Structure of mammalian RNA polymerase II elongation complex bound by α-amanitin and study of mammalian transcription termination and 3' end processing

Dissertation

for the award of the degree

'Doctor rerum naturalium'

of the Georg-August-Universitaet Goettingen

Within the graduate program

'Molecular Biology of Cells'

Of the Georg-August University School of Science (GAUSS)

Submitted by

Xiangyang Liu

From Shanxi, China

Goettingen 2019

Members of Thesis committee

Prof. Dr. Patrick Cramer
Department of Molecular Biology
Max Plank Institute for Biophysical Chemistry, Goettingen

Prof. Dr. Markus Zweckstetter
German Center for Neurodegenerative Diseases
Max Plank Institute for Biophysical Chemistry, Goettingen

Prof. Dr. Jochen Hub
Theoretical Biophysics
Saarland University, Saarbruecken

Members of the Examination board

Prof. Dr. Patrick Cramer (1st Referee)
Department of Molecular Biology
Max Plank Institute for Biophysical Chemistry, Goettingen

Prof. Dr. Henning Urlaub (2nd Referee)
Bioanalytical mass spectrometry
Max Plank Institute for Biophysical Chemistry, Goettingen

Further Members of the Examination board

Prof. Dr. Gregor Eichele
Department of genes and behavior
Max Plank Institute for Biophysical Chemistry, Goettingen

Prof. Dr. Wolfgang Wintermeyer
Ribosome dynamics
Max Plank Institute for Biophysical Chemistry, Goettingen

Affidavit

I, Xiangyang Liu, hereby declare that my dissertation entitled 'Structure of mammalian RNA polymerase II elongation complex bound by α-amanitin and study of mammalian transcription termination and 3' end processing' has been written independently and with no other sources and aids than quoted. This dissertation or parts thereof have not been submitted elsewhere for any academic award or qualification. The electronic version of this dissertation is congruent to the printed versions both in content and in format.

Goettingen, 30th of August 2019

.....

Xiangyang Liu

Acknowledgements

Three years ago, I was so excited to get the PhD position in Patrick's lab. I came to Germany for the first time with my two big suitcases. Everything was so exciting but challenging. Time flies, now I am here for almost four years. I enjoyed my time so much and I think I am so lucky to work in such an amazing lab with so many nice people.

First of all, I would like to thank Patrick Cramer for such an amazing science atmosphere and all the kind instructions, and most of all, for helping me to be more confident. In the last four years, I learnt a lot from you. It's not only about how to be a scientist, but also about how to be a good manager to motivate people. You deal with a lot of things per day but still take every detailed issue from everybody seriously. You never blame me because of my mistakes because you said students are allowed to make mistakes and we learn from our mistakes.

I would also like to thank all of my colleagues for both the help in the lab and in life. Thanks Janine Bluemel for taking care of all the administration issues, which I need to live and work in Germany and also for taking care of even some personal issues sometimes. Thanks Carrie Bernecky and Clemens for helping to start the EM experiment in the lab and also Carrie for the very patient instructions in mammalian Pol II purification. Thanks Sarah Sainsbury for a lot of discussions and instructions for experiment details. Thanks Seychell Vos, Lucas Farnung and Goran Kokic for the help in biochemistry and protein purifications and also for the help at the beginning of my PhD. I would also like to thank Christoph Wigge, Dimitry Tegunov and Christian Dienemann for the Electron microscope support, thank you for setting up the microscope and maintaining the machine. Everything is so user friendly, which made my work much easier and faster. I would also like to thank Isaac Fianu, Ying Chen, Sara Osman, Felix Wagner, Paulina Seweryn, Sandra schilbach and all the other colleagues for all the discussions, advices, ideas and most importantly, happy life.

I would like to thank Ulrich Steuerwald, Juergen Wawrzinek and Hauke Hillen for all the kind assistance in crystallography, you are so patient to give me all the detailed trainings. I would also like to thank our administration team: Kerstin Maier, Kirsten Backs, Thomas Schulz and Petra Rusfor the safety instructions and all the material support. Thanks Ute Neef for taking care of the insect cells which makes my protein expression much easier, thanks Angelika Kruse and Manuela Wenzel for the cleaning and support. I would like to thank people from the system biology lab: Anna Sawicka for offering the cDNA for PCR, Bjoern Schwalb, Kristina Zumer, Gabreil Villamil and Saskia Gressel for a lot of discussions. Thanks all the lab members for such amazing work and atmosphere.

I would like to thank Henning Urlaub and his lab members for support in mass spectrometry for both protein identification and cross linking MS, especially my collaborator, Ralf Pflanz. I would also like to thank my thesis committee members Markus Zweckstetter and Jochen Hub for supervising my PhD and all the ideas and input. I would like to thank the GGNB office for taking care of the program and for organizing the courses and phD training. I would also

like to thank my rotation student Alexander Helmut Rotsch for giving me the chance to supervise him and learn how to be a good supervisor.

I would thank all my friends in Germany and in China. Thanks my forever best friend Zhenghan Di. You are the only one who would keep talking to me when I am not happy, even if it was already midnight in China. Thanks to Wanwan Ge, you are like an angel and always bring people happiness and positive energy. I would also like to thank Yuan Yan, my lovely friend for all the happy time and nice dinners during my thesis writing time.

I would like to thank my parents for growing me up and giving me chance to get good education, and for supporting me to get my PhD in Germany. I would also like to thank my sister for helping me to take care of our parents and little brother during the time I was not in China. I am also very grateful to my brother, you are young, but always bring me braveness and happiness, you are going start your university education this year and wish you the best in everything!

Publications

Liu X, Farnung L, Wigge C, Cramer P. Cryo-EM structure of a mammalian RNA polymerase II elongation complex inhibited by α-amanitin. The Journal of Biological Chemistry 293, 7189-7194 (2018)

The following sections were taken from Liu *et al.*:

Chapter 1:

Summary

Methods

2.2.4 Formation of Elongation complex

2.2.5 Electron microscopy

2.2.6 Model building and refinement

2.2.7 Transcription assay

Results

3.2 Pol II elongation complex formation, activity inhibition by α-amanitin and Cryo-EM grids preparation

3.3 Pol II EC-hGdown1-α-amanitin complex data processing

3.4 Pol II EC-hGdown1-α-amanitin complex overall structure analysis and comparison with yeast EC- α-amanitin complex

3.5 Specificity of α-amanitin binding pocket in mammalian

3.6 α-amanitin resistance caused by binding pocket mutations

Discussion

The second paragraph

The following figures and tables were taken from Liu *et al*.

Figure 1.4. Pol II elongation complex (EC) formation, *in vitro* RNA extension assay, and exemplary micrograph and 2D classes of the dataset

Figure 1.5. Cryo-EM data processing

Figure 1.6. Local resolution of the cryo-EM density map

Summary

Chapter 1

RNA polymerase II (Pol II) is the central enzyme that transcribes eukaryotic protein-coding genes to produce mRNA. The mushroom toxin α-amanitin binds Pol II and inhibits transcription at the step of RNA chain elongation. Pol II from yeast binds α-amanitin with micromolar affinity, whereas metazoan Pol II enzymes exhibit nanomolar affinities. Here, I present the high resolution cryo-EM structure of α-amanitin bound to and inhibited by its natural target, the mammalian Pol II elongation complex. The structure revealed that the toxin is located in a pocket previously identified in yeast Pol II but forms additional contacts with metazoan-specific residues, which explains why its affinity to mammalian Pol II is ~3000 times higher than for yeast Pol II. The work provides the structural basis for the inhibition of mammalian Pol II by the natural toxin α-amanitin and highlights that cryo-EM is well suited to studying interactions of a small molecule with its macromolecular target.

Chapter 2

Transcription termination is coupled to pre-mRNA 3' processing. In mammals, more than twenty protein factors are involved in these processes. The definition of the cleavage site needs not only protein factors but also specific cis sequence elements on pre-mRNA. The best known cis elements include the polyadenylation signal (PAS, featuring the base sequence AAUAAA), the upstream elements (USE, featuring the base sequence UGUA) and downstream elements (DSE, characteristically GU/U rich), which are bound by the cleavage and polyadenylation (CPSF) complex, the cleavage factor I (CFI) complex and the cleavage stimulation factor (CstF) complex respectively. Other termination/3' processing factors include cleavage factor II (CFII), polyadenylation polymerase (PAP), polyadenylate-binding nuclear protein 1 (PABPN1), Pol II carboxy terminal domain (CTD), symplekin (SYMPK), as well as some kinases and phosphatases and other factors. Based on the functional differences, CPSF complex is divided into two modules: the polymerase module, which is composed of CPSF160, WDR33, CPSF30 and Fip1, and the nuclease module which is composed of CPSF100 and CPSF73. The polymerase module binds specifically to PAS site while the nuclease module is responsible for the cleavage of pre-mRNA. CPSF73 is the endonuclease. After cleavage, a polyadenylic acid tail (poly(A) tail) is added to the 3' end of RNA by PAP. In this work, I managed the expression and purification of the sub-complexes CFI, CFII, CstF, CPSF polymerase module and CPSF nuclease module plus symplekin. The CstF complex can be crystallized but the diffraction of the crystal was not good enough to solve the structure yet. The CPSF polymerase module and nuclease module plus symplekin can form a stable complex which is suitable for cryo-EM structure analysis. From the initial data processing, the extra density in addition to the polymerase module can be seen. However, the resolution of the density map needs to be improved by further processing.

Table of contents

Chapter 1

1 Introduction

1.1 The central dogma and RNA polymerases

Genetic information defines various species and their characteristics. In living organisms, genetic information is normally stored in the form of DNA sequences. To make the genetic information function in the organisms and to keep the species characteristic constant, two more biopolymers are necessary, which are RNA and protein. The DNA sequence is used as a template for the synthesis of RNA by a process named transcription, and RNA sequence directs the synthesis of proteins by a process named as translation. This flow of genetic information is the basic outline of the central dogma of molecular biology (Crick, 1970). However, the central dogma also includes DNA replication, where the DNA molecule can replicate itself to provide genetic material for progeny (Meselson and Stahl, 1958). Further supplementary to the central dogma are special forms of information transfer from RNA to DNA and from RNA to RNA, which normally happens in viruses and are called reverse transcription and RNA replication respectively (Ahlquist, 2002; Temin and Mizutani, 1970). Additionally, prions can propagate themselves in host cells which are the only protein to protein information encoding known so far (Prusiner, 1991).

RNA polymerase, abbreviated RNAP, is one kind of enzyme that synthesizes RNA from a DNA template. RNAP exists in both viruses and living organisms. Depending on the species, RNAP might be a single subunit enzyme (Cermakian et al., 1997) or a protein complex with several subunits (Werner and Grohmann, 2011). In Prokaryotes and archaea, there is only one kind of RNAP that transcribes all kinds of RNAs, whereas in eukaryotes, there are three different kinds of RNA polymerases (Roeder and Rutter, 1969; Sentenac, 1985). RNA polymerase I (Pol I) is responsible for the transcription of ribosome RNA (rRNA)(Grummt, 2003) except for 5S rRNA, which is transcribed by RNA polymerase II (Pol II). RNA polymerase III (Pol III) synthesizes small RNAs like U6 spliceosomal small nuclear RNA (snRNA), transfer RNA (tRNAs), adenovirus-associated RNA (VA-RNA) and 7SK RNAs (Geiduschek and Kassavetis, 2001). Pol II is responsible for synthesizing all protein-coding RNAs and most non-coding RNAs, including small nuclear RNAs (snRNAs), small nucleolar RNAs (snoRNAs), microRNAs (miRNAs), cryptic unstable transcripts (CUTs) and stable unannotated transcripts (SUTs)(Liu et al., 2013). In eukaryotes, the protein coding genes are transcribed by Pol II as precursor message RNAs (pre-mRNAs) which need to be further processed to become mature mRNA. Further processing of pre-mRNA includes 5’ capping, 3’ polyadenylation and intron splicing (Hirose and Manley, 2000). The mature mRNA is more stable and can be exported from nuclei to the cytoplasm for translation. In plants, there are two more RNA polymerases, Pol IV and Pol V, which are thought to generate non-coding RNA transcripts and mediate gene-silencing processes (Ream et al., 2009; Zhang et al., 2007). As mentioned before, bacteria has only one type of RNA polymerase which is responsible for the transcription of all kinds of RNAs (Darst et al., 1989). RNAP in archaea varies in different species and shares similar

features with both eukaryotic RNAP and bacterial RNAP. Most of the RNAP known so far in archaea are composed of more than ten subunits and with an overall shape that is quite similar to their eukaryotic counterpart (Hirata et al., 2008). On the other hand, the archaeal RNAP is normally responsible for the transcription of all kinds of RNAs in cells, which is similar to the bacterial RNAP.

The bacterial RNAP is the simplest one among the three kingdoms of life, the core enzyme is composed of five subunits: two copies of α, ß, ß' and ω (Ebright, 2000; Mathew and Chatterji, 2006). The sixth subunit σ is thought to be a complementary subunit which helps to recognize the promoter and start promoter-specific transcription (Kang et al., 1997). RNAP structure varies in different archaeal species. In the known archaeal RNAPs, the protein complex consists of 11 to 13 subunits depending on the species. Taking the *Sulfolobus solfataricus* RNAP as an example, the crystal structure of *S. solfataricus* RNAP was solved in 2008 (Hirata et al., 2008) and it consists of 13 subunits which include RpoA', A'', RpoB, D, K, L, F, H, E, G, N, P and Rpo13. Rpo13 exists only in some archaea species. The overall shape is similar to Pol II in eukaryotes with a 'crab' like shape. RpoE/F makes up the stalk of the polymerase. However, there is one difference between Pol II and archaea RNAP. In archaea RNAP, the biggest subunit RpoA is divided into two polypeptides, which are encoded by two different genes and connected by the 'foot' domain, while in Pol II, the corresponding subunit Rpb1 is a single subunit encoded by one gene.

As mentioned before, there are three different RNA polymerases in eukaryotes which are responsible for the transcription of different types of RNA. All three RNA polymerases (Pol I, Pol II and Pol III) contain a 'conservation core' which was conserved from bacteria to eukaryotes. Yeast Pol I is a 14-subunit complex with a molecular weight of 590kDa. With regards to subunit composition, Pol I contains a Pol II like core which is composed of five subunits (A190, A135, AC40, AC19 and A12.2), five common subunits (Rpb5, Rpb6, Rpb8, Rpb10 and Rpb12) which are the same as Pol II and two specific heterodimeric sub-complexes: A14–A43 and A49–A34.5. A190 and A135 are the two biggest subunits which are corresponding to Rpb1 and Rpb2 respectively (Engel et al., 2013). Pol III is the largest of the three RNA polymerases, which contains 17 subunits and has a total molecular weight of 700kDa. C160 and C128 are the two biggest subunits which correspond to Rpb1 and Rpb2 respectively. The other subunits include the core subunits ABC27, ABC23, ABC14.5, ABC10α and ABC10β which are common between Pol I, Pol II and Pol III, subunits AC40 and AC19 shared by Pol I and Pol III and subunits C25, C17, C11, C53, C37, C82, C34, C31. C53 and C37 form the TFIIF similar complex and C11 is a termination factor for Pol III transcription (Han et al., 2018). Pol II is the best studied RNAP both in yeast and in mammals, which might be attributed to the fact that it is responsible for the transcription of all protein coding genes. Pol II is a 500kDa complex which is composed of the ten-subunit core (Rpb1 to Rpb12) and the Rpb4/7 stalk. Rpb1 and Rpb2 form a clamp with other subunits arraying around the periphery (Cramer et al., 2000). The active center is located in the Rpb1-Rpb2 cleft with a divalent ion of Mg^{2+} for activity (Armache et al., 2003). The Rpb4/7 stalk is highly flexible in Pol II. The structure of mammalian Pol II is quite similar to its yeast counterpart except for

some residue differences (Bernecky et al., 2016b). Different groups of factors are necessary for initiation, elongation and termination during Pol II transcription.

Human Gdown1 is the product of the *POLR2M* gene, the molecular weight is 42 kDa. In some early studies, Gdown1 was thought to be the 13th subunit of Pol II in metazoans because it is associated tightly with Pol II during purification (Hu et al., 2006). The study of Gdown1 also showed that Gdown1 holds the paused Pol II by competing for the same binding position on the initiation complex with TFIIF (Wu et al., 2012). However, the Gdown1 'paused' Pol II can be released by mediator and mediator dependent regulation is enforced by Gdown1 (Jishage et al., 2012).

1.2 α-amanitin - the cyclic Octapeptide from toxic mushrooms

1.2.1 Research history of amanitin

Macroscopic life is composed of fungi, plants, animals and human beings. Human beings are at the higher level of food chains, which means human beings always have more access to food choices. However, not all 'plants in the big garden' are edible. All animals have unique ways to fight for survival, as do plants. Plants cannot move and do real 'fighting', but they also found their own ways to protect themselves during evolution, just like some of the mushrooms. Even though they have pretty colors and can offer animals and people nutrition and energy, they secrete highly toxic chemicals which may cause severe physical injuries or even death after ingestion (Wieland, 1986).

The study of mushroom toxins started from the beginning of 20th century (Ford, 1907), when Hermann Schlesinger and William W. Ford tried to purify the toxic factors from amanita phalioides mushrooms. They managed to purify a heat-resistant substance and named it Amanita-toxin. Even with very crude methods, they managed to purify the toxin to a content of about 10% purity. With preliminary chemical studies, they identified the toxic chemical as an 'aromatic phenol combined with an amine group that it readily forms an indol or pyrrol ring' instead of a beta proteid, a glucoside, or an alkaloid. In 1937, Feodor Lynen and Ulrich Wieland succeeded in crystallizing the Amanita-toxin (Wieland and Hallermayer, 1941). Rudolf Hallermayer also described the crystallization of amanitin in his PhD thesis in 1940. However, because of the high toxicity and low purity, it took another fifty years for people to finally solve the amanitin structure and learn its toxicological mechanism (Wienland and Faulstich, 1991).

1.2.2 The structure of α-amanitin

The amatoxins form a family. The early method defined the name based on electrophoresis. The neutral compound was called α-amanitin and the acidic one was named β-amanitin (Wieland, 1948). γ- and ε-amanitin was also discovered and isolated afterwards as well as some non-poisonous components like amanullin and amaninamide (Buku et al., 1980; Cochet-Meilhac and Chambon, 1974).

α-amanitin is a cyclic peptide which is composed of eight amino acid residues (Hatzoglou et al., 1985). The linkage of 6-hydroxytrytophan and cysteine forms an inner ring (Michelot and Labia, 1988). There are several modified amino acid side chains within the α-amanitin molecule, which include hydroxyl proline at position 4, 4,5-dihydroxy-isoleucine and 6-hydroxy-2-mercapto-L- Tryptophan (Figure 1.1). These modified amino acid residue side chains help α-amanitin bind to RNA polymerase and inhibit transcription in cells (Wieland et al., 1983; Zanotti et al., 1989). These three modified side chains significantly affect the binding affinity of α-amanitin to different RNA polymerases (Figure 1.1).

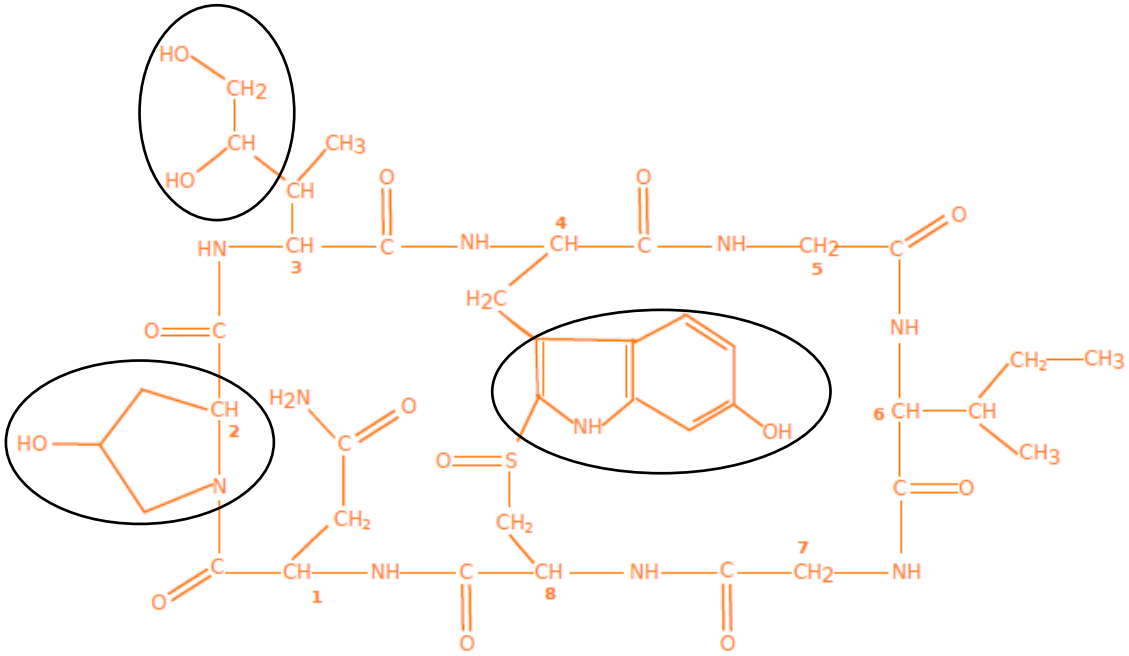


Figure 1.1: schematic diagram of α-amanitin. The three side chains which are important for RNAP binding and activity are highlighted with black circles.

1.2.3 The toxicity of α-amanitin and Pol II

α-amanitin is found in several species of mushrooms from the mushroom genus *Amanita.* Ingestion of mushrooms that contain α-amanitin results in four stages of toxicity symptoms (Mas, 2005; Yilmaz et al., 2015). The first stage of symptoms normally appears 8 to 10 hours after the intake. In this stage, the patient suffers from severe digestive system reactions like nausea and vomiting. Normally a pseudo-recovery stage comes after the first stage which shows almost no symptoms. This makes the diagnosis and emergency treatment difficult because this pseudo-recovery might mislead both the patient and the clinician. However, in the third stage, which normally appears on day three after ingestion, liver and kidney failure becomes obvious, which could be attributed to affected enterohepatic circulation. If no therapy is executed from this stage, the patient would die of massive liver necrosis and kidney failure in 5 to 12 days (Mas, 2005).

The toxicity of α-amanitin comes from its specific binding to Pol II and the inhibition of transcription in cells (Fiume and Stirpe, 1966). The binding affinity of α-amanitin to Pol I and Pol III is much weaker than to Pol II. Pol I is totally insensitive to it, and Pol III is inhibited only at a very high concentration in animals. The binding affinity also varies between virus, bacterias, yeasts and mammals, with the binding affinity more than one thousand times higher in mammals than in yeast (Cochet-Meilhac et al., 1974; Wienland and Faulstich, 1991)(Table 1.1).

Species	RNA-polymerase II	III
	M	M
Various mammals	3×10^{-9}	$1-4 \times 10^{-5}$
HeLa cells	3×10^{-9}–10^{-8}	$1-4 \times 10^{-5}$
Amphibian (*Xenopus laevis*)	5×10^{-8}	2×10^{-5}
Fruitfly (*Drosophila melanogaster*)	3×10^{-8}	
Slimy mushroom	3×10^{-8}	
Plants (corn, soja, wheat)	5×10^{-8}	
Yeast (*Saccharomyces cerevisiae*)	10^{-6}	
Champignon Common meadow mushroom (*Agaricus bisbosporus*)	7×10^{-6}	
Deadly Agaric Death cup (*Amanita phalloides*)	2×10^{-4}	

Table 1.1: α-amanitin binding affinity varies in different organisms. The binding affinity to Pol II is ~1000 times higher in mammals than in yeast (see the red rectangle box). Table was adapted from T. Wieland and H. Faulstich., Fifty years of amanitin, 1991

1.3 Transcription elongation and α-amanitin inhibition in eukaryotes

1.3.1 An overview of transcription cycle

In Eukaryotes, transcription commences with the recognition of the promoter by initiation factors. The assembly of the initiation factors and RNA polymerase forms the pre-initiation complex (Sainsbury et al., 2015). In Pol II transcription initiation, the general initiation factors play important roles. The general transcription factors include TFIIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH. TFIID itself is a big protein complex with a total size of 1.2 MDa. It is composed of TATA box binding protein (TBP) and 13–14 TBP associated factors (TAFs)(Bieniossek et al., 2013). To initiate the transcription, TBP binds to the promoter DNA and bends the DNA by 90 degrees. The whole TFIID factor is responsible for the specific recognition of promoters and DNA bending for initiation (Louder et al., 2016). TFIIA is Pol II transcription specific and helps the binding of TBP to DNA (Hoiby et al., 2007). The opening of double stranded DNA needs the cooperation of TFIIB, TFIIE and TFIIH. The structure of the TFIIB-Pol II complex

elucidated that TFIIB functions in Pol II recruitment, DNA bending and opening, initiation of RNA synthesis and transition from initiation to elongation (Sainsbury et al., 2013). The promoter opening of DNA requires TFIIE and TFIIH. TFIIE is composed of the TFIIE α and TFIIE β subunits and is responsible for the anchoring of the TFIIH kinase module (CAK) to the preinitiation complex. TFIIE also facilitates the recruitment of TFIIH to the initiation complex and stimulates the activity of TFIIH (Miwa et al., 2016). TFIIH is a complex of 10 subunits and consists of both ATPase and kinase activity. The ATPase activity offers energy during DNA opening by hydrolysis of ATP (Schilbach et al., 2017). TFIIF is a three-subunit protein complex that associates with Pol II. TFIIF influences selection of transcription start site, stabilizes the initiation complex (ITC) and assists early RNA synthesis (Robert et al., 1998). After the assembly of the ITC, transcription starts.

However, before going into processive elongation, Pol II would normally suffer from a 'promoter-proximal' pausing, which means most of the initiation factors have left and Pol II stops at the promoter-proximal region (Adelman and Lis, 2012). The 'paused' Pol II is normally stabilized by the protein complexes DRB sensitivity-inducing factor (DSIF) and negative elongation factor (NELF)(Vos et al., 2018b). The formation of the activated Pol II elongation complex requires two more elongation factors and one kinase, which are the Pol II associated factor (PAF), SPT6 and the positive transcription elongation factor (P-TEFb). PAF is a protein complex composed of 6 subunits (Paf1, Rtf1, Ski8, Cdc73, CTR9 and Leo1) in human (Vos et al., 2018a). In the pause-release transition, PAF complex takes the place of NELF on Pol II, and the elongation factor Spt6 binds to the CTD linker of RPB1 and helps to release the paused Pol II. The release of Pol II also needs P-TEFb, which is a cyclin-dependent kinase composed of CDK9 and cyclin T. The phosphorylation of both CTD and elongation factors stimulate Pol II release and elongation (Vos et al., 2018a).

During elongation, Pol II walks along the DNA template and transcribes pre-mRNA with the binding of super elongation factors (Luo et al., 2012). After walking over the poly(A) signal, Pol II suffers from another pause, where the elongation factors are replaced by termination factors (Glover-Cutter et al., 2008). The cleavage at the 3' end of the pre-mRNA induces the termination mechanism, which results in the release of both Pol II and RNA from the template DNA (Richardson, 1993), a process known as transcription termination.

1.3.2 Nucleotide addition cycle and α-amanitin inhibition

As mentioned in the last paragraph, during transcription elongation, Pol II moves along the DNA template and synthesizes a complementary pre-mRNA chain. Extension of the RNA chain is achieved by the nucleotide addition cycle (NAC) (Cramer, 2007). NAC is a highly coordinated process of several elements in the active center of Pol II, which includes the bridge helix, the trigger loop (Wang et al., 2006) and the central magnesium ions. The trigger loop is a highly mobile loop that undergoes folding to catalyze the extension of the RNA chain, and is also important for the translocation of nucleic acids to the next DNA template position after catalysis (Epshtein et al., 2002; Kettenberger et al., 2004; Landick, 2004).

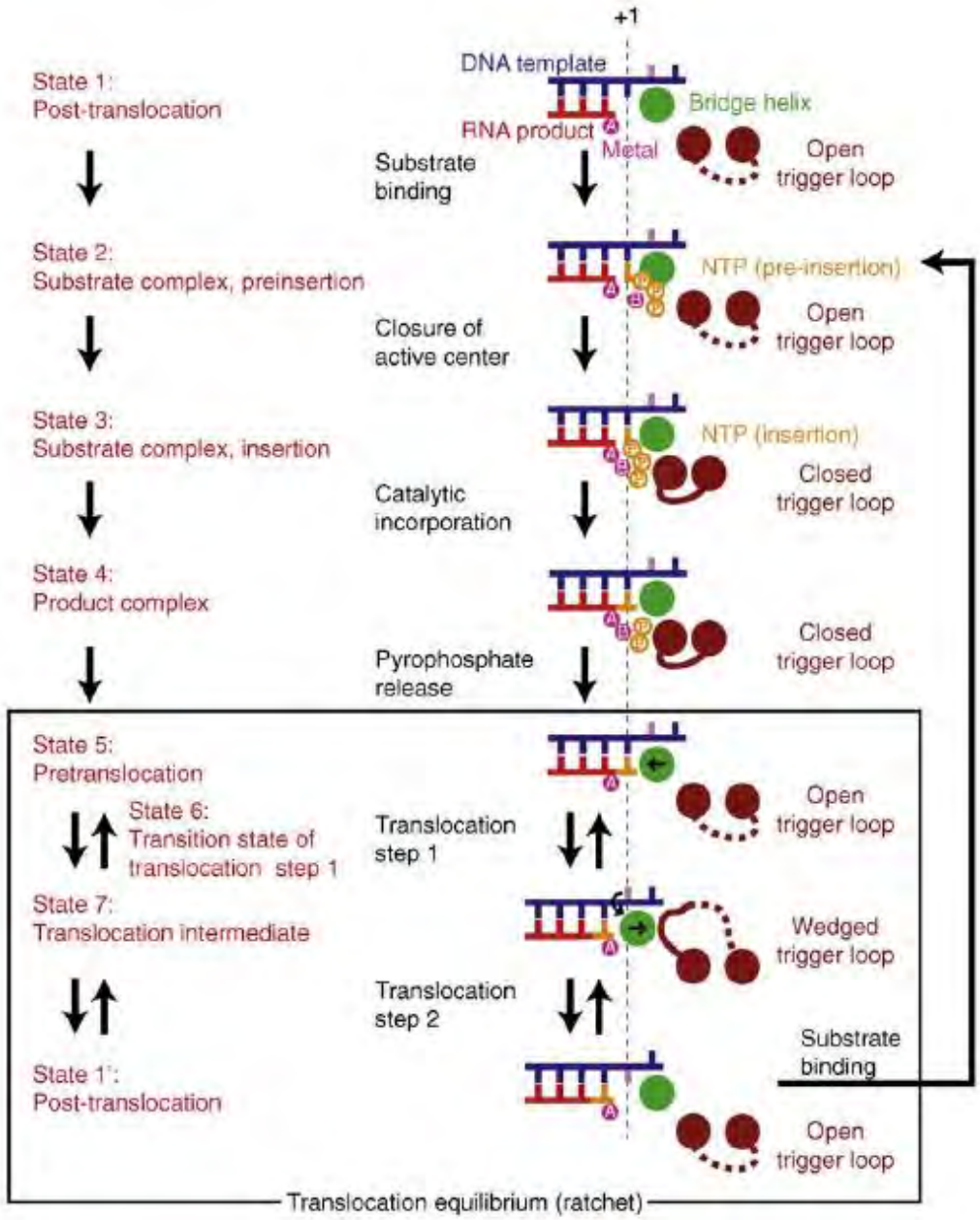


Figure 1.2: Nucleotide addition cycle. Diagram was adapted from Florian Brueckner., et al 2008

During the NAC, a nucleoside triphosphate binds to the transcribing elongation complex (EC), which is formed by Pol II, DNA and the elongating RNA (Gnatt et al., 2001). The insertion and catalytic addition of the nucleotide to the 3' end of the elongating RNA would lead to the formation of a pyrophosphate ion. The release of the pyrophosphate leads to the pre-translocation, which means that the newly added nucleotides at the 3'-terminal side still stays at the substrate site and a new free nucleotide is not allowed to incorporate. To free the active center out for the next NTP binding, the DNA and RNA molecule slide along Pol II and translocate with the help of bridge helix and trigger loop (Naji et al., 2008). However, if Pol II was bound by α-amanitin at its active center at this stage, the small cyclic peptides would trap the bridge helix and the trigger loop and prevent Pol II translocation, which ends

up with an abortive transcription and death of cells because of transcription deficiency (Figure 1.2)(Brueckner and Cramer, 2008).

