 mM IPTG. For the expression of pET-28a(+)-Pap1 and pETDuet–His-Rna15–Rna14, induction was done at 16 °C over night and the cells were harvested the next day. For the expression of pOPINB-Hrp1, induction was done at 37 °C for 3 hours after which the cells were harvested. For the expression of pRSFDuet–His-Pcf11–Clp1, cells were grown in baffled flasks at 37 °C in media containing 0.5% glucose and 1 mM MgSO4 to an OD_{600} of 2.0. Protein expression was induced by adding 500 ml of fresh 2xTY media containing IPTG (final 1 mM IPTG, 30 °C, 3 h).

6.3.2 Bacterial expression of Isotopically labelled proteins

The Yth1 residues 118-170 (ZnF 4 and 5) and residues 118-208 (ZnF 4, 5 and the C-terminal end) were expressed as 15N, 15N-13C labelled proteins in *E. coli* using MOPS minimal medium (Section 6.1.2.3). 100 ml of 10x K-MOPS stock was diluted in 800 ml MQ H_2O. The media was then autoclaved, cooled and 100 ml of the following components were added: 20 mM NH_4Cl, 4 mM KPi, 1% w/v glucose, 10 mg thiamine, 2 mg D-biotin, 2 mg choline chloride, 2 mg folic acid, 2 mg niacinamide, 2 mg D-phantothenic acid, 2 mg pyridoxal, 2 mg riboflavin. For ^{15}N labeling $^{15}NH_4Cl$ was used in H. For ^{13}C labeling ^{13}C-glucose was used.

6.3.3 Baculovirus mediated insect cell over-expression

Once the incorporation of the gene cassettes into pBIG1 or the polygene cassettes into PBIG2 series of vectors have been assessed by restriction enzyme digestion or sequencing (section 6.1.9, 6.2.2.3 and 6.2.2.5), the plasmids are transformed into EmBacY cells (Section 6.1.6).

6.3.3.1 Bacmid preparation

The transformation colonies were grown on bacmid plates for ~ 24 hours. The bacmid incorporation can be verified by classical blue white selection. A few white colonies were picked and used to inoculate 5ml LB medium supplemented with kanamycin, gentamycin. chloramphenicol was additionally added for vectors containing the pIDC backbone and spectinomycin for vectors containing the pIDS backbone. Cultures were grown overnight and cells harvested by centrifugation at 3500rpm for 15 minutes. The supernatant was discarded and the cells were resuspended in 250 µl buffer P1 (50 mM Tris.Cl pH 8, 10 mM EDTA. P1 buffer containing blue lysis indicator is avoided. Cell lysis was carried out by adding 250 µl buffer P2 (200 mM NaOH, 1% SDS), gently inverting the tube ~ 5 times and incubating for 5 min at rom temperature. Adding 350 µl of buffer N3 (4.2 M GnHCl, 0.9 M KAc pH 4.8) results in the formation of thick white precipitate. The mixture was centrifuged for 10 minutes at 20,000 rcf. The resulting supernatant was carefully transferred into a tube containing 800 µl of isopropanol. The tube is mixed well by inverting 4 to 5 times and place on ice for 20 mins. The solution is centrifuged at maximum speed, supernatant removed and 500 µl of 70 % ethanol added to the pellets. The tube was inverted several times to wash the pellet. The mixture is centrifuged for 10 mins at maximum speed and the supernatant removed. The 70 % ethanol wash step is repeated again. This time, the supernatant ethanol is removed using sterile pipette tips inside a flow hood. From this step onwards, all samples are handled under sterile conditions. Air-dry the pellets for less than 2 minutes once the ethanol is completely removed. Care is taken not to over dry the pellets. The pelleted DNA is suspended in 40 µl sterile elution buffer (10 mM Tris pH 8.5). In some instances, a white pellet can be seen along with the DNA pellet. It is likely to be precipitated salt. Resuspended bacmids were transfected into *Sf9* cells on the same day after verifying the presence of all the GoIs by PCR. The remaining bacmids after transfections were stored in the cold room at 4 °C for several months until they are discarded.

6.3.3.2 Verifying clones by PCR

Performing bacmid PCR using gene specific primers is an approach to check the bacmids for the presence of the desired gene expression or polygene cassettes. PCR was performed as described in 6.2.1.1. The following are the gene specific primers used to verify bacmids constructs generated in this dissertation.

Protein	Primers
Cft1	ATGAATGTTTATGATGATGTTCTGG
	TTATTTACCCTGACACAGGCTAC
Pfs2	ATGGATGGTCATAATCAGAATCAG
	CAGGCAGGGTGCTGC
Yth1	ATGAGCCTGATTCATCCGGA
	CACCGTTAATAATGGCGTTC
Fip1	ATGAGCAGCAGCGAAGATG
	TTATTTGCTATTCTGGTTCTGATT
Pap1	ATGAGCAGCCAGAAAGTTTTT
	TTAATTAACATCAACTGCTGCGG
Mpe1	ATGAGCAGCACCATCTTTTAT
	CGGGCTTGCATCTGCC
Cft2	ATGACCTATAAATACAATTGTTGTG
	AATTTTGGCCAGCATATCTGTC
Ysh1	ATGGAACGTACCAATACCACC
	ACACAGCGGTGTAACCAGATTA
Pta1	CCGTCCCACCATCGGGCGCGGATCCATGAGCAGCGCAGAAATG
	GCTTGTCGAGACTGCAGGCTCTAGATTATTTCAGACGATCCAGC

Pti1	CCGTCCCACCATCGGGCGCGGATCCATGACCGATCCGCGTCGTC GCTTGTCGAGACTGCAGGCTCTAGATTAAATAATGTACTCTTTGCGAAAATTTTC C
Ref2	CCGTCCCACCATCGGGCGCGGATCCATGAGCGCTCCGGTTCCG GCTTGTCGAGACTGCAGGCTCTAGATTATTTCTCAAACTGAGGATGGCTCC
Swd2	CCGTCCCACCATCGGGCGCGGATCCATGACCACCGTTAGCATTAAC GCTTGTCGAGACTGCAGGCTCTAGATTATTCATCATACACATAGAAATCAAC
Gl7	CCGTCCCACCATCGGGCGCGGATCCATGGATAGCCAGCCGGTTG GCTTGTCGAGACTGCAGGCTCTAGATTATTTCTTTTTACGACCGCCTG
Ssu72	CCGTCCCACCATCGGGCGCGGATCCATGCCTAGCCATCGTAATAG GCTTGTCGAGACTGCAGGCTCTAGATTAATAATAGCTCGGTGCATAC

Table 6.19: Gene specific primer used in the verification of CPF clones described in section 5.3.1.1.

6.3.3.3 Sf9 cell transfections

All the following steps were performed in sterile conditions in a laminar flow hood. *Sf9* cells in log phase were diluted to a cell count of 0.5 x10^6 cells/ml. 2 ml of the diluted cells were added to each well of a 6-well corning tissue culture plate (Corning Costar, Sigma). The cells were left to adhere to the plates for around 20 minutes. In the meantime, a DNA master mix was prepared for use in five out of the six wells. Cells in the sixth well were treated as positive control for transfection. The five well master mix contains 10 µg of bacmid DNA in total. The master mix is prepared by adding the appropriate volume of bacmid DNA and warm *Sf9* media to a final volume of 980 µl. 20 µl of Fugene transfect HD reagent (promega) is mixed well and pipetted into the DNA media mixture. This mixture is incubated for 10 mins before adding 200 µl per well. The plates are placed into a box along with a damp tissue paper and incubated at 27 °C for 48 to 72 hours. The cells are checked at every 24 hours to monitor the viability and

fluorescence from YFP marker. At 72 hours, the supernatant from all the five wells were removed and pooled together in a 50 ml falcon tube. The supernatant was diluted with *Sf9* media up to 50 ml and FBS was added to a final concentration of 20% v/v. FBS is used to stabilize the virus. The final mixture (hereafter referred to as P1 virus) is sterile filtered using a 50 ml 0.45 µm steriflip. The tube is covered with an aluminium foil and stored in the cold room until further use.

6.3.3.4 Primary virus amplification

100 ml of *Sf9* cells at concentration of 0.5 $x10^6$ cells/ml were grown until 2 $x10^6$ cells/ml in an 500ml Erlenmeyer flask. 5 ml of P1 virus was added to these cells. Cell growth and YFP fluorescence were monitored for every 24 hours. In case the concentration doubled to 4 $x10^6$ cells/ml after 24 hours, additional media was added to dilute the cells back to 2 $x10^6$ cells/ml. Usually, the cell growth arrests after 24 or 48 hours and the YFP fluorescence begin to appear. The supernatant (hereafter referred to as P2 virus) is harvested before the viability starts dropping below 90%. This step is performed around 48 to 72 hours post infection. An ideal condition for harvesting P2 virus is cell count 2 or 3 $x10^6$ cells/ml, 90% cell viability and > 80% YPF fluorescence. P2 virus is harvested by centrifuging the *Sf9* cultures in a refrigerated tabletop centrifuge (Rotana 460 R) at 2000 g for 10 mins. The supernatant is filtered using two 50 ml 0.45 µm steriflip. The resulting P2 virus is immediately used up to infect larger cultures.

6.3.3.5 Protein over-expression

Before going ahead to infect large-scale (~ 2 L) *Sf9* cultures for protein-expression, a small-scale (50 ml) time course expression test is carried out. 50 ml of *Sf9* cells at concentration of 2 $x10^6$ cells/ml (viability > 90%) growing in a 200 ml Erlenmeyer flask are infected using 0.5 ml P2 virus. Samples containing 10^7 cells in total are taken at every 24 hour interval from the growing *Sf9* cell culture. The cells harvested by centrifugation are flash frozen in liquid nitrogen and stored at -80 °C. Cells are taken every 24 hours until the 96-hour time point. Then, the protein(s) of interest are isolated from the cell pellets by carrying out Streptactin or Ni-NTA pull down depending on the

affinity tag present in the construct. The pellets are resuspended in 1 ml of lysis buffer (100 mM HEPES pH 8, 300 mM NaCl, 5% glycerol, 1mM TCEP) and a tiny scoop of glass beads (GE) are added to the cell suspension. Cell lysis is carried out by vortexing for two min and the lysate clarified by ultracentrifugation at 4 °C for 30 minutes. The supernatant is mixed with 20 µl pre-equilibrated Strep or His resins and incubated for 1 hour at 4 °C in an end-over end rotor. The unbound proteins are removed by centrifugation at 600 g for 10 mins at 4 °C. The resins are washed twice with 1 ml wash buffer (100 mM HEPES pH 8, 300 mM NaCl, 1mM TCEP) and the bound-proteins eluted by adding 20 µl of 4x LDS sample buffer. 12 µl of the eluted proteins are analyzed by SDS-PAGE as described in 6.1.11. After analyzing the protein gels, a decision is made about the best time to harvest the infected *Sf9* cells so that the protein expression is maximal. Then, 500 ml of *Sf9* cells at concentration of 2 $x10^6$ cells/ml (viability > 90%) growing in a 2L roller bottle flask is infected with 5 ml of P2 virus. Cell growth parameters are monitored every 24 hours. At the appropriate time (when cell growth is arrested, YFP fluorescence and protein expression is maximum), the Sf9 cell culture is harvested by centrifuging at 4000 rpm for 20 min in 2L bottles. The cell pellets are washed in pre-chilled PBS and flash frozen in liquid nitrogen and stored at -80 °C.

6.4 Protein purification

6.4.1 Polymerase module

Cell pellets from 2 l *Sf9* cells were resuspended in ~120 ml of buffer A (50 mM HEPES pH 7.9, 300 mM NaCl, 1 mM TCEP) supplemented with 50 µg/ml RNAse, 50 µg/ml DNAse and EDTA-free protease inhibitors (PI) (Roche). The cells were lysed by sonication using a 10mm tip (big tip) on a VC 750 ultrasonic processor (Sonics) (3 min total ON time, 70% amplitude, 5 sec ON time and 10 sec OFF time). The resuspended cells were split into two 60 ml halves in a 100 ml kimax glass beaker (Kimble Glass Products, Fisher Scientific) and were sonicated separately. Lysate was cleared by ultracentrifugation at 100,000g (4 °C) for 30 min in a pre-chilled JA25.50 rotor in a Avanti J26 XPI (Beckman Coulter). The clarified lysate was then incubated at 4 °C for 2 hours with 2 ml of StrepTactin Sepharose HP resin (GE) pre-equilibrated with buffer A, on an end over end rotor. Beads were loaded onto a gravity column and washed with 150 ml buffer A prior to elution with buffer E (buffer E is essentially buffer A supplemented with 6 mM of desthiobiotin). Elution was carried out in 2 ml volume per fraction, ten fractions in total. 2m of buffer E were added, the StrepTactin resins were mixed well with it and incubated for 5 min before elution. The elution fractions (20 ml) were pooled together and diluted with 50 mM HEPES pH 7.9 to reduce the salt concentration to 150 mM NaCl. The sample was then filtered using a 0.45 µm membrane filter (Merk) and loaded onto a 1 ml Mono Q 5/50 GL column (GE Healthcare). Elution was performed over a 45 CV gradient from 300–450 mM NaCl. This separated polymerase modules with and without Pap1 (Figure 2.2). The four-subunit complex (without Pap1) was the most abundant species. Individual peak fractions from the anion exchange step were concentrated using a 50 kDa centrifugal filters (Millipore) and was subsequently purified by size exclusion chromatography using a Superose 6 3.2/300 Increase column (GE Healthcare). The size exclusion column was pre-equilibrated with buffer A (Figure 2.3). Purified polymerase module complexes were either used immediately for making cryo-EM grids, or was concentrated using a 50 kDa centrifugal filter (Millipore) to 10–20 µM concentrations. The protein was then flash frozen in liquid nitrogen and stored at -80 °C for future use in biochemical experiments. The same protocol was used to purify polymerase module variant carrying Fip1 Δ 145-170 (Figure 3.9a).

6.4.2 Pap1

BL21(DE3) Star cells containing pET-28a-His-Pap1 from 2 l cultures were resuspended in ~ 70 ml buffer B (50 mM HEPES pH 8.0, 1 M NaCl, 20 mM imidazole, 5% w/v glycerol, 1 mM TCEP, 2 µg/ml DNase I, 2 µg/ml RNase A, PI). The cells were lysed by sonication using a 10 mm tip on a VC 750 ultrasonic processor (Sonics) (10 min total ON time, 100% amplitude, 5 sec ON time and 20 sec OFF time). The glass beaker containing the cell suspension was checked every 3 minutes to ensure that the samples were not over-heated during sonication. If the beaker was found to become warm, the ice bucket carrying the glass beaker was replaced with fresh ice. The lysate was clarified as described in section 6.4.1 and was mixed with 1 ml bed volume of Ni-NTA resin. The resins were pre-equilibrated with buffer C (50 mM HEPES pH 8.0, 500 mM NaCl, 20 mM imidazole, 1 mM TCEP). The resin-lysate mixture was incubated for 2 hours at 4 °C on an end over end rotor (9rpm). The resin was then washed two times with 50 ml of buffer C. Beads were transferred to a gravity column before elution with buffer D (50 mM HEPES pH 8.0, 500 mM NaCl, 300 mM imidazole). Elution was carried in 1ml buffer volume. The elution fractions were collected instantaneously after the buffer was added; there was no incubation time involved. Fractions containing Pap1 were exchanged into buffer F (50 mM HEPES pH 8.0, 100 mM NaCl, 0.5 mM TCEP), and loaded onto a 1 ml HiTrap Heparin HP (GE). Proteins were eluted over a 30 CV gradient up to 100% buffer G (50 mM HEPES pH 8.0, 1 M NaCl, 0.5 mM TCEP). The samples from heparin step were polished using a Superdex 200 10/300 column equilibrated in buffer H (50 mM HEPES pH 8.0, 150 mM NaCl, 1 mM TCEP) (Figure 2.4c). Pap1 was then concentrated to 36 µM using a 30kDa centrifugal filter (Millipore), flash-frozen in liquid nitrogen and stored at -80 °C.

6.4.3 CF IA

BL21(pLysS) Rosetta expressing Rna14–Rna15 from 4l of cultures were suspended in buffer I (50 mM HEPES pH 8.0, 250 mM NaCl, 30 mM imidazole, 5% w/v glycerol, 0.5 mM TCEP, 2 µg/ml DNase I (Sigma), PI). Lysed cells were separated from the cell debris by ultracentrifugation at 75,000g for 30 min. A 5 ml HisTrap HP column (GE) was equilibrated with buffer J (20 mM HEPES pH 8.0, 250 mM NaCl, 0.5 mM TCEP). The clarified lysate was filtered using a 0.45 µm membrane filter (Millipore) and loaded onto the column. Elution was carried out using a 10 CV gradient up to 500 mM

imidazole. Elution fractions were pooled, concentrated using a 50 kDa centrifugal filter (Millipore) at 4000g and polished on a Superdex 200 16/600 column (GE). Buffer J was used in the size exclusion step. Peak fractions were pooled and concentrated to ~ 11.2 mg /ml or 100 µM. Aliquots were flash-frozen in liquid nitrogen and stored at -80 °C.

Cells expressing His-Pcf11–Clp1 from 6l culture were suspended in 200 ml buffer K (50 mM Tris pH 8.0, 250 mM NaCl, 5% w/v glycerol, 1 mM TCEP, 2 µg/ml DNase I, 2 µg/ml RNase A, 1 tablet of PI per 50 ml). Cells were split in two 100 ml volumes and lysed as described earlier. The lysate was separated from cell debris by ultracentrifugation (JA 25.50, 76,000g, 40 min, 4 °C). A 5 ml HisTrap FF crude column (GE) was equilibrated with buffer L (20 mM Tris pH 8.0, 250 mM NaCl, 1 mM TCEP). The clarified lysate was filtered using a 0.45 µm membrane filter (Millipore) and loaded onto the column. A 50 ml wash was done using 20 mM Tris pH 8.0, 20 mM imidazole, 250 mM NaCl, 1 mM TCEP and the proteins were eluted by a 10 CV gradient elution up to 500 mM imidazole.

30 mg of Rna14–Rna15 after the size exclusion step was mixed with 30 mg of Pcf11–Clp1 from the HisTrap column. The mixture was incubated at 4 °C overnight. The complex mixture was diluted to 100 mM NaCl by adding 20 mM Tris (no salt) up to 40 ml volume. There were signs of some precipitation. The precipitants were separated by by centrifugation at 8000g for 10 min. The complex was then filtered using a 0.45 µm membrane filter (Millipore) and loaded onto a 5 ml HiTrap Q HP column (GE). The column was washed with 25 ml of 20 mM Tris pH 8.0, 100 mM NaCl, 1 mM TCEP. Elution was carried out using a 6 CV gradient up to 1 M NaCl. The elution fractions were analyzed by SDS-PAGE. The central fractions containing the complex were polished on a Superdex 200 26/600 column (GE). The size exclusion step was run in buffer L. CF IA was concentrated to 10 mg/ml or 24 µM and stored at -80 °C for further use.

6.4.4 CF IB

Cells from a 2l culture were suspended in 60 ml of buffer M (50 mM HEPES pH 8.0, 300 mM NaCl, 20 mM imidazole, 0.5 mM TCEP supplemented with 10% w/v glycerol, 2 µg/ml DNase I, 2 µg/ml RNase A, PI). Cells were lysed as described in section 6.4.2. The lysate was clarified as described earlier. The lysate was mixed with 1 ml bed

volume of Ni-NTA beads and incubated in the cold room at 9 rpm on an end-over end rotor for an hour. The beads were pre-equilibrated in buffer M. Then the resins were washed with 100 ml buffer M followed by 50 ml of buffer N (50 mM HEPES pH 8.0, 300 mM NaCl, 30 mM imidazole, 0.5 mM TCEP). Elution was performed in a gravity column and with buffer O (50 mM HEPES pH 8.0, 150 mM NaCl, 300 mM imidazole, 0.5 mM TCEP). The eluted Protein was then loaded onto a 1 ml HiTrap Heparin column (GE) at 1 ml per min. The column was washed with 5 ml of buffer O containing 250 mM NaCl. CF IB was eluted using a 20 CV salt gradient step up to 1M NaCl. Fractions containing Hrp1 were pooled and concentrated with a 10 kDa centrifugal filter. CF IB was polished on a Superdex 200 10/300 column (GE). The size exclusion step was run in buffer J. Hrp1 was concentrated in a 30kDa centrifugal filter and stored at -80 °C for future biochemical assays.