In 2002, the first crystal structure of yeast Pol II-α-amanitin was solved and the binding pocket of α-amanitin on Pol II was defined (Bushnell et al., 2002). In 2008, Florian Brueckner and Patrick Cramer solved the crystal structure of yeast Pol II elongation complex inhibited by α-amanitin (Brueckner and Cramer, 2008). In this structure, the trigger loop was locked by α-amanitin in a very special translocation intermediate state, which was defined as 'wedged trigger loop'. The wedged trigger loop helped to elucidate the translocation process in NAC. The binding pocket was also better defined in this structure. However, the binding affinity of α-amanitin to mammalian Pol II is more than 1000 times higher than yeast (Table 1.1)(Wienland and Faulstich, 1991), and also, animals and human beings are normally the targets that the mushrooms need to protect themselves from. As yeasts and mushrooms both belong to the fungus family, yeast should not be the natural target of α-amanitin. To figure out why the binding affinity is much higher in its natural target and whether the binding position is the same in its natural target, we decided to solve the structure of mammalian Pol II bound by α-amanitin in this study. As cryo-EM is a well-known method for solving protein structures nowadays, we were also curious whether it is possible to solve a small molecule bound to a protein complex binding pocket by solving the structure to near atomic resolution.

2 Materials and Methods

2.1 Materials

Name	Composition	Application
100 × PI	1mM leupeptin 2mM pepstatin A 100mM phenylmethylsulfonyl fluoride 280mM benzamidine	Protein purification
hGdown1 Lysis buffer	50mM Hepes pH7.5 10mM Imidazole 300mM NaCl 1mM $CaCl_2$ 10% (V/V) glycerol 1 × PI 1mM DTT	hGdown1 purification
hGdown1 wash buffer 1	50mM Hepes pH7.5 30mM Imidazole 300mM NaCl 1mM $CaCl_2$ 10% (V/V) glycerol 1 × PI 1mM DTT	hGdown1 purification

hGdown1 wash buffer 2	50mM Hepes pH7.5 50mM Imidazole 300mM NaCl 1mM $CaCl_2$ 10% (V/V) glycerol 1 × PI 1mM DTT	hGdown1 purification
hGdown1 Elution buffer	50mM Hepes pH7.5 30mM Imidazole 300mM NaCl 1mM $CaCl_2$ 10% (V/V) glycerol 1 × PI 1mM DTT	hGdown1 purification
0M HepR Buffer	50mM Tris-HCl, pH7.9@4°C 1mM EDTA pH8.0 10uM $ZnCl_2$ 10% (V/V) glycerol 1 × PI	Pol II purification
0.6M HepR Buffer	0.6M Ammonium sulfate 50mM Tris-HCl, pH7.9@4°C 1mM EDTA pH8.0 10uM $ZnCl_2$ 10% (V/V) glycerol	Pol II purification

	1 × PI	
0.15M HepR Buffer	0.15M Ammonium sulfate 50mM Tris-HCl, pH7.9@4°C 1mM EDTA pH8.0 10uM $ZnCl_2$ 10% (V/V) glycerol 1mM DTT 1 × PI	Pol II purification
0.2M HepR Buffer	0.2M Ammonium sulfate 50mM Tris-HCl, pH7.9@4°C 1mM EDTA pH8.0 10uM $ZnCl_2$ 10% (V/V) glycerol 1mM DTT 1 × PI	Pol II purification
0.4M HepR Buffer	0.4M Ammonium sulfate 50mM Tris-HCl, pH7.9@4°C 1mM EDTA pH8.0 10uM $ZnCl_2$ 10% (V/V) glycerol 1mM DTT 1 × PI	Pol II purification
0.5M HepR Buffer	0.5M Ammonium sulfate	Pol II purification

	50mM Tris-HCl, pH7.9@4°C	
	1mM EDTA pH8.0	
	10uM $ZnCl_2$	
	10% (V/V) glycerol	
	1mM DTT	
	1 × PI	
S-300 Buffer	5mM Hepes pH7.25@25°C	Pol II purification
	150mM NaCl	
	10uM $ZnCl_2$	
	10mM DTT	
	1 × PI	
Dilution buffer	50mM Tris-HCl pH7.6	Pol II purification
	1mM EDTA pH8.0	
	10uM $ZnCl_2$	
	2mM DTT	
	1 × PI	
Transcription buffer	20mM Na-Hepes pH7.5	Transcription assay
	60mM $(NH_4)_2SO_4$	
	8mM $MgSO_4$	
	10μM $ZnSO_4$	
	10mM DTT	
	10% (v/v) glycerol)	
Stop Buffer	50mM EDTA	Transcription assay

	6.4M Urea, 1-fold TBE (Sigma-Aldrich)	
Template DNA	5'-GATCAAGCTCAAGTACTTAAGCCT GGTCTATACTAGTACTGCC-3'	EC formation
Non-template DNA	5'-GGCAGTACTAGTATTCTAGTATTG AAAGTACTTGAGCTTGATC-3'	EC formation
RNA	5'-UAUAUGCAUAAAGACCAGGC-3'	EC formation

20% denaturing Urea gel	Urea	8M
	TBE buffer	1x
	Bis:Acrylamide 19:1	20%
	TEMED	10µL to 10ml
	APS	0.025% (w/v)

2.2 Methods

2.2.1 Expression and purification of human Gdown1 (hGdown1)

Purification of hGdown1 was performed as described before (Bernecky et al., 2016a). Gene-optimized hGdown1 (Life Technologies) was cloned into pOPINB (N-terminal His6 tag and 3C protease site). The vector was transformed to BL21(DE3)RIL competent cells and plated on LB agar plate and cultured overnight in 37°C incubator. Single colony was picked and cultured in LB medium with kanamycin and chloramphenicol overnight at 37°C while shaking at 160rpm. The overnight *E.coli* cells were cultured in 2L LB medium (with kanamycin and chloramphenicol) at 37°C for 3 to 4 hours till the OD600 arriving to 0.6 to 0.8, then the protein was expressed by inducing with 0.5mM IPTG for 3 to 4 hours at 37°C. The cells were harvested at a speed of 6000rpm for 15 minutes. The supernatant was gently discarded and

the pellet was re-suspended in hGdown1 lysis buffer, frozen in liquid nitrogen, and kept at 80°C for purification.

For purification, the re-suspended cells were transferred to a metal beaker for sonication with power 20%, 0.6 on, 0.4 off settings for 10 minutes. The sonicated material was transferred to 2 centrifugation tubes and spun down for 30 minutes at 4° with Beckman A27 rotor and a speed of 15,000rpm. The supernatant was transferred to a new tube and filtered with 0.8uM filter. The filtered supernatant was loaded to 5ml HisTrap™ High Performance column (GE Healthcare) which was pre-equilibrated with hGdown1 lysis buffer. The column was washed with 10 CV of hGdown1 lysis buffer, 5CV of hGdown1 wash buffer 1 and eluted with 5CV of hGdown1 elution buffer. The eluted protein was mixed with TEV protease (1:10 ratio of TEV and protein) and dialyzed to hGdown1 lysis buffer overnight. The next day the protein was centrifuged at 27,000rpm for 10 minutes to remove the possible precipitation after cleavage. The supernatant was loaded to Ni column which was equilibrated with lysis buffer beforehand. The flow through was collected, the column was washed with hGdown1 wash buffer 2 and the washed buffer was also collected. The protein was eluted from Ni column and loaded to monoS (GE Healthcare) column. Column was washed for 10CV with wash buffer 1 and eluted with a NaCl gradient from 0M to 1M. The peak fractions were identified with SDS-PAGE. The target protein was pooled, concentrated and loaded to gel filtration. Column Superdex 200 10/300 GL (GE Healthcare) was used for gel filtration. The peak fractions were identified again with SDS-PAGE and the target fractions were pooled and concentrated to 2 to 3mg/ml with Amicon Ultra-15 Centrifugal Filter Unit (10 kDa MWCO) (Merck KGaA, Germany). The final protein solution was centrifuged at maximum speed for 2 minutes and aliquoted as 5ul aliquots, frozen in liquid nitrogen and stored at -80°C ready for use.

2.2.2 Purification of *Sus scrofa* Pol II

Sus scrofa Pol II was purified essentially as described for the bovine Pol II preparation (Hodo and Blatti, 1977; Thompson et al., 1990). 500g frozen pig thymus were crashed into pieces with a hammer. The broken pieces were added to a pre-chilled Warning blender with 1L 0M HepR buffer and homogenized on high speed for 3 minutes. The homogenized material was centrifuged with SLA-1500 rotor at 11,000rpm for 20 minutes at 4°C. Unless special emphasis, all the steps below were carried out at 4°C. The supernatant was filtered with 2 layer of miracloth into a chilled glass graduated cylinder and then transferred to a chilled 2L beaker with stirring bar. 5% polyethylenimine (PEI, Sigma-Aldrich) was slowly added to a final concentration of 0.02% while stirring. The stirring was kept at 4°C for at least 10 minutes. Then the precipitated material was transferred to the centrifugation tubes and centrifuged with SLA-1500 rotor at 11,000rpm for 20 minutes. The pellet from the centrifugation was fully re-suspended with 0.15M HepR buffer and centrifuged with SLA-1500 rotor at 11,000rpm for 20 minutes. At the same time, the MacroPrep Q column was washed with 2 column volume (CV) water, 2CV 0M HepR buffer, 3CV 0.6M HepR buffer and equilibrated with 2CV 0.2M HepR buffer. After centrifugation, the supernatant was adjusted to the conductivity of 0.2M HepR buffer and loaded to MacroPrep Q column with a very slow

flow rate (gravity flow). After loading, the column was washed with 3CV of 0.2M HepR buffer before eluting with 3CV of 0.4M HepR buffer. The eluted fraction was precipitated slowly with finely ground ammonium sulfate till saturation while stirring. The stirring was kept for at least one hour at 4°C before centrifuging with F14 rotor at 15,000rpm for 30 minutes. After centrifugation, the supernatant was gently removed and the pellet was dissolved in 0M HepR buffer with 1mM DTT. This was named as 'Ab input'. The conductivity of 'Ab input' was adjusted to match the conductivity of 0.15M HepR buffer and followed by a centrifugation at 15,000rpm for 30 minutes. The 8WG16 (αRPB1 CTD) antibody-coupled Sepharose column was equilibrated with 0.15M HepR buffer and the Ab input was loaded to the antibody column in gravity flow (the beads bed was not allowed to be disturbed during the whole process). The column was washed with 0.5M HepR buffer (also in gravity flow) and then moved to room temperature. The antibody column was kept at room temperature for at least 15 minutes to make sure the resin is at room temperature. Then the protein was eluted with 0.5M HepR buffer plus 50% (v/v) glycerol. The eluted drops were collected with 50ml conical tubes containing 20ml dilution buffer. The elution was fractionated every 5ml, in total 5 fractions were collected. After elution, all the fractions were identified with SDS-PAGE. Fractions with Pol II were collected and loaded to UnoQ column (Biorad) which was equilibrated with 0.1M HepR buffer beforehand. UnoQ column was washed with 5CV of 0.1M HepR buffer and eluted with a linear gradient from 0.1M HepR to 0.5M HepR. The peak fractions were taken and loaded to SDS-PAGE gel. Fractions without pig Gdown1 were pooled, 3-fold molar excess of hGdown1 was added and kept on ice for 2 to 3 hours. Then the sample was loaded to a HiPrep 16/60 Sephacryl S-300 HR column (GE Healthcare, Little Chalfont, United Kingdom). Peak fractions were identified with SDS-PAGE and fractions containing the Pol II-hGdown1 complex were collected and concentrated to a concentration of 2-3 mg/ml using an Amicon Ultra-15 Centrifugal Filter Unit (100 kDa MWCO) (Merck KGaA, Germany). Sample aliquots were snap frozen in liquid nitrogen and stored at -80 °C prior to use. The typical yield is about 2-4 mg from ~500 g pig thymus.

2.2.3 SDS-PAGE

SDS-PAGE was performed by using pre-cast NuPAGE Bis-Tris 4-12% gels (Invitrogen). 4x NuPAGE LDS loading buffer (Invitrogen) was added to the Protein sample to a final concentration of 1x. The samples were boiled at 95°C for 5 to 10 minutes and loaded carefully to the wells of the gel. At least one well of one gel should be loaded with protein marker (precision plus protein^{TM} Dual Color Standards, BIO-RAD). Gels were run in either 1xMES or 1xMOPS buffer (diluted from NuPAGE 20 x stock, Invitrogen. For small proteins, MOPS buffer has better resolution) for 30 to 60 minutes at 200V. After running, gels were taken out and stained with InstantBlue (Expedeon). The destaining of the gels was performed with water and the gel was scanned with Epson Perfection V700 Photo Fachbettscanner.

2.2.4 Formation of elongation complex (EC)

The DNA scaffold used for the EC is the same as the one used for the bovine RNA polymerase II-DSIF complex (Bernecky et al., 2016a). A 20nt RNA was used for the formation

of a 9nt DNA-RNA hybrid and 11nt of exiting RNA. The template DNA/RNA was annealed (Brueckner et al., 2007) and a 1.5 fold molar excess of scaffold was added to the Pol II-hGdown1 complex. The sample was incubated on ice for 10 min and subsequently incubated for an additional 15 min at 20 °C while shaking at 550rpm. Then the non-template DNA was added and the sample was kept at 20°C for another 20 minutes. The complex was crosslinked with 3mM BS3 (Thermo Scientific, final concentration) on ice for 30 min. The crosslinking reaction was quenched with 50mM ammonium bicarbonate and applied to a Superdex 200 increase 10/300 GL column (GE Healthcare) pre-equilibrated with S-300 buffer. The peak containing the complex was pooled and concentrated to a concentration of 473 µM. A 1.5-fold molar excess of α-amanitin was added to the elongation complex. The sample was incubated on ice for 20 min and then loaded directly to the grids.

2.2.5 Electron microscopy

4 µL of the protein complex solution was applied to glow-discharged Quantifoil R2/2 gold grids (Quantifoil) and plunged into liquid ethane after blotting with a FEI Vitrobot Mark IV (FEI, Hillsboro, USA). Images were acquired on a FEI Titan Krios, operated at 300 keV and equipped with a Gatan K2 Summit direct electron detector and a Quantum GIF. Micrographs were collected automatically with the software package EPU (FEI) at a nominal magnification of 130k (1.07 Å per pixel) in counting mode. The dose rate was 3.8 e-/pixel/s. Three images were acquired per foil hole. Each micrograph was collected with a total dose of 35 electrons per square angstrom over a 10-second exposure, fractionated into 40 frames (0.25 s each). Defocus values ranged from −1 to −3 µm. Micrograph frames were aligned and corrected with MmotionCcorr2 (Zheng et al., 2017). Unless otherwise noted, data processing was performed using RELION 2.1 (Fernandez-Leiro and Scheres, 2017). Contrast transfer function (CTF) parameters were estimated using Gctf (Zhang, 2016). Initial 2D classes were calculated from 2,909 manually selected particles from 37 micrographs. The initial 2D classes were used as templates for auto-picking. After manual inspection of all 2,049 micrographs, a total of 207,410 particles were obtained. Two rounds of 2D classification were performed and bad particles were removed. The resulting data set of 134,512 particles was used for further refinement and focused classification refinement in 3D. The *Bos taurus* Pol II structure (EMDB accession code EMD-3219) (Bernecky et al., 2016a) was low-pass filtered to 40 Å as an initial model for 3D refinement. Initial 3D refinement followed by movie processing and particle polishing yielded a reconstruction at an overall resolution of 3.4 Å (gold-standard Fourier shell correlation criterion 0.143, RELION 2.1). Focused 3D classification without image alignment was performed on the α-amanitin binding pocket, the Pol II stalk (RPB4-RPB7) and upstream DNA, followed by global 3D refinement.

2.2.6 Model building and refinement

Model building was based on the previously published bovine Pol II structure (PDB accession code 5FLM)(Bernecky et al., 2016a). The model was manually fitted in COOT (Emsley et al., 2010). The α-amanitin molecule was taken from a *Saccharomyces cerevisiae* α-amanitin-bound Pol II structure (PDB accession code 2VUM)(Brueckner and Cramer, 2008). The α-amanitin molecule was rigid body fitted into the density. The structure was refined in real

space with special restraints to the nucleic acids and α-amanitin using PHENIX (Torices and Muñoz-Pajares, 2015).

2.2.7 Transcription assay

Template DNA and RNA were mixed at a molar ratio of 1:1 and annealed as described (Brueckner et al., 2007). The template annealed DNA-RNA was mixed with Pol II-hGdown1 complex at a molar ratio of 1:2 and incubated at 28 ℃ for 10 min. Non-template DNA was added and incubated at 28 ℃ for an additional 10 min. The elongation complex was mixed with α-amanitin or buffer (control) at the same molar ratio used for the complex formation. The sample was subsequently incubated on ice for 20 min. 100 μM UTP was added to both control and experimental reactions. The reaction was incubated in transcription buffer at 28 ℃ and samples were taken at indicated time points. The reaction was stopped by adding stop buffer to the reaction. The product RNA was separated using a 20% denaturing urea polyacrylamide gel (300V) and visualized using a GE Typhoon FLA 9500 (GE Healthcare).

3 Results

3.1 Purification of *Sus scrofa* Pol II

Sus scrofa Pol II was purified essentially as described for the bovine Pol II preparation, except that pig thymus instead of bovine thymus was used (Method). Briefly, thymus was homogenized, and the supernatant was filtered. After polyethyleneimine precipitation, Pol II was purified with a MacroPrepQ column, followed by ammonium sulfate precipitation and an affinity column with 8WG16 (αRPB1 CTD) antibody-coupled Sepharose, a UnoQ anion exchange column, and finally a Sephacryl S-300 HiLoad sizing column (Figure 1.3). The typical yield was 2~4 mg from ~500 g of thymus. The fractions from UnoQ column were strictly selected to avoid pig Gdown1 contamination, then hGdown1 which was expressed and purified from *E.coli* was combined with Pol II and purified by gel filtration. Incubation of Pol II and hGdown1 on ice can form a stable Pol II-hGdown1 complex (Figure 1.4A).

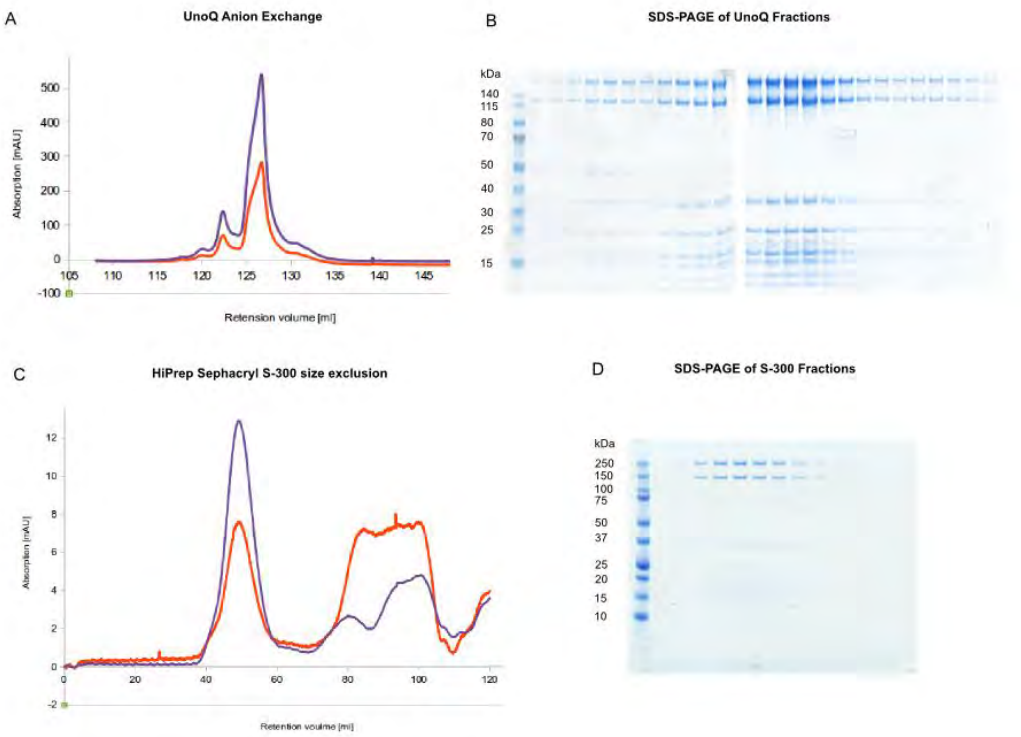


Figure 1.3: Purification of *Sus. Scrofa* Pol II from pig thymus and formation of Pol II-hGdown1 complex. A, chromatogram of UnoQ column, Pol II was concentrated by UnoQ with a high peak coming out within the elution gradient. B, SDS-PAGE of UnoQ fractions, the volume increases from left to right, the earlier Pol II fractions including pig Gdown1 were trashed. C, chromatogram of HiPrep Sephacryl S-300 column. The volume of the column is 120ml. Pol II comes out around 50ml. D, SDS-PAGE of the gel filtration fractions. hGdown1 was bound to Pol II stably after incubation.

3.2 Pol II elongation complex formation, assay of activity inhibition by α-amanitin and cryo-EM grids preparation

The EC was formed with a DNA-RNA scaffold that was highly similar to a previously used one (Bernecky et al., 2016a). The EC was active in RNA synthesis and was inhibited after α-amanitin addition (Figure 1.4B). The EC sample was cross-linked with BS3, incubated with α-amanitin, and immediately applied to EM grids before flash freezing. Cryo-EM analysis revealed a homogeneous distribution of particles that could be classified easily (Figure 1.4C). 134,512 particle images were extracted with RELION and used for 3D reconstruction, resulting in a cryo-EM density map at a nominal resolution of 3.4 Å (Figure 1.5)

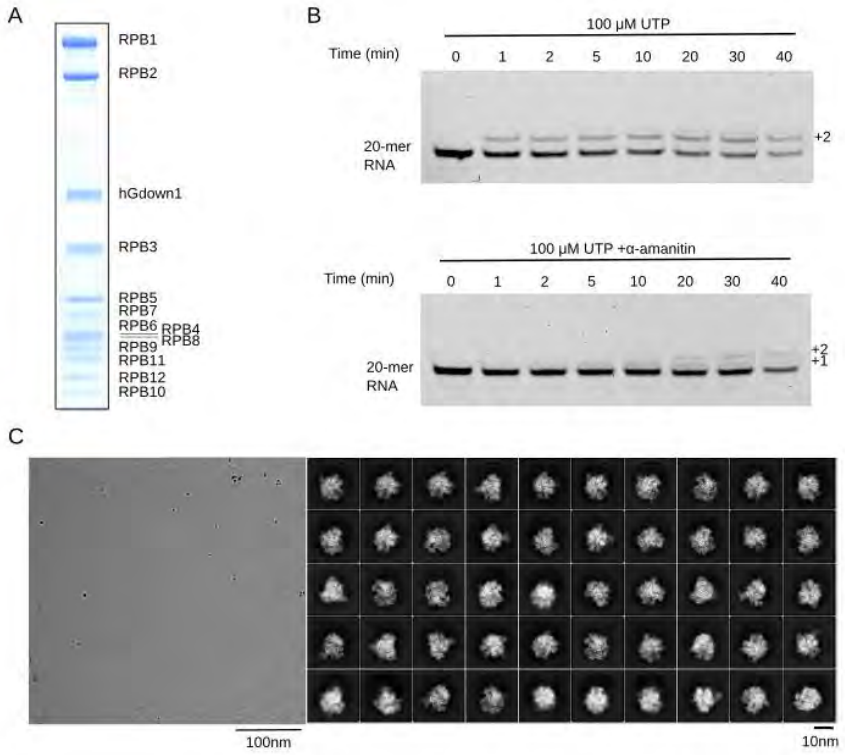


Figure 1.4: Pol II elongation complex (EC) formation, in vitro RNA extension assay, and exemplary micrograph and 2D classes of the dataset. A, SDS-PAGE analysis of the Pol II-hGdown1 complex. B, the reconstituted Pol II-hGdown1 EC is active in RNA extension and inhibited by α-amanitin. In the absence of α-amanitin (upper panel), two uridine residues were incorporated into the RNA of the scaffold upon incubation with 100mM UTP, as expected from the presence of two templating adenine bases downstream. In the presence of α-amanitin (lower panel), nucleotide addition is slowed down, and addition of only one uridine residue was observed, as expected from impaired Pol II translocation. C, representative micrographs and 2D classes generated from the cryo-EM data set.

3.3 Pol II EC-hGdown1-α-amanitin complex data processing

Particles were extracted from micrographs with RELION. Three rounds of 2D classification were executed to sort out the bad particles. Good particles were saved for further data processing. Bovine Pol II (EMD-3219) was filtered to 40 Å as an initial model for 3D refinement and 3D classification. Focus classifications and refinements were used for upstream DNA, α-amanitin pocket and RPB4/7 respectively to improve the model (Figure 1.5). Local resolution was measured along with angular distribution. The resolution of the α-amanitin pocket was about 3 Å (Figure 1.6), which was higher than the overall resolution because of stability.

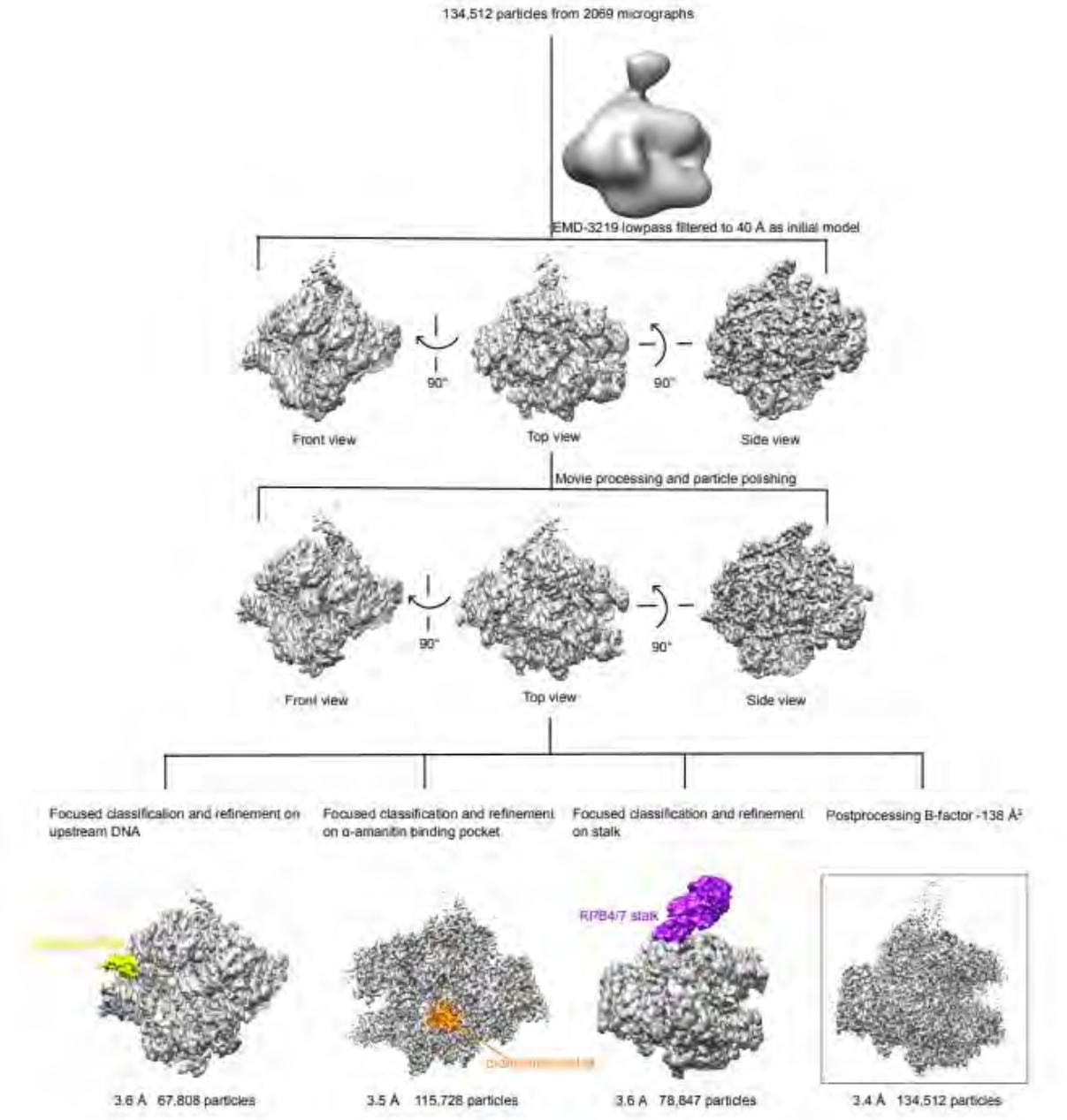


Figure 1.5: Cryo-EM data processing. The structure of bovine Pol II (EMD-3219) was low-pass filtered to 40 Å and used as the initial reference model. Semi-automatically picked particles were used for 3D refinement. Data processing with 3D refinement, movie processing and particle polishing gave a final reconstruction at a nominal resolution of 3.4 Å. Focused classifications and refinements were performed on upstream DNA, α-amanitin and its binding pocket, and the Pol II stalk sub-complex RPB4-RPB7.

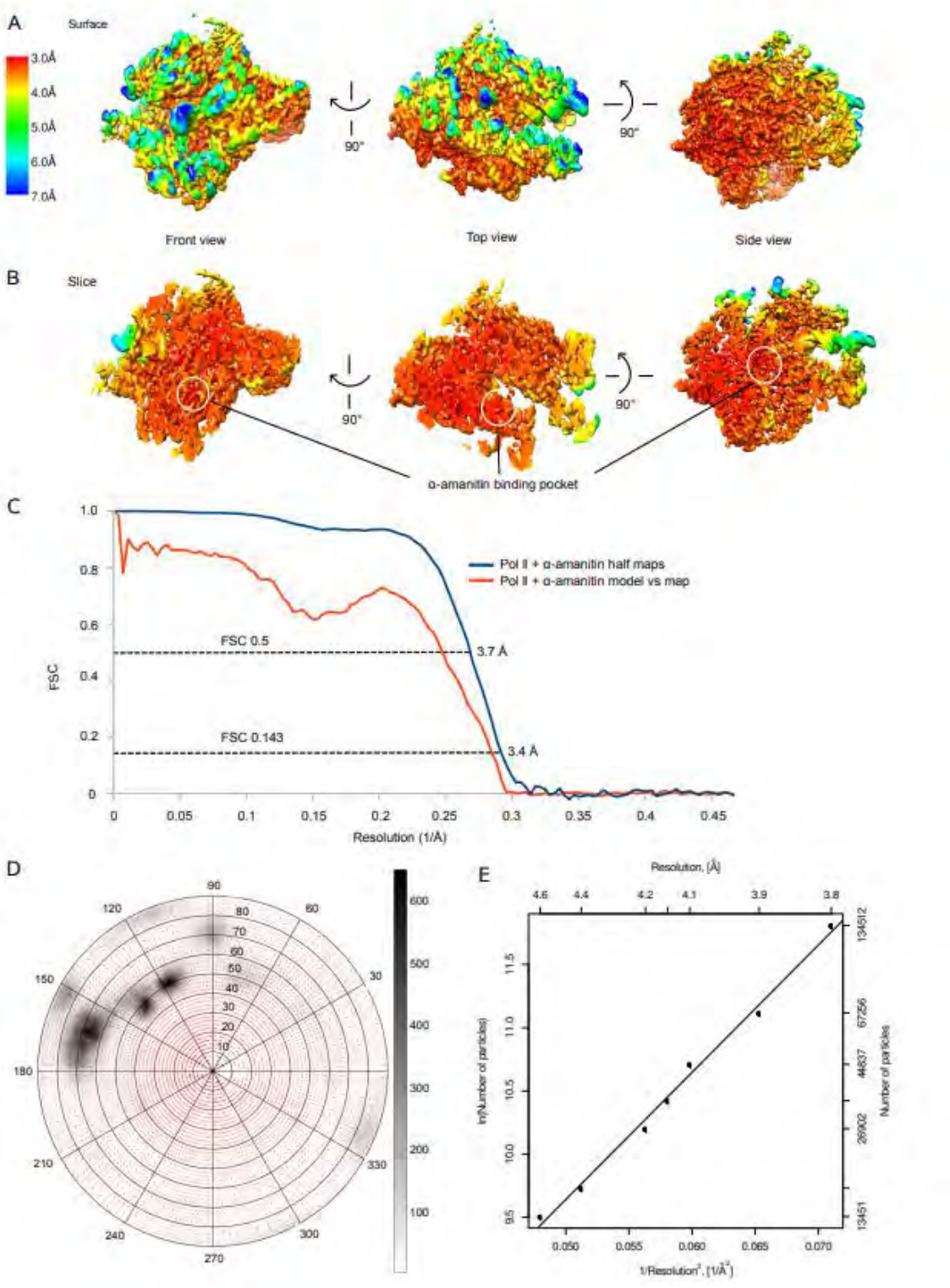


Figure 1.6: Local resolution of the cryo-EM density map. A, Three views of a surface representation of the final cryo-EM density map colored according to local resolution. B, The same views as in 'A' but sliced open to reveal the very high resolution at the active center of the polymerase and around the α-amanitin binding pocket. C, FSC plots for the cryo-EM reconstruction and for the model versus the cryo-EM reconstruction. D, angular distribution map of single particle images. Black shading indicates the number of particles assigned to a

given view, while red dots indicate represented views. E, Resolution versus number of particles plot using random particle subsets with logarithmic and squared reciprocal axes. The slope of the linear fit indicates an overall B-factor of 101 $Å^2$.

3.4 Overall structure analysis of mammalian Pol II EC-α-amanitin complex and comparison with yeast EC-α-amanitin complex

To obtain an atomic model of the mammalian Pol II EC α-amanitin complex, we placed the previously refined bovine Pol II structure into the density and adjusted it locally (Bernecky et al., 2016a). There was no density for hGdown1, which apparently dissociated from the complex. The region around the Pol II active center, including α-amanitin and its binding pocket, was well resolved, with an estimated local resolution of 3.0 Å (Figure 1.6). There were no other significant additional densities observed. We could build an atomic model for α-amanitin and define its chemical interactions with Pol II (Figure 1.8 and Table 1.2). The structure was finished by manual adjustments and real-space refinement. The structure of the Pol II EC is highly similar to the previously determined structure of the bovine counterpart (Bernecky et al., 2016a).

Pig Pol II differs from bovine Pol II in only five residues: RPB1 Glu1968, RPB5 Glu32 and Asp46, RPB6 Ser126, and RPB9 Phe11. The EC adopts the post-translocation state with a straight bridge helix, different from the slightly bent bridge helix observed in the yeast Pol II-α-amanitin crystal structure, which is thought to reflect a translocation intermediate (Brueckner and Cramer, 2008). The trigger loop adopts a conformation that most closely resembles the 'wedged' conformation previously observed in the yeast EC bound by α-amanitin (Brueckner and Cramer, 2008). However, residue Leu1104 (Leu1081 in yeast), which forms a wedge behind the bridge helix in the yeast structure (Brueckner and Cramer, 2008), protrudes 2 Å less in between the bridge helix and the polymerase cleft module, essentially not forming a wedge anymore, and consistent with the observed straight bridge helix. We refer to this slightly altered trigger loop conformation as 'unwedged' because it is likely that it is adopted after the wedged conformation and before the addition of the next nucleotide.

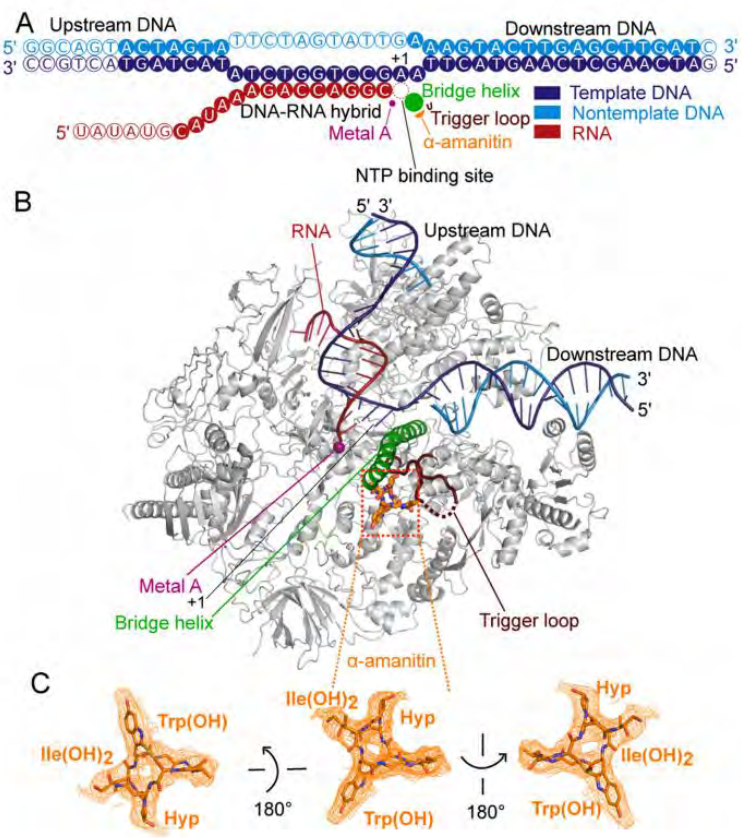


Figure 1.7: Cryo-EM structure of mammalian Pol II EC bound by α-amanitin. A, nucleic acid scaffold is depicted schematically. Filled and unfilled circles represent modeled and not modeled nucleotides, respectively. The nucleotide-binding site (red dashed circle), bridge helix (green), the catalytic metal ion A (pink), trigger loop (brown), and α-amanitin (orange) are indicated. The color code is used throughout. B, Overview of the structure. Pol II is shown as a silver ribbon model, and other elements are colored as in A. C, electron density for α-amanitin (orange mesh) in three different views. Important contact moieties with Pol II are indicated. Nitrogen, oxygen, and sulfur atoms are blue, red, and yellow, respectively.

3.5 Specificity of α-amanitin binding pocket in mammalian

The position and binding pocket of α-amanitin is as observed in the yeast EC (Figure 1.7) (Brueckner and Cramer, 2008). Most contacts between α-amanitin and yeast Pol II observed in the EC are conserved in the mammalian complex, as expected by the high conservation of residues involved in binding the toxin (Figure 1.8). Conserved contacts are also formed by His1108 (yeast His1085) in the trigger loop of Pol II.

Three differences in α-amanitin-Pol II contacts are observed. First, the side chain of RPB1 residue Ser782, conserved over mammals and other metazoan species, forms an additional hydrogen bond with the hydroxyl group in the indole ring of the tryptophan in α-amanitin (Figure 1.3; figure 1.4A). Ser782 lines the bottom of a cage, formed by the universally conserved Pol II residues Arg749, Ile779, and Gln783, for the indole ring of α-amanitin (Figure 1.8, A and B). The yeast residue corresponding to mammalian Ser782 is Ala759 and cannot form this hydrogen bond. Second, Asn792 forms an additional hydrogen bond with

its side chain to the backbone carbonyl group of 4,5-dihydroxyisoleucine in α-amanitin (Figure 1.8B, C; Figure1.9). This contact is not present in the yeast Pol II-α-amanitin complex, because the yeast counterpart of mammalian Asn792 is Ser769, and the observed hydrogen bond is thus not possible. There is a third residue in the amanitin-binding pocket that differs, Asn742 (Figure 1.8, A and B; Figure1.9), which corresponds to Val719 in yeast, but this is unlikely to contribute strongly to the difference in affinity because in both structures these residues form van der Waals contacts with the side chain of isoleucine in α-amanitin.

Thus, compared with the yeast structure, two additional hydrogen bonds are formed between α-amanitin and the mammalian EC. It is known that two additional hydrogen bonds can give rise to enthalpy changes that account for changes in dissociation constants by 3 orders of magnitude (Hubbard and Kamran Haider, 2001; Klebe, 2015).We therefore suggest that the two additional hydrogen bonds account for the much higher affinity of mammalian Pol II for the toxin. This interpretation is supported by known biochemical data obtained with amanitin derivatives that lack certain functional groups (Baumann et al., 2008; Kinghorn, 1987). In particular, alkylation of the hydroxyl group in the indole ring is predicted to prevent hydrogen bond formation and is known to decrease toxicity and inhibitory potential of amanitin (Kinghorn, 1987).

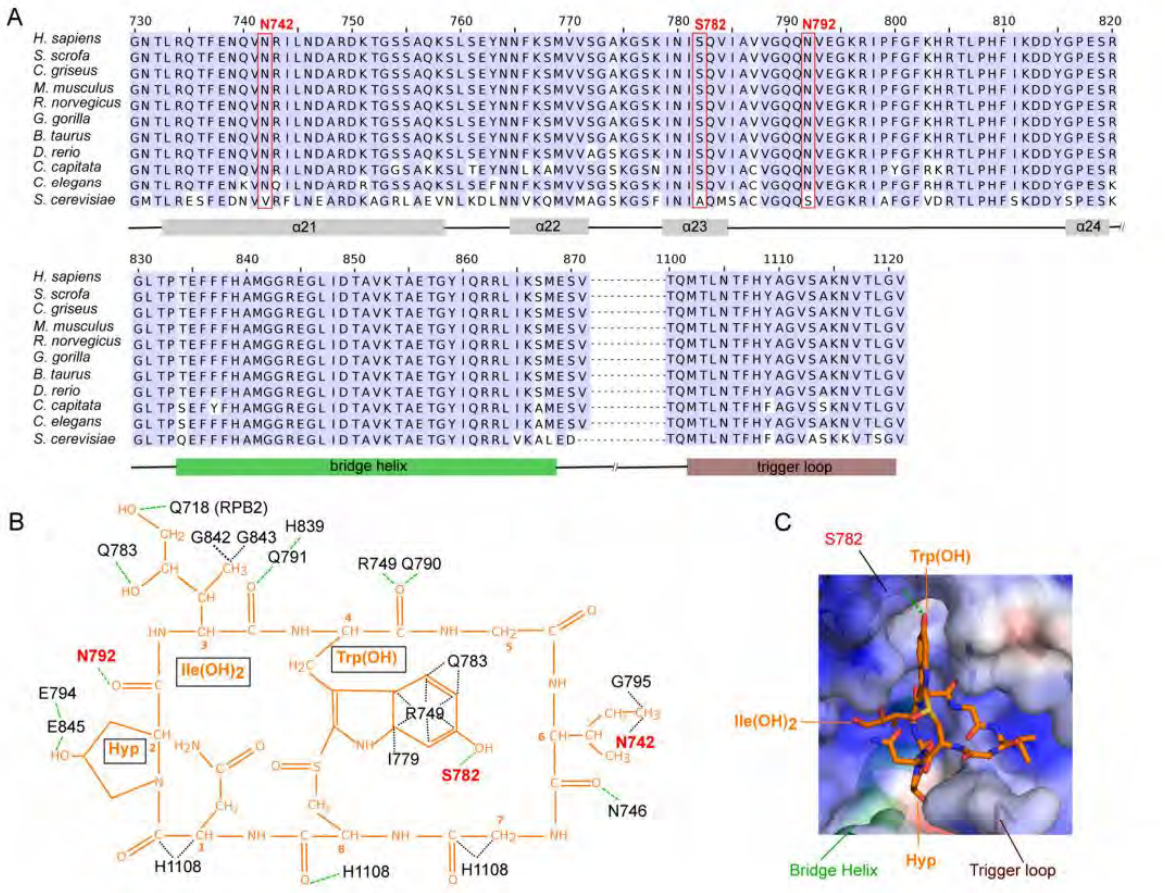


Figure 1.8: Interaction analysis of mammalian Pol II with α-amanitin. A, sequence alignment of residues forming the α-amanitin-binding pocket in RPB1 between various

metazoan species and the yeast *S.cerevisiae* (bottom row). The red boxes indicate amino acid residues that form metazoan-specific interactions with α-amanitin. Helices α21 to α24, bridge helix, and trigger loop are indicated at the bottom of the sequence alignment. B, schematic overview of Pol II-α-amanitin interactions. The chemical structure of α-amanitin is shown in orange. RPB1 residues conserved over eukaryotes are labeled in black, whereas metazoan-specific α-amanitin-interacting residues are labeled in red. The green dashed lines indicate hydrogen bonds, whereas black dashed lines show other interactions. C, surface representation of the amanitin-binding Pol II pocket. Positively and negatively charged surfaces are in blue and red, respectively. The bridge helix, trigger loop, and RPB1 residue Ser782 are indicated.

α-Amanitin residue	α-Amanitin atom	Pol II residue (RPB1)	Pol II atom	Length	Present in yeast Pol II EC–α-amanitin complex
				Å	
2	OD(D)	Glu^{845}	OE1(A)	3.1	Yes
2	O(A)	Asn^{792}	N(D)	3.6	Yes
2	O(A)	Asn^{792}	ND2(D)	2.8	No
3	OD(A)	Gln^{718} (RPB2)	NE2(D)	2.8	Yes
3	OG(A)	Gln^{783}	NE2(D)	3.5	Yes
3	O(A)	Gln^{791}	NE2(D)	3.6	Yes
4	O(A)	Arg^{749}	NE	3.6	Yes
4	O(A)	Gln^{790}	N(D)	2.5	Yes
4	OH2	Ser^{782}	OG	2.5	No
6	O(A)	Asn^{746}	ND2(D)	3.7	Yes
8	O(A)	His^{1108}	NE2(D)	3.6	Yes

Table 1.2: Hydrogen bonds between α-amanitin and *S. scrofa* Pol II

3.6 α-amanitin resistance caused by binding pocket mutations

The structure also suggests the molecular basis for α -amanitin resistance arising from mutations in the binding pocket in Pol II enzymes from mice (Bartolomei and Corden, 1987; Bartolomei and Corden, 1995) and *Drosophila* (Chen et al., 1993). Modeling shows that mutation I779F in mouse RPB1 leads to a steric clash that likely prevents α-amanitin from binding (Figure 1.9B). The additional mouse mutations L745P and R749P likely destabilize helix 21, which forms part of the binding pocket (Figure 1.9B). The Drosophila melanogaster Rpb1 mutations N792D and N793D are predicted to disrupt hydrogen bonds between Pol II and α-amanitin, thereby decreasing affinity (Chen et al., 1993).

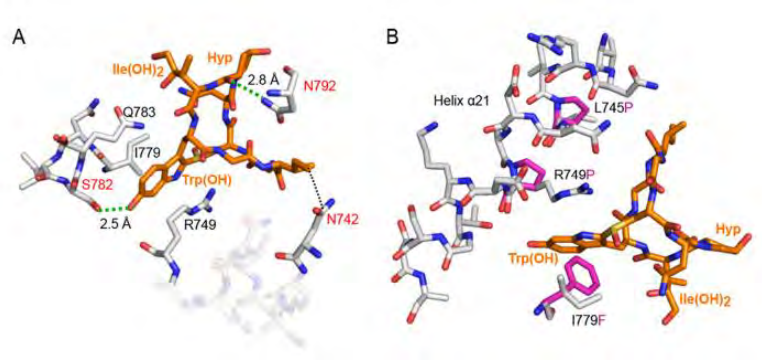


Figure 1.9: Extra hydrogen bonds in mammalian and binding pocket mutation analysis. A, two metazoan-specific hydrogen bonds are indicated with green dashed lines, and the corresponding bond lengths are indicated between α-amanitin and mammalian RPB1. B, modeling of site-specific mutations in the α-amanitin binding pocket that confer resistance to α-amanitin in *Mus musculus*. The Pol II model is shown with gray sticks, whereas the mutated amino acids are shown with magenta sticks.

4. Discussion

In the structure, no density was shown for hGdown1, it might fall off during elongation complex formation or during freezing. However, we need it to make Pol II more homogeneous. Because from the previous experience, even the same fraction from the same gel filtration peak showed a mixture of pol II monomer and dimer, which made the EM processing difficult. However, with hGdown1 binding, the dimer almost disappears and Pol II dimer shows very homogeneous distribution on the grids. The reason is not clear so far, which might need a further study in the future.

More than one century after the discovery of α-amanitin (Ford, 1907), we now provide an atomic model of its structure in complex with its natural target, the mammalian Pol II EC. This work provides the structural basis of mammalian Pol II inhibition by α-amanitin. Whereas insights into the mechanism of transcription inhibition by α-amanitin were already derived from structures of the yeast Pol II (Bushnell et al., 2002) and the yeast EC (Brueckner and Cramer, 2008), our current work additionally provides a molecular explanation for the long-standing observation that α-amanitin has a much higher affinity for mammalian Pol II, compared with the yeast enzyme. Most notably, we observe two additional, well defined hydrogen bonds that are possible in mammalian Pol II enzymes, but not in yeast Pol II, explaining the tighter binding of the toxin to the former. Together with recent studies (Gao et al., 2016; Wong et al., 2017), our work also shows that cryo-EM can now be used to study the detailed interactions of small molecules with proteins, as required for drug design. We note that such applications of cryo-EM still often require that the target molecule or complex has a critical size. In the future, further developments of cryo-EM will, however, likely remove this limitation such that the inhibition of target molecules and complexes of lower molecular weight by small molecules can also be studied.

	Sus scrofa Pol II EC bound by α-amanitin (EMDB-3981; PDB 6EXV)
Data collection and processing	
Magnification	130,000x
Voltage (kV)	300
Electron exposure ($e^-/Å^2$)	35
Defocus range (μm)	1.0-3.0
Pixel size (Å)	1.07
Symmetry imposed	C1
Initial particle images (no.)	207,410
Final particle images (no.)	134,512
Map resolution (Å)	3.4
FSC threshold	0.143
Map resolution range (Å)	3.0-7.0
Refinement	
Initial model used (PDB code)	5FLM
Model resolution (Å)	3.7
FSC threshold	0.5
Model resolution range (Å)	3.0-7.0
Map sharpening *B* factor ($Å^2$)	-138
Model composition	
Non-hydrogen atoms	32,710
Protein residues	3,907
Ligands	α-amanitin (1)
B factors ($Å^2$)	
Protein	53.97
Ligand	56.58
R.m.s. deviations	
Bond lengths (Å)	0.007
Bond angles (°)	0.983
Validation	
MolProbity score	1.77
Clashscore	5.2
Poor rotamers (%)	0.5
Ramachandran plot	
Favored (%)	91.88
Allowed (%)	8.06
Disallowed (%)	0.05

Table 1.3: Cryo-EM data collection, refinement and validation statistics

Chapter2

1 Introduction

1.1 transcription termination

As introduced in chapter1, transcription starts with the recognition of promoter sequences by initiation factors, then both RNAP and initiation factors bind to the promoter and initiate the transcription. In eukaryotes, the transcription of Pol II would suffer from a promoter-proximal pausing before releasing to the gene body (Rougvie and Lis, 1990). The RNAP, stimulated by elongation factors, walks along double strand DNA and produces RNA. When elongation complex encounters termination signals encoded in the DNA sequence, transcription terminates to avoid interfering with the neighboring transcriptional units and to promote RNAP recycling (Kuehner et al., 2011; Richard and Manley, 2009).

Transcription termination means that both RNAP and the transcript dissociate from the template DNA and the transcription of current unit is finished (Porrua et al., 2016; Porrua and Libri, 2015). There are two main reasons for keeping the processivity of the elongation: the interactions between RNAP, elongation factors and nucleotides (Kuehner et al., 2011), and the DNA:RNA hybrid which is 8 nucleotides in length and is maintained during the elongation process (Kireeva et al., 2000; Komissarova et al., 2002). So to dismantle the elongation complex, there are two main processes. Firstly, the abolishment of the interactions between RNAP and elongation factors, which means termination/3' processing factors bind to RNAP or RNA and replace the elongation factors (Mandel et al., 2008). The second important process is the separation of the 8nt DNA:RNA hybrid which stabilizes the elongation complex. Thus, a helicase is necessary to open the DNA:RNA hybrid and cause the collapse of the elongation complex (Porrua and Libri, 2013).

Mechanism of transcription termination is different in different organisms and also varies between Pol I, Pol II and Pol III. A short introduction follows for transcription termination in bacterial, Pol I and Pol III and Pol II respectively.

1.1.1 Transcription termination in bacterial

In bacteria, there are two different termination pathways depending on whether it is factor dependent or it relies only on the signal in the template DNA. The later was named intrinsic termination while the former was named Rho-dependent termination, as the factor Rho is necessary in this pathway (reviewed in Roberts, 2019).

For intrinsic termination, the signal in the template DNA or the product RNA is important and consists of a GC rich hairpin followed by a run of U (d'Aubenton Carafa et al., 1990). In the early termination process, RNAP pauses and an unstable DNA:RNA hybrid is formed. At the same time, the 'U' sequence is synthesized (Gusarov and Nudler, 1999). The synthesis process provides enough time for the formation of the hairpin, which might have several

roles in the termination process (Roberts, 2019). Firstly, the hairpin might initiate the dissociation of the DNA:RNA hybrid with the help of the U tract. Secondly, the hairpin might help to push the bacterial RNAP forward without nucleotide addition, which ends up with the release of RNA and RNAP, and the dissociation of the DNA:RNA hybrid (Gusarov and Nudler, 1999). The third hypothesis is that the formation of the hairpin causes a conformational change in RNAP, which might result in the destabilization of the elongation complex and transcription termination (Lang et al., 1998). This is the allosteric model (Epshtein et al., 2010). While the key point for the first two models is the dissociation of DNA:RNA hybrid, the central idea for the third model is the conformational change in RNAP.

In Rho-dependent pathway, the factor Rho is strictly necessary for termination (Banerjee et al., 2006). Rho is a ring-shaped, homo-hexameric complex, which has RNA binding, translocase and ATP hydrolysis activities. The active form of Rho is an open ring which allows RNA binding to the center of the ring (Roberts, 1969). The RNA binding site is featured by C rich and G poor sequences. Once bound to RNA, the Rho motor translocates towards the 3' end and ultimately catches up with RNAP to dislodge it from DNA (Kuehner et al., 2011).

Termination in eukaryotes is different but also shows conservations with bacteria, for example, the dissociation of DNA:RNA hybrid is important for termination in both bacteria and eukaryotes (Komissarova et al., 2002), which would be introduced as follows.

1.1.2 Transcription termination of Pol I and Pol III

Pol I transcribes the ribosomal RNAs (rRNAs) and Pol III transcribes non-coding RNAs, such as tRNAs, U6 spliceosomal snRNAs etc. For Pol I termination in mammals, the termination signal 'Sal box' is important to stop the elongation and release the RNA chain. The featured sequence for 'Sal box' is AGGTCGACCAGA/TT/ ANTCCG in mouse (Grummt et al., 1985; Kuhn et al., 1988). ‘Sal box’ is recognized by transcription termination factor for Pol I (TTF-1) (Bartsch et al., 1988; Evers et al., 1995). Termination occurs 11bp upstream of 'Sal box' with the help of Pol I and transcript release factor (PTRF)(Mason et al., 1997). Rnt1 is the RNA cleavage factor (Kufel et al., 1999). Some studies showed that Pol I might have similar termination mechanisms like Pol II, such as the torpedo model (Kawauchi et al., 2008). However the detailed mechanism for Pol I termination is not well understood so far.

Pol III can terminate transcription by itself. C11 is one of the subunits of Pol III which mediates the cleavage activity and re-initiation (Whitehall et al., 1994). Subunits C37/C53 can reduce the elongation rate of Pol III after termination signal and lead to release of Pol III and transcripts (Landrieux et al., 2006). The most obvious termination signal is the T stretch 40bp downstream of the mature 3' end of RNAs. Sequences surrounding T tract can also affect the termination efficiency (Cozzarelli et al., 1983).

1.1.3 Transcription termination of Pol II

As mentioned before, Pol II transcribes not only protein coding genes (mRNAs) but also non-coding RNAs (ncRNAs). There are different pathways for mRNAs and ncRNAs termination in yeast and humans, as follows.

1.1.3.1 Pol II termination pathways in yeast

Transcription termination is well studied in yeast. Based on different types of termination factors and product RNAs, there are two different pathways for Pol II termination (Kim et al., 2006): the sen1-dependent pathway (Creamer et al., 2011; Jamonnak et al., 2011), which is normally for ncRNAs, and the Poly(A) dependent pathway, which is for mRNAs (Logan et al., 1987; Whitelaw and Proudfoot, 1986).

The main step in the sen1 dependent pathway is the separation of the DNA:RNA hybrid as discussed before (Steinmetz and Brow, 1996). The hypothesis is that sen1 works as a helicase and unwinds the DNA:RNA hybrid with the help of RNA binding factors, Nrd1 and Nab1 (Arigo et al., 2006; Thiebaut et al., 2006). These three factors comprise the Nrd1-Nab1-Sen1 (NNS) complex which works on ncRNA termination, the pathway is also named as NNS pathway. The disruption of DNA:RNA hybrid by sen1 is ATP-dependent and causes the dissociation of the whole elongation complex and results in Pol II and RNA release from the template DNA (Porrua et al., 2012)(Figure 2.1B).

Poly(A) dependent pathway is a bit more complex, because poly(A) dependent termination is coupled by the pre-mRNA 3’ processing and more protein factors, along with sequence elements on pre-mRNA are involved (Edwalds-Gilbert et al., 1993; Plant et al., 2005). The yeast poly(A) signal is composed of at least three cis elements: the AU-rich efficiency elements (EE) (Guo et al., 1995; Irniger and Braus, 1994; Zhao et al., 1999), the A rich positioning elements (PE)(Guo and Sherman, 1995, 1996) and the U-rich elements located upstream (UUE) or downstream (DUE) of the cleavage site (Heidmann et al., 1994). The cleavage site is featured by a pyrimidine followed by multiple adenosines Y(A)n and the cleavage occurs at the 3’ end of one adenosine (Heidmann et al., 1992; Heidmann et al., 1994). Poly(A) signal is recognized by termination complexes which include cleavage and polyadenylation factor (CPF), cleavage factor 1A and 1B (CFIA and CFIB). Ysh1 is one of the subunits of the CPF complex and it is responsible for the cleavage of pre-mRNA (Garas et al., 2008). The cleavage of pre-mRNA splits the molecule into two pieces, one composed of the 3' end and the other of the 5' end of the pre-mRNA. The 3' end RNA is the target ‘mRNA’, which is polyadenylated at the 3’ end by the polymerase of polyadenylation 1 (Pap1), with the help of 3’ processing factors (Ezeokonkwo et al., 2012). After 3’ polyadenylation, the mature mRNA is transported to cytoplasm for translation. Unlike the 3’ end of the pre-mRNA, the 5' end is still associated with the paused elongation complex, the 5’ end is degraded by Rat 1 exonuclease, which forms a complex with Rail1 and Rtt103 (Kim et al., 2004b; Xiang et al., 2009). There are two main models describing how Pol II is released from template DNA: allosteric model and torpedo model (Richard and Manley, 2009) (Luo et al., 2006). The hypothesis for allosteric model is that the cleavage of the pre-mRNA and binding of termination factors cause a conformational change in the elongation complex, which ends with the release of Pol II, elongation factors and RNA from the template DNA (Kim et al., 2004a; Zhang et al., 2005). For torpedo model, the exonuclease Rat 1 is the main factor (Kim et al., 2004b). The hypothesis is that after the pre-mRNA cleavage, the exposed 5' end of the pre-mRNA is degraded by Rat1 assisted by Rail1 and Rtt103 (Dengl and Cramer, 2009;

Pearson and Moore, 2013). The exonuclease 'chews' along the RNA until it collides with Pol II. The collision causes the collapse of the elongation complex and releases of Pol II from template DNA (Figure 2.1A). However, it is still under debate if the collision can generate enough force to cause the termination (Dengl and Cramer, 2009). Moreover, it may be that a combination of the two models holds true and the termination occurs with both conformational change of the elongation complex and the collision of Rat1 with Pol II (Luo et al., 2006).