6.4.5 Rna14/Rna15 Y21A, Y61A, F63A

2l of *Sf9* cells expressing pBIG1a Rna14/Rna15 were resuspended in buffer P (50 mM HEPES pH 8, 250 mM NaCl, 30 mM imidazole, 5% w/v glycerol, 0.5 mM TCEP, 2 µg/ml DNAse I and PI). The cells were sonicated for 3 minutes at 70% amplitude (4 °C, 5 ON, 10 OFF) using a 10 mm tip on a VC 750 ultrasonic processor (Sonics). The lysate was clarified by ultracentrifugation (4 °C, 18000 rpm, JA 25.50, 35 min). 1 ml bed volume Ni-NTA resin was pre-equilibrated in buffer P and mixed with lysate. The binding reaction was carried at 4 °C for 90 mins in an end over end shaker (9 rpm). The resins were separated by spinning them at 600 g for 5 min. Resins containing bound Rna14/Rna15 is washed with 120 ml of buffer P. Elution is carried by buffer containing 500 mM imidazole. Ten fractions each containing 2 ml elution buffer were collected. All the fractions were pooled together and concentrated in a 30kDa centrifugal filter and exchanged into buffer J during this process. The concentrated samples were filtered using a 0.45 µm membrane filter (Merk) and loaded onto a HiLoad 16/600 Superdex 200 pg (GE) column at 1 ml per min. The mutant complex runs around the same elution volume as the wild type Rna14/Rna15 purified from *E.coli*. The fractions were analyzed by SDS-PAGE. Rna14/Rna15 mutant complex was concentrated up to 24 µM and stored at - 80°C.

6.4.6 Yth1 and Fip1 constructs for SEC-MALS and NMR

The following were the constructs purified: Fip1 residues 180-220, Yth1 residues 118-170 (ZnF 4 and 5) and Yth1 residues 118-208 (ZnF 4, 5 and the C-terminal end). Cells overexpressing pGEX-6P-2 vector containing the above constructs were produced as described in 6.3.1 and 6.3.2. All the constructs were purified using a similar protocol as described in this section. Cells from 1l of culture was suspended in ~ 60 ml of buffer Q (50 mM Tris pH 8, 250 mM NaCl, 1mM TCEP, 2 µg/ml DNAse / RNAse and EDTA-free protease inhibitors). The cells were sonicated and the lysate clarified as described in section 6.4.2. 3 ml of Glutathione Sepharose® 4B (GE) were pre-equilibrated with buffer Q. The lysate and the resins were mixed, binding reaction was carried at 4 °C for 90 mins in an end over end rotor (9 rpm). The resins were transferred to a gravity column after the binding reaction. 2 ml of buffer Q supplemented with glutathione (1g per 50 ml buffer, pH adjusted to 8) were used to elute the GST tagged Yth1/Fip1 constructs. In total, ten elution fractions were collected. The elution fractions were pooled together and the 3C site between GST and the proteins are cleaved overnight using PreScission Protease. 200 µl of protease (5 mg/ml stock) was used to cleave ~ 35 ml of GST tagged proteins at ~ 1 mg/ml. The overnight cleavage reaction contained EDTA-free protease inhibitor as well. The overnight cleavage reaction was concentrated using a 3kDa centrifugal filter and the buffer was exchanged to the size exclusion buffer during this step. Size exclusion can either be done using 20 mM Bis-Tris pH 6, 150 mM NaCl, 0.1 mM TCEP as the buffer or alternatively simply by using PBS. HiLoad 26/600 Superdex 75 pg (GE) was used. Yth1 constructs containing residues ZnF 4 and 5 elutes around ~ 225 ml, ZnF 4, 5 and the C-terminal elutes around ~ 190 ml. The Fip1 180 - 220 construct elutes at ~ 285 ml.

6.4.7 Core-CPF

Core-CPF contains a StrepII-tag on the Pfs2 subunit (similar to polymerase module). So the purification strategy was very similar to that of the polymerase module as described in section 6.4.1. However, there were a few modifications made to the protocol. Firstly, the lysis buffer for core-CPF was 50 mM HEPES pH 7.9, 150 mM NaCl, 1 mM TCEP, 1mM $Mg(OAc)_2$, 50 µg/ml RNAse and DNAse, and EDTA-free protease inhibitors. It is to be noted that 18 tablets of protease inhibitors were added to 250 ml of lysis buffer. Second, after the shallow gradient elution on the monoQ (GE), all the peak fractions

were combined together unlike the polymerase module purification where three different peaks were purified separately afterwards. Interestingly, core-CPF did not separate into three diferent Pap1 varying species on monoQ. The pooled fractions from monoQ were concentrated and loaded onto a Superose 6 Increase 10/300 GL (GE). The final size exclusion buffer was 20 mM HEPES pH 7.9, 150 mM NaCl, 1 mM TCEP, 1mM $Mg(OAc)_2$.

6.4.8 Recombinant CPF

In order to produce recombinant full CPF, a co-infection strategy was used. The protein expression protocol differs from the routinely used procedure described in section 6.3.3.5. P2 virus corresponding to a six-subunit phosphatase module (Ref2 subunit with a StrepII-tag) and another P2 virus corresponding to an eight-subunit core-CPF (containing no affinity tags) were mixed in 1:3, 1:1 or 3:1 v/v ratio. 10 ml of the combined P2 virus was then used to infect 500 ml of Sf9 cells (2 $x10^6$ cells/ml, viability > 90%) growing in a 2L roller bottle flask. Cells are harvested after 72 hours post infection as described in 6.3.3.5.

CPF lysis buffer: 200 mM HEPES pH8, 200 mM KCl, 0.5 mM $Mg(OAc)_2$, 1mM TCEP, 10% w/v glycerol, 2 µg/ml DNaseI (add fresh), 2 µg/ml RNaseA (add fresh) and PI(14 tablets per 120 ml).

CPF wash buffer: 50 mM HEPES pH 8,150 mM KCl, 0.5 mM $Mg(OAc)_2$, 1 mM TCEP

CPF elution buffer: 1.2mg/ml D-desthiobiotin in wash buffer

2L worth of *Sf9* cell pellets were suspended in ~ 120ml CPF lysis buffer (up to 60ml lysis buffer for every 1L of cells harvested). Sonication and clarification of lysate was done as described in section 6.4.1. The clarified lysate was incubated with 2 ml (bed volume) of pre-equilibrated StrepTactin Sepharose HP resin (GE). Binding reaction was performed for 2 hours at 4 °C in an end over end shaker (9 rpm). The resins are separated from the unbound lysate by centrifugation at 600g for 10 mins. The resins were washed with 200 ml CPF wash buffer in a gravity column. The Strep elution step

was done at room temperature. 3ml CPF elution buffer was added to the resin, mixed well using a 2 ml glass pipette tip and the elution slurry was incubated for 5 mins. Ten elution fractions each containing 3 ml volume was collected. The eluates were then pooled together. The ~ 30 ml recovery solution was filtered by passing it through a 0.45 µm membrane filter (Merk). The solution was then loaded onto a 1 ml resource Q column (GE) pre-equilibrated with CPF wash buffer. Elution was carried out over a 100 CV gradient up to 1M salts. The elution fractions were analysed by SDS-PAGE. Such a shallow gradient on an anion-exchange column results in the separation of CPF with varying amounts of Pap1. Only those fractions containing stoichiometric amounts of Pap1 were pooled together and concentrated using a 50 kDa centrifugal filter upto ~ 50 to 100 µl volume. The CPF preparation was polished by running a superose 6 3.2/300 Increase column (GE). The size exclusion column was pre-equilibrated with CPF wash buffer. Purified CPF was used immediately for making negative stain EM grids. The samples used in biochemical assays were from concentrated CPF that had been flash frozen in liquid nitrogen and stored at -80 °C. To test whether rCPF could survive freezing and thawing, 50 µl of recovery from the resQ column were frozen. The frozen sample was allowed to thaw at room temperature and subjected to size exclusion chromatography as described earlier.

Native CPF was purified as described earlier (Schreieck et al. 2014; Easter 2014) .

6.4.9 Nuclease module subunits

Cft2 and Ysh1/Mpe1 were purified as described earlier (Hill et al. 2019).

6.5 RNA production

The DNA template to be used in *in vitro* transcription (IVT) was obtained as described in section 6.2.10. The PCR to make the DNA template was carried out in a 150 µl volume (scaled up in comparison to table 6.3). The PCR products were purified by QIAquick PCR Purification Kit (Qiagen).

IVT was set-up as follows:

Component name	Amount
Nuclease free water	up to 500 µl
10x transcription buffer	50 µl
rNTP (25 mM)	125 µl
DNA template	~ 5 - 10 µg
Pyrophosphatase (100 U/µl)	5 µl
RiboLock (40 U/µl)	5 µl
T7 RNA Pol (0.5 mg/ml)	100 µl
$MgCl_2$ (400 mM)	50 µl

Table 6.20: Setting up IVT

IVT reaction was carried out for 1 hour at 37 °C. The reaction was stopped by adding 500 µl of 2X RNA loading dye to the mix. The products were run on a preparative 6% urea-acrylamide gel (Section 6.1.10.2). The RNA band of interest was identified by UV shadowing (Hassur and Whitlock 1974). The band was excised and RNA extracted from the gel using electro-elution in a Whatman Elutrap electroelution system kit (Thermofisher) running on 1x TBE (made with nuclease free H_20). Elution fractions were collected every one-hour and the final time point taken at 5 or 6 hours after the beginning. The fractions were analyzed by running a 6% ureapolyacrylamide gel. Fractions containing pure RNA were concentrated using a 15 ml 3 kDa centrifugal filter, flash frozen in liquid nitrogen and stored in - 80°C until future use.

6.6 Biochemical assays

6.6.1 *In vitro* cleavage and polyadenylation assays

6.6.1.1 Polyadenylation assays of 5'-FAM-CYC1-pc

A 42 nt sequence from the 3′ UTR of *CYC1* was chemically synthesized with a 5′ 6-FAM fluorophore (IDT). This was the substrate RNA used in all the assays described in chapter 3 and 4. *In vitro* assays were carried out in a polyadenylation buffer that consisted of 10 mM HEPES pH 7.9, 150 mM KOAc, 2 mM $Mg(OAc)_2$, 0.05 mM EDTA, 2% PEG (v/v) and 1 mM DTT. Sub-stocks of the purified proteins were made from their frozen stock in a standard buffer (50 mM HEPES pH 7.9, 150 mM NaCl, 1 mM TCEP). The sub-stock concentration was 500 nM for polymerase module, Pap1 and CPF, and 4 µM for the accessory factors CF IA and Rna14/Rna15. The reaction was assembled as follows:

Component name	Per reaction / time point
Nuclease free water	up to 20 µl
5X polyadenylation buffer	4 µl
5′-FAM-*CYC1* (4 µM)	2 µl
DTT (10 mM)	2 µl
RiboLock (10 U/µl)	1 µl
Accessory factors (4 µM)	2 µl
Polymerase (500 nM)	2 µl
ATP (20 mM)	2 µl

Table 6.21: Setting up *in vitro* polyadenylation assays

If a reaction with ten time points was to be performed (Figure 2.7), a 220 µl reaction mix was made. The assays were performed at 30 °C on a heat block. The reaction time point 'zero' was taken before the addition of ATP. At every time point, 20 µl of the reaction solution was removed from the master mix and mixed with 20 µl of 2x RNA loading dye. Products were analyzed by running on a 15% w/v ureapolyacrylamide gel

for 30 minutes at 400 V. The gel was imaged on a Typhoon FLA-7000 laser scanner (GE Healthcare).

For assays with accessory factors (Figures 4.1, 4.2, 4.3, 4.4 and 4.5), the master mix containing 5′-FAM-*CYC1* was pre-incubated with equal molar amounts of accessory factors for 5 minutes before the addition of polymerase. For assays comparing the activity of Pap1 and the polymerase module variants, purified proteins were first resolved by 4–12% SDS-PAGE (Invitrogen) and visualized by staining with InstantBlue™ (Expedion) (Figure 2.5). 8 µl of 150 nM amounts of polymerase module variants (with and without Pap1) and isolated Pap1 were loaded per lane of the gel (Figure 2.5a). 3 µl of 500 nM amount of polymerase module variant and 3 µl of 1000 nM amount of isolated Pap1 were loaded per lane (Figure 2.5b). Intensity of Pap1 bands were analyzed by Image J. After normalising for the intensity of Pap1, the proteins were further used in polyadenylation assay described in Figure 6 and 7.

6.6.1.2 Cleavage and polyadenylation assays of CYC13′-UTR, 5′-CYC1-pc and 5′-CYC1-pc-A_{30}

The cleavage and polyadenylation assays shown in chapter 5 were performed in a very similar manner to section 6.6.1.1, barring a few modifications. The concentrations of the RNA substrate were 100 nM in all the assays, different from the earlier ones in section 6.6.1.1. Secondly, the cleavage and polyadenylation reactions were stopped by adding 4 µl of STOP solution (130 mM EDTA, 5% (w/v) SDS, 12 mg/ml proteinase K made in assay buffer) to 20 µl time reaction mix. The STOP reaction mix was incubated at 37 °C for 10 min, and on ice for 20 min before adding 24 µl of 2x RNA loading dye to it. 18 µl of the final sample were loaded per well on a 15% w/v ureapolyacrylamide gel and run for 30 minutes at 400 V. The RNA used in the cleavage and polyadenylation assays were not fluorescently labelled. So the gels were visualized by staining with SyBr Green II and were imaged on a ChemiDoc XRS+ (BioRad). Third, the order of addition of the reaction components was different from that of table 6.21. In these assays, the substrate RNA was added in the end (before ATP). In case of cleavage only assays as shown in appendix 8.13, reaction was started by the addition of RNA in the end.

6.6.2 *In vitro* pull downs

Bait proteins and complexes containing a StrepII-tag were diluted to a concentration of 1.5 µM in pull-down buffer (10 mM HEPES pH 7.9, 150 mM NaCl, 0.5 mM $Mg(OAc)_2$) containing 0.05% Tween-20. 100 µl of bait protein was mixed with 40 µl bed volume of StrepTactin resins (GE) and this mixture was topped up to 1 ml with pull-down buffer. The negative control reaction did not contain any bait protein added to the resins; instead only 100 µl buffer was added. The binding reaction was carried out for 1 hour at 4°C. The unbound proteins (supernatant) were separated from the resins by centrifugation at 600g for 5 min. The supernatant was removed; the pelleted resins were washed with 1 ml of pull-down buffer. The wash step was repeated again. 1 ml of pull-down buffer was added and the resins were mixed well. The slurry was split into four tubes. Each tube contained 10 µl bait-loaded resins in 250 µl pull-down buffer. 20 µL of untagged prey proteins at 4 µM concentration were added to each of the four tubes. The second binding reaction was performed for 60 min at 4°C. The supernatant was removed by centrifugation; resins were washed four times with 1 ml pull-down buffer as described earlier. Proteins were eluted from the resins by the addition of 4x LDS and incubation at 95°C for 2 min. The elution samples were analyzed by SDS-PAGE. A 4%–12% gradient gel was used to run the samples in MOPS-SDS buffer at 200 V for 50 min. Proteins were visualized by staining with InstantBlue™ (Expedeon).

6.6.3 EMSA

A 96-well plate was used to assemble the protein-RNA mixture. In a single well, 8 µl of purified protein was incubated with 1 µl of 500 nM RNA and 1 µl of 10x orange loading dye (0.4% w/v orange G, 1 mM EDTA pH 8, 50% v/v glycerol). In a similar manner, six wells in a row were filled with protein-RNA mix. The protein-RNA mix was incubated on ice for 20 minutes. The final concentrations of proteins were 0 nM, 125 nM, 250 nM, 500 nM, 1 µM and 2 µM in each of the wells. The protein was diluted to various concentrations from the stock using a buffer (50 mM HEPES pH 7.9, 150 mM NaCl, 1 mM TCEP) containing 1 U RiboLock Inhibitor (ThermoScientific). The sample from each of the wells was analyzed by running a native-polyacrylamide gel (6% [w/v] 19:1 acrylamide: bisacrylamide, 1xTBE) for 40 minutes at 100 V. The gel was imaged on a Typhoon FLA-7000 laser scanner (GE Healthcare).

6.6.4 Protein-RNA pull downs

30 µl of StrepTactin Sepharose HP resin (GE) were washed with 1 ml of nuclease free H_20, followed by 1 ml of polyadenylation buffer. Washes were separated from the StrepTactin resins by centrifuging at 600g for 5 min. To the resins, 56 µl of 268 nM purified CPF was added. The slurry was topped up to 300 µl with polyadenylation buffer. The binding reaction was performed for 45 min at 4 °C. The unbound CPF was removed by centrifugation and the resins washed once with 1 ml polyadenylation buffer. Care was taken not to leave the resins dry at any point of time during the entire procedure. Then 300 µl of the cleavage reaction mixture was added to the resins bound to CPF and the cleavage reaction performed for 10 min at 30 °C.

Component name	Total amount added
1X polyadenylation buffer	454.2 µl
RiboLock (40 U/µl)	15 µl
DTT (60 mM)	30 µl
CF IA (2.4 µM)	75 µl
CF IB (10.4 µM)	17.4 µl
CYC1 3′-UTR (7.2 µM)	8.4 µl

Table 6.22: Preparing the cleavage reaction mix

The slurry is then centrifuged at 600g for 10 mins and supernatant removed. The resins are washed three times with 1 ml of polymerase buffer containing 0.01 % Tween-20. 20 µl of 2x RNA loading dye are added to the resins to elute the protein-RNA complex. The control pull-down was performed alongside where there was no CPF used. 56 µl of polyadenylation buffer was used in the first bait-binding step instead.

6.6.5 Analytical Size Exclusion Chromatography

6.6.5.1 Cross-linking of polymerase module

12 µl of polymerase module (with two copies of Pap1) at 8.4 mg/ml (22 µM) was mixed with 88 µl of buffer J and 25 µl of BS3 at 10 mM concentration. The mixture was incubated on ice for 30 min. The reaction was quenched by addition of 25 µl of NH_4HCo_3. The cross-linked sample was loaded onto a Superose 6 Increase 3.2/300 column using a 100 µl loop. The column was run at a flow rate of 0.06 ml/min.

6.6.5.2 Interaction between Rna14/Rna15

40 µl of purified polymerase module (without Pap1) at 29 µM was mixed with 50 µl of Rna14/Rna15 at 90 µM and 20 µl of buffer J. The mixture was incubated on ice for 15 minutes. The mixture was loaded onto a Superose 6 Increase 3.2/300 column in a 100 µl loop. The column was run at a flow rate of 0.06 ml/min. The control experiment consisted of performing size exclusion chromatography of the following two samples: 40 µl of polymerase module (without Pap1) at 29 µM mixed with 70 µl of buffer J; and 50 µl of Rna14/Rna15 at 90 µM mixed with 60 µl of buffer J.

The ternary complex of polymerase module, Rna14/Rna15 subunits and 5′-FAM-*CYC1*-pc were prepared as follows: 50 µl of purified polymerase module (without Pap1) at 40 µM was mixed with 25 µl of purified Rna14/Rna15 at 100 µM, 10 µl of 5′-FAM-*CYC1*-pc at 100 µM and 25 µl of 50 mM HEPES pH 7.9, 175 mM NaCl, 1 mM TCEP. The mixture was incubated on ice for 30 min. Size exclusion chromatography was performed as mentioned in the previous paragraph.