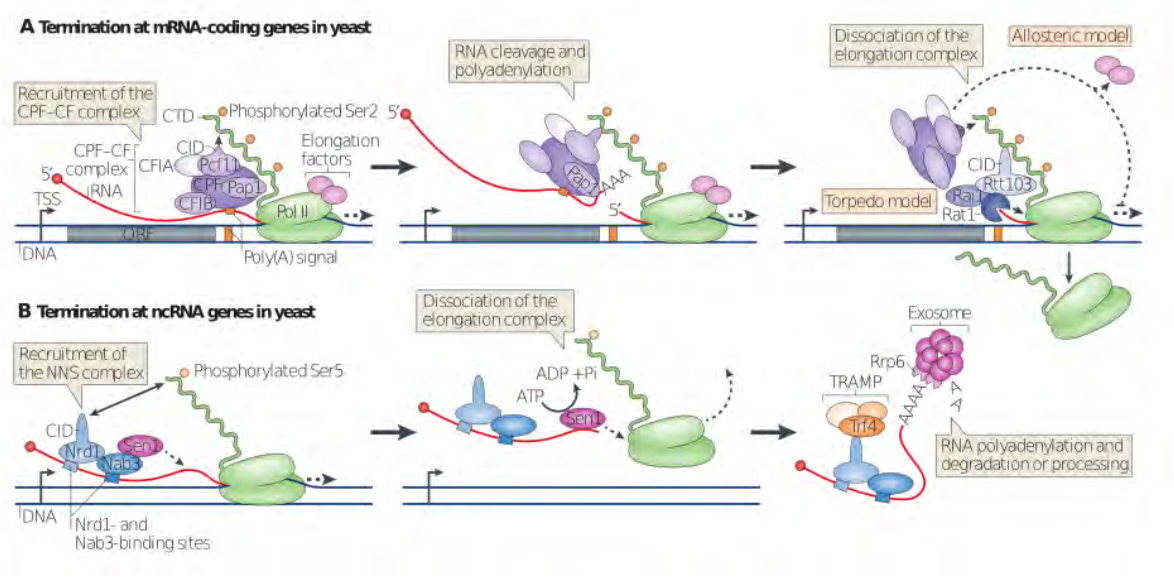


Figure 2.1: Transcription termination pathways in yeast. A, CPF-CF pathway is used for the termination of protein coding genes, in which CPF-CF complex, along with some other termination factors is recruited to Ser2 phosphorylated CTD. After pre-mRNA is cleaved at the cleavage site by Ysh1, there are two hypothetical models for how Pol II is released from template DNA, torpedo model and allosteric model. B, termination of noncoding RNAs is executed by Nrd-Nab3-Sen1 (NNS) pathway, the dissociation of DNA:RNA hybrid by Sen1 causes dissociation of elongation complex. Diagram was adapted from Jason N. Kuehner et al., MCB, 2011

1.1.3.2 Termination pathways in metazoans

Transcription termination in mammalian Pol II is similar to its yeast counterpart but with some differences. For example, the NNS termination pathway for ncRNAs is not conserved in human, as senataxin (SETX), the inferred homologue of sen1 in human, shows different function (Moreira et al., 2004). However, another pathway executed by integrator complex and ARS2 was well studied in snRNA termination, which needs the function of NELF (Gruber et al., 2012; Hallais et al., 2013; Narita et al., 2007)(Figure 2.2B). There is no helicase in the complex and the termination occurs by the exchange of elongation factors to termination factors. However, the detailed mechanism of this pathway has not yet been described.

Termination for pre-mRNAs is similar to its yeast counterpart, but there are more factors participating and the sequence elements on pre-mRNA are more conserved. There are more

than 20 protein factors participating in 3' processing in human cells (Mandel et al., 2008; Xiang et al., 2014). Depending on the functional differences, they are divided into different complexes, which includes cleavage and polyadenylation specificity factor (CPSF), cleavage stimulation factor (CstF), cleavage factor I and II (CFI and CFII), Symplekin (SYMPK) and polymerase for polyadenylation (PAP)(Christofori and Keller, 1988; Gilmartin and Nevins, 1989; Takagaki et al., 1989). There are several cis elements on the pre-mRNA recognized by the 3' processing factors. Firstly, the highly conserved polyadenylation signal (PAS), featuring AAUAAA, is normally 10 to 35 nucleotides upstream of the cleavage site (Beaudoing et al., 2000; Hu et al., 2005; Pauws et al., 2001). PAS site is specifically recognized and bound by CPSF complex. Secondly, the downstream elements (DSE) featuring GU/U rich sequence, is 30 nucleotides downstream of the cleavage site and bound by CstF complex (Chou et al., 1994; Gil and Proudfoot, 1987; McLauchlan et al., 1985). DSE is not as conserved as PAS site and varies in different genes (McLauchlan et al., 1985). Thirdly, the upstream element (USE), which is composed of multiple UGUA motifs and is positioned 40 to 100 nucleotides upstream of the cleavage site. USE is bound specifically by CFI complex (Hu et al., 2005). The cleavage site is featured by 'CA' and cleavage normally occurs between 'C' and 'A' (Chen et al., 1995).

The allosteric model and torpedo model were also widely accepted in the termination of mammalian protein coding genes (Figure 2.2A). The overall idea is similar as in yeast. In torpedo model, the cleavage of pre-mRNA is executed by CPSF73, the homologue of Ysh1 (Mandel et al., 2006). Afterwards, the mature mRNA is exported to the cytoplasm for translation. At the same time, the 5'end RNA was degraded by XRN2, which is the exonuclease in human (homolog of Rat1)(West et al., 2004). SETX functions in promoting XRN2-dependent termination (Skourti-Stathaki et al., 2011)(Figure 2.2A).

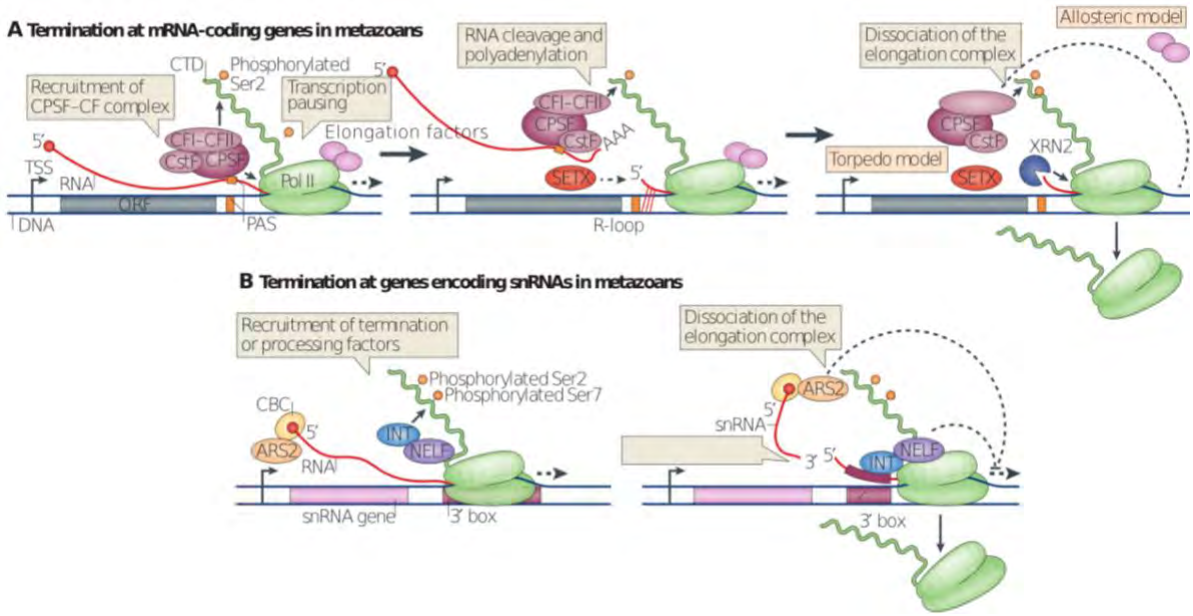


Figure 2.2: Transcription termination in metazoans. A, Poly(A) dependent pathway in protein coding genes in metazoans. 70% of mammalian genes have the highly conserved

AAUAAA site, which is recognized by CPSF complex associated with other factors. The torpedo model and allosteric model also exist in metazoans and are highly conserved in yeast and human. B, termination in noncoding RNAs in metazoans is different than in yeast. NNS pathway has not been discovered in mammals so far. Instead, ARS2 and integrator complex execute the termination of noncoding RNAs in mammals. The diagram was adapted from from Jason N. Kuehner et al., MCB, 2011

1.2 3' end processing

The polyadenylation at the 3' end occurs in most of the protein coding RNAs, as the poly(A) tail is required for mRNA maturing. However, the histone mRNA is an exception, the pre-mRNA of histones is cleaved after a stem-loop structure and the upstream RNA is not polyadenylated (Dominski et al., 2005; Marzluff et al., 2008). For the other protein coding genes, a poly(A) tail is added to the 3' end by the PAP (Wahle, 1991). In mammals, the length of the poly(A) tail is normally ~250 nt. The length of poly(A) tail is determined by a crosstalk between PABPN1, CPSF and PAP (Kuhn et al., 2009; Wahle, 1995). There are several functions of poly(A) tail, which includes the protection of mRNA from degradation, localization of mature RNA in the cells, transportation of mRNA from nucleus to cytoplasm and the translation efficiency (Preiss and Hentze, 1998).

After the RNA cleavage, the PAP adds the 250nt poly(A) tail to the 3' mRNA by using ATP (Balbo and Bohm, 2007; Martin et al., 2000). In metazoans, there are at least four different PAPs, including PAP, Neo-PAP, star-PAP and TPAP (Chan et al., 2011; Edmonds, 1990). The canonical PAP is the most well studied one and it is conserved between yeast and human (Raabe et al., 1991; Wahle, 1991). PAP belongs to the DNA polymerase ß family and the structure study reveals a three-globular-domain organization (Edmonds, 1990). The active site hides between the three domains and opens upon substrate binding (Balbo et al., 2007; Bard et al., 2000; Martin et al., 2004). The C terminal extension of PAP exists only in higher eukaryotes and is enriched with serine and threonine (Martin and Keller, 1996). The serine and threonine region is the target for posttranslational modifications, which is related with PAP activity modulation (Zhao and Manley, 1996). In 3' processing, PAP was shown to associate with CPSF complex for its function (Takagaki et al., 1990).

1.3 Termination/3' end processing factors in human

In human cells, transcription termination and pre-mRNA 3' end processing are two different processes. However, termination is normally coupled by the pre-mRNA 3' end processing and they share the necessary protein factors (Bentley, 2005). Comparing with co-transcriptional capping and splicing, which occur at the beginning and in the middle of the transcription cycle, respectively, 3' end processing normally happens at the end of transcription and is coupled with termination (Bentley, 2014). There is a big machinery which is responsible for termination and 3' end processing. According to early biochemistry identification, these factors can be grouped into four sub-complexes: CPSF complex, CstF complex, CFI and CFII (Takagaki et al., 1989) and single subunit PAP. A short introduction about these complexes as follows.

CPSF complex

In human, CPSF complex is composed of at least 6 subunits, which include CPSF160, CPSF30, WDR33 (Shi et al., 2009), Fip1 (Kaufmann et al., 2004), CPSF100 and CPSF73 (Murthy and Manley, 1992; Wahle, 1991). CPSF160 is the scaffold protein and is composed of tandem WD40 repeats clustered into three major ß-propellers (Neuwald and Poleksic, 2000). Based on the functional differences, CPSF complex is divided into two modules, the polymerase module (CPSF160, CPSF30, WDR33, Fip1) (Clerici et al., 2017; Clerici et al., 2018; Sun et al., 2018) and the nuclease model (CPSF100 and CPSF73). Polymerase module binds directly to PAS site via CPSF30 and the N terminal WD40 domain of WDR33 (Sun et al., 2018). The zinc fingers in CPSF30 are responsible for making contacts with RNA and other proteins (Barabino et al., 1997; Sun et al., 2018). CPSF73 is the endonuclease for pre-mRNA cleavage whose function is Zn^{2+} dependent. CPSF100 and CPSF73 form a dimer and they share high sequence homology, however, CPSF100 is endonuclease deficient because it lacks the zinc-binding domain (Mandel et al., 2006; Ryan et al., 2004). Some previous work also included symplekin as a part of the CPSF complex (Sullivan et al., 2009), however, in this work, symplekin would be introduced separately.

CstF complex

In humans, CstF complex is composed of three subunits: CstF77, CstF64 and CstF50 (Gilmartin and Nevins, 1991; Takagaki et al., 1990; Takagaki et al., 1989). CstF complex binds to DSE and stimulates cleavage in 3' processing (MacDonald et al., 1994). To current knowledge, CstF complex assembles with two copies of each subunit in cells (Bai et al., 2007; Legrand et al., 2007) and associates with Pol II during elongation and termination. CstF77 works as a bridge by interacting with both CstF50 and CstF64 (Takagaki and Manley, 2000). CstF64 binds to the terminal proline region of CstF77 (Hockert et al., 2010) and binds to RNA via its N terminal RNA recognition motif (RRM)(Perez Canadillas and Varani, 2003; Takagaki et al., 1992; Takagaki and Manley, 1997). CstF50 has no counterpart in yeast. N-terminal part of CstF50 is responsible for the dimerization of the whole complex (Moreno-Morcillo et al., 2011; Takagaki and Manley, 2000).

Symplekin (SYMPK)

Symplekin is a big protein with the molecular weight of 141kDa. It is thought to be a scaffold protein which bridges CPSF complex and CstF complex (Keon et al., 1996). In cells, symplekin is tightly associated with CPSF100 and CPSF73 (Hofmann et al., 2002; Sullivan et al., 2009), so it might also stimulate the activity of CPSF73 (Sullivan et al., 2009). For this reason, symplekin is often considered a part of the CPSF complex. Symplekin forms a stable complex with SSU72 (Ghazy et al., 2009; Xiang et al., 2010), which is a Ser5 phosphatase of Pol II CTD, and stimulates its phosphatase activity.

CFI and CFII

The CFI complex is assembled as a heterotetramer with a dimer of the small subunit, CFIm25, and a dimer of the big subunit which can be CFIm59, CFIm68, or CFIm72 (Ruegsegger et al., 1996; Ruegsegger et al., 1998). CFIm59 and CFIm68 are encoded by two

paralogous genes, and CFIm72 is an isoform of CFIm68 (Ruepp et al., 2011). The three big subunits may be functionally redundant because CFIm68 and CFIm25 are capable of reconstituting CFI activity in vitro (Ruegsegger et al., 1998). CFI complex binds specifically to USE and assists the selection of poly(A) site (Li et al., 2011; Yang et al., 2011).

CFII is composed of Pcf11 and Clp1 in human. Human Pcf11 is twice as big as its yeast counterpart and they have sequence similarities only within the N terminal CTD interaction domain (CID domain) (de Vries et al., 2000), which binds specifically to Ser2 phosphorylated CTD during termination (Meinhart and Cramer, 2004). The middle domain of Pcf11 binds to both Pol II CTD and RNA, so this domain might bridge the CTD to pre-mRNA. In yeast, the CFIA complex is composed of Rna14, Rna15 (the homology of CstF complex) and Pcf11, Clp1 (Gordon et al., 2011). However, in human CstF complex and CFI are two different complexes and no evidence shows that they can form a complex so far. Human Clp1 is an active 5'-OH polynucleotide kinase and interacts with both CPSF and CFI (Weitzer and Martinez, 2007).

Other factors involved in termination and 3' end processing

There are some other factors involved in 3' processing that are not a part of the above-mentioned complexes, such as polyadenylate-binding nuclear protein 1 (PABPN1), Senataxin (SETX), Retinoblastoma-binding protein 6 (RBBP6) and some kinases and phosphatases during termination, such as CDK12/Cyclin K and Protein phosphatase 1 (PP1). PABPN1 is thought to bind to PAP together with poly(A) binding protein (PABP) and assist its function (Blobel, 1973). In early studies, it was also shown that PABPN1 binds to poly(A) tail and works as a ruler to decide the length of poly(A) tail together with CPSF complex and PAP (Kuhn et al., 2009). SETX is a huge protein in human with a molecular weight of 303kDa. It is believed to be a yeast sen1 homologue (Sariki et al., 2016), however, no helicase activity was demonstrated for SETX up to now. The exact function of SETX in termination is also not clear yet. RBBP6 is the homolog of yeast Mpe1 but the two proteins share very low sequence homology and the function of RBBP6 in 3' end processing is not clear (Di Giammartino et al., 2014; Wagschal et al., 2012). The kinases and phosphatases are mostly working on CTD modifications, for example, CDK12/Cyclin K was found to peak at the 3’ end of genes (Bosken et al., 2014), which means it functions in the late elongation and termination and might be functionally overlapping with CDK7 and CDK9. SSU72 is a Ser5 phosphatase which specifically removes Ser5 phosphorylation during termination (Ganem et al., 2003).

Pol II C terminal domain (CTD) and phosphorylation

Rpb1, the largest subunit of Pol II, has the long extended C terminal domain (CTD). CTD consists of consensus repeats Tyr1-Ser2-Pro3-Thr4-Ser5-Pro6-Ser7, with the repeating number of 52 in human and 26 in yeast (Brickey and Greenleaf, 1995). CTD is unstructured and the function is not fully understood. During the whole transcription process, CTD is a target for a wide range of post-translational modifications, of which the best known is phosphorylation (Eick and Geyer, 2013).

Pol II is recruited to promoters in a dephosphorylated form. After recruitment, Ser5 is phosphorylated by CDK7, which is part of the transcription factor IIH (TFIIH)(Fisher, 2019). Ser2 phosphorylation is executed by CDK9/Cyclin T (Ctk1 in yeast). CDK9/Cyclin K composes the positive transcription elongation factor (P-TEFb) complex (Bacon and D'Orso, 2019), which plays an important role in Pol II release during promoter-proximal pausing (Vos et al., 2018a). With Pol II moving to the end of the gene, another kinases act on CTD, such as CDK12/Cyclin K and CDK13/Cyclin K (Bartkowiak et al., 2010). Even though the clear function of these kinases is not clear yet, the concentration of CDK12 and CDK13 peaks at the end of the gene (Bosken et al., 2014).

At different transcription stages, CTD has different phosphorylation states, which might be correlated with different protein factors being recruited to Pol II (Buratowski, 2009; Egloff and Murphy, 2008). The ratio of Ser2 phosphorylation and Ser5 phosphorylation (Ser5-P to Ser2-P) is becoming lower as Pol II goes from 5' end to 3' end of genes (Vasiljeva et al., 2008). Proteins involved in early transcription events, such as capping, prefer to bind to Ser5-P. However, the 3' processing factors, prefer to bind to Ser2-P, which is a CTD-modification enriched during late stages of transcription (Ahn et al., 2004). Ser7 phosphorylation (Ser7-P) is another CTD modification involved in the recruitment of the integrator complex to snRNA encoding genes (Egloff, 2012). However, the function in mRNA transcription is not clear. Recent studies show that Ser7-P is present at the promoter region of protein coding genes and the phosphorylation level increases towards the 3' region (Kim et al., 2010). The function of Ser7-P in termination remains unclear. Dephosphorylation of Thr1 of CTD is also thought to be important in termination, which is executed by SSU72 in human (Mayer et al., 2012).

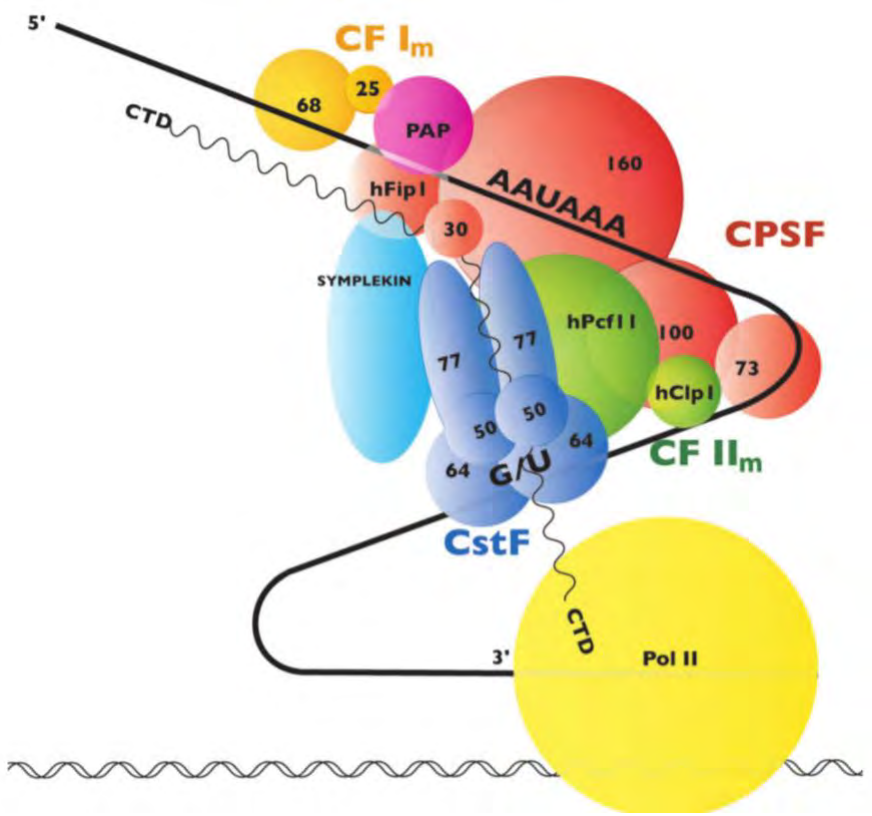


Figure 2.3: A cartoon representation of termination/3' processing factors in human Pol II transcription. CPSF complex, CstF complex, CFI, CFII and Pol II are colored in red, blue, orange, green and orange, respectively. PAP is colored in magenta. Pol II is colored in yellow. CTD is shown as a wavy line extending from Pol II. PAS and downstream element sequences are indicated above the RNA. The cartoon is adapted from C.R.Mandel et al., cellular and molecular life sciences, 2008

1.4 Pre-mRNA 3' processing in humans and aims of this work.

As mentioned above, there are two steps in 3' polyadenylation. Firstly, the pre-mRNA is cleaved at the cleavage site, which is defined by the 3' processing factors and the three elements in the pre-mRNA. There might also be some auxiliary elements on the pre-mRNA, located further downstream or upstream from the cleavage site (Zhao et al., 1999). CPSF73 is the endonuclease which performs the pre-mRNA cleavage (Mandel et al., 2006). The second step is the polyadenylation process, in which the polymerase PAP adds a poly(A) tail to the 3' end of the RNA by using ATP. There are at least four different types of PAP in metazoans, including PAP, Neo-PAP, Star-PAP, and tPAP (Edmonds, 1990). PAP is most widespread in human. The length of poly(A) tail is determined by the crosstalk between the poly(A) binding protein, PAP and CPSF (Kuhn et al., 2009). Some studies showed that the length of poly(A) tail is determined by how many copies of poly(A) binding protein bind to the poly(A) tail. PABPN1 is the poly(A) binding protein in the nucleus and one copy of PABPN1 binds to ~30nt of RNA (Wahle, 1995). Also, the binding of PABPN1 facilitates the function of PAP (Wahle, 1995). Thus, the coordination of PAP and PABPN1 determines the

length of poly(A) tail. The length of poly(A) is correlated with the stability of both mRNA and protein.

For many years, the structure study of termination/3' processing factors was limited to truncated subunits or domains, which includes the HAT-N domain of CstF77 (Bai et al., 2007), dimerization domain of CstF50 (Moreno-Morcillo et al., 2011), Pcf11 CID domain (Meinhart and Cramer, 2004) etc. There are also some crystal structures of protein complexes like CFIm68/25 (Yang et al., 2011). But for most sub-complexes, the subunits arrangement inside the complex is not clear because the crystallization of big proteins or protein complexes is difficult, or because the complex itself is too dynamic for structural studies. In the last few years, with the improvement of cryo-EM in both hardware and software, the structures of big complexes were possible to be solved to near atomic resolution without crystallization. In the 2017 and 2018, the CPF polymerase module from yeast (Casanal et al., 2017) and CPSF polymerase module from human (Clerici et al., 2017; Clerici et al., 2018; Sun et al., 2018) were solved by both crystallography and EM respectively. The polymerase module of CPSF (in human) and CPF (in yeast) are quite similar: the three ß-propellers of CPSF160 (BPA, BPB and BPC) are organized like a trefoil with WDR33 sitting on top of BPA and BPC (Casanal et al., 2017; Sun et al., 2018). The structures also revealed that CPSF polymerase module recognizes specifically the PAS site. WDR33 N-terminal domain and CPSF30 have direct interaction with the AAUAAA sequence (Sun et al., 2018). This is a breakthrough for the study of 3' processing and can be used as a starting point for future studies.

So far, there are still quite some questions to be answered about 3' processing and termination, which includes how the pre-mRNA cleavage site is defined, how the cleavage occurs, what the working mechanism of CPSF73 and PAP is, and how the 3' processing is coordinated with termination. All these questions are challenging for structure studies as the termination and 3' processing might be very dynamic processes. This means it will be challenging to lock the complex in a stable state suitable for structure analysis.

In this study, I am trying to reveal the mechanism of termination/3' processing by using in vitro reconstitutions, cryo-EM, crystallography and biochemical assays. To get enough proteins for in vitro studies, I expressed and purified all canonical factors involved in 3' processing/termination. These factors were expressed and purified as defined complexes, such as CstF complex, CFI and CFII. The CPSF complex was divided into 2 modules: the polymerase module (CPSF160, WDR33, Fip1 and CPSF30) and nuclease module (Symplekin, CPSF100, CPSF73). These two modules were expressed and purified individually. Whether symplekin is one of the subunits of CPSF complex is still under debate. However, earlier studies showed that symplekin forms a stable complex with CPSF100 and CPSF73 (Sullivan et al., 2009). Thus, symplekin was expressed with the CPSF nuclease module and this strategy worked.

The structure of CPSF polymerase module was solved in the last two years. However, the endonuclease responsible for the pre-mRNA cleavage, CPSF73, was not included in the complex. Even earlier studies showed that CPSF73 alone had the nuclease activity (Mandel et al., 2006). It is still unknown how CPSF73 works in the full CPSF complex, and what the stimulation mechanism of the nuclease activity is, and also how the polymerase module and nuclease module coordinate during the cleavage process.

In this work, I assembled the full CPSF complex (including symplekin) with purified CPSF polymerase and nuclease modules, and I analyzed the complex by both negative staining and cryo-EM. In parallel, I tried to investigate the cleavage activity of the complex in vitro and compared it with CPSF73 activity alone. However the cleavage activity assay turned out to be tricky and more experiments need to be done. Combining both structural studies and biochemistry, I am trying to understand the role for CPSF complex in 3' processing. This work would show the progress so far.

In parallel, I attempted to crystallize the CstF complex, which was previously not possible due to its flexibility.

2 Materials and Methods

2.1 Materials

2.1.1 Bacterial strains and cell lines

Species	Strain/cell lines	Genotype /origin	Supplier
E. coli	XL1-Blue	recA1 endA1 gyrA96 thi-1 hsdR17 supE44 relA1 lac [F' proAB lacIqZΔM15 Tn10 (Tetr)]	Agilent
E. coli	BL21-Codon Plus(DE3)-RIL	*E. coli* B F- ompT hsdS(rB- mB-) dcm+ Tetr E. coli gal λ (DE3) endA Hte [argU ileY leuW Camr]	Agilent
E. coli	DH10EMBacY	F- mcrA Δ(mrr-hsdRMS-mcrBC) φ80dlacZΔM15 ΔlacX74 endA1 recA1 deoR Δ(ara, leu)7697 araD139 galU galK λ- rpsL nupG / bMON14272‡ yfp+/ pMON7124	Geneva Biotech
Spodoptera frugiperda	Sf21 (IPLB-Sf21-AE)	immortalized pupal ovarian tissue cells[261]	Thermo Fisher Scientific
Spodoptera frugiperda	Sf9	immortalized pupal ovarian tissue cells, clonal isolate of parental cell line IPLB-Sf21-AE[261]	Thermo Fisher Scientific
Trichoplusia ni	Hi5 (High Five) (BTI-TN-5B1-4)	immortalized pupal ovarian tissue cells[262]	Expression system

2.1.2 Chemicals and kits

Name	**Application**	**Company**
General chemicals	Buffers etc.	Merck, Roth, Sigma-Aldrich
Enzyme additives and other reagents	Cloning	Fermentas, New England Biolabs (NEB), Promgea
Plasmid preparation kit	Plasmid extraction from *E.coli*	QIAGEN
Gel extraction kit	Gel extraction of linearized vector or PCR product	QIAGEN

2.1.3 Additives for *E. coli* and insect cell culture.

Additives	Application	1000x Stock
Ampicillin	Antibiotic for *E.coli* culture	100 mg/mL in ddH_2O
Kanamycin	Antibiotic for *E.coli* culture	50 mg/mL in ddH_2O
Chloramphenicol	Antibiotic for *E.coli* culture	30 mg/mL in ethanol
Spectinomycin	Antibiotic for *E.coli* culture	50 mg/mL in ddH_2O
Gentamycin	Antibiotic for *E.coli* culture	10mg/mL in ddH_2O
Streptomycin	Antibiotic for *E.coli* culture	30 mg/mL in ddH_2O
IPTG	expression induction	1M in ddH_2O (the working concentration varies from 0.5 to 1mM
X-Gal	blue-white selection	150 mg/mL in DMSO

2.1.4 Buffers and solutions

Buffer	Composition/Description (Supplier)	Application
4x SDS-PAGE loading dye	45% (v/v) glycerol, 280mM Tris pH 6.8 at 20°C, 8% (w/v) SDS, 10% (v/v) β-mercaptoethanole, 0.03% (w/v) bromophenol blue	SDS-PAGE
gel electrophoresis running buffer	20x NuPAGETM MES/MOPS SDS running buffer (Invitrogen)	SDS-PAGE
gel staining	InstantBlue (Expedeon)	Coomassie staining
PCR master mix	2x Phusion® High-Fidelity PCR Master Mix (NEB)	PCR
6x DNA loading dye	Gel Loading Dye, Purple (NEB)	agarose gel electrophoresis
10x TAE	50mM EDTA pH 8.0 at 20°C, 2.5M Tris-acetate	agarose gel electrophoresis
NEBufferTM 3.1	50mM Tris-HCl pH 7.9 at 25°C, 100mM NaCl, 10mM $MgCl_2$, 0.1mg/mL BSA (NEB)	restriction endonuclease digest

CutSmart® buffer	20mM Tris-acetate pH 7.9 at 25°C, 50mM potassium acetate, 10mM magnesium acetate, 0.1 mg/mL BSA (NEB)	restriction endonuclease digest
T4 polymerase buffer	10 mM Tris-HCl pH 8.0 at 25°C, 500mM NaCl, 100mM $MgCl_2$, 10mM DTT	LIC cloning
T4 DNA ligase buffer	50mM Tris-HCl pH 7.5 at 25°C, 10mM $MgCl_2$, 1mM ATP, 10mM DTT (NEB)	ligation
P1	50mM Tris-HCl pH 8.0 at 25°C, 10mM EDTA, 100µg/mL RNase A (QIAGEN)	Bacmid isolation
P2	200mM NaOH, 1% SDS (QIAGEN)	Bacmid isolation
N3	4.2M Gu-HCl, 0.9M potassium acetate pH 4.8 (QIAGEN)	Bacmid isolation
DPBS	138mM NaCl, 2.7mM KCl, 8.1mM Na_2HPO_4 pH 6.9, 1.47mM KH_2PO_4 pH 6.9	Insect cell culture
X-treme GENETM 9	supplied in 80% ethanol, final concentration 1.5 µL/mL	transfection agent for insect cells
T7 RNA polymerase Storage Conditions(NEB)	50mM Tris-HCl, 100mM NaCl, 20mM β-ME, 1 mM EDTA, 50% Glycerol, 0.1% Triton® X-100 pH 7.9 @ 25°C	In vitro transcription
5x Transcription buffer	200mM Tris-HCl, 30mM $MgCl_2$, 5mM DTT.10 mM spermidine (pH 7.9 @ 25°C)	In vitro transcription
NTP set	100mM ATP, UTP, CTP, GTP	In vitro transcription
10xTBE buffer (Sigma-Aldrich)	890mM Tris, 890mM boric acid, 20mM EDTA.	Urea gel
2x RNA Loading buffer (NEB)	6.4M Urea, 1xTBE buffer, 50mM EDTA	Urea gel

Resuspension buffer	50mM Tris-HCl pH8.0@20°C, 10mM EDTA (pH8.0), 50mM Glucose, 0.01mg/ml Dnase free Rnase A	Maxiprep
Lysis solution	0.2M NaOH, 1% SDS	Maxiprep
Neutralization solution	4M KOAc pH5.5 (pH with acetic acid)	Maxiprep
3M sodium acetate pH 5.2		RNA extraction
Phenol-Chloroform-Isoamyl alcohol (25:24:1)		DNA Precipitation
Gibco® Sf-900TM III SFM	low-hydrolysate, serum-free, protein-free, animal origin-free insect cell culture medium/Thermo Fisher Scientific	Sf9 / Sf21 culture (growth and maintenance of suspension and monolayer cultures; baculovirus production and propagation)
ESF921TM	serum-free, protein-free insect cell culture media, supplemented with L-glutamine and Kolliphor® P188 / Expression Systems	Hi5 culture (growth and maintenance of suspension cultures;
uranyl formate solution	2% (w/v) uranyl formate in ddH_2O	negative staining
LB	1% (w/v) tryptone, 0.5% (w/v) yeast extract, 0.5% (w/v) NaCl (1.5% (w/v) agar for solid plates)	*E. coli* culture

2.1.5 cDNAs origins of 3' processing factors and corresponding yeast genes

Gene name	cDNA origin	cDNA vector selection Marker	Yeast homolog
CPSF160 (CPSF1)	Harvard medical school	Kanamycin	Cft1/Yhh1

	database (HsCD00045496)		
WDR33 (1-572aa)	gblock from IDT	--	Pfs2
CPSF30(CPSF4)	Harvard medical school database HsCD00367627	Spectinomycin	Yth1
Fip1	gBlock from IDT	--	Fip1
CPSF100(CPSF2)	Harvard medical school database HsCD00379114	Spectinomycin	Ctf2/Ydh1
CPSF73(CPSF3)	Harvard medical school database HsCD00334392	Chloramphenicol	Ysh1(Brr5)
symplekin(SYMPK)	Harvard medical school database HsCD00045464	Chloramphenicol	Pta1
PAP	gBlock from IDT	--	Pap1
PPP1CB	Harvard medical school database HsCD00005414	Kanamycin	Glc7
PPP1CC	Harvard medical school database HsCD00005169	Ampicillin	
SSU72	gBlock from IDT	--	Ssu72
Pcf11	MRCPPU	Ampicillin	Pcf11
Clp1	Harvard medical school database HsCD00378275	Spectinomycin	Clp1
CstF77	Harvard medical school database HsCD00339748	Ampicillin	Rna 14
CstF64	Harvard medical school database	Ampicillin	Rna15

	HsCD00331178		
CstF50	Harvard medical school database HsCD00322461	Chloramphenicol	--
CFIM68(CPSF6)	Amplified from genome cDNA	--	--
CFIM25(CPSF5)	Harvard medical school database HsCD00323128	Chloramphenicol	--
CDK12	Harvard medical school database HsCD00021466	Ampicillin	CDK12
CyclinK(CCNK)	Harvard medical school database HsCD00327466	Ampicillin	CyclinK(CCNK)
PABPN1	Harvard medical school database HsCD00330743	Chloramphenicol	--

2.1.6 Buffers for protein purification

Name of buffers	**Composition**
Lysis Buffer	20mM Hepes-NaOH pH7.4@20 °C, 300mM NaCl, 30mM Imidazole, 10% glycerol, 1mM/5mM DTT (or 0.5mM TCEP), 1x protease inhibitor
Ni Elution Buffer	20mM Hepes-NaOH pH7.4@20°C, 300mM NaCl, 500mM Imidazole, 10% glycerol, 1mM/5mM DTT (or 0.5mM TCEP), 1x protease inhibitor
Amylose Elution Buffer	20mM Hepes-NaOH ph7.4@20°C, 300mM NaCl, 30mM Imidazole, 117mM maltose, 10% glycerol, 1mM/5mM DTT (or 0.5mM TCEP), 1x protease inhibitor
Gel Filtration Buffer	20mM Hepes-NaOH ph7.4@20°C, 300mM NaCl, 10% glycerol,

	1mM/5mM DTT (or 0.5mM TCEP),

2.2 Methods

2.2.1 Polymerase Chain Reaction (PCR)

Polymerase chain reaction (PCR) was used to amplify DNA fragments from different cDNA templates. 50ul reaction was set up including 10 to 250ng template DNA, 0.5µM forward and reverse primer respectively, 200µM of dNTP mix and 1U of Phusion or Q5 polymerase (NEB). Thermo cycling was set up with 3 minutes of denaturation at 95 or 98 °C depending on the polymerase, 30 seconds denaturation (95 or 98°C), 30 seconds primer annealing (The primers were designed to have an annealing temperature of 55 to 60°C). Extension time was set according to the length of the target DNA, typically 30s for 1kb. The whole PCR program includes 35 cycles and finalized with 10 minutes elongation at 72 °C.