6.7 Electron microscopy studies

6.7.1 Negative stain electron microscopy

The following were the samples assessed by negative stain electron microscopy in this dissertation: the three polymerase module variants and recombinant CPF. Uranyl acetate (1.5% w/v) solution was used as a stain in making the samples. The stain was filtered using a 0.22 μM syringe filter before use in grid making. The final concentration of the protein samples while making the grids were as follows:
130 nM polymerase module (with one Pap1), 100 nM polymerase module (without Pap1), 100 nM polymerase module (With two Pap1s) and 30 nM rCPF. Continuous carbon grids CF400-Cu-UL (EMS) were used to make sample for negative stain EM. The grids were glow-discharged for 30 s before 3 μl of freshly purified proteins (at the aforementioned concentrations) were added to the surface of the grids. The proteins were allowed to adsorb on the grids for 45 seconds. Using a pointed tip of a whatman filter paper, the protein droplet is blotted. The surface of the grid containing the adsorbed proteins is immersed on to a 40 μl drop of uranyl acetate placed on a clean strip of parafilm tape (Partec Premium). After 15 seconds, the grid is moved on to a nearby 40 μl drop and immersed for 15 more seconds. The drop of uranyl acetate is removed by blotting with a pointed tip of a whatman filter paper. The grids are now ready for imaging in the microscope. Micrographs were acquired on a Tecnai Spirit microscope (FEI) operating at 120 keV. Images were recorded on a Ultrascan 1000 CCD camera (Gatan). The polymerase module data set was collected at 21,000x magnification. A total of 133 and 120 micrographs were acquired for the polymerase module without Pap1 and with Pap1 respectively. For the polymerase module, particle picking was done using E2boxer swarm software (Tang et al. 2007) and subsequent data processing was done using Relion 1.4 (Scheres 2012). In order to clean the data sets, several rounds of reference-free 2D classification was performed. Finally, the 2-D classes shown in Figure 2.9a contained 6068 particles for the polymerase without Pap1 and 3363 particles for the sample polymerase with Pap1. A featureless spherical blob was used as the initial model to perform 3D refinement (Figure 2.9b).

Whereas for rCPF, a larger data set consisting of 472 micrographs were collected at −1.78 μm defocus and with a total electron dose of ~ 30 e-/Å^2 over 1 s. All the data

processing including particle picking were done in Relion 2.0. Owing to inherent sample heterogeneity, I did not preceded further with data processing beyond preliminary 2D classification.

6.7.2 Electron cryomicroscopy

Cryo-EM analyses of four different samples are presented in this dissertation - polymerase module (without Pap1), polymerase module (with Pap1), cross-linked polymerase module (with two Pap1s) and a ternary complex of polymerase module, Rna14/Rna15 and *CYC1* RNA.

6.7.3 Sample screening

Ultrastable gold supports (UltraAuFoil R1.2/1.3) grids were used for preparing all the samples (C. J. Russo and Passmore 2014). Only protein complexes freshly eluted from size exclusion column were vitrified on cryo-EM grids. No frozen materials were used. Aliquots of 3 µl of sample were applied onto glow-discharged gold grids. There was no specified waiting time between sample application and blotting. Blotting was done for 3 s with a blot force of -10 and using filter papers that were pre-incubated in the Vitrobot chamber 40 minutes prior to the start of the grid making session. The grids were plunge-frozen in liquid ethane using an FEI Vitrobot MKIII, at 100% humidity and 4 °C. The grids were stored in liquid nitrogen until further use. Preliminary screening of the sample was done in a FEI Tecnai G2-F20 using a model 626 70° tilt cryo-holder at liquid nitrogen temperatures. During screening, I was looking for grids that resulted in micrographs with well-separated and homogenous particles. Replica of such grids were used in collecting larger data sets on a FEI Titan Krios.

6.7.4 Data acquisition and image processing

For the preliminary cryo-EM analysis of polymerase module (without Pap1), micrographs were collected on FEI Titan Krios microscope (MRC LMB, Krios 2) operated at 300 keV equipped with a Falcon-II direct electron detector (FEI). Data was collected at a magnification of 47000x and a calibrated pixel size of 1.77 Å. Images were recorded over 2.3 seconds with a total dose of ~ 40 e– /Å2 in linear mode. Two

data sets were collected with 722 micrographs in the first one and 651 in the second. Both the data sets were collected in the same microscope under similar imaging conditions. The data sets were processed separately and the particles were merged together after several rounds of 2D classification of the individual data sets (Figure 11). A mask diameter of 180 Å and a box size of 130 pixels were used.

The structure determination of polymerase module was performed together with Ana Casañal. For subsequent cryo-EM studies of polymerase module, a large data set consisting of 4,227 micrographs were collected on a FEI Titan Krios equipped with a K2 detector. An energy filter (Gatan Imaging Filter) with a slit width of 20 eV was used to improve the signal to noise in the final images. 3,413 micrographs were collected over three two-day session at MRC LMB (Krios 2) with a calibrated pixel size of 1.4 Å, and 814 micrographs were collected at Diamond Light Source (DLS) electron Bio-Imaging Centre (eBIC) with a calibrated pixel size of 1.27 Å. Movies (consisting of 20 frames) were recorded in superresolution mode over 16 s (0.8 s per frame). The total electron dose was 45 e– /Å^2. Images were collected over a defocus range of 1.5–3.5 μm. All data processing was performed in RELION-2 (Kimanius et al. 2016). MotionCor 2 embedded within RELION-2 was used to correct for the beam-induced motion (Zheng et al. 2016). A 3D model of full CPF from an earlier study in the lab (Casañal et al. 2017b), was low-pass filtered to 40 Å and was used as an initial model to perform 3D refinements. The two data sets were processed individually until the 3D refine step, the individual maps were compared in *Chimera* (Pettersen et al. 2004) to calculate a scaling factor (appendix 8.3). The scaling factor was then used to re-scale the micrographs collected at Diamond to 1.4 Å pixel size in RELION-2. The re-scaled data set was processed further. RELION-2 was used to perform particle polishing and per-frame B-factor weighting independently for each dataset. These shiny particles from the original LMB and re-scaled Diamond data sets were combined resulting in 460,167 particles in total. This was subjected to further data processing as described in Figure 2.13. The final polymerase module map had an overall resolution of 3.5 Å. The resolution estimation was based on the gold standard Fourier shell correlation (FSC) at 0.143. The FSC is obtained from comparisons between two independently refined half-sets. Local resolution map was estimated in RELION-2. It is worth noting that a small data set using Quantifoil R1.2/1.3 grids coated with graphene oxide was also collected in addition to the above two data sets. It was rationalized that the use of such supports

would enable us to get more rare views in the 3D reconstruction. However it did not result in additional views.

For the sample of polymerase module containing Pap1, and cross-linked polymerase module containing two copies of Pap1, data was collected on a FEI Titan Krios microscope operated at 300 keV with a Falcon-III direct electron detector (FEI).

A total of 852 micrographs containing polymerase module with Pap1 were acquired at a magnification of 47000x corresponding to a calibrated pixel size of 1.77 Å, and in linear mode. The total electron dose was ~ 35 e– /Å2. 628 micrographs were chosen for further data processing after manually removing the bad micrographs after CTF estimation. Gctf embedded in RELION-2 was used in the CTF estimation (Zhang 2016). A mask diameter of 200 Å and a box size of 140 pixels were used in the data processing. A total of 216,375 particles were extracted and subjected to 2D classification.

A total of 116 micrographs of cross-linked polymerase module containing two copies of Pap1 were collected at a magnification of 59000x corresponding to a calibrated pixel size of 1.33 Å. Movies (consisting of 75 frames) with 60 s exposure time and a total electron dose of ~ 25 e– /Å2 were recorded while the camera was operating in counting mode. 50,471 particles were extracted from the micrographs and subjected to reference free 2-D classification with a mask diameter of 220 Å and a box size of 150 pixels.

For the sample of the ternary complex, 644 micrographs were collected on a FEI Titan Krios microscope (MRC LMB, Krios 1) operated at 300 keV with a Falcon-III direct electron detector (FEI). Images were collected at a magnification of 59000x corresponding to a calibrated pixel size of 1.37 Å and in linear mode with a total electron dose of ~ 52 e– /Å2. 132,717 particles were extracted from the micrographs and subjected to reference free 2-D classification with a mask diameter of 260 Å and a box size of 220 pixels.

6.7.5 Model building and refinement of Cft1-Pfs2-Yth1 subunits

Homology models of Cft1 and Pfs2 from *Phyre2* (Kelley et al. 2015) along with the crystal structure of DDB1 and DDB2 subunits (PDB 3EI3) were used in building Cft1

and Pfs2 atomic models. In total, Cft1 and Pfs2 homology models contained four beta-propellers. Each beta-propeller domains were fitted into the map and refined using jiggle fit and morphing in Coot (Emsley et al. 2010). At this point, it was uncertain where Yth1 and Fip1 residues could fit into the map. The shape of the remaining un-modeled density resembled that of a zinc finger domain. The ZnF2 from the crystal structure of the ZnF 2 and 3 of CPSF30 (PDB 2RHK) was used in molecular replacement. This resulted in assignment of the un-modeled density to Yth1 ZnF 1 and 2. However, there was no additional density where Fip1 residues could be modeled. The final model consisting of Cft1-Pfs2-Yth1 was refined with Refmac v.5.8 (Nicholls, Long, and Murshudov 2012; Brown et al. 2015) and phenix.refine (Afonine et al. 2012). The model was validated with MolProbity (V. B. Chen et al. 2010).

6.8 Biophysical techniques

6.8.1 Protein and nucleic acid quantification

UV spectrometry on a NanoDrop ND- 1000 spectrophotometer (Thermo Scientific) was used to measure protein and nucleic acid concentrations. Absorbance at 280 nm for proteins and 260 nm for nucleic acids were used to estimate concentrations along with the extinction coefficients presented in the following table:

Protein	**Extinction coefficient (M^{-1} cm-1)**	**Molecular mass (kDa)**
Pap1	63830	65
Polymerase module+0Pap1	245700	273
Polymerase module+1Pap1	309530	337
Polymerase module+2Pap1	373360	402
Polymerase module+1Cft2	330960	369
CF IA	417710	348
Rna14/Rna15	291080	225
Pcf11/Clp1	126630	122
Core-CPF	477070	571
Full CPF	676360	860
Ysh1/Mpe1	82280	137

Table 6.23: Extinction coefficients of protein complexes studied in this dissertation

RNA	Extinction coefficient (mM^{-1} cm^{-1})	Length (nucleotides)
CYC1 3′-UTR	3272	259
5′-*CYC1*-pc	2295	181
5′-*CYC1*-pc-A_{30}	2757	211
5′-UTR *MFA2* 3′-UTR	5310	417
GCN4 3′-UTR	*3804*	300

Table 6.24: Extinction coefficients of RNA used in this dissertation

6.8.2 Tandem mass spectrometry

The protein complex being studied is analyzed by SDS-PAGE and the individual subunits are separated. The gel is stained using InstantBlue™ (Expedion) for 30 minutes. The gel was placed on a sterile Saran wrap (pre-cleaned with 70% ethanol), the desired band of interest was excised using fresh blades. The bands are stored in a 1.5 ml eppendorf tube with a few µl of milliQ H_2O. Farida Begum from the LMB mass spectrometry facility performed all further steps. The band was destained, reduced, alkylated and digested in-gel by trypsin. Liquid chromatography coupled with tandem mass spectrometry was performed. The acquired data was searched against a protein database using the Mascot search engine. The data was analyzed using Scaffold (Proteome software).

6.8.3 Cross-linking mass spectrometry

Gianluca Degliesposti from the LMB mass spectrometry facility performed all the cross-linking mass spectrometry experiments. The cross-linking reaction was performed using purified polymerase module (with and without Pap1) or polymerase module (with and without Rna14/Rna15) at a concentration of 1 mg/ml proteins resuspended in 50 mM HEPES pH 7.9 buffer. For the former samples, BS3 (Creative Molecules, Canada) was used at 600 µM concentration to chemically cross-link lysine, whereas the latter samples were cross-linked by N-hydroxysuccinimide (NHS) esters disuccinimidyl dibutyric urea (BuUrBu also known as DSBU). The cross-linking reaction was

performed at 45 min at 37 °C; the reaction quenched by addition of 50 mM ammonium carbonate and incubation for 15 min. The cross-linked samples were freeze-dried, reduced and alkylated. Trypsin digestion was carried out over night at 37 °C. The digests were run on a Superdex Peptide 3.2/300 (GE Healthcare) and the collected fractions were exchanged into 2% v/v acetonitrile and 2% v/v formic acid. Next, nano-scale capillary LC-MS/MS was carried out using an Ultimate U3000 HPLC (ThermoScientific Dionex, San Jose, USA). A C18 Acclaim PepMap100 nanoViper (ThermoScientific Dionex), trapped and separated the peptides. A gradient of acetonitrile was used to elute the peptides. The column outlet was connected to a hybrid dual pressure linear ion trap mass spectrometer (Orbitrap Velos, ThermoScientific) through a nano-flow electrospray ionization source. With a resolution of 30,000, data dependent analysis was carried out for the full MS spectrum and ten MS/MS. MS spectra were obtained across a m/z range of 300–2000 and a threshold energy of 35 for collision-induced dissociation.

MS data were analysed with XQuest (Rinner et al. 2008) and MeroX (Götze et al. 2015) according to the XL reagent used (BS3 --> XQuest, DSBU --> MeroX. Searches were carried out against a database containing all the polymerase module, Rna14 and Rna15 sequences plus their reversed sequences used as decoy for the calculation of false discovery rate. Each MS/MS spectra was manually inspected and validated.

6.8.4 HDX

Sarah Maslen from the LMB mass spectrometry facility performed all the HDX experiments. Deuterium exchange reactions of polymerase module (with and without Pap1), complex of polymerase module with Rna14/Rna15 were initiated by diluting the complexes in D2O (99.8% D2O ACROS, Sigma, UK) in 50 mM HEPES pH 7.9, 150 mM NaCl, 1mM TCEP to give a final D2O percentage of ~95%. Deuterium labeling was carried out at 23°C and at four time points, 0.3 s (3 s on ice), 3 s, 30 s and 300 s in triplicate. Chilled 2.4% v/v formic acid in 2 M guanidinium hydrochloride is used to quench the reaction. The samples are immediately frozen in liquid nitrogen and stored at −80°C for further analysis. The samples were thawed and proteolytic digestion was performed with pepsin. The digested peptides were separated by reverse phase HPLC (Acquity UPLC BEH C18 column 1.7 μm, 100 mm x 1 mm, Waters, UK). Peptide detection was carried out over a over a m/z of 300–2000 on a SYNAPT G2-Si HDMS

mass spectrometer (Waters, UK). The source temperature of the spectrometer was 80°C and spray voltage 2.6 kV. Spectra were collected in positive ion mode. Peptide identification was performed as described in Silva et al, 2005. DynamX software was used for peptide identification and analysis. Back-exchange of deuterium from the amide into the buffer was not corrected for. Only relative (not absolute) deuterium incorporation levels were measured. The HD exchange in a peptide could stem from either a single amide or multiple amides within the peptide.

6.8.5 NMR

Conny Yu from the NMR facility performed all the NMR experiments and the SEC-MALS. Isotopically-labeled proteins were overexpressed in M9 minimal medium containing 1.7g/L yeast nitrogen base (Sigma Y1251), 1g/L $^{15}NH_4Cl$ and for 3D-experiments additionally [^{13}C]-glucose. All experiments, including binding studies, were carried out in the same buffer (50 mM HEPES pH 7.4, 150 mM NaCl, 5% D_2O). For backbone assignment experiments, data were collected at 278 K using Bruker Avance II+ 700 MHz for optimal signal-to-noise and protein stability. Assignment was transferred to 298K for binding studies, by following the chemical shift perturbation in 5K intervals.

^{1}H-^{15}N BEST–TROSY (Band Selective Excitation short Transients Transverse Relaxation Optimized SpectroscopY) experiments were collected with an in-house optimized pulse sequence (Favier and Brutscher 2011; Schanda, Van Melckebeke, and Brutscher 2006).

Assignment of backbone amide peaks was completed using standard triple resonance spectra: HNCO, HN(CA)CO, HNCA, HN(CO)CA, HNCACB and CBCACONH. An HN(COCA)NNH spectra was collected to confirm backbone assignment. Backbone datasets were acquired using non-uniform sampling protocols and processed with qMDD using compressed sensing reconstruction (Kazimierczuk and Orekhov 2011).

All spectra were processed using *Topspin 3.5* and analyzed with *NMRFAM-Sparky 1.412*, with assignment aided by *MARS* (Jung and Zweckstetter 2004).

Weighted chemical shift perturbations were calculated using the equation $\sqrt{(\Delta\delta^{1}H)^{2}+(0.2(\Delta\delta^{15}N)^{2})}$ with $\Delta\delta^{1}H$ and $\Delta\delta^{15}N$ being the chemical shift differences between free and bound states. Secondary structure was derived from ^{13}C chemical shifts of Ca, Cb and Co using TALOS (Shen et al. 2009; Shen and Bax 2013).

6.8.6 SEC MALS

100 µl of Yth1:Fip1 complex at 4 mg/ml was injected onto a Superdex 75 increase 10/300gl (GE) column. The run was carried at 0.5 ml/min. An Agilent 1200 series LC system with an on-line Dawn Helios ii (Wyatt), a QELS+ module (Wyatt) and an Optilab rEX differential refractive index detector (Wyatt) was used. QELS module was used to replace the detector 12 in the light scattering cell.

6.9 Bioinformatics

6.9.1 Sequence analysis

Balaji Santhanam from the Babu lab performed the sequence analysis of the interaction surface between Pfs2 and Cft1. Orthologs of Pfs2 and Cft1 were identified by *PSI-BLAST* and *JackHMMER* (Johnson, Eddy, and Portugaly 2010). *MSAProbs* (Liu, Schmidt, and Maskell 2010) was used to make multiple sequence alignments that were then refined based on *PSI-BLAST* and *JackHMMER* profiles and *DSSP* (Kabsch and Sander 1983; Touw et al. 2015) secondary structures assignments. The equivalent interaction residues between Cft1 and pfs2 orthologs in humans were identified based on the multiples sequence alignment file. The BLOSUM6 scores for the interaction mediating residues in yeast and humans were analysed. Any score above zero was regarded as conserved. The interaction was visualized using *Cytoscape* (Cline et al. 2007; Shannon et al. 2003).

6.9.2 Structural analysis

Electrostatic surface potentials were calculated by solving the Poisson– Boltzmann equation at pH 7. The APBS plugin in PyMOL was used for the calculation (Baker et al. 2001). Hydrophobic surface representation was performed using the color_h.py script obtained from the Laboratory of Supramolecular Crystallography, Institute for Protein Research, Osaka University.

All solvent atoms and ligands were removed from Pap1 and Cft1 structures (PDB: 1FAO, 6EOJ). Docking was performed with HADDOCK 2.2 (van Zundert et al. 2016). For preparation of ambiguous interactions restraints, the Cft1 residues K609, K1330, K1325 and Pap1 residues K 183 and K536 were specified as active. For both molecules, passive residues were defined automatically within a 6.5 Å radius of active residues.

Structures used in Figures were rendered in *PyMOL* (Schrödinger LLC). Figures of EM maps were rendered in *Chimera* (Pettersen et al. 2004).

7 References

Abe, A, Y Hiraoka, and T Fukasawa. 1990. “Signal Sequence for Generation of MRNA 3’ End in the Saccharomyces Cerevisiae GAL7 Gene.” *The EMBO Journal* 9 (11): 3691–97.

Afonine, Pavel V., Ralf W. Grosse-Kunstleve, Nathaniel Echols, Jeffrey J. Headd, Nigel W. Moriarty, Marat Mustyakimov, Thomas C. Terwilliger, Alexandre Urzhumtsev, Peter H. Zwart, and Paul D. Adams. 2012. “Towards Automated Crystallographic Structure Refinement with *Phenix.Refine*.” *Acta Crystallographica Section D Biological Crystallography* 68 (4): 352–67. https://doi.org/10.1107/S0907444912001308.

Agirrezabala, Xabier, and Mikel Valle. 2015. “Structural Insights into TRNA Dynamics on the Ribosome.” *International Journal of Molecular Sciences* 16 (5): 9866–95. https://doi.org/10.3390/ijms16059866.

Aibara, Shintaro, James M. B. Gordon, Anja S. Riesterer, Stephen H. McLaughlin, and Murray Stewart. 2017. “Structural Basis for the Dimerization of Nab2 Generated by RNA Binding Provides Insight into Its Contribution to Both Poly(A) Tail Length Determination and Transcript Compaction in Saccharomyces Cerevisiae.” *Nucleic Acids Research* 45 (3): 1529–38. https://doi.org/10.1093/nar/gkw1224.

Amrani, N, M Minet, M Le Gouar, F Lacroute, and F Wyers. 1997. “Yeast Pab1 Interacts with Rna15 and Participates in the Control of the Poly(A) Tail Length in Vitro.” *Molecular and Cellular Biology* 17 (7): 3694–3701.

Amrani, N, M Minet, F Wyers, M E Dufour, L P Aggerbeck, and F Lacroute. 1997. “PCF11 Encodes a Third Protein Component of Yeast Cleavage and Polyadenylation Factor I.” *Molecular and Cellular Biology* 17 (3): 1102–9.

Anderson, J T, S M Wilson, K V Datar, and M S Swanson. 1993. “NAB2: A Yeast Nuclear Polyadenylated RNA-Binding Protein Essential for Cell Viability.” *Molecular and Cellular Biology* 13 (5): 2730–41.

Baejen, Carlo, Phillipp Torkler, Saskia Gressel, Katharina Essig, Johannes Söding, and Patrick Cramer. 2014. “Transcriptome Maps of MRNP Biogenesis Factors Define Pre-MRNA Recognition.” *Molecular Cell* 55 (5): 745–57. https://doi.org/10.1016/J.MOLCEL.2014.08.005.

Bai, Yun, Thierry C Auperin, Chi-Yuan Chou, Gu-Gang Chang, James L Manley, and Liang Tong. 2007. “Crystal Structure of Murine CstF-77: Dimeric Association and Implications for Polyadenylation of MRNA Precursors.” *Molecular Cell* 25 (6):

863–75. https://doi.org/10.1016/j.molcel.2007.01.034.

Baker, N A, D Sept, S Joseph, M J Holst, and J A McCammon. 2001. "Electrostatics of Nanosystems: Application to Microtubules and the Ribosome." *Proceedings of the National Academy of Sciences of the United States of America* 98 (18): 10037–41. https://doi.org/10.1073/pnas.181342398.

Balbo, Paul B., and Andrew Bohm. 2007. "Mechanism of Poly(A) Polymerase: Structure of the Enzyme-MgATP-RNA Ternary Complex and Kinetic Analysis." *Structure* 15 (9): 1117–31. https://doi.org/10.1016/j.str.2007.07.010.

Bandaranayake, Ashok D., and Steven C. Almo. 2014. "Recent Advances in Mammalian Protein Production." *FEBS Letters* 588 (2): 253–60. https://doi.org/10.1016/j.febslet.2013.11.035.

Barabino, W Hübner, A Jenny, L Minvielle-Sebastia, and W Keller. 1997. "The 30-KD Subunit of Mammalian Cleavage and Polyadenylation Specificity Factor and Its Yeast Homolog Are RNA-Binding Zinc Finger Proteins." *Genes & Development* 11 (13): 1703–16. https://doi.org/10.1101/GAD.11.13.1703.

Barabino, Martin Ohnacker, and Walter Keller. 2000a. "Distinct Roles of Two Yth1p Domains in 3′-End Cleavage and Polyadenylation of Yeast Pre-MRNAs." *The EMBO Journal* 19 (14): 3778–87. https://doi.org/10.1093/emboj/19.14.3778.

Barabino, Silvia M.L., Martin Ohnacker, and Walter Keller. 2000b. "Distinct Roles of Two Yth1p Domains in 3′-End Cleavage and Polyadenylation of Yeast Pre-MRNAs." *The EMBO Journal* 19 (14): 3778–87. https://doi.org/10.1093/emboj/19.14.3778.

Bard, J, A M Zhelkovsky, S Helmling, T N Earnest, C L Moore, and A Bohm. 2000. "Structure of Yeast Poly(A) Polymerase Alone and in Complex with 3'-DATP." *Science (New York, N.Y.)* 289 (5483): 1346–49.

Barilla, D., B A Lee, and N J Proudfoot. 2001. "Cleavage/Polyadenylation Factor IA Associates with the Carboxyl-Terminal Domain of RNA Polymerase II in Saccharomycescerevisiae." *Proceedings of the National Academy of Sciences* 98 (2): 445–50. https://doi.org/10.1073/pnas.021545298.

Barnwal, Ravi Pratap, Susan D Lee, Claire Moore, and Gabriele Varani. 2012. "Structural and Biochemical Analysis of the Assembly and Function of the Yeast Pre-MRNA 3' End Processing Complex CF I." *Proceedings of the National Academy of Sciences of the United States of America* 109 (52): 21342–47. https://doi.org/10.1073/pnas.1214102110.

Beaudoing, E, S Freier, J R Wyatt, J M Claverie, and D Gautheret. 2000. "Patterns of

Variant Polyadenylation Signal Usage in Human Genes." *Genome Research* 10 (7): 1001–10.

Beilharz, T. H., and T. Preiss. 2007. "Widespread Use of Poly(A) Tail Length Control to Accentuate Expression of the Yeast Transcriptome." *RNA* 13 (7): 982–97. https://doi.org/10.1261/rna.569407.

Bennetzen, J L, and B D Hall. 1982. "The Primary Structure of the Saccharomyces Cerevisiae Gene for Alcohol Dehydrogenase." *The Journal of Biological Chemistry* 257 (6): 3018–25.

Berkovits, Binyamin D., and Christine Mayr. 2015. "Alternative 3′ UTRs Act as Scaffolds to Regulate Membrane Protein Localization." *Nature* 522 (7556): 363–67. https://doi.org/10.1038/nature14321.

Bieniossek, Christoph, Tsuyoshi Imasaki, Yuichiro Takagi, and Imre Berger. 2012. "MultiBac: Expanding the Research Toolbox for Multiprotein Complexes." *Trends in Biochemical Sciences* 37 (2): 49–57. https://doi.org/10.1016/j.tibs.2011.10.005.

Bienroth, S, W Keller, and E Wahle. 1993. "Assembly of a Processive Messenger RNA Polyadenylation Complex." *The EMBO Journal* 12 (2): 585–94.

Brimacombe, Richard, and Wolfgang Stiege. 1985. "ARTICLE Structure and Function of Ribosomal RNA." *Biochem. J.* Vol. 229.

Brown, Alan, Fei Long, Robert A. Nicholls, Jaan Toots, Paul Emsley, and Garib Murshudov. 2015. "Tools for Macromolecular Model Building and Refinement into Electron Cryo-Microscopy Reconstructions." *Acta Crystallographica Section D Biological Crystallography* 71 (1): 136–53. https://doi.org/10.1107/S1399004714021683.

Brune, C., Sarah E Munchel, Nicole Fischer, Alexandre V Podtelejnikov, and Karsten Weis. 2005. "Yeast Poly(A)-Binding Protein Pab1 Shuttles between the Nucleus and the Cytoplasm and Functions in MRNA Export." *RNA* 11 (4): 517–31. https://doi.org/10.1261/rna.7291205.

Bucheli, M. E., X. He, C. D. Kaplan, C. L. Moore, and S. Buratowski. 2007. "Polyadenylation Site Choice in Yeast Is Affected by Competition between Npl3 and Polyadenylation Factor CFI." *RNA* 13 (10): 1756–64. https://doi.org/10.1261/rna.607207.

Butler, J S, and T Platt. 1988. "RNA Processing Generates the Mature 3' End of Yeast CYC1 Messenger RNA in Vitro." *Science (New York, N.Y.)* 242 (4883): 1270–74.

Casañal, A. Kumar, C.H. Hill, A.D. Easter, P. Emsley, G. Degliesposti, Y. Gordiyenko, et al. 2017a. "Architecture of Eukaryotic MRNA 3′-End Processing Machinery."

Science 358 (6366). https://doi.org/10.1126/science.aao6535.

Casañal, Ananthanarayanan Kumar, Chris H Hill, Ashley D Easter, Paul Emsley, Gianluca Degliesposti, Yuliya Gordiyenko, et al. 2017b. "Architecture of Eukaryotic MRNA 3'-End Processing Machinery." *Science (New York, N.Y.)* 358 (6366): 1056–59. https://doi.org/10.1126/science.aao6535.

Chambers, Adam C., Mine Aksular, Leo P. Graves, Sarah L. Irons, Robert D. Possee, and Linda A. King. 2018. "Overview of the Baculovirus Expression System." In *Current Protocols in Protein Science*, 5.4.1-5.4.6. Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/cpps.47.

Chan, Serena L, Ina Huppertz, Chengguo Yao, Lingjie Weng, James J Moresco, John R Yates Iii, Jernej Ule, James L Manley, and Yongsheng Shi. 2014. "CPSF30 and Wdr33 Directly Bind to AAUAAA in Mammalian MRNA 39 Processing." https://doi.org/10.1101/gad.250993.114.

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

Chen, and Hyman. 1998. "A Specific RNA-Protein Interaction at Yeast Polyadenylation Efficiency Elements." *Nucleic Acids Research* 26 (21): 4965–74.

Chen, Jie, and Claire Moore. 1992. "Separation of Factors Required for Cleavage and Polyadenylation of Yeast Pre-MRNA." *Molecular and Cellular Biology*.

Chen, Vincent B., W. Bryan Arendall, Jeffrey J. Headd, Daniel A. Keedy, Robert M. Immormino, Gary J. Kapral, Laura W. Murray, Jane S. Richardson, and David C. Richardson. 2010. "*MolProbity* : All-Atom Structure Validation for Macromolecular Crystallography." *Acta Crystallographica Section D Biological Crystallography* 66 (1): 12–21. https://doi.org/10.1107/S0907444909042073.

Cheng, Hailing, Xiaoyuan He, and Claire Moore. 2004. "The Essential WD Repeat Protein Swd2 Has Dual Functions in RNA Polymerase II Transcription Termination and Lysine 4 Methylation of Histone H3." *Molecular and Cellular Biology* 24 (7): 2932–43.

Christofori, Gerhard, and Walter Keller. 1988. "3′ Cleavage and Polyadenylation of MRNA Precursors in Vitro Requires a Poly(A) Polymerase, a Cleavage Factor, and a SnRNP." *Cell* 54 (6): 875–89. https://doi.org/10.1016/S0092-8674(88)91263-9.

Clerici, Marcello, Marco Faini, Ruedi Aebersold, and Martin Jinek. 2017. “Structural Insights into the Assembly and PolyA Signal Recognition Mechanism of the Human CPSF Complex.” *ELife* 6 (December). https://doi.org/10.7554/eLife.33111.

Clerici, Marcello, Marco Faini, Lena M. Muckenfuss, Ruedi Aebersold, and Martin Jinek. 2018. “Structural Basis of AAUAAA Polyadenylation Signal Recognition by the Human CPSF Complex.” *Nature Structural & Molecular Biology* 25 (2): 135–38. https://doi.org/10.1038/s41594-017-0020-6.

Cline, Melissa S, Michael Smoot, Ethan Cerami, Allan Kuchinsky, Nerius Landys, Chris Workman, Rowan Christmas, et al. 2007. “Integration of Biological Networks and Gene Expression Data Using Cytoscape.” *Nature Protocols* 2 (10): 2366–82. https://doi.org/10.1038/nprot.2007.324.

Cretu, Constantin, Jana Schmitzová, Almudena Ponce-Salvatierra, Olexandr Dybkov, Evelina I. De Laurentiis, Kundan Sharma, Cindy L. Will, Henning Urlaub, Reinhard Lührmann, and Vladimir Pena. 2016. “Molecular Architecture of SF3b and Structural Consequences of Its Cancer-Related Mutations.” *Molecular Cell* 64 (2): 307–19. https://doi.org/10.1016/J.MOLCEL.2016.08.036.

Curinha, Ana, Sandra Oliveira Braz, Isabel Pereira-Castro, Andrea Cruz, and Alexandra Moreira. 2014. “Implications of Polyadenylation in Health and Disease.” *Nucleus (Austin, Tex.)* 5 (6): 508–19. https://doi.org/10.4161/nucl.36360.

Cyril Dominguez, And Rolf Boelens, and Alexandre M. J. J. Bonvin. 2003. “HADDOCK: A Protein−Protein Docking Approach Based on Biochemical or Biophysical Information.” https://doi.org/10.1021/JA026939X.

Das, Kalyan, Li-Chung Ma, Rong Xiao, Brian Radvansky, James Aramini, Li Zhao, Jesper Marklund, et al. 2008. “Structural Basis for Suppression of a Host Antiviral Response by Influenza A Virus.” *Proceedings of the National Academy of Sciences of the United States of America* 105 (35): 13093–98. https://doi.org/10.1073/pnas.0805213105.

Derti, Adnan, Philip Garrett-Engele, Kenzie D. MacIsaac, Richard C. Stevens, Shreedharan Sriram, Ronghua Chen, Carol A. Rohl, Jason M. Johnson, and Tomas Babak. 2012. “A Quantitative Atlas of Polyadenylation in Five Mammals.” *Genome Research* 22 (6): 1173–83. https://doi.org/10.1101/gr.132563.111.

Dheur, Sonia, Keith R Nykamp, Nicolas Viphakone, Maurice S Swanson, and Lionel Minvielle-Sebastia. 2005. “Yeast MRNA Poly(A) Tail Length Control Can Be Reconstituted in Vitro in the Absence of Pab1p-Dependent Poly(A) Nuclease Activity.” *The Journal of Biological Chemistry* 280 (26): 24532–38.

https://doi.org/10.1074/jbc.M504720200.

Dheur, Sonia, Le Thuy Anh Vo, Florence Voisinet-Hakil, Michèle Minet, Jean-Marie Schmitter, François Lacroute, Françoise Wyers, and Lionel Minvielle-Sebastia. 2003. "Pti1p and Ref2p Found in Association with the MRNA 3' End Formation Complex Direct SnoRNA Maturation." *The EMBO Journal* 22 (11): 2831–40. https://doi.org/10.1093/emboj/cdg253.

Dichtl, B, and W Keller. 2001. "Recognition of Polyadenylation Sites in Yeast Pre-MRNAs by Cleavage and Polyadenylation Factor." *The EMBO Journal* 20 (12): 3197–3209. https://doi.org/10.1093/emboj/20.12.3197.

Dichtl, Diana Blank, Martin Sadowski, Wolfgang Hübner, Stefan Weiser, and Walter Keller. 2002. "Yhh1p/Cft1p Directly Links Poly(A) Site Recognition and RNA Polymerase II Transcription Termination." *The EMBO Journal* 21 (15): 4125–35.

Doktycz, Mitchel J, Frank W Larimer, Miro Pastrnak, and Audrey Stevens. 1998. "This Contribution Is Part of the Special Series of Inaugural Articles by Members of the Comparative Analyses of the Secondary Structures of Synthetic and Intracellular Yeast MFA2 MRNAs." *Biochemistry National Academy of Sciences*. Vol. 95.

Easter, Ashley. 2014. "The Structure and Function of Cleavage and Polyadenylation Factor from Saccharomyces Cerevisiae." *Doctoral Thesis*, no. May.

Edmonds, M. 1990. "Polyadenylate Polymerases." *Methods in Enzymology* 181: 161–70.

Edmonds, M, M H Vaughan, and H Nakazato. 1971. "Polyadenylic Acid Sequences in the Heterogeneous Nuclear RNA and Rapidly-Labeled Polyribosomal RNA of HeLa Cells: Possible Evidence for a Precursor Relationship." *Proceedings of the National Academy of Sciences of the United States of America* 68 (6): 1336–40.

Edmonds, Mary, and Richard Abrams. 1960. "Polynucleotide Biosynthesis: Formation of a Sequence Adenylate Units from Adenosine Triphosphate by an Enzyme from Thymus Nuclei." *The Journal of Biological Chemistry*. Vol. 235.

Edmonds, Mary, and Mary Ann Winters. 1976. "Polyadenylate Polymerases." In , 149–79. https://doi.org/10.1016/S0079-6603(08)60069-0.

Edmund P. Walsh, Douglas J. Lamont, and Kenneth A. Beattie, and Michael J. R. Stark*. 2002. "Novel Interactions of Saccharomyces Cerevisiae Type 1 Protein Phosphatase Identified by Single-Step Affinity Purification and Mass Spectrometry†." https://doi.org/10.1021/BI015815E.

Egli, Christoph M, Christoph Springer, and Gerhard H Braus. 1995. "A Complex Unidirectional Signal Element Mediates GCN4 MRNA 3 End Formation in

Saccharomyces Cerevisiae." *Molecular and Cellular Biology*. Vol. 15.

Eick, Dirk, and Matthias Geyer. 2013. "The RNA Polymerase II Carboxy-Terminal Domain (CTD) Code." *Chemical Reviews* 113 (11): 8456–90. https://doi.org/10.1021/cr400071f.

Emsley, P., B. Lohkamp, W. G. Scott, and K. Cowtan. 2010. "Features and Development of *Coot*." *Acta Crystallographica Section D Biological Crystallography* 66 (4): 486–501. https://doi.org/10.1107/S0907444910007493.

Ezeokonkwo, Chukwudi, Mohamed A Ghazy, Alexander Zhelkovsky, Pei-Chun Yeh, and Claire Moore. 2012. "Novel Interactions at the Essential N-Terminus of Poly(A) Polymerase That Could Regulate Poly(A) Addition in Saccharomyces Cerevisiae." https://doi.org/10.1016/j.febslet.2012.03.036.

Ezeokonkwo, Chukwudi, Alexander Zhelkovsky, Rosanna Lee, Andrew Bohm, and Claire L Moore. 2011. "A Flexible Linker Region in Fip1 Is Needed for Efficient MRNA Polyadenylation." https://doi.org/10.1261/rna.2273111.

Fasken, Milo B., Murray Stewart, and Anita H. Corbett. 2008. "Functional Significance of the Interaction between the MRNA-Binding Protein, Nab2, and the Nuclear Pore-Associated Protein, Mlp1, in MRNA Export." *Journal of Biological Chemistry* 283 (40): 27130–43. https://doi.org/10.1074/jbc.M803649200.

Favier, Adrien, and Bernhard Brutscher. 2011. "Recovering Lost Magnetization: Polarization Enhancement in Biomolecular NMR." *Journal of Biomolecular NMR* 49 (1): 9–15. https://doi.org/10.1007/s10858-010-9461-5.

Fitzgerald, M, and T Shenk. 1981. "The Sequence 5'-AAUAAA-3'forms Parts of the Recognition Site for Polyadenylation of Late SV40 MRNAs." *Cell* 24 (1): 251–60.

Garas, M., B. Dichtl, and W. Keller. 2008. "The Role of the Putative 3' End Processing Endonuclease Ysh1p in MRNA and SnoRNA Synthesis." *RNA* 14 (12): 2671–84. https://doi.org/10.1261/rna.1293008.

Gavin, Anne-Claude, Markus Bösche, Roland Krause, Paola Grandi, Martina Marzioch, Andreas Bauer, Jörg Schultz, et al. 2002. "Functional Organization of the Yeast Proteome by Systematic Analysis of Protein Complexes." *Nature* 415 (6868): 141–47. https://doi.org/10.1038/415141a.

Giammartino, Dafne Campigli, and James L Manley. 2014. "New Links between MRNA Polyadenylation and Diverse Nuclear Pathways." *Molecules and Cells* 37 (9): 644–49. https://doi.org/10.14348/molcells.2014.0177.

Gibson, Daniel G, Lei Young, Ray-Yuan Chuang, J Craig Venter, Clyde A Hutchison, and Hamilton O Smith. 2009. "Enzymatic Assembly of DNA Molecules up to

Several Hundred Kilobases." *Nature Methods* 6 (5): 343–45. https://doi.org/10.1038/nmeth.1318.

Gordon, James M B, Sergei Shikov, Jason N Kuehner, Melissa Liriano, Eunhee Lee, Walter Stafford, Mathias Bach Poulsen, Celia Harrison, Claire Moore, and Andrew Bohm. 2011. "Reconstitution of CF IA from Overexpressed Subunits Reveals Stoichiometry and Provides Insights into Molecular Topology." https://doi.org/10.1021/bi200964p.

Götze, Michael, Jens Pettelkau, Romy Fritzsche, Christian H. Ihling, Mathias Schäfer, and Andrea Sinz. 2015. "Automated Assignment of MS/MS Cleavable Cross-Links in Protein 3D-Structure Analysis." *Journal of The American Society for Mass Spectrometry* 26 (1): 83–97. https://doi.org/10.1007/s13361-014-1001-1.

Graber, J. H., C. R. Cantor, S. C. Mohr, and T. F. Smith. 1999. "Genomic Detection of New Yeast Pre-MRNA 3'-End-Processing Signals." *Nucleic Acids Research* 27 (3): 888–94. https://doi.org/10.1093/nar/27.3.888.

Green, Deanna M, Kavita A Marfatia, Emily B Crafton, Xing Zhang, Xiaodong Cheng, and Anita H Corbett. 2002. "Nab2p Is Required for Poly(A) RNA Export in Saccharomyces Cerevisiae and Is Regulated by Arginine Methylation via Hmt1p." *The Journal of Biological Chemistry* 277 (10): 7752–60. https://doi.org/10.1074/jbc.M110053200.