2.2.2 Agarose Gel Electrophoresis and Gel Extraction

The DNA fragments generated from PCR or restriction enzyme digested vector need to be separated by electrophoresis in 1% (w/v) agarose gels. 1g agarose was dissolved in 100ml 1x TAE buffer by heating in the microwave. 2ul of SYBRTMSafe DNA Gel Stain (Invitrogen) was added. The 'agarose solution' was poured into the gel chamber and left at room temperature for at least half hour to solidify. The prepared gel was covered with 1xTAE buffer and ready for sample loading. The PCR product or linearized vector was mixed with an appropriate amount of 6x DNA Loading Dye (NEB) and loaded onto the gel together with a 1 kb DNA Ladder size standard (NEB)(or 100bp marker based on the size of the product). The gel was run at 120V for 20 to 30 minutes till a sufficient separation of different fragments and then imaged with a UV imager. Bands with the target size was cut and extracted according to the gel extraction kit protocol (QIAGEN). Briefly, 3 volumes of buffer QG was added to the gel and incubated at 50°C till it dissolved completely (check in between and invert up and down for a few times). Then 1 gel volume of isopropanol was added and mixed thoroughly. The dissolved DNA was applied to the column and washed with 750ul PE buffer. The column was spun down at the maximum speed for 1 minute to remove the ethanol before the DNA fragments were eluted with 50ul ddH_2O. Normally the water for elution was heated up to 65°C to improve the elution efficiency.

2.2.3 Preparation of chemically competent *E.coli* cells

Cells from old aliquot or commercialized origin were plated on LB agar plate with appropriate antibiotics and cultured overnight at 37°C incubator. The next day, single colony was picked and cultured in LB medium with corresponding antibiotics at 37°C overnight (16 to 18 hours). 5ml $MgCl_2$ and appropriate antibiotics were added to 1 liter pre-warmed SOB

medium. The overnight *E.coli* culture was added to the medium at a 1: 250 dilution ratio. The cells were cultured at 37°C while shaking till the OD600 reached to 0.5 to 0.6. Then the cell culture was transferred to the 250ml conical wide-mouth centrifuge tubes (Thermoscientific) which were pre-chilled on ice. The cells were incubated on ice for 10 min in the centrifugation tubes and then spun down at 3000g for 10 min at 4°C. The supernatant was carefully discarded and the pellet was re-suspended in 75ml of inoue buffer per 250ml cells. The re-suspended cells were incubated on ice for 10 min. Again the cells were spun down at 3000g at 4°C for 10 min. The supernatant was carefully discarded and the cells were re-suspended in 10ml of inoue buffer per 250ml cells. 700ul DMSO was added drop by drop to the 10ml cells while shaking. The cells were kept on ice for 10 min and aliquoted as 100ul aliquots in 1.5ml tubes (Eppendorf), snap-frozen in liquid nitrogen and stored at -80°C for using. To make sure the transformation efficiency is high and no contamination was introduced during the preparation process, one quality control by transformation is necessary.

2.2.4 Preparation of electrocompetent *E.coli* cells

The *E.coli* cell culturing was the same as the chemically competent cells except for that the DH10αEMBacY cells were used. The cell pellet was washed two times with pre-chilled sterilized ddH_2O and one time with 10% (v/v) glycerol. Then the cells were re-suspended in 10% glycerol, aliquoted as 100ul fractions in 1.5ml eppendorf tubes, flash-frozen in liquid nitrogen and stored in -80°C.

2.2.5 Ligation-independent cloning (LIC)

Ligation-independent cloning (LIC) is a new strategy for cloning, which is faster and more efficient than the traditional method. LIC cloning depends mostly on T4 polymerase (LIC-qualified, Novagen), which has both an exonuclease and a polymerase activity. In the absence of substrate dNTPs, T4 polymerase 'chews' back the ends of double strand DNA and generates overhangs. However, to prevent the 3'-5' exonuclease activity from continuing indefinitely, Addition of specific dNTPs is necessary, which would restrict the 3'-5' exonuclease processivity to the site of the first matching DNA base on the complimentary strand (Supplementary figure 1A). MacroBac Series-438 vectors were designed to comprise a LIC-compatible site for the insertion of ORFs, which is exposed after cleaved with SspI. The complimentary overhangs permit the annealing of vectors and inserts but prevent internal annealing. The PCR fragment and the linearized vector were treated with T4 polymerase separately. 20ul of reactions were set, which comprised of 150 ng of linearized DNA (vector or insert), 2.5 mM of the respective dNTPs (dCTP for inserts and dGTP for vectors), 5mM DTT, 2µL of 10x T4 DNA Polymerase buffer (NEB) and 2U of T4 polymerase (Novagen). The reaction system was incubated at 25°C for 40 min followed by the enzyme heat inactivation at 75 °C for 20 min. For annealing after T4 polymerase treatment, 2µL (50 to 100ng) of vector and 2µL of insert DNA were mixed (the volume can also be 4ul and 4ul if the inserts are long or the concentration of the fragments are low) and incubated at 25°C for 30 min. The rest of the T4 polymerase treated DNA can be stored at -20°C for future use. Reactions were stopped by adding 1.3µL of 25mM EDTA and incubated for 10 min at 25°C. The

complete reaction volume was directly transformed into XL1-Blue chemically competent cells.

2.2.6 Sequence and Ligation-Independent Cloning (SLIC)

SLIC cloning method is similar to LIC cloning. Both LIC and SLIC need T4 polymerase for its exonuclease and polymerase activity. The only difference is that no dCTP or dGTP would be added into the SLIC reaction. Comparing with LIC, SLIC doesn't have very exact overhangs for annealing. The linearized vector, the insert prepared by PCR, T4 DNA polymerase and corresponding reaction buffer were mixed and kept at room temperature for 10 minutes to generate the overhang and anneal the two fragments. Then the reaction was transferred on ice for another 10 minutes and transformed directly to XL-blue chemically competent cells.

2.2.7 Transformation of chemically competent *E.coli*

Plasmid or DNA ligation reaction was mixed with 100ul of chemically competent *E.coli* cells and incubated on ice for 30 minutes. Cells were heat shocked at 42°C for 70 seconds and kept on ice for another 3 minutes. 1ml of LB was added to the cells and the culture was recovered at 37°C for 45 minutes to 2 hours. After recovery, the cells were spun down at 13,000 rpm for 1min. The supernatant was gently removed and 100ul of fresh LB medium was added to the pellet. The cells were re-suspended and plated on the LB agar plate plus corresponding antibiotics. The plate was cultured at 37°C incubator.

2.2.8 Concatenation of poly-promoter MacroBac Series-438 vectors containing multiple ORFs

After inserting ORF of each gene to 438 serials vector, the next step is to connect these genes on one vector with their own promoters and terminators. Because each gene has its independent promoter and terminator, the order for connection doesn't matter. The acceptor vector would be linearized with SspI, while the donor vector would be digested with PmeI, which would end up with 2 fragments, the one with the target ORF and corresponding promoter and terminator would be used for ligation (supplementary Figure 1B). Again LIC was used for the ligation of acceptor and donor vectors. dCTP was added to the donor vector reaction while dGTP was added to the acceptor reaction. After the treatment with T4 polymerase, the donor and acceptor vector were annealed and transformed into XL1-Blue cells following the standard LIC protocol.

2.2.9 Site-directed mutation correction

QuickChange approach was used to correct the mutations in the target vector which came from the cDNA or were introduced by the cloning process (UV light). Forward and reverse primers with correct sequences were designed to amplify the dsDNA from the same site around the mutation site, one of the two primers need a 5' phosphate group for the later ligation reaction. After amplification, the PCR reaction was treated with Dpn I at 37°C for at least two hours to remove the parental plasmid. The PCR product was then purified by agarose gel extraction. The purified DNA was ligated with T4 ligase (ThermoScientific) and transformed to XL1-Blue cells. For T4 ligation, 10ul reaction was set up with 20 to 100ng DNA fragments plus 1ul 10 x buffer and 1U T4 ligase. The reaction was kept at room temperature

for 10 minutes and transferred to ice for 3 minutes and then transformed to the competent cells.

2.2.10 Introduction of the multi ORFs into baculovirus shuttle vectors (bacmid Preparation)

The blue white screening method is a classical screening method. The method is based on the β-galactosidase gene for its α-complementation function. The β-galactosidase in the host *E.coli* strain was not active because the deletion of the first 41 amino acids, however this can be remedied by the expression of the first 59 amino acids by introducing one vector, the short peptides is named as α-peptide and the rescue of β-galactosidase activity is called α-complementation. X-gal is colorless, however, within the induction of IPTG, X-gal can be cleaved to form a blue pigment 5,5'-dibromo-4,4'-dichloro-indigo, which would make the whole colony look bright blue. For the blue white screening design, multiple cloning sites were introduced to the α-peptide coding area. If the target fragment was successfully introduced to the vector, the expression of α-peptide was destroyed and the β-galactosidase was inactivated, then the colony would end up with a white color, which can be differentiated from the negative blue colonies.

The final 438 series vector which includes all target ORFs was transformed into the DH10Multibac cells. This *E.coli* strain features the respective viral genome on a bacmid vector, and the transformed vector can be transferred to the genome by gene transposition. 1µg of the construct plasmid was added to 200µL electrocompetent DH10Mutibac cells and kept on ice for 15 minutes. Then the cells were transferred to electroporation cuvette to execute the electroporation (one pulse, 25 µF, 1.8 kV). 1mL LB medium was added and the whole system was transferred to 15mL culture tube to grow for 5 hours to overnight while shaking at 37˚C, because the cells need some time for transposition. After recovery, 25 to 100ul cells (based on the transformation efficiency, which is normally quite efficient) were plated to the X-gal plates. The X-gal plates are normal LB agar plates with 150 ug/mL X-gal, 1mM IPTG and 10ug/mL gentamycin. After 36 to 48 hours incubation, there should be obvious blue/white colonies on the plate. At least three of the white colonies should be picked. The white colonies were plated on a new X-gal plate and cultured for another 24 to 48 hours to exclude false-positive colonies. The white colony was cultured at 37°C overnight in 4~6ml LB medium plus gentamycin for bacmid preparation. For bacmid extraction, the miniprep kit was used and the first several steps were the same as miniprep. After adding N3 and spun down for 10min at 15,000rpm. The supernatant was transferred to a new 1.5ml tube and 700µL isopropanol was added. After mixing with votex, the tube was incubated at -20˚C for 5 hours or -80°C for 2 hours to precipitate the DNA. Afterwards the tube was taking out and the DNA was spun down by centrifugation at maximum speed for half hour. After centrifugation, the supernatant was trashed while the pellet was washed with 500µL 70% ethanol. After centrifugation for 10 minutes at the maximum speed, the ethanol was carefully removed. 30ul ethanol was left on top of the pellet till transfection.

2.2.11 Bacmid transfection to sf9 cells and V0 production

All transfection steps were operated in Biological Safety Cabinets. The ethanol on top of the bacmid was removed gently and the pellet was left in the hood for 5 to 10 minutes with the lid open for ethanol evaporation. 20µL water was added to the top of the pellet gently and the lid was then closed, pipetting to re-suspend pellet was not allowed because this would shear the bacmid. To dissolve DNA, Incubation with water for 10-20 minutes is necessary. A mastermix can be prepared during the incubation time, which contains 10ul of Xtreme Gene 9 transfection agent and 100ul Gibco® Sf-900TM III SFM for each bacmid transfection. 200ul of sf9 media was added to the dissolved bacmid DNA plus 100µL of transfection agent master mix. The whole reaction system was incubated for 60 minutes. Again pipetting up and down was not allowed, the medium and mastermix should be added to the bacmid gently. During the incubation time, the sf9 cells were prepared in a 6 well plate. Each well was either filled with 3ml sf9 cells with the density of $1E^6$ or 3ml medium as control. For one transfection, there should be at least one medium control and one cell control to make sure that both the medium and the cells were not contaminated. After pipetting cells into the wells, the plate was gently shaken manually to make sure that the cells are distributed as 'single-layer' at the bottom of the plate. After one hour, the bacmid mixture was added to the corresponding wells drop by drop. Normally two wells were used for each bacmid strain (Supplementary figure 2). The plates were incubated at 27°C for 48 to 72 hours. Cells were checked with fluorescent microscope to track the 'green cells' because successfully transfected cells would express YFP which is visible under fluorescent microscope. The 'green cells' should begin to appear after 48 hours. V0 should be harvested maximum 72 hours after transfection. For harvesting, the supernatant was carefully sucked with pipette tubes and stored in 15ml Falcon tubes. The prepared V0 were marked with date, cell type and name of bacmid etc.

2.2.12 V1 production and virus propagation

To amplify V0 and make the expression more efficient, V1 was made with sf21 cells. 25 mL of sf21 cells with a density of 1.0×10^6 cells/ml was infected with 50ul to 1ml V0 (based on the number of green cells during V0 production, more green cells mean stronger virus). For the culture of sf9 cells and sf21 cells, the flask should have ten times more volume (for example, a 500ml flask can hold maximum 50ml culture, if the volume of the cells was over 50ml after dilution, the extra cells should be either trashed or transferred to a new flask).The cells were cultured at 27°C while shaking at 60rpm. The cell culture should be checked every 24 hours about its density, viability and diameter. The density of the cells should double after 24 hours, but not change so much afterwards, which was named as the day after proliferation (DPA). The density of cells should be kept at 1.0×10^6 cells/ml every day. After the viability dropped below 88 percent, the virus should be harvested, which was normally 48 to 72 hours after DPA depending on the activity of the virus. V1 was harvested by centrifugation at 250g for 15 minutes at 4 °C. The supernatant was transferred to a new 50ml Falcon and stored at 4°C for expression. The virus should be marked with name, date and cell type. The pellet can be kept at -20°C for pull down assay of the expression.

2.2.13 Protein expression in Hi5 cells

600ml Hi5 cells were cultured to a density of 1.0×10^6 cells/ml in a 3L flask, 300ul to 1ml V1 (based on the viability of V1) was added to the Hi5 cells and cultured at 27°C while shaking at 60rpm. The cells need to be checked every 24 hours for their viability, diameter and activity. Normally the DPA reached the next day after transfection, which means more medium should be added to keep the density at 1.0×10^6 cells/ml. For the next few days, the population activity of the cells should be more than 90%. The cells should be harvested if the activity dropped below 88% to make sure the protein expression was at the peak and the protein was not degraded because of cell death. However, if the density of cells was not double in the next day or the cells kept dividing, one should think about changing the infection volume of V1. For some proteins which are not so stable and can be degraded easily, the cells should be harvested in 24 to 48 hours to avoid degradation.

2.2.14 Protein expression in *E.coli*

Vector with target gene was transformed into *E.coli* BL21 competent cells and plated to LB plate with corresponding antibiotics. Single colony was picked after 16 to 18 hours and cultured in LB medium with corresponding antibiotics at 37°C overnight. 2L of LB plus antibiotics was prepared and the overnight culture was added to the flask to an OD600 of ~0.2. The *E.coli* was cultured at 37°C for 3 to 4 hours till OD600 arrived to 0.6 ~ 0.8. For Zinc finger protein, 0.2mM $ZnCl_2$ was added and incubated at 37°C for another 15 to 30 minutes. Afterwards, the flask was taken out from 37°C shaker and cooled down on ice for 20 minutes. Then 0.25mM to 1mM IPTG was added and the expression was performed at 37°C for 3 hours or 18°C overnight while shaking at 160rpm. The cells were harvested by centrifugation at 4000 rpm for 15 min. The pellet was re-suspended in Lysis buffer, flash frozen in liquid nitrogen and kept at -80°C for purification.

2.2.15 General purification of protein complexes

One of the subunits of the protein complex was designed to have an N terminal His-MBP tag. The purification was done by two rounds of affinity purification (Ni and maltose) and one round of gel filtration. The harvested cells with their lysis buffer were taken out from -80°C and thawed in water bath which was kept at room temperature. The thawed cell suspension was transferred to a metal beaker for sonication (3 min at 30 % output with ON = 0.6 s and OFF = 0.4 s). The sonicated lysate was transferred to oak ridge centrifugation tubes for rotor A27 and spun down at 27,000 rpm for 30 minutes. The supernatant was transferred to ultra-centrifuge tubes for Ti-45 rotor and spun down at 45,000 rpm for 1 hour. Then the supernatant was collected and filtered with syringe filter for loading. HisTrap HP 5 mL column (GE healthcare) was washed with water for 10 column volumes and equilibrated in lysis buffer ready for loading. The supernatant was loaded to Ni column with peristaltic pump at a slow flow rate. At the same time, the Amylose column (Home made with amylose resin from NEB) was equilibrated in lysis buffer with Äkta system (GE healthcare) at the flow rate of 1ml/min. After loading, the Ni column was transferred to Äkta system, washed with 50ml of lysis buffer and then connected with amylose column (Supplementary figure 3). The protein was eluted to amylose column with Ni elution buffer. In this step, the target protein

bound to the amylose column while the random protein came out. After elution, the Ni column was disconnected and the amylose column was washed with lysis buffer till the baseline reached. The protein was eluted from amylose column with amylose elution buffer. The elution was collected with 96 well plates with 1ml volume fractions. The peak fractions were picked and loaded to the SDS-PAGE gel. According to the gel, target fractions were pooled and collected for further purification. TEV and lamda phosphatase were added to the protein at a ratio of 1:5 and 1:20 respectively. The whole solution was dialyzed overnight with Thermo Scientific SnakeSkin 7K MWCO dialysis bag to 1L lysis buffer plus 1mM $MnCl_2$ (as the lamda phosphatase need Mn^{2+} to be active). The protein was taken out from the dialysis bag the next morning and loaded to the pre-equilibrated Ni column with peristaltic pump. This step would help to separate TEV and MBP from target protein. Because TEV and MBP bound to the Ni column (there is one his tag on TEV) while the target protein came out from the flow through after his-MBP cleavage overnight. Lysis buffer was used to wash the column to make sure that all the target protein comes out from the column. Bradford solution was used to track the flow and to decide where to stop the collection. The column was eluted with Ni elution buffer afterwards and the elution was collected (because some tags were not cleavable and some protein binds to Ni column itself, both the flow through and elution should be collected for later SDS-PAGE identification). Fractions from overnight dialysis, flow through and elution were loaded to SDS-PAGE to figure out the fractions of the target protein. If the MBP-His tag was not cleavable, the target protein would come out from the elution together with TEV-MBP, which needs to be separated by further gel filtration afterwards. The target protein from the reverse Ni step would be concentrated with Amicon Ultra-15 centrifugal filter unit (100-kDa molecular mass cut-off) (Merck) to 0.9ml, spun down by centrifugation at 15,000rpm for 10 minutes and loaded to Superose 6 increase 10/300 column (GE healthcare) which was equilibrated with gel filtration buffer beforehand. The peak fractions from gel filtration were identified by SDS-PAGE gel, the target protein fractions were pooled together and concentrated with Amicon Ultra-15 centrifugal filter unit (100-kDa molecular mass cut-off). The concentration was checked from time to time during concentration. The final concentration of the protein complex was controlled to 30~50uM. The protein was aliquoted as 6ul aliquots and flash frozen in liquid nitrogen. The aliquots were stored at -80°C marked with date, name and concentration etc.

2.2.16 Mxiprep

The RNA for in vitro cleavage assay was produced with T7 RNA polymerase (NEB) by in vitro transcription. To get enough template DNA for transcription, the vector carrying the DNA template needs to be prepared by maxiprep. The *E.coli* cells which carried the target vector were cultured in 250ml LB with corresponding antibiotics at 37 °C overnight. The overnight culture was spun down at 4000rpm for 15 minutes with F14 rotor. The pellet was re-suspended in 12ml resuspension buffer and transferred to a 50ml falcon tube, 12ml lysis buffer was added and the tube was inverted 4 to 8 times. To make the lysis of the cells more efficient, the falcon was left at room temperature for 5 minutes. The reaction was quenched with 12ml neutralizing solution and inverted 4 to 8 times. The tube was centrifuged for 30

minutes at 4000rpm at 4°C. The supernatant was transferred to a fresh 50ml Falcon with 25ml 25:24:1 phenol:chloroform:isoamylalcohol (Merck)(all the operations with organic solvent were performed in the hood). The solution was inverted 4 times and spun down for 30 minutes at 4000rpm at 4°C. The supernatant was transferred to a new 50ml falcon with 25ml chloroform. The solution was inverted 2 times and spun down for 30 minutes at 4000rpm at 4°C. Again the supernatant was carefully pipetted out and transferred to a new 50ml falcon with 25ml 100% ethanol. 2.5ml 3M sodium acetate pH5.2 was added and the whole tube was transferred to -80°C for one hour. The DNA was then spun down for 30 minutes at 4000rpm at 4°C. The supernatant was discarded and the pellet was washed with 70% ethanol and spun down at 4000rpm for 15 minutes. The pellet was dried for 5 minutes and re-suspended in 1ml RNase free water. The concentration was checked with nanodrop and marked on the tube for future use.

2.2.17 Template DNA linearization by Hind III

Template vector for in vitro transcription needs to be linearized with Hind III to avoid supercoil in transcription. 1mg vector was digested in 1ml reaction system (150ul of 20,000U/ul Hind III-HF with 1× Cutsmart buffer, from NEB) at 37°C overnight. The next day phenol chloroform reaction was used to remove remaining restriction enzyme and isolate the linearized DNA. 100µL RNase free NaAc 3M was added to the 1ml reaction digestion followed by 1mL Phenol-Chloroform-Isoamyl alcohol (25:24:1). The solution was transferred to phase lock tube and spun at maximum speed for 10 minutes at 4°C. The upper aqueous phase was transferred to a fresh tube with 1000µL chloroform. The solution was Vortexed and transferred to phase lock tube and spun down at 4°C for 10 minutes at 15,000 rpm. Again aqueous phase was transferred to fresh tube. 700µL isopropanol (0.7 vol) was added and Incubated at -20°C for 1 hour to precipitate DNA. The DNA was pelleted by centrifugation and re-suspended in 70µL RNase free ddH_2O. The concentration was checked and marked on the tube for future use.

2.2.18 RNA production by in vitro transcription

1mL in vitro transcription reaction include 200µL of 5x Transcription buffer, 1µL Triton X-100 (1% w/v solution), 24µL $MgCl_2$ (1M) , 40µL ATP (100mM stock), UTP, CTP, GTP respectively, 100µg linearized DNA and 17.4µL T7 RNA polymerase (T7 RNA polymerase, transcription buffer and NTP set were all from NEB. All the other materials were RNase free). The reaction was kept at 37°C overnight (maximum 16 hours) while shaking at 350rpm. The RNA was precipitated the next day. Each in vitro transcription reaction was split into two tubes (500µL per tube). 80ul of 0.5M EDTA pH 8.0 was added to each tube to dissolve MgPP formed during the in vitro transcription. 35µL 5M NaCl and 430µL isopropanol was added to precipitate the nucleotides. To make the precipitation more efficient, the reaction was normally incubated at -80°C for 2 hours. After thawing, the solution was centrifuged at 13000 xg for 30 minutes at 4°C. The pellet was washed with 1mL 70% (v/v) ethanol and spun down for 10 minutes at 13000xg. The final RNA pellet was dried in the hood and re-suspended in 100µL water. 1 to 5ul of RNA was mixed with 2× RNA loading buffer and loaded to 10% urea gel to check the transcription efficiency. The gel was run in 1x TBE buffer

and stained with methyl blue. The RNA bands can be seen in blue after staining. After making sure that the product was as expected, the full RNA product was mixed with 2x RNA loading buffer and was loaded to 10% urea gel to run for 40 minutes at 300V. Target RNA band was cut out by using UV shadowing. The RNA gel was pushed through syringe to get small pieces. 1ml 0.3M sodium acetate pH 5.2 was added to the gel and kept at -20°C for 2 hours. The gel was spun down at the maximum speed for 30 minutes and the supernatant was kept on ice. 1ml 0.3M sodium acetate pH 5.2 was added to the pellet gel again and the tube was kept at 37°C for 2 to 3 hours while shaking at 1000rpm. The gel was spun down at the maximum speed for 30 minutes and the supernatant was kept again. This RNA extraction process was repeated 4 to 5 times. RNA concentration was checked at every round, the extraction can be stopped if the concentration was very low. In the end, all the supernatants were pooled together and were precipitated with isopropanol (0.7 vol). The reaction was kept at -80°C for two hours and spun down at maximum speed for 30 minutes. The pellet was washed with 70% ethanol once, dried in the hood and dissolved in water. RNA aliquots were kept at -80°C ready for use.

2.2.19 CPSF+symplekin complex preparation

CPSF complex polymerase module and symplekin-CPSF100/73 complex were purified separately as the general His-MBP purification protocol (Supplementary Figure 3). After eluted from amylose column, the two sub-complexes were mixed, TEV and lamda phosphatase were added. The whole protein solution was transferred to SnakeSkin 7K MWCO dialysis bag and dialyzed to lysis buffer (with 1mM $MnCl_2$ for lamda phosphatase activity) overnight. The protein was taken out from the dialysis bag the next morning and loaded to Histrap HP 5ml column equilibrated with lysis buffer. The protein complex bound back to the Ni column because the His-MBP tag on CPSF160 was not cleavable, while the SYMPK complex would come out from the flow through because the tag on SYMPK was cleaved nicely by TEV. The column was washed with lysis buffer till almost no protein coming out from the flow through (Using coomassie blue to track). Then the protein complex was eluted from the column with Ni elution buffer, together with TEV and MBP. In the later step, the target protein complex was concentrated with Amicon Ultra-15 centrifugal filter unit (100-kDa molecular mass cut-off) and loaded to Superose 6 increase 10/300 gel filtration column for size exclusion. TEV protease and lamda phosphatase can be removed in two processes: concentration and gel filtration, as the concentrator would keep only the proteins with the molecular weight of more than 100kDa, part of MBP and TEV can be removed in this step. The protein complex and TEV/MBP can be nicely separated by gel filtration because they have very big difference of molecular weight. 50ul of the peak fraction was dialyzed with Thermo Scientific Slide-A-Lyzer 20K MWCO MINI Dialysis Device for at least 4 hours. The dialysis buffer was gel filtration buffer without glycerol and the target of dialysis was to remove the glycerol in the buffer. The 'dialysis cup' needs to be washed with ddH_2O for 20 to 30 minutes before sample loading.