Groner, N Hynes, and S Phillips. 1974. "Length Heterogeneity in the Poly (Adenylic Acid) Region of Yeast Messenger Ribonucleic Acid." *Biochemistry* 13 (26): 5378–83.

Groner, and Stephen L Phillips. 1975. "Polyadenylate Metabolism in the Nuclei and Cytoplasm of Saccharomyces Cerevisiae." *The Journal of Biological Chemistry*. Vol. 250.

Gross, S., and C. Moore. 2001. "Five Subunits Are Required for Reconstitution of the Cleavage and Polyadenylation Activities of Saccharomyces Cerevisiae Cleavage Factor I." *Proceedings of the National Academy of Sciences* 98 (11): 6080–85. https://doi.org/10.1073/pnas.101046598.

Guéguéniat, Julia, Adrien F. Dupin, Johan Stojko, Lionel Beaurepaire, Sarah Cianférani, Cameron D. Mackereth, Lionel Minvielle-Sébastia, and Sébastien Fribourg. 2017. "Distinct Roles of Pcf11 Zinc-Binding Domains in Pre-MRNA 3′-End Processing." *Nucleic Acids Research* 45 (17): 10115–31. https://doi.org/10.1093/nar/gkx674.

Gunderson, S I, K Beyer, G Martin, W Keller, W C Boelens, and L W Mattaj. 1994.

"The Human U1A SnRNP Protein Regulates Polyadenylation via a Direct Interaction with Poly(A) Polymerase." *Cell* 76 (3): 531–41.

Guo, Z, P Russo, D F Yun, J S Butler, and F Sherman. 1995. "Redundant 3' End-Forming Signals for the Yeast CYC1 MRNA." *Proceedings of the National Academy of Sciences of the United States of America* 92 (10): 4211–14.

Harrap, K.A. 1972. "The Structure of Nuclear Polyhedrosis Viruses: II. The Virus Particle." *Virology* 50 (1): 124–32. https://doi.org/10.1016/0042-6822(72)90352-2.

Hassur, Steven M., and Howard W. Whitlock. 1974. "UV Shadowing—A New and Convenient Method for the Location of Ultraviolet-Absorbing Species in Polyacrylamide Gels." *Analytical Biochemistry* 59 (1): 162–64. https://doi.org/10.1016/0003-2697(74)90020-7.

He, Xiaoyuan, Asad U Khan, Hailing Cheng, Donald L Pappas, Michael Hampsey, and Claire L Moore. 2003. "Functional Interactions between the Transcription and MRNA 3' End Processing Machineries Mediated by Ssu72 and Sub1." *Genes & Development* 17 (8): 1030–42. https://doi.org/10.1101/gad.1075203.

Hector, R. E., Keith R Nykamp, Sonia Dheur, James T Anderson, Priscilla J Non, Carl R Urbinati, Scott M Wilson, Lionel Minvielle-Sebastia, and Maurice S Swanson. 2002. "Dual Requirement for Yeast HnRNP Nab2p in MRNA Poly(A) Tail Length Control and Nuclear Export." *The EMBO Journal* 21 (7): 1800–1810. https://doi.org/10.1093/emboj/21.7.1800.

Heidmann, S, B Obermaier, K Vogel, and H Domdey. 1992. "Identification of Pre-MRNA Polyadenylation Sites in Saccharomyces Cerevisiae." *Molecular and Cellular Biology* 12 (9): 4215–29.

Heidmann, S, C Schindewolf, G Stumpf, and H Domdey. 1994. "Flexibility and Interchangeability of Polyadenylation Signals in Saccharomyces Cerevisiae." *Molecular and Cellular Biology* 14 (7): 4633–42. https://doi.org/10.1128/MCB.14.7.4633.

Helmling, Steffen, Alexander Zhelkovsky, and Claire L Moore. 2001. "Fip1 Regulates the Activity of Poly(A) Polymerase through Multiple Interactions." *MOLECULAR AND CELLULAR BIOLOGY* 21 (6): 2026–37. https://doi.org/10.1128/MCB.21.6.2026-2037.2001.

Henry, M, C Z Borland, M Bossie, and P A Silver. 1996. "Potential RNA Binding Proteins in Saccharomyces Cerevisiae Identified as Suppressors of Temperature-Sensitive Mutations in NPL3." *Genetics* 142 (1): 103–15.

Higgs, D. R., S. E. Y. Goodbourn, J. Lamb, J. B. Clegg, D. J. Weatherall, and N. J.

Proudfoot. 1983. “α-Thalassaemia Caused by a Polyadenylation Signal Mutation.” *Nature* 306 (5941): 398–400. https://doi.org/10.1038/306398a0.

Higuchi, Russell, Barbara Krummel, and Randall Saiki. 1988. “A General Method of *in Vitro* Preparation and Specific Mutagenesis of DNA Fragments: Study of Protein and DNA Interactions.” *Nucleic Acids Research* 16 (15): 7351–67. https://doi.org/10.1093/nar/16.15.7351.

Hill, Chris H, Ananthanarayanan Kumar, Mathias Girbig, Mark Skehel, and Lori A Passmore. 2019. “Activation of the Endonuclease That Defines MRNA 3 0 Ends Requires Incorporation into an 8-Subunit Core Cleavage and Polyadenylation Factor Complex Correspondence.” *Molecular Cell* 73. https://doi.org/10.1016/j.molcel.2018.12.023.

Hirose, Yutaka, and James L. Manley. 1998. “RNA Polymerase II Is an Essential MRNA Polyadenylation Factor.” *Nature* 395 (6697): 93–96. https://doi.org/10.1038/25786.

Hockert, J Andrew, Hsiang-Jui Yeh, and Clinton C MacDonald. 2010. “The Hinge Domain of the Cleavage Stimulation Factor Protein CstF-64 Is Essential for CstF-77 Interaction, Nuclear Localization, and Polyadenylation.” *The Journal of Biological Chemistry* 285 (1): 695–704. https://doi.org/10.1074/jbc.M109.061705.

Hollerer, Ina, Kerstin Grund, Matthias W Hentze, and Andreas E Kulozik. 2014. “MRNA 3’end Processing: A Tale of the Tail Reaches the Clinic.” *EMBO Molecular Medicine* 6 (1): 16–26. https://doi.org/10.1002/emmm.201303300.

Hsin, J.-P., and J. L. Manley. 2012. “The RNA Polymerase II CTD Coordinates Transcription and RNA Processing.” *Genes & Development* 26 (19): 2119–37. https://doi.org/10.1101/gad.200303.112.

Hu, Jun, Carol S Lutz, Jeffrey Wilusz, and Bin Tian. 2005. “Bioinformatic Identification of Candidate Cis-Regulatory Elements Involved in Human MRNA Polyadenylation.” *RNA (New York, N.Y.)* 11 (10): 1485–93. https://doi.org/10.1261/rna.2107305.

Huez, G., G. MARBAIX, E. HUBERT, Y. CLEUTER, D. GALLWITZ, E. WEINBERG, and R. DEVOS. 1978. “Functional Stabilisation of HeLa Cell Histone Messenger RNAs Injected into Xenopus Oocytes by 3′- OH Polyadenylation.” *Nature* 271 (5645): 572–73. https://doi.org/10.1038/271572a0.

Huez, Georges, Gerard Marbaix, Evelyne Hubert, Yvette Cleuter, Madeleine Leclercq, Hubert Chantrenne, Hermona Soreq, Uri Nudel, and Uriel Z Littauer. 1975. “Readenylation of Polyadenylate-Free Globin Messenger RNA Restores Its

Stability in Vivo." *Eur. J. Biochem*. Vol. 59.

Humphrey, Tim, Gerhard Christofori2, Vlatka Lucijanic, and Walter Keller2. 1987. "Cleavage and Polyadenylation of Messenger RNA Precursors in Vitro Occurs within Large and Specific 3' Processing Complexes." *The EMBO Journal*. Vol. 6.

Hyman, Linda E, Stephanie H Seiler, John Whoriskey, and Claire L Moore. 1991. "Point Mutations Upstream of the Yeast ADH2 Poly(A) Site Significantly Reduce the Efficiency of 3'-End Formation." *Molecular and Cellular Biology*. Vol. 11.

Iglesias, N., E. Tutucci, C. Gwizdek, P. Vinciguerra, E. Von Dach, A. H. Corbett, C. Dargemont, and F. Stutz. 2010. "Ubiquitin-Mediated MRNP Dynamics and Surveillance Prior to Budding Yeast MRNA Export." *Genes & Development* 24 (17): 1927–38. https://doi.org/10.1101/gad.583310.

Irniger, S, and G H Braus. 1994. "Saturation Mutagenesis of a Polyadenylation Signal Reveals a Hexanucleotide Element Essential for MRNA 3' End Formation in Saccharomyces Cerevisiae." *Proceedings of the National Academy of Sciences of the United States of America* 91 (1): 257–61.

Irniger, S, C M Egli, and G H Braus. 1991. "Different Classes of Polyadenylation Sites in the Yeast Saccharomyces Cerevisiae." *Molecular and Cellular Biology* 11 (6): 3060–69.

Ishida, T., and K. Kinoshita. 2007. "PrDOS: Prediction of Disordered Protein Regions from Amino Acid Sequence." *Nucleic Acids Research* 35 (Web Server): W460–64. https://doi.org/10.1093/nar/gkm363.

Jin, Wenxing, Yi Wang, Chao-Pei Liu, Na Yang, Mingliang Jin, Yao Cong, Mingzhu Wang, and Rui-Ming Xu. 2016. "Structural Basis for SnRNA Recognition by the Double-WD40 Repeat Domain of Gemin5." *Genes & Development* 30 (21): 2391–2403. https://doi.org/10.1101/gad.291377.116.

Johnson, L Steven, Sean R Eddy, and Elon Portugaly. 2010. "Hidden Markov Model Speed Heuristic and Iterative HMM Search Procedure." *BMC Bioinformatics* 11 (1): 431. https://doi.org/10.1186/1471-2105-11-431.

Jung, Young-Sang, and Markus Zweckstetter. 2004. "Mars – Robust Automatic Backbone Assignment of Proteins." *Journal of Biomolecular NMR* 30 (1): 11–23. https://doi.org/10.1023/B:JNMR.0000042954.99056.ad.

Kabsch, Wolfgang, and Christian Sander. 1983. "Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features." *Biopolymers* 22 (12): 2577–2637. https://doi.org/10.1002/bip.360221211.

Kaufmann, Isabelle, Georges Martin, Arno Friedlein, Hanno Langen, and Walter Keller.

2004. “Human Fip1 Is a Subunit of CPSF That Binds to U-Rich RNA Elements and Stimulates Poly(A) Polymerase.” *The EMBO Journal* 23 (3): 616–26. https://doi.org/10.1038/sj.emboj.7600070.

Kazimierczuk, Krzysztof, and Vladislav Yu. Orekhov. 2011. “Accelerated NMR Spectroscopy by Using Compressed Sensing.” *Angewandte Chemie International Edition* 50 (24): 5556–59. https://doi.org/10.1002/anie.201100370.

Keller, W, S Bienroth, K M Lang, and G Christofori. 1991. “Cleavage and Polyadenylation Factor CPF Specifically Interacts with the Pre-MRNA 3’ Processing Signal AAUAAA.” *The EMBO Journal* 10 (13): 4241–49.

Kelley, Lawrence A, Stefans Mezulis, Christopher M Yates, Mark N Wass, and Michael J E Sternberg. 2015. “The Phyre2 Web Portal for Protein Modeling, Prediction and Analysis.” *Nature Protocols* 10 (6): 845–58. https://doi.org/10.1038/nprot.2015.053.

Kelly, Sara W. Leung, Luciano H. Apponi, Anna M. Bramley, Elizabeth J. Tran, Julia A. Chekanova, Susan R. Wente, and Anita H. Corbett. 2010. “Recognition of Polyadenosine RNA by the Zinc Finger Domain of Nuclear Poly(A) RNA-Binding Protein 2 (Nab2) Is Required for Correct MRNA 3′-End Formation.” *Journal of Biological Chemistry* 285 (34): 26022–32. https://doi.org/10.1074/jbc.M110.141127.

Kelly, S. M., S. A. Pabit, C. M. Kitchen, P. Guo, K. A. Marfatia, T. J. Murphy, A. H. Corbett, and K. M. Berland. 2007. “Recognition of Polyadenosine RNA by Zinc Finger Proteins.” *Proceedings of the National Academy of Sciences* 104 (30): 12306–11. https://doi.org/10.1073/pnas.0701244104.

Kennedy, Sarah A, Monica L Frazier, Mindy Steiniger, Ann M Mast, William F Marzluff, and Matthew R Redinbo. 2009. “Crystal Structure of the HEAT Domain from the Pre-MRNA Processing Factor Symplekin.” *Journal of Molecular Biology* 392 (1): 115–28. https://doi.org/10.1016/j.jmb.2009.06.062.

Kessler, M F Henry, E Shen, J Zhao, S Gross, P A Silver, and C L Moore. 1997. “Hrp1, a Sequence-Specific RNA-Binding Protein That Shuttles between the Nucleus and the Cytoplasm, Is Required for MRNA 3’-End Formation in Yeast.” *Genes & Development* 11 (19): 2545–56. https://doi.org/10.1101/GAD.11.19.2545.

Kessler, M M, Jing Zhao, and Claire L Moore. 1996. “Purification of the Saccharomyces Cerevisiae Cleavage/Polyadenylation Factor I.”

Kim Guisbert, Karen S., Hao Li, and Christine Guthrie. 2006. “Alternative 3′ Pre-MRNA Processing in Saccharomyces Cerevisiae Is Modulated by Nab4/Hrp1 In

Vivo." Edited by Nick J Proudfoot. *PLoS Biology* 5 (1): e6. https://doi.org/10.1371/journal.pbio.0050006.

Kimanius, Dari, Björn O Forsberg, Sjors HW Scheres, and Erik Lindahl. 2016. "Accelerated Cryo-EM Structure Determination with Parallelisation Using GPUs in RELION-2." *ELife* 5 (November). https://doi.org/10.7554/eLife.18722.

Kiselev, N A, M B Sherman, and V L Tsuprun. 1990. "Negative Staining of Proteins." *Electron Microscopy Reviews* 3 (1): 43–72. https://doi.org/10.1016/0892-0354(90)90013-I.

Krishnamurthy, Shankarling, Xiaoyuan He, Mariela Reyes-Reyes, Claire Moore, and Michael Hampsey. 2004. "Ssu72 Is an RNA Polymerase II CTD Phosphatase." *Molecular Cell* 14 (3): 387–94. https://doi.org/10.1016/S1097-2765(04)00235-7.

Kühn, Uwe, Miriam Gündel, Anne Knoth, Yvonne Kerwitz, Sabine Rüdel, and Elmar Wahle. 2009. "Poly(A) Tail Length Is Controlled by the Nuclear Poly(A)-Binding Protein Regulating the Interaction between Poly(A) Polymerase and the Cleavage and Polyadenylation Specificity Factor." *Journal of Biological Chemistry* 284 (34): 22803–14. https://doi.org/10.1074/jbc.M109.018226.

Kyburz, Andrea, Martin Sadowski, Bernhard Dichtl, and Walter Keller. 2003. "The Role of the Yeast Cleavage and Polyadenylation Factor Subunit Ydh1p/Cft2p in Pre-MRNA 3¢-End Formation." https://doi.org/10.1093/nar/gkg478.

Laishram, Rakesh S. 2014. "Poly(A) Polymerase (PAP) Diversity in Gene Expression – Star-PAP vs Canonical PAP." *FEBS Letters* 588 (14): 2185–97. https://doi.org/10.1016/J.FEBSLET.2014.05.029.

Lambert, Nicole, Alex Robertson, Mohini Jangi, Sean McGeary, Phillip A Sharp, and Christopher B Burge. 2014. "RNA Bind-n-Seq: Quantitative Assessment of the Sequence and Structural Binding Specificity of RNA Binding Proteins." *Molecular Cell* 54 (5): 887–900. https://doi.org/10.1016/j.molcel.2014.04.016.

Lee, S. D., and C. L. Moore. 2014. "Efficient MRNA Polyadenylation Requires a Ubiquitin-Like Domain, a Zinc Knuckle, and a RING Finger Domain, All Contained in the Mpe1 Protein." *Molecular and Cellular Biology* 34 (21): 3955–67. https://doi.org/10.1128/MCB.00077-14.

Leeper, Thomas C., Xiangping Qu, Connie Lu, Claire Moore, and Gabriele Varani. 2010. "Novel Protein–Protein Contacts Facilitate MRNA 3′-Processing Signal Recognition by Rna15 and Hrp1." *Journal of Molecular Biology* 401 (3): 334–49. https://doi.org/10.1016/J.JMB.2010.06.032.

Leitner, Alexander, Thomas Walzthoeni, and Ruedi Aebersold. 2014. "Lysine-Specific

Chemical Cross-Linking of Protein Complexes and Identification of Cross-Linking Sites Using LC-MS/MS and the XQuest/XProphet Software Pipeline." *Nature Protocols* 9 (1): 120–37. https://doi.org/10.1038/nprot.2013.168.

Li, Jinyou, Qi-En Wang, Qianzheng Zhu, Mohamed A. El-Mahdy, Gulzar Wani, Mette Prætorius-Ibba, and Altaf A. Wani. 2006. "DNA Damage Binding Protein Component DDB1 Participates in Nucleotide Excision Repair through DDB2 DNA-Binding and Cullin 4A Ubiquitin Ligase Activity." *Cancer Research* 66 (17): 8590–97. https://doi.org/10.1158/0008-5472.CAN-06-1115.

Lidschreiber, Michael, Ashley D Easter, Sofia Battaglia, Juan B Rodríguez-Molina, Ana Casañal, Manuel Carminati, Carlo Baejen, et al. 2018. "The APT Complex Is Involved in Non-Coding RNA Transcription and Is Distinct from CPF." *Nucleic Acids Research* 46 (21): 11528–38. https://doi.org/10.1093/nar/gky845.

Lingner, Joachim, Isabelle Radtke, Elmar Wahles, and Walter Keller. 1991. "Purification and Characterization of Poly(A) Polymerase from Saccharomyces Cereuisiae." Vol. 266.

Liu, Y., B. Schmidt, and D. L. Maskell. 2010. "MSAProbs: Multiple Sequence Alignment Based on Pair Hidden Markov Models and Partition Function Posterior Probabilities." *Bioinformatics* 26 (16): 1958–64. https://doi.org/10.1093/bioinformatics/btq338.

Lutz, C S, and J C Alwine. 1994. "Direct Interaction of the U1 SnRNP-A Protein with the Upstream Efficiency Element of the SV40 Late Polyadenylation Signal." *Genes & Development* 8 (5): 576–86. https://doi.org/10.1101/GAD.8.5.576.

Mandel, Corey R., Syuzo Kaneko, Hailong Zhang, Damara Gebauer, Vasupradha Vethantham, James L. Manley, and Liang Tong. 2006. "Polyadenylation Factor CPSF-73 Is the Pre-MRNA 3'-End-Processing Endonuclease." *Nature* 444 (7121): 953–56. https://doi.org/10.1038/nature05363.

Manley, J L. 1983. "Accurate and Specific Polyadenylation of MRNA Precursors in a Soluble Whole-Cell Lysate." *Cell* 33 (2): 595–605. https://doi.org/10.1016/0092-8674(83)90440-3.

Manley, James L. 1988. "Polyadenylation of MRNA Precursors." *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression* 950 (1): 1–12. https://doi.org/10.1016/0167-4781(88)90067-X.

Masson, Glenn R., Meredith L. Jenkins, and John E. Burke. 2017. "An Overview of Hydrogen Deuterium Exchange Mass Spectrometry (HDX-MS) in Drug Discovery." *Expert Opinion on Drug Discovery* 12 (10): 981–94.

https://doi.org/10.1080/17460441.2017.1363734.

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

Mcdevitt, Michael A, Gregory M Gilmartin/'^, Westley H Reeves, and Joseph R Nevins. 1988. "Multiple Factors Are Required for Poly(A) Addition to a MRNA 3' End."

Mclaughlin, Calvin S, Jonathan R Warner, Mary Edmonds, Hiroshi Nakazato, and Maurice H Vaughan. 1973. "Polyadenylic Acid Sequences in Yeast Messenger Ribonucleic Acid." *Journal of Biological Chemistry*. Vol. 248.