2.2.20 CPSF+SYMPK complex negative staining grids preparation and checking

Before making cryo grids, the sample was normally checked by negative staining. For making negative staining grids, 2% uranyl formate (UF) solution was used. One big drop of ddH_2O (1ml) and three drops of UF solution (25ul for each drop) were prepared on the plastic film. 4ul of protein was applied to the glow-discharged side of the grid and kept on tweezer for 1 minute, then the grid was washed with water for 1 minute with the protein drop side touching the water surface, the grid was taken out from the water drop and stained with the 3 UF solution drops one by one (20 seconds each). Then the grid was kept on bench for one minute with the drop on. The drop was sucked from the other side of the grid with filter paper. The grid was kept on the bench for several minutes for drying and then transferred to the CM120 (Phillips) for checking. The negative staining grids can also be kept in the grid box for future screening (the storage of the grid should avoid light and can be stored maximum one month). Only samples with correct particle size and good distribution would be used for further cryo-EM analysis.

2.2.21 CPSF+SYMPK complex cryo-EM grids preparation and data analysis with Glacios

After dialysis, the concentration of the protein was normally diluted 2 to 3 times. The concentration of the gel filtration peak fraction was around 0.5mg/ml, after dialysis, the concentration was ~0.25mg/ml, which was a proper concentration for cryo grids making. QUANTIFOIL Au 2/2 grids were used for freezing. After glow discharge, 4ul of sample was loaded to the glow discharged side of the grid, after waiting for 10 seconds, the grid was blotted for 8.5s with blotting force 6 and dropped into liquid ethane. The grids were prepared by using Vitrobot Mark IV (FEI, Hillsboro, OR). The temperature of the Vitrobot chamber was set to 4°C and the humidity with 100%. The grid was kept in liquid nitrogen and screened with Glacios (Thermoscientific). Grids with thin ice and good particle distribution were used for further data collection.

Initial cryo-dataset was collected with Glacios operated at 200keV and equipped with Falcon3 camera. Micrographs were collected automatically with the software package EPU (FEI) at the magnification of 120K (1.23 Å per pixel) in linear mode. The dose rate was 45.17e−/pixel/s. three images were acquired per foil hole. Each micrograph was collected with a total dose of 29.86 electrons per square angstrom over a 1.52s exposure, fractionated into 30 frames. Defocus values ranged from 1.25 to 3µm. Warp (Dimitry Tegunov, 2018) was used for micrograph alignment, motion correction and CTF correction, as well as particle picking. The data processing was performed using cryoSPARC (Punjani et al., 2017).

2.2.22 CPSF+SYMPK complex Titan Krios data collection and processing

The data was collected on a FEI (Thermo fisher Scientific) Titan Krios, operated at 300 keV, and equipped with a Gatan K3 Summit direct electron detector and a Quantum GIF. Micrographs were collected automatically with the software package Serial EM

(Mastronarde, 2003) at a nominal magnification of 81K (1.05 Å per pixel) in counting mode. The dose rate was 30.25e−/pixel/s. Three images were acquired per foil hole. Each micrograph was collected with a total dose of 27.5 electrons per square angstrom over a 1.491 exposure, fractionated into 40 frames. Defocus values ranged from 1 to 3µm. Micrograph frames were aligned and corrected with Warp (Dimitry Tegunov, 2018), CTF correction and particle picking were also performed with Warp (Dimitry Tegunov, 2018). Initial data analysis was performed with cryoSPARC (Punjani et al., 2017), while further data processing was performed with RELION 3.0.6 (Zivanov et al., 2018).

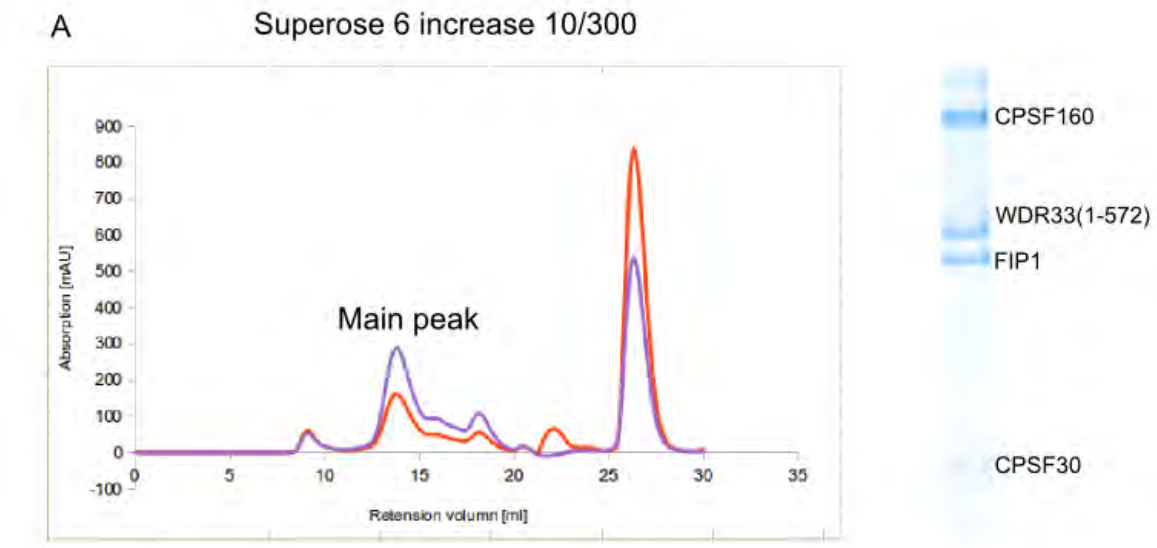
2.2.23 CstF complex crystallization

CstF complex was expressed in insect cells and purified as the normal purification protocol (Supplementary Figure 3). The protein was concentrated to 10mg/ml, 200ul was taken out and kept on ice. Then the rest was concentrated to 20mg/ml. Both samples were sent for setting up drops. The drops were set up with Gryphon crystallization robot on intelli plates. 12 commercial crystallization Kits were used, which includes 9 from QIAGEN (AmSO4, Classics, Classics II, Classics Lite, JCSG+, PEG, PEGs II, pHclear, pHclear II), 2 from JENA BIO SCIENCE (Wizard 1+2, Wizard 3+4) and 1 from HAMPTON RESEARCH (Index). The drops were checked with Rock imager. The crystals were picked out from the drop, flash-frozen into liquid nitrogen and sent to Synchrotron in Sweden for diffraction check.

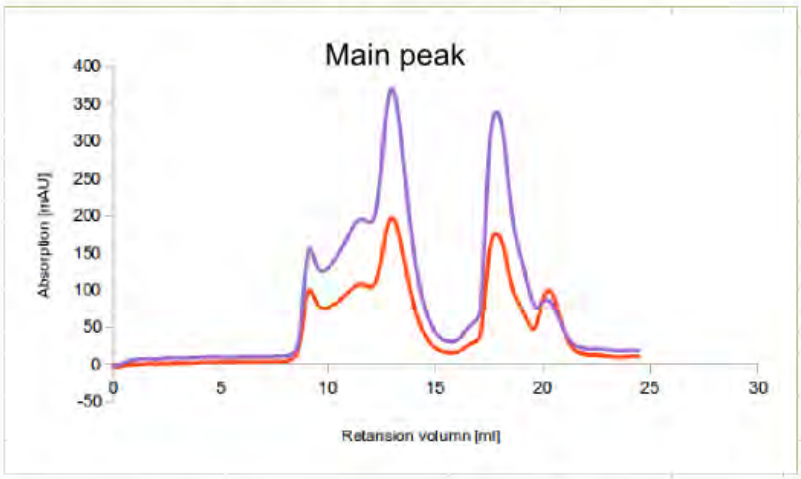
3 Results

3.1 Purification of termination/3' processing factors (subcomplexes)

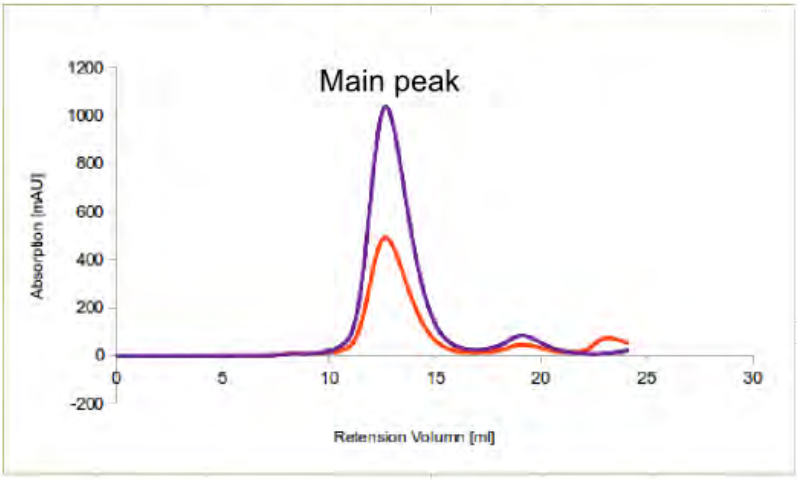
Based on current literature, the factors were purified as part of multi-subunit complexes. Five well known complexes were expressed in insect cells and purified following the purification flowchart (Supplementary figure 3), which includes the Pcf11-Clp1 (CFII) complex, CPSF polymerase module, CPSF nuclease module with symplekin, CstF complex and CFIm68/25 (CFI) complex. CFII and CstF complexes were initially designed as one complex based on the study in yeast, in which Rna14-Rna15-Pcf11-Clp1 form the CFIA complex (Gordon et al., 2011). In this combined construct all subunits were combined in one vector and either CstF77 or Pcf11 was tagged. However, such approach was not successful because only subcomplexes containing the tagged subunits could be purified (data not show), this was also shown in early studies (Takagaki et al., 1989). Thus, CFII and CstF complexes were redesigned and purified separately. One of the subunits of the complex was tagged with the His-MBP at the N-terminus. A TEV site was added between the target protein and the His-MBP tag to make sure the tag can be removed in the final purification step. However, the cleavage site was not exposed at the surface of the structure, which made the cleavage difficult, such as the case when the tag was installed on CPSF160. Pcf11, CstF77, CPSF160, SYMPK and CFIM68 were tagged in corresponding complex CFII, CstF, CPSF polymerase module, CPSF nuclease module and CFI respectively. The tag on PCF11 was cleavable, but the protein was still bound back to the Ni column during the reverse nickel step, maybe because of some internal Ni binding sequences. The protein was eluted from the Ni column together with MBP and TEV, however, MBP and TEV were removed by gel filtration chromatography step (Figure 2.4).



B


Superose 6 increase 10/300
Main peak
Absorption [mAU]
Retansion volumn [ml]
SYMPK
CPSF100
CPSF73

C


Superose 6 increase 10/300
Main peak
Absorption [mAU]
Retension Volumn [ml]
CSTF77
CSTF64
CSTF50

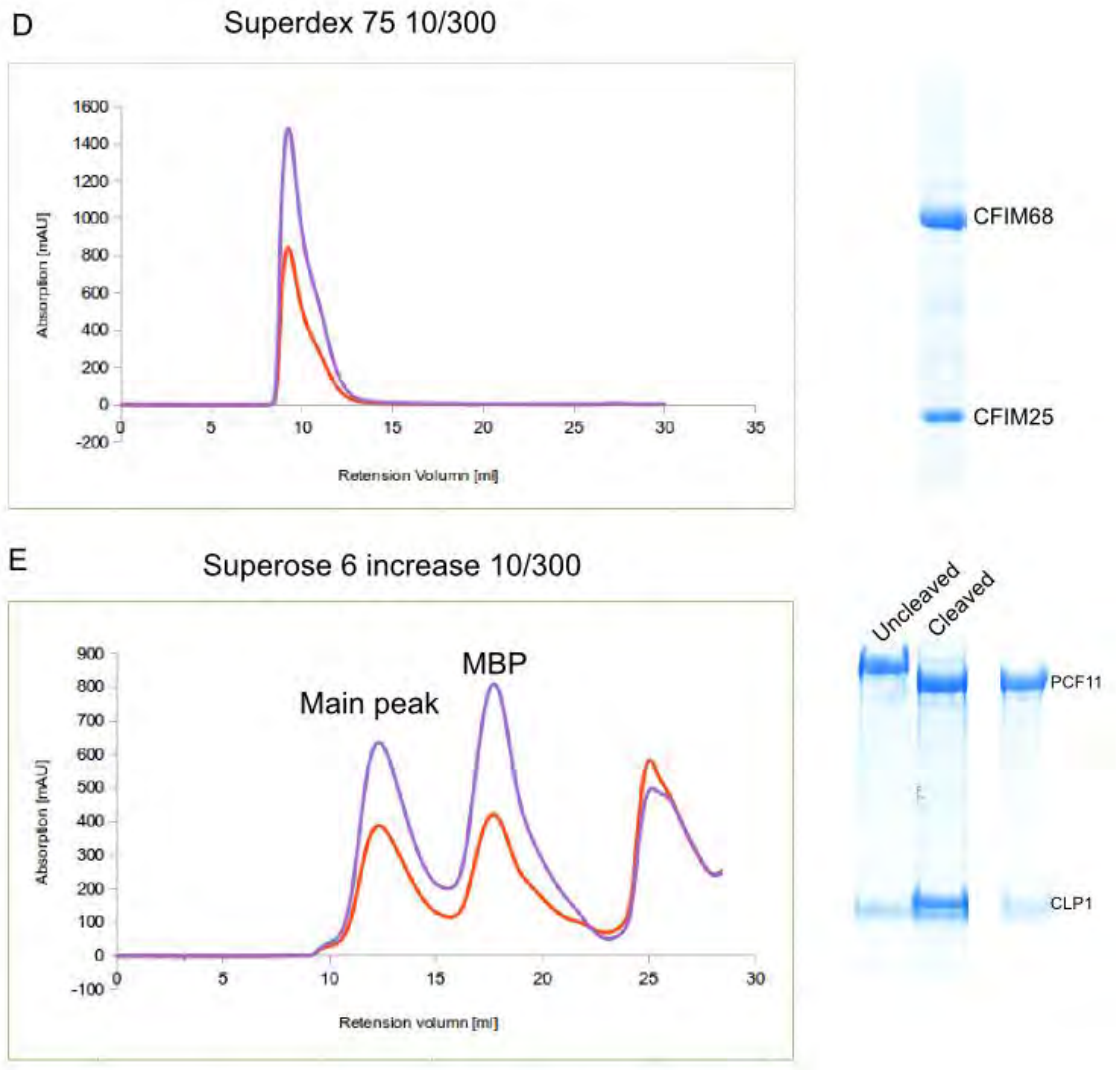


Figure 2.4: Purification of termination/3' processing complexes. A, Purification of CPSF polymerase module, chromatogram of gel filtration is shown on the left and SDS-PAGE gel is shown on the right (this is consistent in B, C, D and E). The target protein peak and subunit names are marked. CPSF160 was tagged with his-MBP at the N terminus but the tag was not cleavable. B, Purification of symplekin-CPSF100/73 complex. Symplekin was tagged at the N-terminus and the tag was removed after the reverse Ni step. The shoulder of the peak might come from protein aggregation. C, CstF complex purification, CstF77 was tagged at the N-terminus and the tag was cleavable. D. CFIm68/25 complex. Superdex 75 column was used for gel filtration because of better resolution. CFIm68 was tagged and the tag was cleavable. E, Purification of Pcf11/Clp1. Pcf11 was tagged, the protein bound back to Ni column even though the tag was cleavable (right panel). So there is one peak after the main peak, which contains MBP and TEV.

3.2 CPSF polymerase module and nuclease module containing symplekin form a stable complex

After purification of the complexes, both pull down and gel filtration were executed to check the interactions between different complexes. By checking the interactions, I investigated which complexes are stable enough for structure analysis. After the initial screening, I managed to assemble the full CPSF complex containing the symplekin. CPSF polymerase module and nuclease module plus symplekin formed a stable complex both on the gel filtration column and the sucrose gradient (Figure 2.5). This was verified at least three times. In gel filtration, the complex peak shifted to the earlier volume compared with the sub-complex peak. In sucrose gradient, there is also a clear shift on the native gel (Figure 2.5B, right panel)

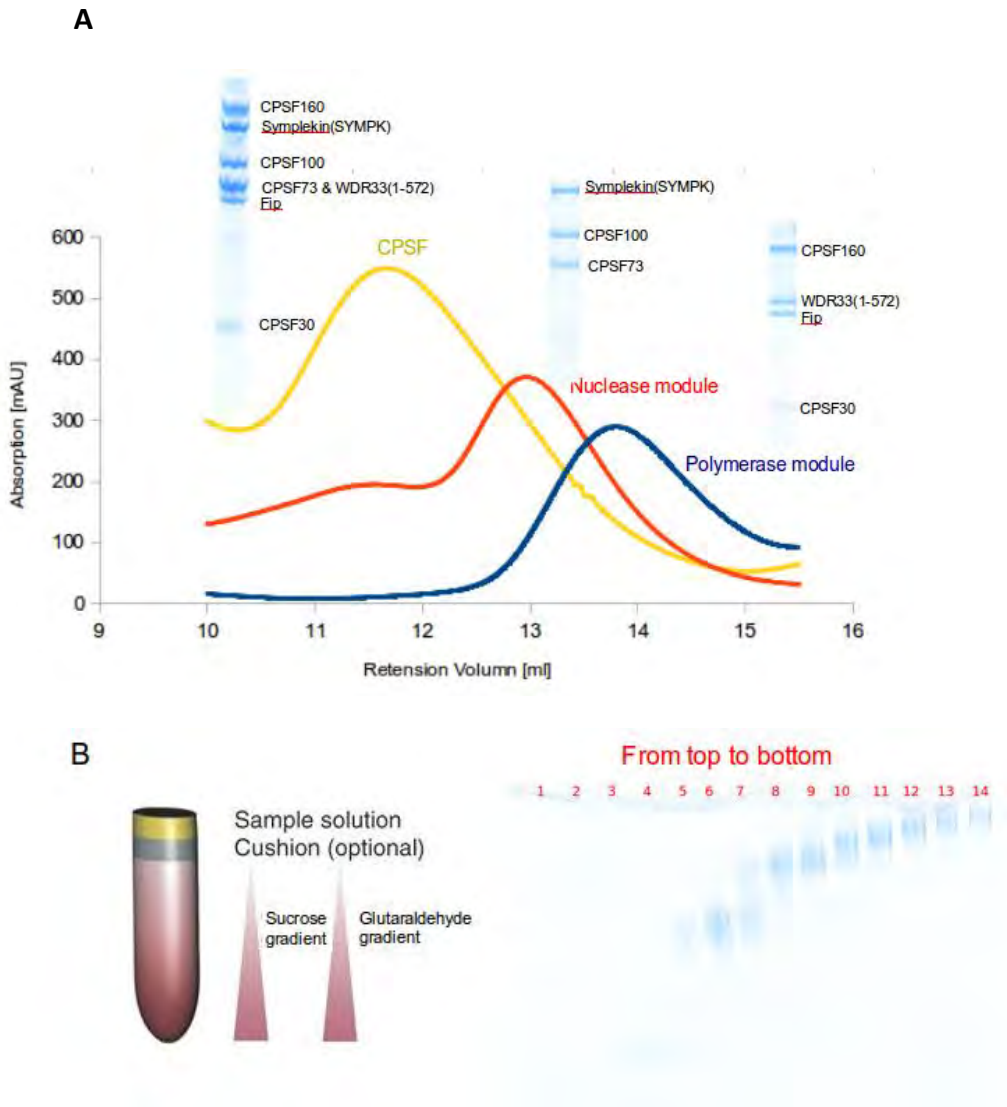


Figure 2.5: CPSF polymerase module and nuclease module plus symplekin form a stable complex. A, the complex peak is in yellow, polymerase module peak is in blue and nuclease peak is in red. The SDS-PAGEs are shown next to the corresponding peaks and the subunits are marked. A truncated version of WDR33 (residue 1 to 572) was expressed in the

polymerase module. B, Sucrose gradient diagram (left panel) and corresponding native gel (right panel). The polymerase module and the nuclease module alone were detected in the upper fractions whereas the complex formed by the two modules was detected in the bottom fractions. Proteins at the very bottom might be aggregated proteins.

3.3 Initial cryo-EM structure analysis of the CPSF-symplekin complex with Glacios

The grids with CPSF-symplekin complex were screened with Glacios. Grid with proper ice thickness and nice particle distribution was used for overnight data collection. 85,216 particles were picked with Warp (Dimitry Tegunov, 2018) with a box size of 330 pixels from the overnight dataset (Figure 2.6A). 2D classification was executed several rounds to remove the bad particles with cryoSPARC (Punjani et al., 2017) (Figure 2.6B). After 2D classification sorting, 68,513 particles were extracted for further processing. Ab-initio reconstruction gave out 3 initial models. These three maps were refined individually with the cryoSPARC 'Homogeneous refinement' function. One class that includes 31.4% particles showed extra density in addition to the polymerase module (Figure 2.6C). The resolution of this map calculated by cryoSPARC was 8.37Å, however, upon visual inspection it was obvious that the map looks worse than 8.37Å and the resolution might be overestimated by cryoSPARC. Even though the resolution was not as high as cryoSPARC estimated, structure of the CPSF polymerase module (PDB 6BM0)(Sun et al., 2018) could be fitted into the density map unambiguously. The extra density was quite obvious even at low threshold. It was difficult to assign the extra density to the additional subunits at this resolution. However, since the particles aligned well both in 2D and 3D and at least one third of the particles contained the additional density aside from the polymerase module, I attempted to improve the resolution and build a model by acquiring a better dataset with Titan Krios.

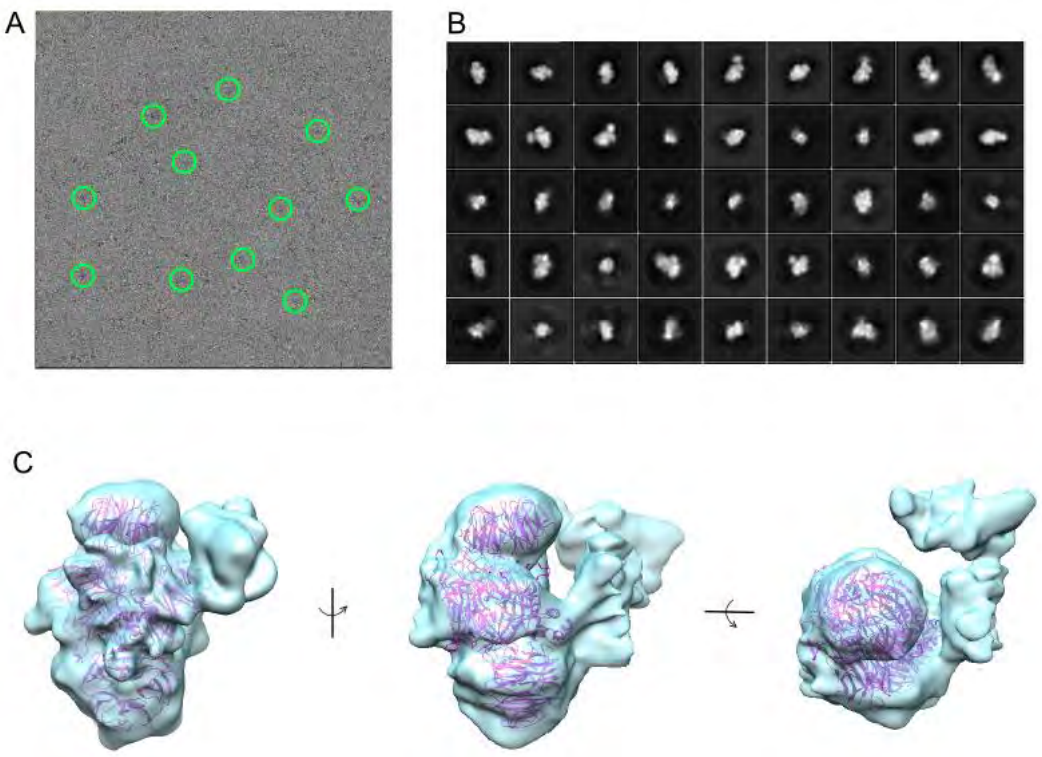


Figure 2.6: Initial cryo-EM data analysis of CPSF-SYMPK complex. A, Typical micrograph, particles are highlighted with green circles. B, Part of the 2D classes from data processing, classes contaminated with gold particles were discarded before another round of 2D classification. C, Three different views of the 3D map which showed extra density in addition to the polymerase module. The 3D map was created with Ab-initio reconstruction and refined with homogeneous refinement in cryoSPARC. Structure of the CPSF polymerase module (PDB code 6BM0 (Sun et al., 2018) in color purple) was fitted into the density map (cyan).

3.4 CstF complex crystallization

The yield of CstF complex expression in insect cells was very good. Around 20mg of protein can be purified from 1 liter of hi5 cells. The purity of the complex was also suitable for a crystallization trial (Figure 2.7B). For the first screening, there were needle crystals in several conditions (Figure 2.7C). These conditions contained 100mM Tris or Hepes (pH from 7.0 to 8.0) and the precipitant of different molecular weights of PEGs (Figure 2.7C). However, even more conditions were set based on the initial conditions, the diffraction of the crystals was not improved at all. Some more optimization is necessary to improve the diffraction of the crystals in the future.

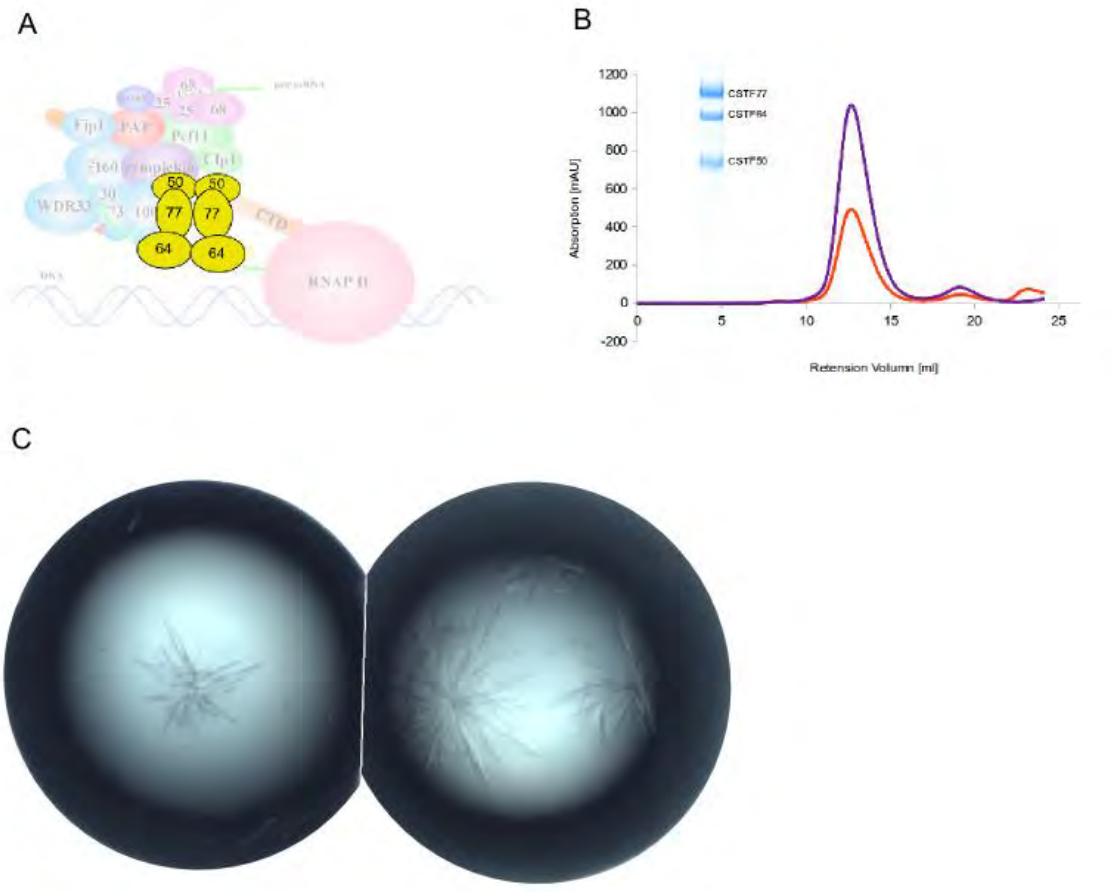


Figure 2.7: Purification and crystallization of CstF complex. A, cartoon showing the position of CstF complex in the termination machinery. CstF exists as a heterodimer and binds specifically to the pre-mRNA DSE in the 3' processing/termination complex. B, gel filtration chromatogram of CstF complex and the corresponding SDS-PAGE gel. The protein purity was reasonable for crystallization. C, initial crystallization screening with the commercial kit. Left panel: 12% PEG4000, 100mM Hepes pH7.5, 100mM sodium acetate trihydrate. Crystals appeared after 3 days. Right panel: 10% PEG6000, 100mM Hepes pH7.0. Crystals appeared after 24 hours.

4 Discussion and future perspectives

4.1 CPSF-symplekin complex - structure and function

4.1.1 First data collection and analysis with Titan Krios

After initial analysis of Glacious data, a dataset from Titan Krios was collected with Serial EM and K3 camera (Gatan). In 64 hours, 8,651 micrographs were collected, and around 1.83 million particles were picked with Warp from these micrographs. The particles were re-extracted with RELION3.0.6 (Zivanov et al., 2018) with 2 times binning. All the processing was executed with RELION3.0.6. The extracted particles were submitted for 2D classification and the 'bad' classes (classes with gold particles or fussy looking ones) were sorted out. Particles in the good classes were saved and used for the second round of 2D classification. This process was repeated several times till the particle set was 'clean enough'. Then the good particles were used for further processing. There are several ways to get an initial model: calculating the initial model with RELION with the particles from the same dataset, using the model obtained from the Glacious dataset (the one shown in Figure 2.6) or using the lowpass filtered human polymerase module map (EMD-7112)(Sun et al., 2018). All these three models were used for 3D classification to see which works best, and the initial model built by RELION was finally picked for further processing because upon visual inspection it looked better than the other two. However, in the later 3D classification and 3D refinement process, the particles showed strong preferred-orientation. Because the information on the missing views was lacking, it was difficult to improve the resolution of the density map. Thus, to improve the density map, the quality of the dataset itself should be improved. The angular distribution diversity can be changed by either tilting the sample stage for a defined degree or by influencing the particle distribution biochemically. Tilting the sample stage to get the lost orientations is normally the first thing to try, because this microscopy setting is easy to manipulate and it is not a very time consuming process. The second method is to chemically crosslink the sample prior to grid preparation or to add detergent to the sample. This may lead to a more diverse angular distribution. The advantage of the second method is that it can solve the problem fundamentally, but the disadvantage is that it is more complicated because we are not sure which kind of crosslinker or detergent would work, and normally a lot of biochemistry experiments need to be done before an improvement can be made. Considering all these issues, we decided to collect another dataset by tilting the stage for 35 degrees in hope to capture the missing particle orientations.

4.1.2 Tilt data collection with Titan Krios and analysis

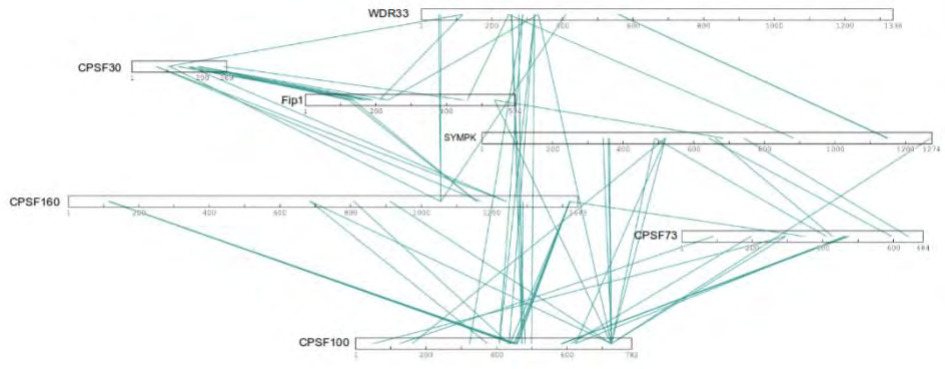
The settings for tilt data collection were the same as the first Titan Krios dataset except that the sample stage was tilted for 35 degrees instead of keeping it at zero degree. The tilt data collection was performed for ~36 hours. The frame alignment, motion correction, CTF correction and particle picking were executed with Warp in parallel with the data collection. 800,000 particles were picked in total by using Warp. The tilted particle dataset was combined with the original untilted particle dataset. The particles were re-extracted with RELION3.0.6. All the data processing was executed with RELION3.0.6. The angular

distribution diversity was improved a bit by combining the tilted and the untilted dataset (as seen from the angular distribution map, data is not shown), however, the processing is still on the way to improve the density map. The density around the polymerase part seems quite flexible, which made it difficult to improve, and also binding of the nuclease module might make the complex prefer to distribute as one orientation. Because from one of the 3D classes which includes only the polymerase module, it is reasonable to get a structure where the secondary structure is visible after 3D refinement, however, the 3D class with 'extra density' is tricky to be improved, maybe more focus classification (on the 'extra density' part) or refinement need to be done in the future. If the tilt dataset turns out to be not possible to get the structure, I would try to remake the grids. Maybe crosslinking with the sample would help.

4.1.3 CPSF100 might work as the bridge between CPSF73, symplekin and the polymerase module

The fresh CPSF-symplekin complex was sent for crosslinking MS analysis (it was done by Ralf Pflanz from Henning Urlaub lab) about the interactions between different subunits, especially the interaction between the known polymerase module and the unknown nuclease module. The crosslinking MS was repeated two times with different concentration of BS3 and DSS. From the result, I found that CPSF100 might be the bridge between the polymerase and nuclease module, because there are quite some crosslinks between CPSF160 and CPSF100 and also between WDR33 and CPSF100, even at very high threshold (Figure 2.8), for example the linking between CPSF160-Lys1420 and CPSF100-Lys450 and the linking between WDR33-Lys109 and CPSF100-Lys499. And these crosslinkings were very repeatable. I tried to map the crosslinking residues to the polymerase module and most of them were at the same side as the extra density (data is not shown), which means the density map is reliable from the crosslinking MS perspective.

A



B

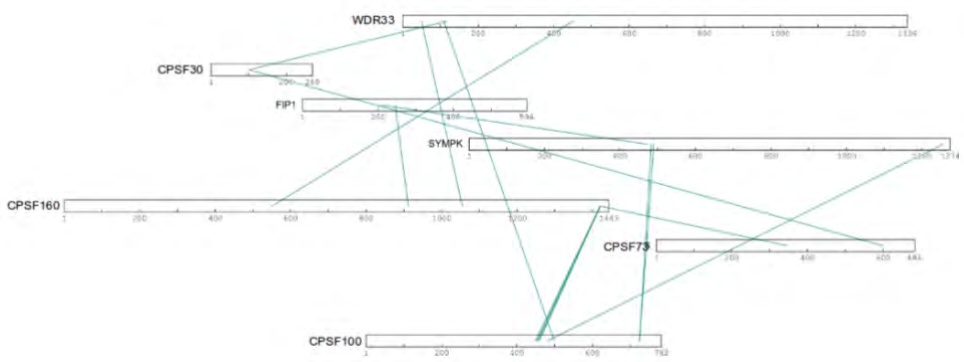


Figure 2.8: Crosslinking MS result of CPSF-symplekin complex. A, crosslinking with 2mM BS3 and 1mM DSS on ice for 2 hours for, B, crosslinking with 1mM BS3 and 0.5mM DSS on ice for 1 hour.

4.1.4 CPSF-symplekin complex cleavage activity

CPSF73 is the endonuclease required for the pre-mRNA 3' cleavage. It took a long time for the identification of the endonuclease responsible for 3' cleavage. In 2006, the crystal structure of CPSF73 was solved by Tong lab (Mandel et al., 2006). In this work they defined CPSF73 as the endonuclease for pre-mRNA 3' cleavage by both domain analysis and activity assays. In addition, they showed that the endonuclease activity of CPSF73 needs both Ca^{2+} and Zn^{2+}. Before the activity assay, CPSF73 was incubated with 5mM Ca^{2+} for 30 minutes to be activated in vitro. However, this is a very high Ca^{2+} concentration compare to the concentration of free Ca^{2+} in mammalian cells, which is normally between 10 to 100nM (Milo, 2017). So it is not clear if the assay can represent the real case in the cells. In vivo, CPSF73 is part of the big 3' processing machinery, the definition of the exact cleavage site and stimulation of CPSF73 activity might need more factors participating. In this work, CPSF73 was assembled into the whole CPSF complex together with symplekin and the complex was tested for pre-mRNA cleavage. In case the complex can indeed cleave the pre-mRNA, several question follow: (i) what is the minimum complex which is active in cleavage?, (ii) is SYMPK-CPAF100/73 complex active?, (iii) which divalent ion is necessary for the cleavage? To answer these questions, I performed some very initial cleavage activity assays with purified components. While CPSF73 only and syplekin-CPSF100/73 complex had very weak cleavage activity, the CPSF-symplekin complex showed increased activity. Thus, even though the activity of the CPSF-symplekin complex was quite weak, it is much stronger than the activity of CPSF73 and the symplekin-CPSF100/73 complex alone (data is not shown). Different time points from 30 minutes to 2 hours were taken and no obvious differences were found between half an hour and 2 hours. 0.5uM Ca^{2+}, Mg^{2+} and Zn^{2+} were used for the assay, respectively, and no difference was seen (data is not shown). This means if CPSF73 was assembled into the 3' processing machinery, it might show some endonuclease activity, and whether divalent ion is necessary in endonuclease activity is not

clear. All these results need to be further verified by more experiments. The structural information might also help to clarify the cleavage activity in the future.

4.2 CstF complex and DSE

In this work, I managed to crystallize human CstF complex, which was thought as very flexible in early studies. However the crystals from the initial screening did not diffract at all. In the later work, I generated two strategies to improve the diffraction of the crystals: limited proteolysis and binding the protein with RNA. For limited proteolysis, the protein was incubated with different kinds of proteases respectively (Proti-Ace Kit - Hampton Research, protease to protein ratio 1:100) on ice overnight and then used for setting up drops. Crystals fished from one of the 'clostripan digested' drops showed some diffraction (~8 Å), however, the crystals need to be further improved. In the future, some constructs can be designed based on the previous experiments and secondary structure prediction. By cutting off some flexible loop inside the protein, hopefully the diffraction of the crystal can be improved. The limited proteolysis fragments can be sent for MS analysis to get some information for constructs design. Moreover, CstF works as a dimer in 3' processing based on most of the literature, the question is: if CstF complex itself is organized as a dimer or it is organized as a dimer when associated with the 3' processing machinery. If CstF complex itself is a heterodimer, then it might also be a good candidate for cryo-EM analysis. Because the molecular weight of the dimer is around 400 kDa, which is neither too small nor too big for EM structure analysis. However, maybe some strategies need to be taken to fix the flexibility problem.

DSE is thought as the binding elements of CstF in canonical model. The conserved sequence YGTGTTYY (Y=pyrimidine) of DSE was defined in 1985 (McLauchlan et al., 1985) and the PAR-CLIP data of CstF64 in 2012 showed the similar result (Martin et al., 2012). Based on these studies, two RNA fragments with different length and different sequences (CUGUCU and UGUGUUUU) were designed for CstF binding tests. Fluorescence anisotropy showed very weak binding affinity even with very low salt concentration buffer (Kd >1uM in 30mM NaCl buffer), the binding affinity of longer RNA is a bit higher than the shorter one (Kd ≈0.8uM) (data not show). This might imply that the binding of CstF complex to RNA depends more on the length of RNA instead of the sequence itself, or this might suggest that the binding of CstF to DSE is quite dynamic. In the future, maybe some studies can be done to detect the conserved RNA secondary structure around PAS site, because the secondary structure of RNA might be the target for the recognition of some protein factors.

4.3 Human Pcf11 and termination

In this work, I managed to purify the human Pcf11-Clp1 complex. So far this is the least well understood factor in mammalian 3' processing complex. Most of the studies were about the yeast CFIA complex (Pcf11-Clp1-Rna14-Rna15). In yeast, Pcf11 has independent function in 3'end processing and termination (Sadowski et al., 2003). It was shown to bind specifically to Ser2-P phosphorylated CTD peptides, the dephosphorylation of Ser2-P by some phosphatase

would lead to the dissociation of termination factors from CTD, which helps to regenerate the initiation competent Pol II (Meinhart and Cramer, 2004).

Pcf11 in humans is twice as big as its yeast homolog and shares CTD-interaction domain at its N terminal. In this work, I purified the full length Pcf11-Clp1 complexe and was planning to assemble it to a Pol II termination complex which includes almost all the termination factors. However, recent study showed that Pcf11 might be a regulatory factor rather than a core subunit of the 3' end processing, which means Pcf11 binds transiently to the complex (Kamieniarz-Gdula et al., 2019; Shi et al., 2009). However, as Pcf11 was shown to interact directly with Ser2-P CTD peptides in yeast, I think it is deserved to check the interaction between human Pcf11 and Ser2-P CTD peptides or the interaction between Pcf11 and Pol II. For the Pol II-Pcf11 interaction, the specific Ser2-P might be important. Very recently we found CDK12/CCNK has very specific Ser2 phosphorylation function, this might help us to form the 'termination complex' in the future. Because Pcf11 might be an important factor to connect termination and 3' end processing. In the future, maybe some in vitro transcription and co-transcriptional 3'end processing experiment can be done to find the stable termination-3' end processing complex which is suitable for structure analysis.

4.4 CFIm68 and SR proteins

The SR (Ser-Arg-rich) protein family is featured by an N termination RNA recognition motif (RRM) and a C terminal RS domain with variable length, RS domain is rich in arginine and serine (Sahebi et al., 2016). For some SR proteins, they have more than one RRM (Shepard and Hertel, 2009). SR proteins mainly function in pre-mRNA processing, especially for pre-mRNA splicing. They also function in various post-splicing activities, which include mRNA nucleus localization, nuclear export and translation (Graveley, 2000; Sanford et al., 2004). SR proteins interact with RNA simultaneously via its RRM and interact with other protein factors via the RS domain (Graveley et al., 1998; Wu and Maniatis, 1993). The phosphorylation regulation on serine of RS domain is an important process for the function of SR proteins. Both hyper- and hypo phosphorylated RS domain are unable to support splicing anymore (Graveley, 2000). SR proteins were first identified by the study of splicing in Drosophila (Tze-Bin Chou, 1987). The definition of SR proteins is based on the presence of a phosphoepitope which can be recognized by the monoclonal antibody mAb104 and the conservation across vertebrates and invertebrates (Roth et al., 1990). There are nine members of human SR protein family which have similar structure organization (figure 2.9). Most SR proteins are enriched in nuclear compartments termed speckles, which can be seen throughout the nucleus (Lafarga et al., 2009). RS domain is responsible for targeting SR proteins to speckles (Spector, 1993). Recently, another group of proteins were found which have similar structure organization as SR proteins but might not be recognized by the monoclonal antibody mAb104. These proteins were named as SR-like proteins (Long and Caceres, 2009).

CFIm68 (also named as CPSF6) is a typical SR-like protein with both the N terminal RRM and C terminal RS domain (Figure 2.8). CFim68 is involved in pre-mRNA 3' processing and binds specifically to the upstream sequence UGUA as mentioned before. The function of CFIm68 is

highly regulated by phosphorylation (Jang et al., 2019). In my purification, I was trying to purify CFIm68 and CFIm25 as a complex, CFIm68 was tagged at the N terminal with His-MBP tag and the complex was expressed in insect cells. Because insect cell expression might introduce some extra phosphorylation to the protein, normally we put lamda phosphatase to the overnight dialysis protein (the amylose elution protein with TEV) to remove the extra phosphorylation. However, I found a hydrogel inside the dialysis bag after overnight dialysis with TEV and lamda phosphatase. To figure out whether the phase separation was from TEV cleavage or from dephosphorylation, the protein was purified again following the same protocol and controls were set up for dialysis (TEV only, Lamda phosphatase only, none and both). I found that lamda phosphatase caused the phase separation because hydrogel was only found in the protein solution with lamda phosphatase but not in the TEV control (data is not shown). This means CFIm68/25 complex was phase separated after dephosphorylation. I repeated the experiment two more times and result was the same, which means this is really not an accident phenomenon and it deserves more attention. In later step, I checked the domain composition of both proteins and found that CFIm68 is a SR like protein and the dephosphorylation of RS domain might cause the phase separation.

The crystal structure of CFIm68/25 was solved in 2011 (Yang et al., 2011), however the RS domain of CFIm68 was not included in the structure, maybe because the RS domain was too flexible and it was cut out for crystallization in this work. However, the RS domain might be an important domain for the function of CFIm68. Consistent with my speculation, recent studies showed that the phosphorylation state of CFIm68 is related with nuclear import and alternative polyadenylation (which would be discussed later(Jang et al., 2019). This study concluded that both the binding of CFIm68 to transportin 3 (TNPO3, a nuclear transport protein) and import to nucleus were independent of phosphorylation, but the hyperphosphorylated CFIm68 might lead to the failure of nuclear import. In addition to this study, very early literature also showed that the speckles in nuclear is not only an apartment of SR proteins and RNAs, but also a location for transcription factors and 3’ processing factors, which means the nuclear speckles might be a ‘processing factory’ of RNAs (Schul et al., 1998; Zeng et al., 1997). The concentrated factors in the compartment might make the RNA processing (including splicing and 3’ end processing) more efficient. However, more work need to be done in the future to clarify the function of the speckles and the relationship between RNA processing and transcription.

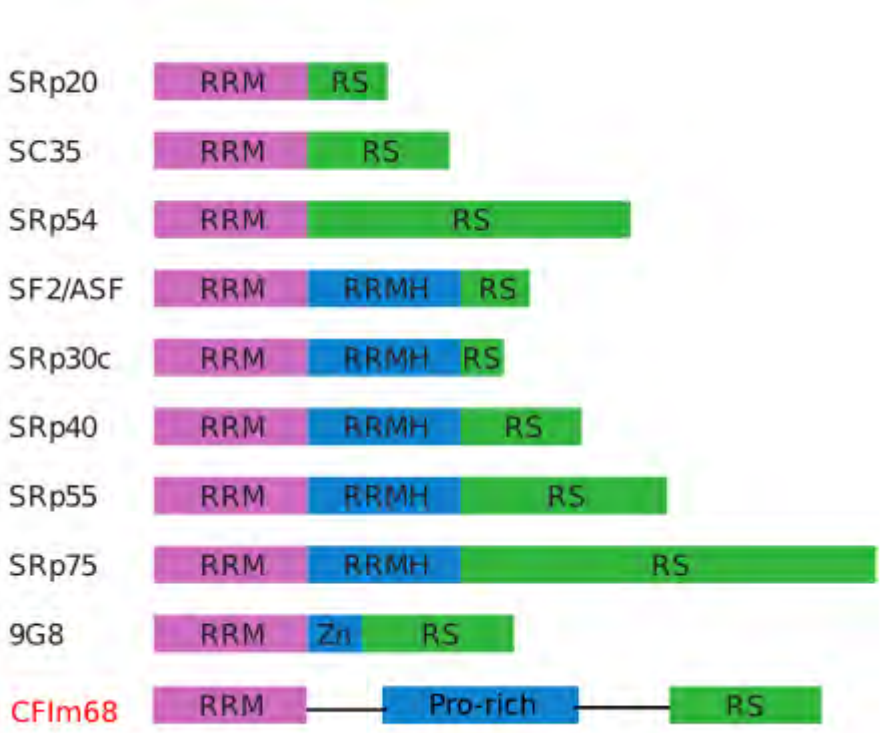


Figure 2.9: SR protein family. There are 9 proteins in SR family which share the similar features with an N terminal RRM domain and C terminal RS domain. CFIm68 is a 3' processing factor which shares similar feature with SR protein family and is named as SR-like protein. The figure was adapted from Shepard and Hertel: Genome Biology 2009.

4.5 CFI complex and alternative cleavage and polyadenylation

In more than 50% of human genes, pre-mRNA cleavage and polyadenylation can occur at multiple locations at the 3' end through a process called alternative polyadenylation (APA) (Tian et al., 2005). APA is a co-transcriptional process that expands mRNA transcript diversity. There are two different types of APA based on the positional difference: In some cases, APA occurs inside the coding region which results in different protein isoforms, this type of APA is referred to as coding region APA (CR-APA); In other cases, APA sites are located in the 3' untranslated region (3' UTR), the coding region is the same but the length of UTR changes, which affects the expression efficiency of the protein, this type of APA is known as UTR-APA (Di Giammartino et al., 2011). APA is highly regulated in different tissues and in tumorigenesis (Ji et al., 2009; Masamha and Wagner, 2018). The function of CR-APA is similar to alternative splicing, which also results in different isoforms of a protein, while regulation of UTR-APA is more complicated, as 3' UTRs often harbor microRNA (miRNA) and protein factors binding sites (Barreau et al., 2005; Fabian et al., 2010). Different lengths of UTRs affect not only the stability but also the transportation, localization and translational efficiency of mRNA. Transcripts with shorter UTR produce higher level of proteins (Mayr and Bartel, 2009; Sandberg et al., 2008), which normally happens in cell proliferation or in pathological cases like cancers, while in normal cell differentiation, 3' UTR tends to be longer (Ji et al., 2009; Wang et al., 2008).

Early studies showed that the knock down of CFI (either CfIm68 or CFIm25) leads to the shorter 3' UTR (Martin et al., 2012; Masamha et al., 2014) and that CFIm functions as an enhancer dependent activator of mRNA 3' processing (Zhu et al., 2018). Recent studies also showed that phosphorylation regulation of CFIm68 RS domain was important for the length control of 3'UTR (Jang et al., 2019). This study showed that the hypophosphorylated or physiologically phosphorylated RS domain of CFIm68 has normal binding affinity with TNPO3, which correlates with normal CFim68 nuclear import and distal or middle PAS usage (long 3' UTR), while the hyperphosphorylated RS domain would cause the interaction defect of CFIm68 and TNPO3. The result is that CFIm68 cannot be imported to the nucleus, and the cells have to use CFIm59 as an alternative PAS recognition factor, which would lead to the proximal PAS usage and shorter 3'UTR. However, the clear mechanism for the long or short 3' UTR regulation is still not clear. In the future, more work should be done to figure out the detailed factors participation, interaction and regulatory mechanism.

Even though APA was well studied during the last several years, we still have very limited knowledge about it. Further studies on the APA are required in the future, with a special emphasis on its regulatory function in tumorgenesis (Masamha and Wagner, 2018).

4.6 Definition of the endonuclease for pre-mRNA cleavage

Transcription is an important part of the central dogma in molecular biology. Since Pol II is responsible for the transcription of all protein coding genes, its structure has been studied for many years. In 2000, the first Pol II structure was solved (Cramer et al., 2000), which opened the door for the structural study of transcription. Perhaps because of its foremost position in the transcription cycle, transcription initiation was quite well studied, including the promoter recognition, binding of initiation factors and Pol II, DNA opening and synthesis of the initial RNA (8~9 nucleotides). More recently, the transition phase in the transcription cycle between the initiation and elongation (also known as the promotor-proximal pausing), was structurally characterized (Vos et al., 2018a; Vos et al., 2018b). However, transcription termination is the least known process in the whole transcription cycle. The regulation of Pol II termination is an important process because it defines the boundaries of the transcription unit and avoids the interfering of transcription between different genes. More importantly, termination guarantees that the Pol II can be recycled timely for the initiation of a new round of transcription.

Pol II termination was well studied in yeast for both non-coding RNAs and mRNAs. There is a big machinery which is responsible for termination and mRNA 3' processing both in yeast and in humans. As the cleavage of pre-mRNA is an important step in polyadenylation and maybe also in termination, scientists put a lot effort to figure out which endonuclease is responsible for the pre-mRNA cleavage. In 2006, Tong. L and colleagues solved the structure of human CPSF73 and yeast CPSF 100(Mandel et al., 2006). They defined CPSF73 as the endonuclease based on the structure analysis of the active site and sequence analysis. However, they also showed that CPSF73 can only exhibit its endonuclease activity, along with exonuclease activity, after incubating with Ca^{2+} at 37°C for half an hour. However the

concentration of Ca^{2+} they used for assay is much higher than the concentration physiologically, so there must be some other mechanisms and protein factors which function to stimulate the nuclease activity of CPSF73 and to define the exact cleavage site in vivo. Very recently, In March of 2019, Lori A. Passmore and her team defined the 8-subunit core that is necessary for the activation of endonuclease in pre-mRNA 3' end processing in yeast (Hill et al., 2019). The 8-subunit core includes Ysh1, Cft2, Mpe1, CFIA, a complex of Rna14, Rna15, Pcf11 and Clp1 and CFIB (Hrp1), which was named CPF core. After association with the CPF core, Ysh1 was active to cleave the pre-mRNA at the cleavage site in vitro. The corresponding factors for Ysh1 and Cft2 in human are CPSF73 and CPSF100 respectively. For CFIA, there are two sub-complexes in human, which are CstF complex and CFII (Pcf11 and Clp1). CFIB corresponds to CFI (CFIm68/25) complex. Even the 3' processing factors are highly conserved in yeast and human. There are some differences, for example, as mentioned before, CFIA complex is composed of Rna14, Rna15, Pcf11 and Clp1 in yeast, however the homolog CstF complex and Pcf11/Clp1 did not prefer to form a complex in vitro from my all my eperiments. They have no interactions in my assays. RBBP6, the human homolog of Mpe1, was not well studied, even though Mpe1 is a very important factor for the pre-mRNA cleavage in yeast. Also, symplekin forms a stable complex with CPSF100 and CPSF73 in our work. The corresponding protein factor of symplekin in yeast is Pta1, which fell off the CPF core and didn't show much importance in cleavage (Hill et al., 2019). In the future, more work should be done to figure out the corresponding 'CPF core' in humans.

In this work so far, I managed to reconstitute the whole CPSF complex plus symplekin, which includes CPSF160, CPSF100, CPSF73, WDR33 (1-572aa), CPSF30, Fip1 and symplekin. The symplekin-CPSF100/73 complex was purified first and then combined with CPSF complex polymerase module. As discussed before, CPSF73 shows some endonuclease activity when assembled in the whole complex. However, it is still uncertain if the CPSF complex combined with the symplekin is sufficient for the full pre-mRNA cleavage activity or additional factors also play a role. Hopefully further biochemical and structural studies can offer more insights into this question.

4.7 Future perspectives

4.7.1 Termination pausing and the disengagement of Pol II from template DNA

The promoter-proximal pausing is well known and studied as introduced in chapter 1. However, there is another pausing, which occurs when Pol II transcribes over the poly(A) tail, is not so well studied. Whether pausing is a prerequisite for termination is still under debate in both yeast and mammals (Creamer et al., 2011; Plant et al., 2005). So far only two sequences were discovered to function in termination pausing: The MAZ element which is found in liver-specific C2 complement gene and featured with the sequence G5AG5 (Ashfield et al., 1994), and CCAAT-box which was found in the adenovirus (Connelly and Manley, 1989). Pausing was thought to occur before termination when Pol II slows down at the termination region. One explanation is that pausing after transcribing over poly(A) signal gives enough time for either the catching up of XRN2 (torpedo model) or the exchange of

elongation factors to termination factors (allosteric model)(Glover-Cutter et al., 2008). If this is the hypothesis, then what is the driving force for pausing and how is the paused elongation complex (EC) released? All these questions need a further study in the future. For now the hypothesis might be that the CPSF complex is bound with elongation complex during elongation and 'scans' for the termination signal, once EC 'walks' over the termination signal, CPSF would bind specifically to both pre-mRNA PAS site and Pol II and cause the initial pausing, then the binding of the other termination factors like CstF would stabilize the pausing. The release of Pol II from template DNA might need more interactions/competition between different factors or Pol II (Nag et al., 2007) and the regulation of kinase/phosphatase of both CTD and termination factors, just like the promoter-proximal Pol II release (Vos et al., 2018a). However, the hypothesis needs to be verified in the future by more experiments. The cleavage of pre-mRNA is a prerequisite for allosteric model, so figuring out the relationship between pre-mRNA cleavage and termination might help for getting the correct termination model.

4.7.2 Termination and re-initiation

In a simplistic view, a gene is a linearized DNA sequence with a promoter at one end and terminator at the other end. In transcription, Pol II initiated from promoter with the help of general factors and terminates at the end of the gene with the termination factors binding. After releasing from the template DNA, Pol II was recycled and the new transcription started (Fuda et al., 2009)(Figure 2.9A). However, in the real cells, it is much more complicated. The main question for this simplified model is how efficient Pol II can be recycled, because there is a big space gap between promoter and terminator. Further studies showed that termination and initiation have a crosstalk based on several discoveries. The first finding is that mutation of poly(A) site impaired termination and also decreased the initiation of the same gene, which implies a relationship between termination and re-initiation, and also the mutation or deletion of termination factors impairs the initiation (Mapendano et al., 2010). The second interesting observation is the direct interaction between initiation factors and termination factors, like the interaction between TFII D and CPSF (Dantonel et al., 1997), the binding of phosphorylated TFII B and CstF complex (Wang et al., 2010). Both these observations imply that initiation and termination didn't occur independently, they must have a crosstalk to make the transcription more efficient. The question is, how? There are three putative models so far. The first, also the most widely accepted model, is the gene-looping. The hypothesis is that the promoter and terminator of the same gene have direct physical contacts or at proximity via initiation or termination factors, which ensured that the released Pol II can be recycled efficiently for the re-initiation (Figure 2.9B)(El Kaderi et al., 2009; Hampsey et al., 2011; O'Sullivan et al., 2004). By using the method of chromosome conformation capture (3C) method (Tolhuis et al., 2002), the study showed that the gene loops are dynamic structures which form upon transcription. The formation of the gene loops need the interaction of initiation factor TFIIB and termination factors Pta1 and CPF, the interaction was regulated by the phosphatase SSu72 (Hampsey et al., 2011). The second possibility is that the active genes are positioned in a sub compartment, where both Pol II,

initiation and termination factors are highly compacted, which makes the transcription and Pol II cycling much more efficient (Osborne et al., 2004; Yao et al., 2007)(Figure 2.9C). These sub-compartments are membraneless organelles which are formed by phase separation. The most well-known membraneless organelle is the nucleolus. In recent years, phase separation was also found in transcriptional regulation (DenesHnisz, 2017). However, even this theme has been demonstrated in several organisms, its study is still at the very early stage. Maybe in the future, more phase separation study of Pol II and transcription would offer more evidence knowledge for this theme. Thirdly, the termination and initiation might occur in long distance but they communicate via extra signals like chromatin structure, chromatin-associated factors or some small non-coding RNAs (Figure 2.9D). In some early studies, people even assumed that the terminator might work as a promoter or vice versa. I speculate that all the transcription and RNA processing factors might concentrate in one compartment in the nucleus, which makes the transcription and pre-mRNA processing much more efficient. However, the gene looping is also necessary for efficient transcription. In the future, more work can be done for the 'transcription efficiency' study.

4.7.3 Termination is a regulatory way for gene expression

Termination is not only a way for genomic partitioning, but also a regulatory mechanism for gene expression. Instead of the conventional concept that transcription starts from promoter and ends at terminator, recent studies showed that the genome is highly transcribed, even the non-coding area. The entirely transcription needs some ways for regulation, termination is one way. Termination occurs not only at the end of one gene but also the beginning and middle of the ORF, which is an important way for transcription regulation. The typical example for termination regulation is the clusters of amino acid biosynthesis genes in bacteria (Merino E, 2005), which is named as premature termination. When enough amino acids exist in the cells, a termination complex would form at the 5' UTR, which would release Pol II from the template DNA before it going to the protein coding region. The regulation complex is normally composed of protein factors and some non-coding RNAs (Naville and Gautheret, 2010). Premature termination or attenuation was also discovered in virus or eukaryotic organisms and was thought as a widespread regulatory way (Kim and Levin, 2011). Defective termination affects both co-transcriptional splicing and RNA synthesis and stability. In the future, more study can be focused on the early termination regulation of gene expression.