McMullan, G., A.R. Faruqi, D. Clare, and R. Henderson. 2014. "Comparison of Optimal Performance at 300keV of Three Direct Electron Detectors for Use in Low Dose Electron Microscopy." *Ultramicroscopy* 147 (December): 156–63. https://doi.org/10.1016/j.ultramic.2014.08.002.

Meinke, Gretchen, Chukwudi Ezeokonkwo, Paul Balbo, Walter Stafford, Claire Moore, and Andrew Bohm. 2008. "Structure of Yeast Poly(A) Polymerase in Complex with a Peptide from Fip1, an Intrinsically Disordered Protein." *Biochemistry* 47 (26): 6859–69. https://doi.org/10.1021/bi800204k.

Millevoi, Stefania, Adrien Decorsière, Clarisse Loulergue, Jason Iacovoni, Sandra Bernat, Michael Antoniou, and Stéphan Vagner. 2009. "A Physical and Functional Link between Splicing Factors Promotes Pre-MRNA 3' End Processing." *Nucleic Acids Research* 37 (14): 4672–83. https://doi.org/10.1093/nar/gkp470.

Minvielle-Sebastia, L., K Beyer, A M Krecic, R E Hector, M S Swanson, and W Keller. 1998. "Control of Cleavage Site Selection during MRNA 3' End Formation by a Yeast HnRNP." *The EMBO Journal* 17 (24): 7454–68. https://doi.org/10.1093/emboj/17.24.7454.

Minvielle-Sebastia, L, P J Preker, and W Keller. 1994. "RNA14 and RNA15 Proteins as Components of a Yeast Pre-MRNA 3'-End Processing Factor." *Science (New York, N.Y.)* 266 (5191): 1702–5.

Minvielle-Sebastia, L, P J Preker, T Wiederkehr, Y Strahm, and W Keller. 1997. "The Major Yeast Poly(A)-Binding Protein Is Associated with Cleavage Factor IA and Functions in Premessenger RNA 3'-End Formation." *Proceedings of the National Academy of Sciences of the United States of America* 94 (15): 7897–7902. https://doi.org/10.1073/PNAS.94.15.7897.

Minvielle-Sebastia, L, B Winsor, N Bonneaud, and F Lacroute. 1991. “Mutations in the Yeast RNA14 and RNA15 Genes Result in an Abnormal MRNA Decay Rate; Sequence Analysis Reveals an RNA-Binding Domain in the RNA15 Protein.” *Molecular and Cellular Biology* 11 (6): 3075–87.

Mizrahi, N, and C Moore. 2000. “Posttranslational Phosphorylation and Ubiquitination of the Saccharomyces Cerevisiae Poly(A) Polymerase at the S/G(2) Stage of the Cell Cycle.” *Molecular and Cellular Biology* 20 (8): 2794–2802.

Moore, C L, and P A Sharp. 1984. “Site-Specific Polyadenylation in a Cell-Free Reaction.” *Cell* 36 (3): 581–91.

Moore, C L, H Skolnik-David, and P A Sharp. 1986. “Analysis of RNA Cleavage at the Adenovirus-2 L3 Polyadenylation Site.” *The EMBO Journal* 5 (8): 1929–38.

Moore, Claire L, and Phillip A. Sharp. 1985. “Accurate Cleavage and Polyadenylation of Exogenous RNA Substrate.” *Cell* 41 (3): 845–55. https://doi.org/10.1016/S0092-8674(85)80065-9.

Moreno-Morcillo, Maria, Lionel Minvielle-Sébastia, Sébastien Fribourg, and Cameron D. Mackereth. 2011. “Locked Tether Formation by Cooperative Folding of Rna14p Monkeytail and Rna15p Hinge Domains in the Yeast CF IA Complex.” *Structure* 19 (4): 534–45. https://doi.org/10.1016/J.STR.2011.02.003.

Müller, Mathias Q., Frank Dreiocker, Christian H. Ihling, Mathias Schäfer, and Andrea Sinz. 2010. “Cleavable Cross-Linker for Protein Structure Analysis: Reliable Identification of Cross-Linking Products by Tandem MS.” *Analytical Chemistry* 82 (16): 6958–68. https://doi.org/10.1021/ac101241t.

Murthy, K G, and J L Manley. 1992. “Characterization of the Multisubunit Cleavage-Polyadenylation Specificity Factor from Calf Thymus.” *The Journal of Biological Chemistry* 267 (21): 14804–11.

Nedea, Eduard, Xiaoyuan He, Minkyu Kim, Jeff Pootoolal, Guoqing Zhong, Veronica Canadien, Timothy Hughes, Stephen Buratowski, Claire L Moore, and Jack Greenblatt. 2003. “Organization and Function of APT, a Subcomplex of the Yeast Cleavage and Polyadenylation Factor Involved in the Formation of MRNA and Small Nucleolar RNA 3’-Ends.” *The Journal of Biological Chemistry* 278 (35): 33000–10. https://doi.org/10.1074/jbc.M304454200.

Nevins, Joseph R, and James E Darnell. 1978. “Steps in the Processing of Ad2 MRNA: Poly(A)+ Nuclear Sequences Are Conserved and Poly(A) Addition Precedes Splicing.” *Cell*. Vol. 15.

Nicholls, Robert A., Fei Long, and Garib N. Murshudov. 2012. “Low-Resolution

Refinement Tools in *REFMAC* 5." *Acta Crystallographica Section D Biological Crystallography* 68 (4): 404–17. https://doi.org/10.1107/S090744491105606X.

Noble, C. G., Philip A. Walker, Lesley J. Calder, and Ian A. Taylor. 2004. "Rna14-Rna15 Assembly Mediates the RNA-Binding Capability of Saccharomyces Cerevisiae Cleavage Factor IA." *Nucleic Acids Research* 32 (11): 3364–75. https://doi.org/10.1093/nar/gkh664.

O'Connor, J P, and C L Peebles. 1992. "PTA1, an Essential Gene of Saccharomyces Cerevisiae Affecting Pre-TRNA Processing." *Molecular and Cellular Biology* 12 (9): 3843–56.

O'Reilly, David R., Lois. Miller, and Verne A. Luckow. 1994. *Baculovirus Expression Vectors : A Laboratory Manual*. Oxford University Press.

Ohnacker, M., Silvia M. L. Barabino, Pascal J. Preker, and Walter Keller. 2000. "The WD-Repeat Protein Pfs2p Bridges Two Essential Factors within the Yeast Pre-MRNA 3'-End-Processing Complex." *The EMBO Journal* 19 (1): 37–47. https://doi.org/10.1093/emboj/19.1.37.

Pancevac, Christina, David C Goldstone, Andres Ramos, and Ian A Taylor. 2010. "Structure of the Rna15 RRM-RNA Complex Reveals the Molecular Basis of GU Specificity in Transcriptional 3'-End Processing Factors." *Nucleic Acids Research* 38 (9): 3119–32. https://doi.org/10.1093/nar/gkq002.

Paule, M R, and R J White. 2000. "Survey and Summary: Transcription by RNA Polymerases I and III." *Nucleic Acids Research* 28 (6): 1283–98.

Paulson, A. R., and L. Tong. 2012. "Crystal Structure of the Rna14-Rna15 Complex." *RNA* 18 (6): 1154–62. https://doi.org/10.1261/rna.032524.112.

Pérez-Cañadillas, José Manuel. 2006. "Grabbing the Message: Structural Basis of MRNA 3′UTR Recognition by Hrp1." *The EMBO Journal* 25 (13): 3167–78. https://doi.org/10.1038/sj.emboj.7601190.

Perez Canadillas, J. M., and Gabriele Varani. 2003. "Recognition of GU-Rich Polyadenylation Regulatory Elements by Human CstF-64 Protein." *The EMBO Journal* 22 (11): 2821–30. https://doi.org/10.1093/emboj/cdg259.

Pettersen, Eric F., Thomas D. Goddard, Conrad C. Huang, Gregory S. Couch, Daniel M. Greenblatt, Elaine C. Meng, and Thomas E. Ferrin. 2004. "UCSF Chimera?A Visualization System for Exploratory Research and Analysis." *Journal of Computational Chemistry* 25 (13): 1605–12. https://doi.org/10.1002/jcc.20084.

Plaschka, Clemens, Pei-Chun Lin, and Kiyoshi Nagai. 2017. "Structure of a Pre-Catalytic Spliceosome." *Nature* 546 (7660): 617–21.

https://doi.org/10.1038/nature22799.

Preiss, T, M Muckenthaler, and M W Hentze. 1998. “Poly(A)-Tail-Promoted Translation in Yeast: Implications for Translational Control.” *RNA (New York, N.Y.)* 4 (11): 1321–31.

Preker, P J, and W Keller. 1998. “The HAT Helix, a Repetitive Motif Implicated in RNA Processing.” *Trends in Biochemical Sciences* 23 (1): 15–16.

Preker, P J, J Lingner, L Minvielle-Sebastia, and W Keller. 1995. “The FIP1 Gene Encodes a Component of a Yeast Pre-MRNA Polyadenylation Factor That Directly Interacts with Poly(A) Polymerase.” *Cell* 81 (3): 379–89. https://doi.org/10.1016/0092-8674(95)90391-7.

Preker, P J, M Ohnacker, L Minvielle-Sebastia, and W Keller. 1997. “A Multisubunit 3’ End Processing Factor from Yeast Containing Poly(A) Polymerase and Homologues of the Subunits of Mammalian Cleavage and Polyadenylation Specificity Factor.” *The EMBO Journal* 16 (15): 4727–37. https://doi.org/10.1093/emboj/16.15.4727.

Proudfoot, N. J., and G. G. Brownlee. 1976. “3′ Non-Coding Region Sequences in Eukaryotic Messenger RNA.” *Nature* 263 (5574): 211–14. https://doi.org/10.1038/263211a0.

Qu, Xiangping, Jose-Manuel Perez-Canadillas, Shipra Agrawal, Julia De Baecke, Hailing Cheng, Gabriele Varani, and Claire Moore. 2007. “The C-Terminal Domains of Vertebrate CstF-64 and Its Yeast Orthologue Rna15 Form a New Structure Critical for MRNA 3′-End Processing.” *Journal of Biological Chemistry* 282 (3): 2101–15. https://doi.org/10.1074/jbc.M609981200.

Reyes-Reyes, Mariela, and Michael Hampsey. 2007. “Role for the Ssu72 C-Terminal Domain Phosphatase in RNA Polymerase II Transcription Elongation.” *Molecular and Cellular Biology* 27 (3): 926–36. https://doi.org/10.1128/MCB.01361-06.

Rinner, Oliver, Jan Seebacher, Thomas Walzthoeni, Lukas N Mueller, Martin Beck, Alexander Schmidt, Markus Mueller, and Ruedi Aebersold. 2008. “Identification of Cross-Linked Peptides from Large Sequence Databases.” *Nature Methods* 5 (4): 315–18. https://doi.org/10.1038/nmeth.1192.

Roguev, A., D Schaft, A Shevchenko, W W Pijnappel, M Wilm, R Aasland, and A F Stewart. 2001. “The Saccharomyces Cerevisiae Set1 Complex Includes an Ash2 Homologue and Methylates Histone 3 Lysine 4.” *The EMBO Journal* 20 (24): 7137–48. https://doi.org/10.1093/emboj/20.24.7137.

Rosado-Lugo, Jesús D, and Michael Hampsey. 2014. “The Ssu72 Phosphatase Mediates

the RNA Polymerase II Initiation-Elongation Transition." *The Journal of Biological Chemistry* 289 (49): 33916–26. https://doi.org/10.1074/jbc.M114.608695.

Rüegsegger, U, D Blank, and W Keller. 1998. "Human Pre-MRNA Cleavage Factor Im Is Related to Spliceosomal SR Proteins and Can Be Reconstituted in Vitro from Recombinant Subunits." *Molecular Cell* 1 (2): 243–53.

Russnak, R, K W Nehrke, and T Platt. 1995. "REF2 Encodes an RNA-Binding Protein Directly Involved in Yeast MRNA 3'-End Formation." *Molecular and Cellular Biology* 15 (3): 1689–97.

Russo, Christopher J, and Lori A Passmore. 2014. "Electron Microscopy: Ultrastable Gold Substrates for Electron Cryomicroscopy." *Science (New York, N.Y.)* 346 (6215): 1377–80. https://doi.org/10.1126/science.1259530.

Russo, P, W Z Li, D M Hampsey, K S Zaret, and F Sherman. 1991. "Distinct Cis-Acting Signals Enhance 3' Endpoint Formation of CYC1 MRNA in the Yeast Saccharomyces Cerevisiae." *The EMBO Journal* 10 (3): 563–71.

Ryan, Kevin, Olga Calvo, and James L Manley. 2004. "Evidence That Polyadenylation Factor CPSF-73 Is the MRNA 3' Processing Endonuclease." *RNA (New York, N.Y.)* 10 (4): 565–73.

Sachs, Alan B., and Ronald W. Davis. 1989. "The Poly(A) Binding Protein Is Required for Poly(A) Shortening and 60S Ribosomal Subunit-Dependent Translation Initiation." *Cell* 58 (5): 857–67. https://doi.org/10.1016/0092-8674(89)90938-0.

Sadowski, M., Bernhard Dichtl, Wolfgang Hübner, and Walter Keller. 2003. "Independent Functions of Yeast Pcf11p in Pre-MRNA 3' End Processing and in Transcription Termination." *The EMBO Journal* 22 (9): 2167–77. https://doi.org/10.1093/emboj/cdg200.

Salisbury, Jesse, Keith W Hutchison, and Joel H Graber. 2006. "A Multispecies Comparison of the Metazoan 3'-Processing Downstream Elements and the CstF-64 RNA Recognition Motif." *BMC Genomics* 7 (1): 55. https://doi.org/10.1186/1471-2164-7-55.

Schanda, Paul, Hélène Van Melckebeke, and Bernhard Brutscher. 2006. "Speeding Up Three-Dimensional Protein NMR Experiments to a Few Minutes." *Journal of the American Chemical Society* 128 (28): 9042–43. https://doi.org/10.1021/ja062025p.

Scheres, Sjors H.W. 2012. "RELION: Implementation of a Bayesian Approach to Cryo-EM Structure Determination." *Journal of Structural Biology* 180 (3): 519–30. https://doi.org/10.1016/J.JSB.2012.09.006.

Schmid, Manfred, Pawel Olszewski, Vicent Pelechano, Ishaan Gupta, Lars M. Steinmetz, and Torben Heick Jensen. 2015. "The Nuclear PolyA-Binding Protein Nab2p Is Essential for MRNA Production." *Cell Reports* 12 (1): 128–39. https://doi.org/10.1016/j.celrep.2015.06.008.

Schmid, Manfred, Mathias Bach Poulsen, Pawel Olszewski, Vicent Pelechano, Cyril Saguez, Ishaan Gupta, Lars M. Steinmetz, Claire Moore, and Torben Heick Jensen. 2012. "Rrp6p Controls MRNA Poly(A) Tail Length and Its Decoration with Poly(A) Binding Proteins." *Molecular Cell* 47 (2): 267–80. https://doi.org/10.1016/j.molcel.2012.05.005.

Schmidt, Karyn, and J. Scott Butler. 2013. "Nuclear RNA Surveillance: Role of TRAMP in Controlling Exosome Specificity." *Wiley Interdisciplinary Reviews: RNA* 4 (2): 217–31. https://doi.org/10.1002/wrna.1155.

Schneider, Caroline A, Wayne S Rasband, and Kevin W Eliceiri. 2012. "NIH Image to ImageJ: 25 Years of Image Analysis." *Nature Methods 2012 9:7*, June.

Schönemann, Lars, Uwe Kühn, Georges Martin, Peter Schäfer, Andreas R. Gruber, Walter Keller, Mihaela Zavolan, and Elmar Wahle. 2014. "Reconstitution of CPSF Active in Polyadenylation: Recognition of the Polyadenylation Signal by WDR33." *Genes & Development* 28 (21): 2381–93. https://doi.org/10.1101/gad.250985.114.

Schreieck, Amelie, Ashley D Easter, Stefanie Etzold, Katrin Wiederhold, Michael Lidschreiber, Patrick Cramer, and Lori A Passmore. 2014. "RNA Polymerase II Termination Involves CTD Tyrosine Dephosphorylation by CPF Subunit Glc7 Europe PMC Funders Group." *Nat Struct Mol Biol* 21 (2): 175–79. https://doi.org/10.1038/nsmb.2753.

Scrima, Andrea, Renata Konícková, Bryan K Czyzewski, Yusuke Kawasaki, Philip D Jeffrey, Regina Groisman, Yoshihiro Nakatani, Shigenori Iwai, Nikola P Pavletich, and Nicolas H Thomä. 2008. "Structural Basis of UV DNA-Damage Recognition by the DDB1-DDB2 Complex." *Cell* 135 (7): 1213–23. https://doi.org/10.1016/j.cell.2008.10.045.

Seckler, James M., Mary D. Barkley, and Patrick L. Wintrode. 2011. "Allosteric Suppression of HIV-1 Reverse Transcriptase Structural Dynamics upon Inhibitor Binding." *Biophysical Journal* 100 (1): 144–53. https://doi.org/10.1016/j.bpj.2010.11.004.

Shannon, Paul, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. 2003. "Cytoscape: A

Software Environment for Integrated Models of Biomolecular Interaction Networks." *Genome Research* 13 (11): 2498–2504. https://doi.org/10.1101/gr.1239303.

Shen, Yang, and Ad Bax. 2013. "Protein Backbone and Sidechain Torsion Angles Predicted from NMR Chemical Shifts Using Artificial Neural Networks." *Journal of Biomolecular NMR* 56 (3): 227. https://doi.org/10.1007/S10858-013-9741-Y.

Shen, Yang, Frank Delaglio, Gabriel Cornilescu, and Ad Bax. 2009. "TALOS+: A Hybrid Method for Predicting Protein Backbone Torsion Angles from NMR Chemical Shifts." *Journal of Biomolecular NMR* 44 (4): 213–23. https://doi.org/10.1007/s10858-009-9333-z.

Shi, Yongsheng, Dafne Campigli, Di Giammartino, Derek Taylor, Ali Sarkeshik, William J Rice, John R Yates Iii, Joachim Frank, and James L Manley. 2009. "Molecular Architecture of the Human Pre-MRNA 3′ Processing Complex." https://doi.org/10.1016/j.molcel.2008.12.028.

Shukla, Arun K., Gerwin H. Westfield, Kunhong Xiao, Rosana I. Reis, Li-Yin Huang, Prachi Tripathi-Shukla, Jiang Qian, et al. 2014. "Visualization of Arrestin Recruitment by a G-Protein-Coupled Receptor." *Nature* 512 (7513): 218–22. https://doi.org/10.1038/nature13430.

Sierra-Gallay, Léna Zig, Ailar Jamalli, and Harald Putzer. 2008. "Structural Insights into the Dual Activity of RNase J." *Nature Structural & Molecular Biology* 15 (2): 206–12. https://doi.org/10.1038/nsmb.1376.

Soucek, Sharon, Anita H. Corbett, and Milo B. Fasken. 2012. "The Long and the Short of It: The Role of the Zinc Finger Polyadenosine RNA Binding Protein, Nab2, in Control of Poly(A) Tail Length." *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* 1819 (6): 546–54. https://doi.org/10.1016/j.bbagrm.2012.03.006.

Steitz, T A, and J A Steitz. 1993. "A General Two-Metal-Ion Mechanism for Catalytic RNA." *Proceedings of the National Academy of Sciences of the United States of America* 90 (14): 6498–6502.

Studier, F W, and B A Moffatt. 1986. "Use of Bacteriophage T7 RNA Polymerase to Direct Selective High-Level Expression of Cloned Genes." *Journal of Molecular Biology* 189 (1): 113–30.

Stumpf, G, and H Domdey. 1996. "Dependence of Yeast Pre-MRNA 3'-End Processing on CFT1: A Sequence Homolog of the Mammalian AAUAAA Binding Factor." *Science (New York, N.Y.)* 274 (5292): 1517–20.

Sullivan, Kelly D, Mindy Steiniger, and William F Marzluff. 2009. “A Core Complex of CPSF73, CPSF100, and Symplekin May Form Two Different Cleavage Factors for Processing of Poly(A) and Histone MRNAs.” *Molecular Cell* 34 (3): 322–32. https://doi.org/10.1016/j.molcel.2009.04.024.

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

Tacahashi, Yoko, Steffen Helmling, and Claire L Moore. 2003. “Functional Dissection of the Zinc Fnger and Fanking Domains of the Yth1 Cleavage/Polyadenylation Factor.” https://doi.org/10.1093/nar/gkg265.

Takagaki, Y, C C MacDonald, T Shenk, and J L Manley. 1992. “The Human 64-KDa Polyadenylylation Factor Contains a Ribonucleoprotein-Type RNA Binding Domain and Unusual Auxiliary Motifs.” *Proceedings of the National Academy of Sciences of the United States of America* 89 (4): 1403–7.

Tang, Guang, Liwei Peng, Philip R. Baldwin, Deepinder S. Mann, Wen Jiang, Ian Rees, and Steven J. Ludtke. 2007. “EMAN2: An Extensible Image Processing Suite for Electron Microscopy.” *Journal of Structural Biology* 157 (1): 38–46. https://doi.org/10.1016/J.JSB.2006.05.009.

Tarun, S Z, and A B Sachs. 1995. “A Common Function for MRNA 5’ and 3’ Ends in Translation Initiation in Yeast.” *Genes & Development* 9 (23): 2997–3007.

Thimiri Govinda Raj, Deepak B., Lakshmi S. Vijayachandran, and Imre Berger. 2014. “OmniBac: Universal Multigene Transfer Plasmids for Baculovirus Expression Vector Systems.” In , 123–30. https://doi.org/10.1007/978-1-62703-691-7_7.

Tian, B., Jun Hu, Haibo Zhang, and Carol S Lutz. 2005. “A Large-Scale Analysis of MRNA Polyadenylation of Human and Mouse Genes.” *Nucleic Acids Research* 33 (1): 201–12. https://doi.org/10.1093/nar/gki158.

Tian, Bin, and Joel H Graber. 2012. “Signals for Pre-MRNA Cleavage and Polyadenylation.” *Wiley Interdisciplinary Reviews. RNA* 3 (3): 385–96. https://doi.org/10.1002/wrna.116.

Touw, Wouter G., Coos Baakman, Jon Black, Tim A. H. te Beek, E. Krieger, Robbie P. Joosten, and Gert Vriend. 2015. “A Series of PDB-Related Databanks for Everyday Needs.” *Nucleic Acids Research* 43 (D1): D364–68. https://doi.org/10.1093/nar/gku1028.

Tsiapalis, Chris M, John W Dorson, and I ? J Bollum. 1975. “Purification of Terminal Riboadenylate Transferase from Calf Thymus Gland.” *The Journal of Biological Chemistry*. Vol. 250.

Tudek, Agnieszka, Marta Lloret-Llinares, and Torben Heick Jensen. 2018. “The Multitasking PolyA Tail: Nuclear RNA Maturation, Degradation and Export.” *Philosophical Transactions of the Royal Society B: Biological Sciences* 373 (1762): 20180169. https://doi.org/10.1098/rstb.2018.0169.

Vagner, S, C Vagner, and I W Mattaj. 2000. “The Carboxyl Terminus of Vertebrate Poly(A) Polymerase Interacts with U2AF 65 to Couple 3’-End Processing and Splicing.” *Genes & Development* 14 (4): 403–13.

Vaňáčová, Štěpánka, Jeannette Wolf, Georges Martin, Diana Blank, Sabine Dettwiler, Arno Friedlein, Hanno Langen, Gérard Keith, and Walter Keller. 2005. “A New Yeast Poly(A) Polymerase Complex Involved in RNA Quality Control.” Edited by Phillip Zamore. *PLoS Biology* 3 (6): e189. https://doi.org/10.1371/journal.pbio.0030189.

Venkatesan, S, N Elango, and R M Chanock. 1983. “Construction and Characterization of CDNA Clones for Four Respiratory Syncytial Viral Genes.” *Proceedings of the National Academy of Sciences of the United States of America* 80 (5): 1280–84. https://doi.org/10.1073/PNAS.80.5.1280.

Viphakone, Nicolas, Florence Voisinet-Hakil, and Lionel Minvielle-Sebastia. 2008. “Molecular Dissection of MRNA Poly(A) Tail Length Control in Yeast.” *Nucleic Acids Research* 36 (7): 2418–33. https://doi.org/10.1093/nar/gkn080.

Vo, L. T. A., M. Minet, J.-M. Schmitter, F. Lacroute, and F. Wyers. 2001. “Mpe1, a Zinc Knuckle Protein, Is an Essential Component of Yeast Cleavage and Polyadenylation Factor Required for the Cleavage and Polyadenylation of MRNA.” *Molecular and Cellular Biology* 21 (24): 8346–56. https://doi.org/10.1128/MCB.21.24.8346-8356.2001.

Voorn, L van der, and H L Ploegh. 1992. “The WD-40 Repeat.” *FEBS Letters* 307 (2): 131–34.

Wahle, E. 1991. “Purification and Characterization of a Mammalian Polyadenylate Polymerase Involved in the 3’ End Processing of Messenger RNA Precursors.” *The Journal of Biological Chemistry* 266 (5): 3131–39.

———. 1995. “Poly(A) Tail Length Control Is Caused by Termination of Processive Synthesis.” *The Journal of Biological Chemistry* 270 (6): 2800–2808. https://doi.org/10.1074/jbc.270.6.2800.

Wales, Thomas E., and John R. Engen. 2006. “Hydrogen Exchange Mass Spectrometry for the Analysis of Protein Dynamics.” *Mass Spectrometry Reviews* 25 (1): 158–70. https://doi.org/10.1002/mas.20064.

Wang, Xiao, and Chuan He. 2014. “Dynamic RNA Modifications in Posttranscriptional Regulation.” *Molecular Cell* 56 (1): 5–12. https://doi.org/10.1016/J.MOLCEL.2014.09.001.

Weissmann, Florian, Georg Petzold, Ryan VanderLinden, Pim J Huis In ’t Veld, Nicholas G Brown, Fabienne Lampert, Stefan Westermann, Holger Stark, Brenda A Schulman, and Jan-Michael Peters. 2016. “BiGBac Enables Rapid Gene Assembly for the Expression of Large Multisubunit Protein Complexes.” *Proceedings of the National Academy of Sciences of the United States of America* 113 (19): E2564-9. https://doi.org/10.1073/pnas.1604935113.

Werner-Allen, Jon W, Chul-Jin Lee, Pengda Liu, Nathan I Nicely, Su Wang, Arno L Greenleaf, and Pei Zhou. 2011. “Cis-Proline-Mediated Ser(P)5 Dephosphorylation by the RNA Polymerase II C-Terminal Domain Phosphatase Ssu72.” *The Journal of Biological Chemistry* 286 (7): 5717–26. https://doi.org/10.1074/jbc.M110.197129.

Winters, Mary Ann, and Mary Edmonds. 1973. “A Poly (A) Polymerase from Calf Thymus.” *The Journal of Biological Chemistry*. Vol. 248.

Wu, Xuebing, and David P. Bartel. 2017. “Widespread Influence of 3′-End Structures on Mammalian MRNA Processing and Stability.” *Cell* 169 (5): 905-917.e11. https://doi.org/10.1016/j.cell.2017.04.036.

Xiang, Kehui, James L Manley, and Liang Tong. 2012. “An Unexpected Binding Mode for a Pol II CTD Peptide Phosphorylated at Ser7 in the Active Site of the CTD Phosphatase Ssu72.” *Genes & Development* 26 (20): 2265–70. https://doi.org/10.1101/gad.198853.112.

Xiang, Kehui, Takashi Nagaike, Song Xiang, Turgay Kilic, Maia M. Beh, James L. Manley, and Liang Tong. 2010. “Crystal Structure of the Human Symplekin–Ssu72–CTD Phosphopeptide Complex.” *Nature* 467 (7316): 729–33. https://doi.org/10.1038/nature09391.

Xiang, Kehui, Liang Tong, and James L. Manley. 2014. “Delineating the Structural Blueprint of the Pre-MRNA 3′-End Processing Machinery.” *Molecular and Cellular Biology* 34 (11): 1894–1910. https://doi.org/10.1128/MCB.00084-14.

Yan, Chuangye, Ruixue Wan, Rui Bai, Gaoxingyu Huang, and Yigong Shi. 2016. “Structure of a Yeast Activated Spliceosome at 3.5 Å Resolution.” *Science (New*

York, N.Y.) 353 (6302): 904–11. https://doi.org/10.1126/science.aag0291.

Yang, Qin, and Sylvie Doublié. 2011. "Structural Biology of Poly(A) Site Definition." *Wiley Interdisciplinary Reviews: RNA* 2 (5): 732–47. https://doi.org/10.1002/wrna.88.

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

Zarkower, David, and Marvin Wickens. 1987. "Formation of MRNA 3' Termini: Stability and Dissociation of a Complex Involving the AAUAAA Sequence." *The EMBO Journal*. Vol. 6.

Zhang, Kai. 2016. "Gctf: Real-Time CTF Determination and Correction." *Journal of Structural Biology* 193 (1): 1–12. https://doi.org/10.1016/j.jsb.2015.11.003.

Zhao, J, M Kessler, S Helmling, J P O'Connor, and C Moore. 1999. "Pta1, a Component of Yeast CF II, Is Required for Both Cleavage and Poly(A) Addition of MRNA Precursor." *Molecular and Cellular Biology* 19 (11): 7733–40.

Zhao, J, M M Kessler, and C L Moore. 1997. "Cleavage Factor II of Saccharomyces Cerevisiae Contains Homologues to Subunits of the Mammalian Cleavage/ Polyadenylation Specificity Factor and Exhibits Sequence-Specific, ATP-Dependent Interaction with Precursor RNA." *The Journal of Biological Chemistry* 272 (16): 10831–38. https://doi.org/10.1074/JBC.272.16.10831.

Zhelkovsky, A., Yoko Tacahashi, Tommy Nasser, Xiaoyuan He, Ulrike Sterzer, Torben Heick Jensen, Horst Domdey, and Claire Moore. 2006. "The Role of the Brr5/Ysh1 C-Terminal Domain and Its Homolog Syc1 in MRNA 3'-End Processing in Saccharomyces Cerevisiae." *RNA* 12 (3): 435–45. https://doi.org/10.1261/rna.2267606.

Zhelkovsky, Kessler, and Moore. 1995. "Structure-Function Relationships in the Saccharomyces Cerevisiae Poly(A) Polymerase." *The Journal of Biological Chemistry*. Vol. 270.

Zheng, Shawn Q., Eugene Palovcak, Jean-Paul Armache, Yifan Cheng, and David A. Agard. 2016. "Anisotropic Correction of Beam-Induced Motion for Improved Single-Particle Electron Cryo-Microscopy." *BioRxiv*, July, 061960. https://doi.org/10.1101/061960.

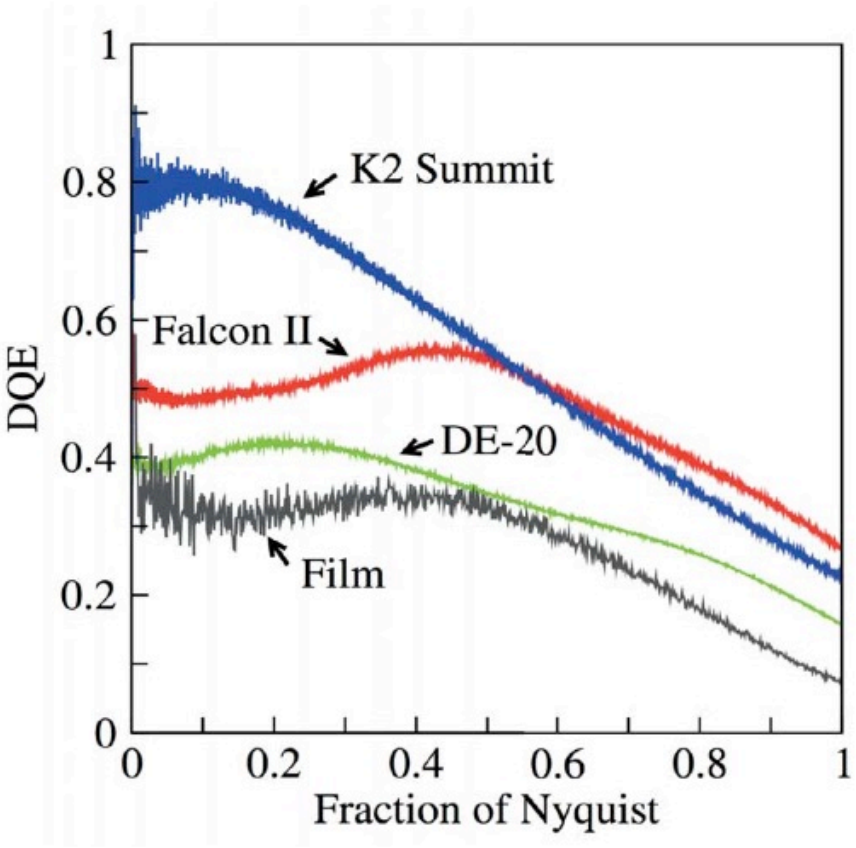
Zundert, G.C.P. van, J.P.G.L.M. Rodrigues, M. Trellet, C. Schmitz, P.L. Kastritis, E.

Karaca, A.S.J. Melquiond, M. van Dijk, S.J. de Vries, and A.M.J.J. Bonvin. 2016. "The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes." *Journal of Molecular Biology* 428 (4): 720–25. https://doi.org/10.1016/J.JMB.2015.09.014.

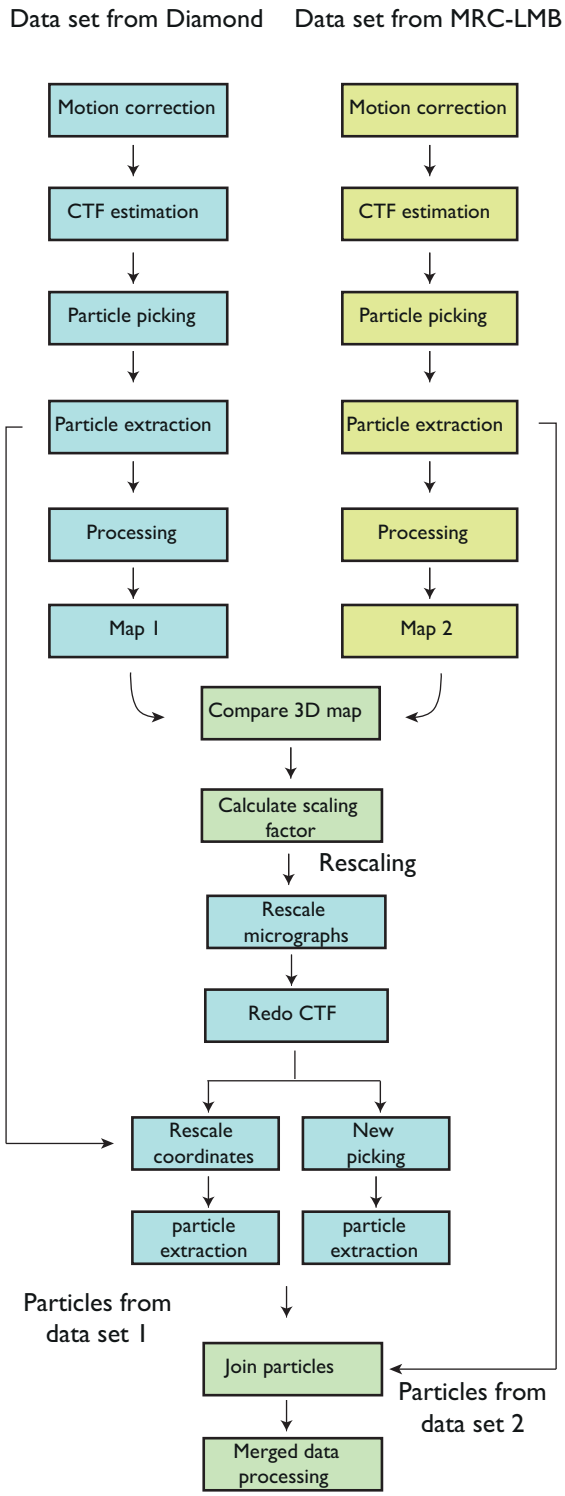
8 Appendix

	Yeast complex	Yeast protein	M.W (kDa)	Human protein	M.W (kDa)	Human complex	Sequence identity / similarity (%)	Essential (Yes/No)
	Polymerase module	Pap1	65	PAPOLA	83	CPSF	40 / 59	Yes
PF I		Pfs2	53	WDR33	146		37 / 57	Yes
		Yth1	25	CPSF30	30		36 / 51	Yes
		Fip1	36	hFip1	67		22 / 34	Yes
CF II		Cft1	153	CPSF160	161		17 / 34	Yes
	Nuclease module	Ysh1	88	CPSF73	77		44 / 65	Yes
		Cft2	96	CPSF100	88		22 / 41	Yes
		Mpe1	50	RBBP6	202		28 / 48	Yes
CF II	Phosphatase module	Pta1	89	Symplekin	141		17 / 37	Yes
		Pti1	47	-	-		-	Yes
		Ref2	60	-	-		-	No
		Swd2	37	WDR82	35		34 / 52	Yes
		Glc7	36	PP1A	38		85 / 93	Yes
		Ssu72	24	SSU72	23		44 / 65	Yes
CF I	Cleavage Factor IA	Rna14	78	Cstf-77		CstF	26 / 42	Yes
		Rna15	33	Cstf-64			27/ 41	Yes
		-	-	Cstf-50			-	Yes
		Clp1	50	hClp1		CF IIm	27/44	Yes
		Pcf11	72	hPcf11			21/32	Yes
	Cleavage Factor IB	Hrp1	60	-	-		-	Yes
		-	-	CF I 25		CF Im	-	Yes
		-	-	CF I 68			-	Yes
		-	-	CF I 59			-	Yes

Appendix 8.1: Yeast and human CPF subunits



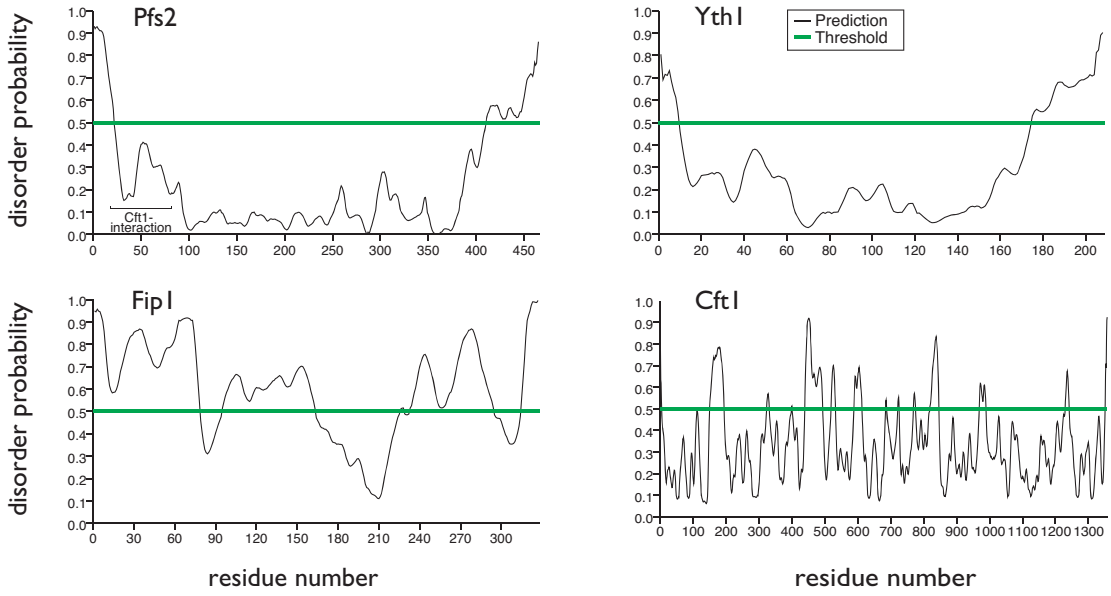
Appendix 8.2: DQE measured as a function of spatial frequency for different detectors. (Adapted from McMullan et al, 2014)



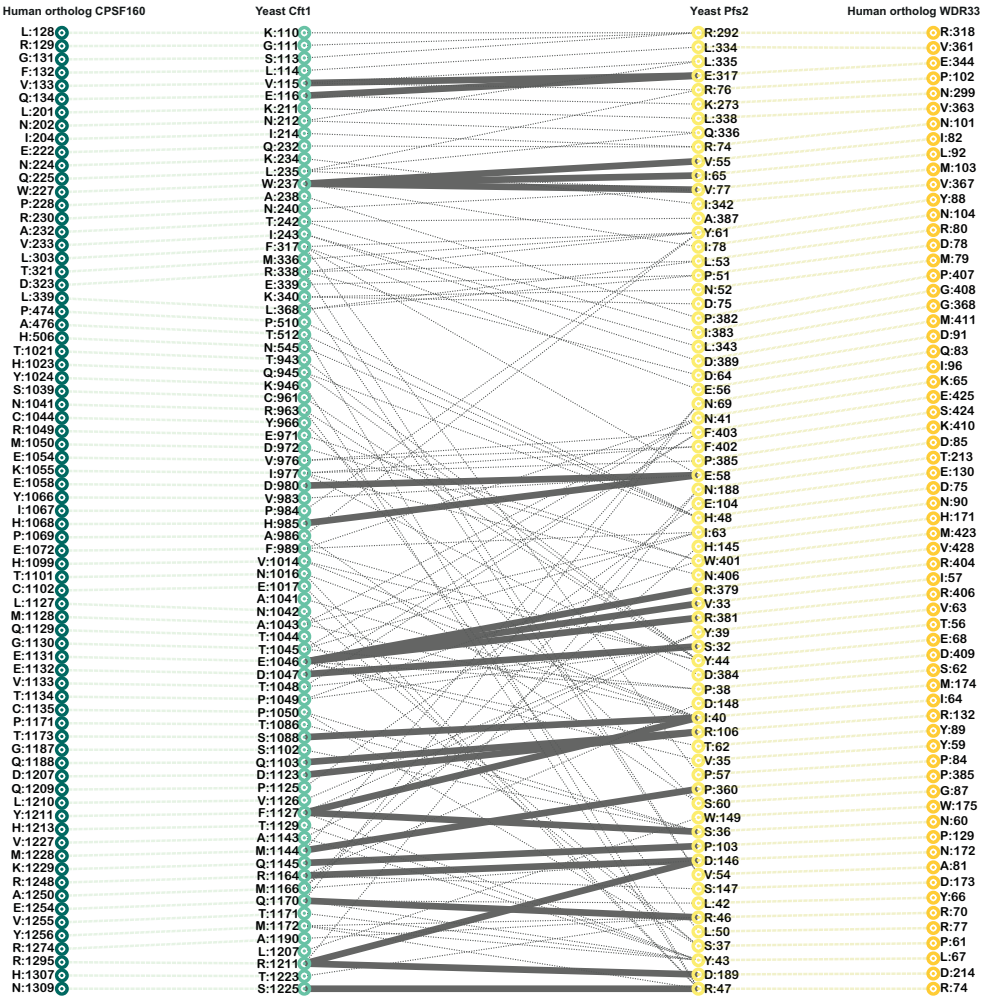
Appendix 8.3: Workflow for merging cryo-EM data sets from different microscopes. The original micrographs from the Diamond data set are rescaled and the particles extracted from these rescaled micrographs are merged with particles from the LMB data set. Further processing can be carried out using standard procedures.

Data Collection	
Pixel size (Å)	1.4
Defocus (μm)	1.5 - 3.5
Voltage (keV)	300
No. of particles	77,917
Electron dose ($e^-/Å^2$)	45
Model composition	
Non-hydrogen atoms	13,745
Protein residues	1,717
Zinc ions	2
Model refinement	
Resolution (Å)	3.5
Average B-factor ($Å^2$)	-120
Fourier shell correlation	0.78
Rfactor	0.38
RMS deviations	
Bonds (Å)	0.0067
Rfactor (°)	1.26
Validation	
Molprobity score	2.92
Clash score, all atoms	11.73
Good rotamers (%)	91.78
Ramachandran plot	
Favoured (%)	86.19
Outliers (%)	2.06

Appendix 8.4: Cryo-EM data collection and statistics of refinement.



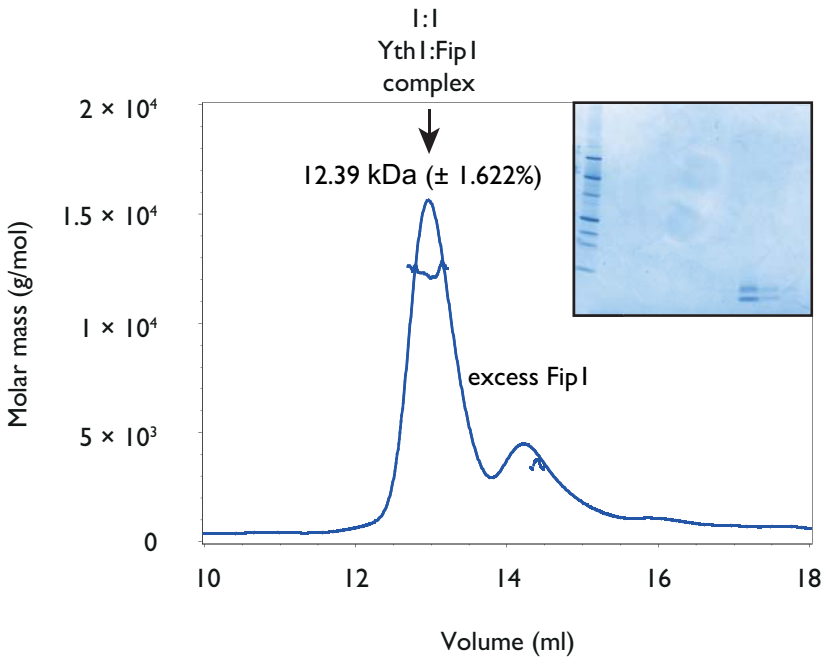
Appendix 8.5: Protein disorder prediction using PrDOS. Plots demonstrating the probability of disorder against amino acid residue for Pfs2, Yth1, Fip1 and Cft1. Plots were generated using PrDOS (Ishida and Kinoshita 2007). The false positive rate is 5%.



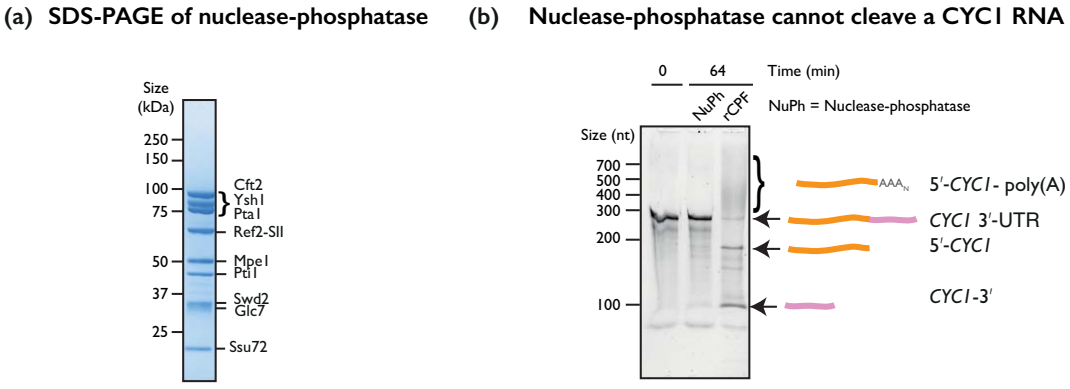
Appendix 8.6: Conservation of yeast Cft1–Pfs2 interactions in human CPSF160–WDR33. Residues mediating Cft1 and Pfs2 interactions are listed. The interactions include salt bridges or hydrogen bonds; van der Waals contacts; and hydrophobic contacts. Shown in orange and green dotted lines are equivalent residues from the human orthologs. Thick and dark grey lines represent the conserved interactions between yeast and humans. Also shown in thin, light grey lines are non-conserved interactions.

Protein 1	Protein 2	Distance (Å)	pAm	pAm + Pap1	
Cft1_K685	Cft1_K689	9.5	✓	×	Intra-protein crosslinks mapped on to the structure
Cft1_K842	Cft1_Y1266	22.2	✓	✓	
Cft1_K926	Cft1_K1002	27.4	✓	×	
Cft1_K930	Cft1_K1002	15.5	✓	✓	
Cft1_K930	Cft1_S1000	13.4	✓	×	
Cft1_K933	Cft1_K1002	9.3	✓	×	
Cft1_Y941	Cft1_K1006	10.01	✓	×	
Cft1_K942	Cft1_S959	10.2	×	✓	
Cft1_K942	Cft1_K962	4.9	✓	×	
Cft1_K942	Cft1_K1006	13.5	×	✓	
Cft1_K942	Cft1_K1011	14.5	✓	✓	
Cft1_S959	Cft1_K1006	9.3	✓	×	
Cft1_K942	Cft1_S1013	14.6	✓	×	
Yth1_K44	Yth1_K54	15	✓	✓	
Yth1_K44	Yth1_K64	23.5	✓	×	
Pap1_K53	Pap1_K128	15	×	✓	
Pap1_K70	Pap1_K109	11.8	×	✓	
Pap1_K183	Pap1_K290	27	×	✓	
Cft1_S1027	Yth1_K54	16	✓	✓	Inter-protein crosslinks mapped on to the structure
Cft1_K1028	Yth1_K54	15.3	✓	✓	
Pfs2_K210	Yth1_K44	32.8	✓	×	
Cft1_K184	Cft1_K189	N/A	✓	×	
Cft1_K184	Cft1_K224	N/A	×	✓	
Cft1_K184	Cft1_K1330	N/A	✓	✓	
Cft1_K186	Cft1_S198	N/A	✓	✓	
Cft1_K189	Cft1_S198	N/A	×	✓	
Cft1_S416	Cft1_K440	N/A	✓	×	
Cft1_S591	Cft1_K609	N/A	✓	✓	
Cft1_Y585	Pap1_K5	N/A	×	✓	
Cft1_K609	Pap1_K183	N/A	×	✓	
Cft1_K1330	Pap1_K183	N/A	×	✓	
Cft1_K1325	Pap1_K536	N/A	×	✓	
Cft1_K1011	Pfs2_K25	N/A	✓	×	
Cft1_K1104	Pfs2_K25	N/A	✓	×	
Cft1_K1104	Pfs2_Y26	N/A	✓	×	
Cft1_K1330	Pfs2_S207	N/A	✓	×	
Cft1_K1330	Pfs2_K210	N/A	×	✓	
Cft1_K1330	Pfs2_Y393	N/A	✓	×	
Cft1_S1027	Yth1_K145	N/A	✓	×	
Cft1_K1028	Yth1_K145	N/A	✓	×	
Cft1_K1330	Yth1_K44	N/A	✓	✓	
Cft1_S591	Fip1_K44	N/A	✓	×	
Cft1_S591	Fip1_K47	N/A	✓	×	
Cft1_S591	Fip1_K148	N/A	✓	✓	
Cft1_K609	Fip1_K126	N/A	×	✓	
Cft1_S1027	Fip1_K148	N/A	×	✓	
Cft1_S1349	Fip1_K148	N/A	✓	×	
Pap1_K34	Pap1_S38	N/A	×	✓	
Pap1_S344	Pap1_K536	N/A	×	✓	
Pap1_K527	Pap1_K536	N/A	×	✓	
Pap1_K432	Pap1_K536	N/A	×	✓	
Pap1_K536	Pap1_K546	N/A	×	✓	
Pap1_K536	Pap1_K549	N/A	×	✓	
Pap1_K536	Pap1_S550	N/A	×	✓	
Pap1_K290	Pfs2_Y393	N/A	×	✓	
Pap1_K290	Pfs2_K396	N/A	×	✓	
Pap1_K183	Yth1_K44	N/A	×	✓	
Pap1_K290	Yth1_K44	N/A	×	✓	
Pap1_K290	Yth1_54	N/A	×	✓	
Pap1_K34	Fip1_K44	N/A	×	✓	
Pap1_K290	Fip1_S121	N/A	×	✓	
Pap1_K290	Fip1_K148	N/A	×	✓	
Pap1_K290	Fip1_K219	N/A	×	✓	
Pap1_K292	Fip1_K148	N/A	×	✓	
Pfs2_Y26	Yth1_K76	N/A	✓	✓	
Pfs2_K244	Yth1_K44	54.4	✓	×	
Pfs2_S453	Yth1_K76	N/A	✓	✓	
Fip1_S38	Fip1_K47	N/A	✓	✓	
Fip1_K44	Fip1_K47	N/A	✓	✓	
Fip1_K44	Fip1_S50	N/A	✓	×	
Fip1_K148	Fip1_K219	N/A	×	✓	
Fip1_S152	Fip1_K219	N/A	✓	×	
Fip1_K219	Yth1_K182	N/A	✓	✓	
Fip1_K219	Yth1_K191	N/A	✓	✓	
Fip1_K219	Yth1_K196	N/A	✓	✓	
Yth1_K182	Yth1_K191	N/A	✓	×	
Yth1_K182	Yth1_K196	N/A	✓	✓	

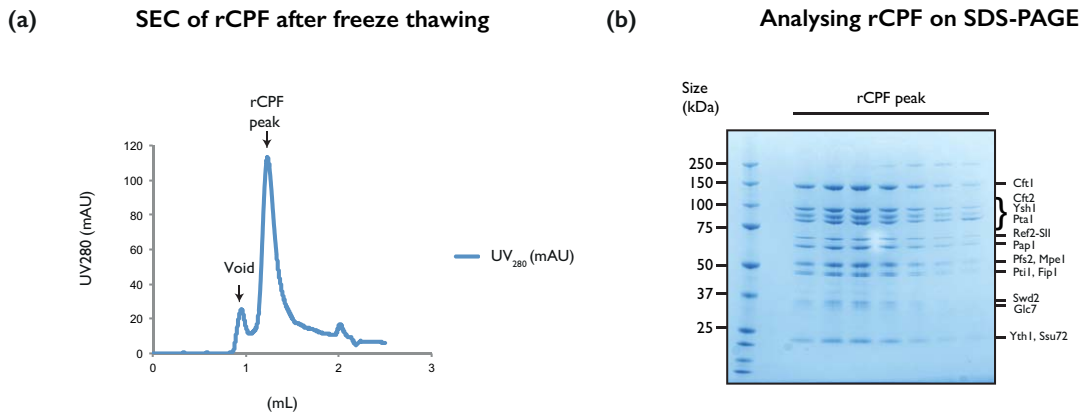
Appendix 8.7: Amino acids involved in protein-protein cross-linking, identified by mass spectrometry.



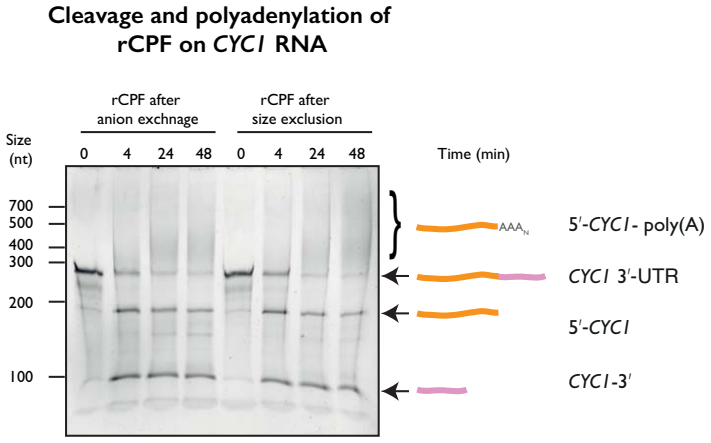
Appendix 8.8: SEC-MALS profile of Yth1/Fip1 complex with excess Fip1. The theoretical molecular masses for Yth1 (118-208) and Fip1 (180-220) are 6.2 and 5 kDa respectively.



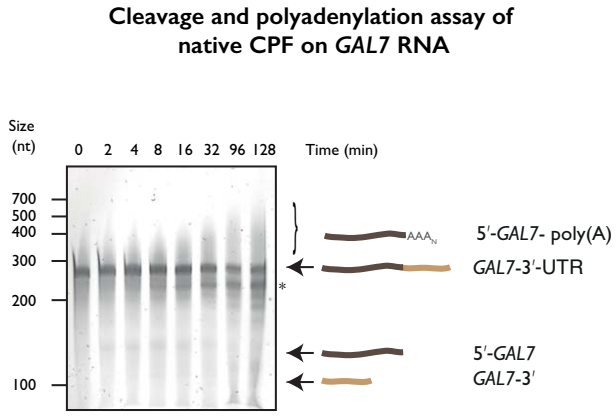
Appendix 8.9: Nuclease-phosphatase module is not sufficient to cleave mRNA 3' end **(a)** SDS-PAGE showing purified nucleo-phosphatase after size exclusion chromatography that was used in the assay. **(b)** Cleavage and polyadenylation assay of purified nucleo-phosphatase analyzed by denaturing urea-PAGE. The control reaction was performed with rCPF. *CYC1* is the substrate RNA. Cleavage products are 5′- *CYC1* (in orange) and *CYC1*-3′ (in pink).



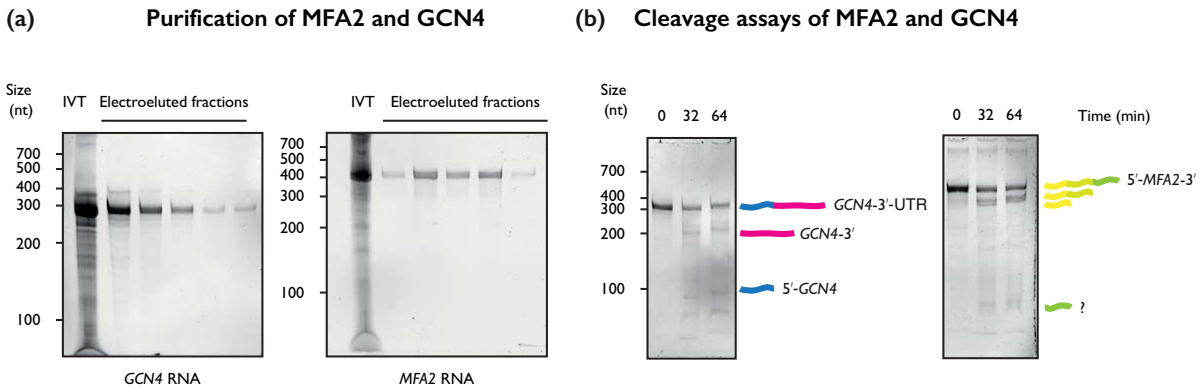
Appendix 8.10: Frozen and thawed rCPF behaves similar to fresh rCPF on a size exclusion column **(a)** Chromatogram of frozen and thawed rCPF run on a Superose 6 Increase 3.2/300 column in SEC buffer. **(b)** The peak fractions were analyzed using SDS-PAGE.



Appendix 8.11: Testing the functional activity of rCPF. Coupled cleavage and polyadenylation assay of purified rCPF analyzed by denaturing urea-PAGE. The activity of the sample after anion exchange column was compared to the sample after size exclusion. CYC1 is the substrate RNA. Cleavage products are 5′-CYC1(in orange) and CYC1-3′ (in pink). The cleaved 5′-CYC1 gets hyper-polyadenylated.



Appendix 8.12: CPF cannot cleave GAL7 3′ UTR. Coupled cleavage and polyadenylation assay of native CPF analyzed by denaturing urea-PAGE. It is found that CPF cannot cleave GAL7 RNA. Cleavage products are 5′-GAL7 (in dark brown) and GAL7-3′ (in light brown). * denotes non-specific cleavage. The reaction contained both CF IA and CF I B, in addition to CPF.



Appendix 8.13: Introducing new substrate RNAs to study 3′end processing. **(a)** The products of the in vitro transcription (IVT) reaction of *GCN4* 3′UTR and *MFA2* RNA are analyzed by denaturing urea-PAGE along with the purified fractions from electroelution. **(b)** Cleavage assay of *GCN4* and *MFA2* are analyzed by denaturing urea-PAGE. *GCN4* undergoes cleavage at it's major poly(A) site to yield downstream (magenta) and upstream (blue) products of 203 nt and 97 nt respectively. Whereas *MFA2* RNA contains multiple cleavage sites and the cleaved RNA products are depicted in yellow and two different shades of green.

The deep Stillness
Seeping into the rocks
The voice of the Cicada

Matsuo Bashō, (1644 - 1694)