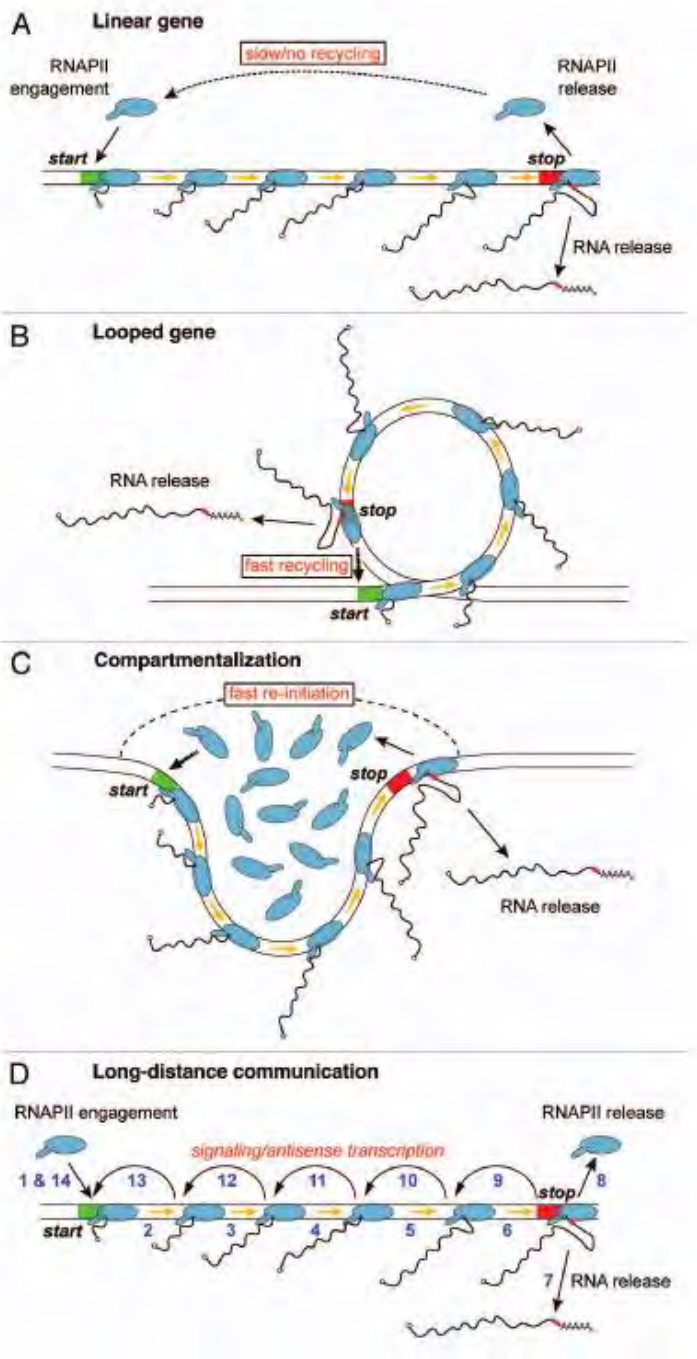
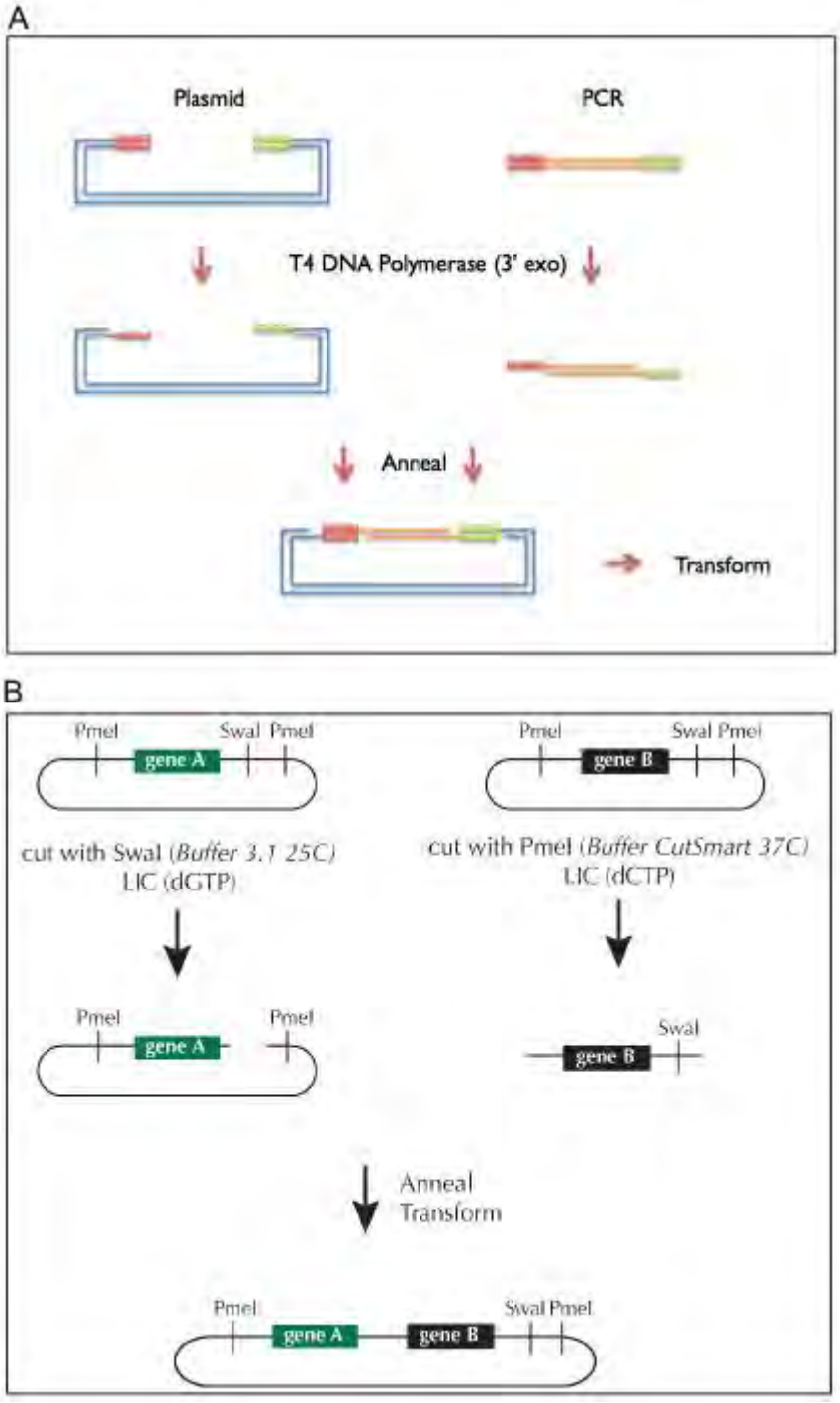
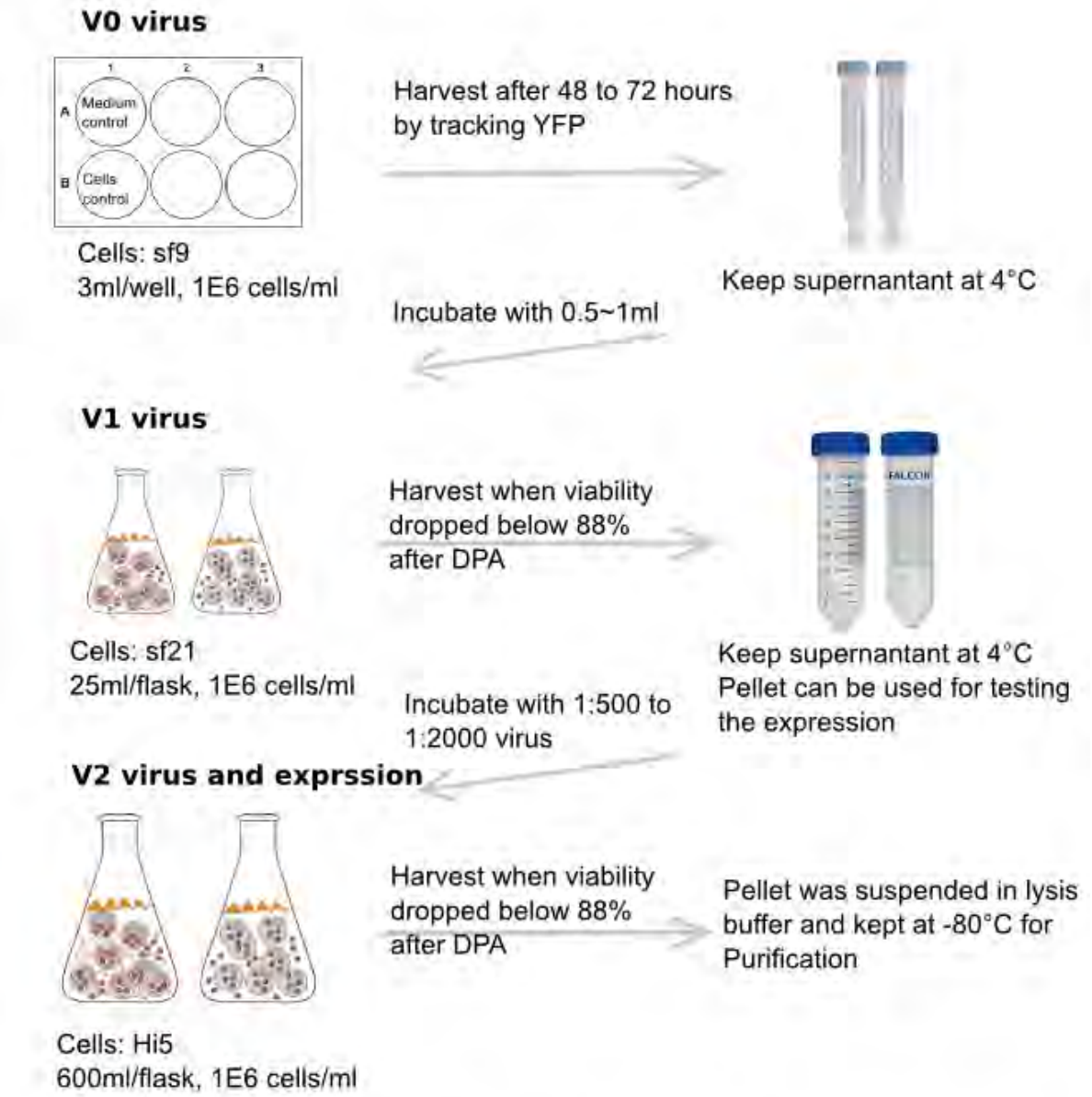


Figure 2.10: Hypothesis for efficient termination and re-initiation. A, Classical linear initiation and termination, the polymerase cannot be recycled efficiently. B, gene looping brings promoter and terminator closer, which is more efficient for transcription. C, there are a lot of factors and polymerase in the small compartment, which made the re-initiation more efficient. D, the gene itself is linear but the initiation and termination have some crosstalk via signal. Diagram adapted from Søren Lykke-Andersenet al., cell cycle 2011

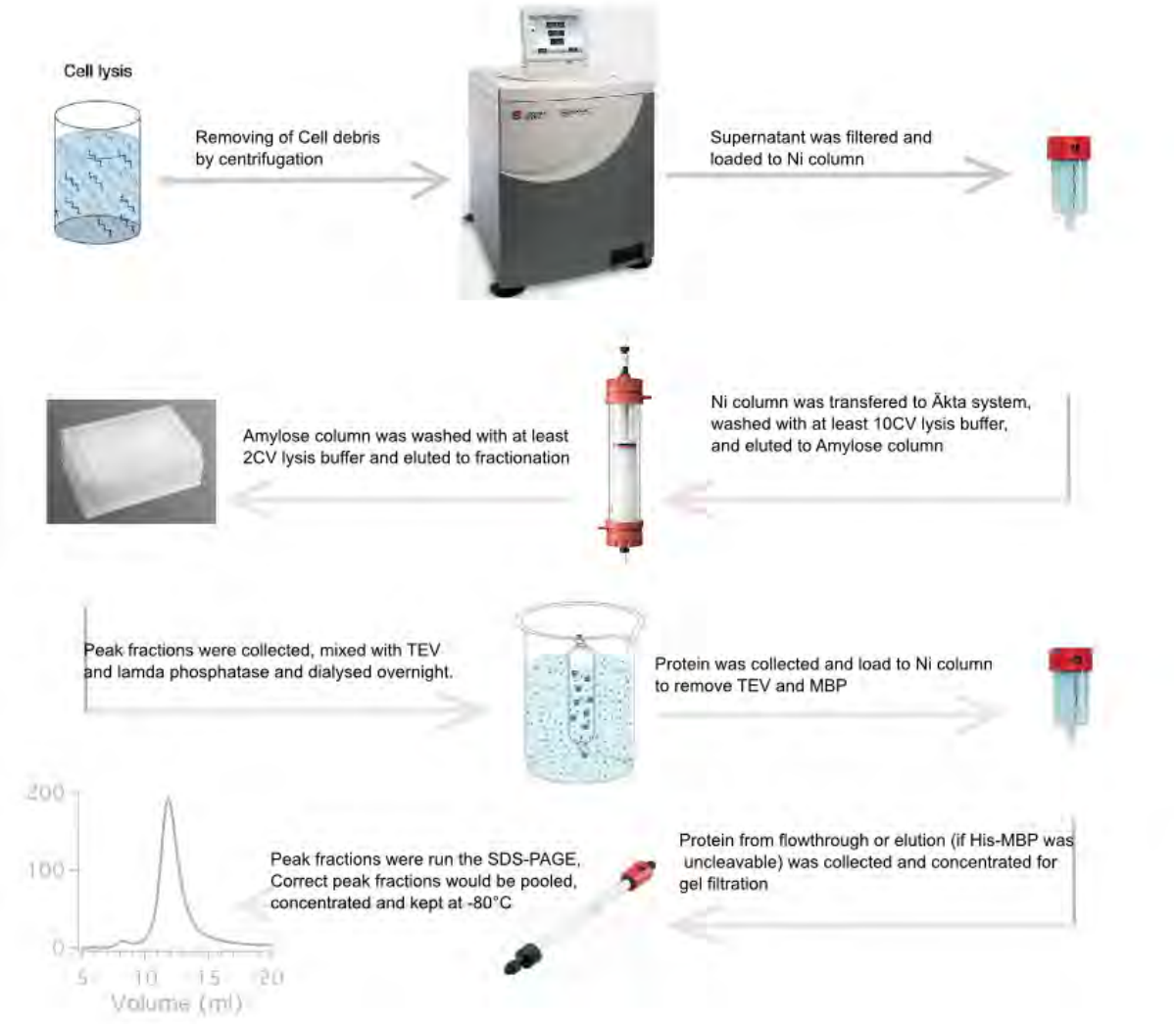
Supplemental materials



Supplementary Figure 1: LIC cloning flowchart, adapted from MacroLab_Vectors_v8 April 3, 2014. A, the linearized vector and PCR product were treated with T4 DNA polymerase separately with corresponding dNTPs (dCTP or inserts and dGTP for vector) to generate the overhangs for annealing. The nicks after annealing can be repaired by *E.coli* cells after transformation. B, the acceptor and doner vector were digested with SwaI and PmeI respectively. The 'LIC' method was used again for the ligation and the annealed vector was transformed to *E.coli* cells.



Supplementary Figure 2: Protein expression in insect cells. V0, V1 and V2 (or expression) are made with sf9, sf21 and hi5 cells respectively. Cells should be checked every day to keep its viability and avoid contamination.



Supplementary Figure 3: Protein purification flowchart by His-MBP tag. The supernatant from the lysate was loaded with peristaltic pump, the wash and elution of both Ni column and amylose column was executed by Äkta system (GE healthcare)

List of Abbreviations

RNAP	RNA polymerase
rRNA	ribosome RNA
snRNA	small nuclear RNA
tRNA	transfer RNA
pre-mRNA	precursor message RNA
Pol II	RNA polymerase II
EC	Elongation complex
ITC	Initiation transcription complex
Hepes	2-[4-(2-hydroxyethyl)piperazin-1-yl]ethanesulfonic acid
β-ME	β-mercaptoethanol
His	Histidine
nt	Nucleotides
kDa	Kilodalton
CTD	RNA polymerase II C-terminal domain
NAC	Nucleotide addition cycle
EM	Electron microscopy
IPTG	isopropyl-β-D-1-thiogalactopyranoside
LB	Lysogeny broth
NTP	nucleotide triphosphate
SDS	sodium dodecyl sulfate
UV	ultra violet
CTF	contrast transfer function
ncRNA	non-coding RNA
3’ UTR	3’ untranslated region

References

Adelman, K., and Lis, J.T. (2012). Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. Nat Rev Genet *13*, 720-731.

Ahlquist, P. (2002). RNA-dependent RNA polymerases, viruses, and RNA silencing. Science *296*, 1270-1273.

Ahn, S.H., Kim, M., and Buratowski, S. (2004). Phosphorylation of serine 2 within the RNA polymerase II C-terminal domain couples transcription and 3' end processing. Molecular cell *13*, 67-76.

Arigo, J.T., Eyler, D.E., Carroll, K.L., and Corden, J.L. (2006). Termination of cryptic unstable transcripts is directed by yeast RNA-binding proteins Nrd1 and Nab3. Molecular cell *23*, 841-851.

Armache, K.J., Kettenberger, H., and Cramer, P. (2003). Architecture of initiation-competent 12-subunit RNA polymerase II. Proc Natl Acad Sci U S A *100*, 6964-6968.

Ashfield, R., Patel, A.J., Bossone, S.A., Brown, H., Campbell, R.D., Marcu, K.B., and Proudfoot, N.J. (1994). MAZ-dependent termination between closely spaced human complement genes. EMBO J *13*, 5656-5667.

Bacon, C.W., and D'Orso, I. (2019). CDK9: a signaling hub for transcriptional control. Transcription *10*, 57-75.

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

Balbo, P.B., and Bohm, A. (2007). Mechanism of poly(A) polymerase: structure of the enzyme-MgATP-RNA ternary complex and kinetic analysis. Structure *15*, 1117-1131.

Balbo, P.B., Toth, J., and Bohm, A. (2007). X-ray crystallographic and steady state fluorescence characterization of the protein dynamics of yeast polyadenylate polymerase. J Mol Biol *366*, 1401-1415.

Barabino, S.M., Hubner, W., Jenny, A., Minvielle-Sebastia, L., and Keller, W. (1997). The 30-kD subunit of mammalian cleavage and polyadenylation specificity factor and its yeast homolog are RNA-binding zinc finger proteins. Genes Dev *11*, 1703-1716.

Bard, J., Zhelkovsky, A.M., Helmling, S., Earnest, T.N., Moore, C.L., and Bohm, A. (2000). Structure of yeast poly(A) polymerase alone and in complex with 3'-dATP. Science *289*, 1346-1349.

Bartkowiak, B., Liu, P., Phatnani, H.P., Fuda, N.J., Cooper, J.J., Price, D.H., Adelman, K., Lis, J.T., and Greenleaf, A.L. (2010). CDK12 is a transcription elongation-associated CTD kinase, the metazoan ortholog of yeast Ctk1. Genes Dev *24*, 2303-2316.

Bartolomei, M.S., and Corden, J.L. (1987). Localization of an alpha-amanitin resistance mutation in the gene encoding the largest subunit of mouse RNA polymerase II. Molecular and Cellular Biology *7*, 586-594.

Bartolomei, M.S., and Corden, J.L. (1995). Clustered α-amanitin resistance mutations in mouse. MGG Molecular & General Genetics *246*, 778-782.

Baumann, K., Zanotti, G., and Faulstich, H. (2008). A β-turn in α-amanitin is the most important structural feature for binding to RNA polymerase II and three monoclonal antibodies. Protein Science *3*, 750-756.

Beaudoing, E., Freier, S., Wyatt, J.R., Claverie, J.M., and Gautheret, D. (2000). Patterns of variant polyadenylation signal usage in human genes. Genome Res *10*, 1001-1010.

Bentley, D.L. (2005). Rules of engagement: co-transcriptional recruitment of pre-mRNA processing factors. Curr Opin Cell Biol *17*, 251-256.

Bentley, D.L. (2014). Coupling mRNA processing with transcription in time and space. Nat Rev Genet *15*, 163-175.

Bernecky, C., Herzog, F., Baumeister, W., Plitzko, J.M., and Cramer, P. (2016a). Structure of transcribing mammalian RNA polymerase II. Nature *529*, 551-554.

Bieniossek, C., Papai, G., Schaffitzel, C., Garzoni, F., Chaillet, M., Scheer, E., Papadopoulos, P., Tora, L., Schultz, P., and Berger, I. (2013). The architecture of human general transcription factor TFIID core complex. Nature *493*, 699-702.

Blobel, G. (1973). A protein of molecular weight 78,000 bound to the polyadenylate region of eukaryotic messenger RNAs. Proc Natl Acad Sci U S A *70*, 924-928.

Bosken, C.A., Farnung, L., Hintermair, C., Merzel Schachter, M., Vogel-Bachmayr, K., Blazek, D., Anand, K., Fisher, R.P., Eick, D., and Geyer, M. (2014). The structure and substrate specificity of human Cdk12/Cyclin K. Nature communications *5*, 3505.

Brickey, W.J., and Greenleaf, A.L. (1995). Functional studies of the carboxy-terminal repeat domain of Drosophila RNA polymerase II in vivo. Genetics *140*, 599-613.

Brueckner, F., and Cramer, P. (2008). Structural basis of transcription inhibition by α-amanitin and implications for RNA polymerase II translocation. Nature Structural & Molecular Biology *15*, 811-818.

Brueckner, F., Hennecke, U., Carell, T., and Cramer, P. (2007). CPD damage recognition by transcribing RNA polymerase II. Science (New York, N.Y.) *315*, 859-862.

Buku, A., Wieland, T., Bodenmuller, H., and Faulstich, H. (1980). Amaninamide, a new toxin of Amanita virosa mushrooms. Experientia *36*, 33-34.

Buratowski, S. (2009). Progression through the RNA polymerase II CTD cycle. Molecular cell *36*, 541-546.

Bushnell, D.A., Cramer, P., and Kornberg, R.D. (2002). Structural basis of transcription: α-Amanitin-RNA polymerase II cocrystal at 2.8 A resolution. Proceedings of the National Academy of Sciences *99*, 1218-1222.

Casanal, A., Kumar, A., Hill, C.H., Easter, A.D., Emsley, P., Degliesposti, G., Gordiyenko, Y., Santhanam, B., Wolf, J., Wiederhold, K.*, et al.* (2017). Architecture of eukaryotic mRNA 3'-end processing machinery. Science *358*, 1056-1059.

Cermakian, N., Ikeda, T.M., Miramontes, P., Lang, B.F., Gray, M.W., and Cedergren, R. (1997). On the evolution of the single-subunit RNA polymerases. J Mol Evol *45*, 671-681.

Chan, S., Choi, E.A., and Shi, Y. (2011). Pre-mRNA 3'-end processing complex assembly and function. Wiley Interdiscip Rev RNA *2*, 321-335.

Chen, F., MacDonald, C.C., and Wilusz, J. (1995). Cleavage site determinants in the mammalian polyadenylation signal. Nucleic acids research *23*, 2614-2620.

Chen, Y., Weeks, J., Mortin, M.A., and Greenleaf, A.L. (1993). Mapping mutations in genes encoding the two large subunits of Drosophila RNA polymerase II defines domains essential for basic transcription functions and for proper expression of developmental genes. Molecular and Cellular Biology *13*, 4214-4222.

Chou, Z.F., Chen, F., and Wilusz, J. (1994). Sequence and position requirements for uridylate-rich downstream elements of polyadenylation signals. Nucleic acids research *22*, 2525-2531.

Christofori, G., and Keller, W. (1988). 3' cleavage and polyadenylation of mRNA precursors in vitro requires a poly(A) polymerase, a cleavage factor, and a snRNP. Cell *54*, 875-889.

Clerici, M., Faini, M., Aebersold, R., and Jinek, M. (2017). Structural insights into the assembly and polyA signal recognition mechanism of the human CPSF complex. eLife *6*.

Clerici, M., Faini, M., Muckenfuss, L.M., Aebersold, R., and Jinek, M. (2018). Structural basis of AAUAAA polyadenylation signal recognition by the human CPSF complex. Nat Struct Mol Biol *25*, 135-138.

Cochet-Meilhac, M., and Chambon, P. (1974). Animal DNA-dependent RNA polymerases. 11. Mechanism of the inhibition of RNA polymerases B by amatoxins. Biochimica Et Biophysica Acta *353*, 160-184.

Cochet-Meilhac, M., Nuret, P., Courvalin, J.C., and Chambon, P. (1974). Animal DNA-dependent RNA polymerases. 12. Determination of the cellular number of RNA polymerase B molecules. Biochimica Et Biophysica Acta *353*, 185-192.

Connelly, S., and Manley, J.L. (1989). A CCAAT box sequence in the adenovirus major late promoter functions as part of an RNA polymerase II termination signal. Cell *57*, 561-571.

Cramer, P. (2007). Gene transcription: extending the message. Nature *448*, 142-143.

Cramer, P., Bushnell, D.A., Fu, J., Gnatt, A.L., Maier-Davis, B., Thompson, N.E., Burgess, R.R., Edwards, A.M., David, P.R., and Kornberg, R.D. (2000). Architecture of RNA polymerase II and implications for the transcription mechanism. Science (New York, N.Y.) *288*, 640-649.

Creamer, T.J., Darby, M.M., Jamonnak, N., Schaughency, P., Hao, H., Wheelan, S.J., and Corden, J.L. (2011). Transcriptome-wide binding sites for components of the Saccharomyces cerevisiae non-poly(A) termination pathway: Nrd1, Nab3, and Sen1. PLoS Genet *7*, e1002329.

Crick, F. (1970). Central dogma of molecular biology. Nature *227*, 561-563.

d'Aubenton Carafa, Y., Brody, E., and Thermes, C. (1990). Prediction of rho-independent Escherichia coli transcription terminators. A statistical analysis of their RNA stem-loop structures. J Mol Biol *216*, 835-858.

Dantonel, J.C., Murthy, K.G., Manley, J.L., and Tora, L. (1997). Transcription factor TFIID recruits factor CPSF for formation of 3' end of mRNA. Nature *389*, 399-402.

Darst, S.A., Kubalek, E.W., and Kornberg, R.D. (1989). Three-dimensional structure of Escherichia coli RNA polymerase holoenzyme determined by electron crystallography. Nature *340*, 730-732.

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

DenesHnisz, K., Richard A.Young, Arup K.Chakraborty, Phillip A.Sharp (2017). A Phase Separation Model for Transcriptional Control. Cell *169*, 13-23.

Dengl, S., and Cramer, P. (2009). Torpedo nuclease Rat1 is insufficient to terminate RNA polymerase II in vitro. J Biol Chem *284*, 21270-21279.

Di Giammartino, D.C., Li, W., Ogami, K., Yashinskie, J.J., Hoque, M., Tian, B., and Manley, J.L. (2014). RBBP6 isoforms regulate the human polyadenylation machinery and modulate expression of mRNAs with AU-rich 3' UTRs. Genes Dev *28*, 2248-2260.

Di Giammartino, D.C., Nishida, K., and Manley, J.L. (2011). Mechanisms and consequences of alternative polyadenylation. Molecular cell *43*, 853-866.

Dimitry Tegunov, P.C. (2018). Real-time cryo-EM data pre-processing with Warp. bioRxiv.

Dominski, Z., Yang, X.C., and Marzluff, W.F. (2005). The polyadenylation factor CPSF-73 is involved in histone-pre-mRNA processing. Cell *123*, 37-48.

Ebright, R.H. (2000). RNA polymerase: structural similarities between bacterial RNA polymerase and eukaryotic RNA polymerase II. J Mol Biol *304*, 687-698.

Edmonds, M. (1990). Polyadenylate polymerases. Methods Enzymol *181*, 161-170.

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

Egloff, S. (2012). Role of Ser7 phosphorylation of the CTD during transcription of snRNA genes. RNA Biol *9*, 1033-1038.

Egloff, S., and Murphy, S. (2008). Cracking the RNA polymerase II CTD code. Trends Genet *24*, 280-288.

Eick, D., and Geyer, M. (2013). The RNA polymerase II carboxy-terminal domain (CTD) code. Chem Rev *113*, 8456-8490.

El Kaderi, B., Medler, S., Raghunayakula, S., and Ansari, A. (2009). Gene looping is conferred by activator-dependent interaction of transcription initiation and termination machineries. J Biol Chem *284*, 25015-25025.

Emsley, P., Lohkamp, B., Scott, W.G., and Cowtan, K. (2010). Features and development of Coot. Acta Crystallographica. Section D, Biological Crystallography *66*, 486-501.

Engel, C., Sainsbury, S., Cheung, A.C., Kostrewa, D., and Cramer, P. (2013). RNA polymerase I structure and transcription regulation. Nature *502*, 650-655.

Epshtein, V., Mustaev, A., Markovtsov, V., Bereshchenko, O., Nikiforov, V., and Goldfarb, A. (2002). Swing-gate model of nucleotide entry into the RNA polymerase active center. Molecular cell *10*, 623-634.

Ezeokonkwo, C., Ghazy, M.A., Zhelkovsky, A., Yeh, P.C., and Moore, C. (2012). Novel interactions at the essential N-terminus of poly(A) polymerase that could regulate poly(A) addition in Saccharomyces cerevisiae. FEBS Lett *586*, 1173-1178.

Fernandez-Leiro, R., and Scheres, S.H.W. (2017). A pipeline approach to single-particle processing in RELION. Acta Crystallogr D Struct Biol *73*, 496-502.

Fisher, R.P. (2019). Cdk7: a kinase at the core of transcription and in the crosshairs of cancer drug discovery. Transcription *10*, 47-56.

Fiume, L., and Stirpe, F. (1966). Decreased RNA content in mouse liver nuclei after intoxication with alpha-amanitin. Biochimica Et Biophysica Acta *123*, 643-645.

Ford, H.S.a.W.W. (1907). On the chemical properties of Amanita-toxin. J. Biol. Chem *3*, 279-283.

Fuda, N.J., Ardehali, M.B., and Lis, J.T. (2009). Defining mechanisms that regulate RNA polymerase II transcription in vivo. Nature *461*, 186-192.

Ganem, C., Devaux, F., Torchet, C., Jacq, C., Quevillon-Cheruel, S., Labesse, G., Facca, C., and Faye, G. (2003). Ssu72 is a phosphatase essential for transcription termination of snoRNAs and specific mRNAs in yeast. EMBO J *22*, 1588-1598.

Gao, Y., Cao, E., Julius, D., and Cheng, Y. (2016). TRPV1 structures in nanodiscs reveal mechanisms of ligand and lipid action. Nature *534*, 347-351.

Garas, M., Dichtl, B., and Keller, W. (2008). The role of the putative 3' end processing endonuclease Ysh1p in mRNA and snoRNA synthesis. RNA *14*, 2671-2684.

Geiduschek, E.P., and Kassavetis, G.A. (2001). The RNA polymerase III transcription apparatus. J Mol Biol *310*, 1-26.

Ghazy, M.A., He, X., Singh, B.N., Hampsey, M., and Moore, C. (2009). The essential N terminus of the Pta1 scaffold protein is required for snoRNA transcription termination and Ssu72 function but is dispensable for pre-mRNA 3'-end processing. Mol Cell Biol *29*, 2296-2307.

Gil, A., and Proudfoot, N.J. (1987). Position-dependent sequence elements downstream of AAUAAA are required for efficient rabbit beta-globin mRNA 3' end formation. Cell *49*, 399-406.

Gilmartin, G.M., and Nevins, J.R. (1989). An ordered pathway of assembly of components required for polyadenylation site recognition and processing. Genes Dev *3*, 2180-2190.

Gilmartin, G.M., and Nevins, J.R. (1991). Molecular analyses of two poly(A) site-processing factors that determine the recognition and efficiency of cleavage of the pre-mRNA. Mol Cell Biol *11*, 2432-2438.

Glover-Cutter, K., Kim, S., Espinosa, J., and Bentley, D.L. (2008). RNA polymerase II pauses and associates with pre-mRNA processing factors at both ends of genes. Nat Struct Mol Biol *15*, 71-78.

Gnatt, A.L., Cramer, P., Fu, J., Bushnell, D.A., and Kornberg, R.D. (2001). Structural basis of transcription: an RNA polymerase II elongation complex at 3.3 A resolution. Science *292*, 1876-1882.

Gordon, J.M., Shikov, S., Kuehner, J.N., Liriano, M., Lee, E., Stafford, W., Poulsen, M.B., Harrison, C., Moore, C., and Bohm, A. (2011). Reconstitution of CF IA from overexpressed subunits reveals stoichiometry and provides insights into molecular topology. Biochemistry *50*, 10203-10214.

Graveley, B.R. (2000). Sorting out the complexity of SR protein functions. RNA *6*, 1197-1211.

Graveley, B.R., Hertel, K.J., and Maniatis, T. (1998). A systematic analysis of the factors that determine the strength of pre-mRNA splicing enhancers. EMBO J *17*, 6747-6756.

Gruber, J.J., Olejniczak, S.H., Yong, J., La Rocca, G., Dreyfuss, G., and Thompson, C.B. (2012). Ars2 promotes proper replication-dependent histone mRNA 3' end formation. Molecular cell *45*, 87-98.

Grummt, I. (2003). Life on a planet of its own: regulation of RNA polymerase I transcription in the nucleolus. Genes Dev *17*, 1691-1702.

Guo, Z., Russo, P., Yun, D.F., Butler, J.S., and Sherman, F. (1995). Redundant 3' end-forming signals for the yeast CYC1 mRNA. Proc Natl Acad Sci U S A *92*, 4211-4214.

Guo, Z., and Sherman, F. (1995). 3'-end-forming signals of yeast mRNA. Mol Cell Biol *15*, 5983-5990.

Guo, Z., and Sherman, F. (1996). 3'-end-forming signals of yeast mRNA. Trends Biochem Sci *21*, 477-481.

Gusarov, I., and Nudler, E. (1999). The mechanism of intrinsic transcription termination. Molecular cell *3*, 495-504.

Hallais, M., Pontvianne, F., Andersen, P.R., Clerici, M., Lener, D., Benbahouche Nel, H., Gostan, T., Vandermoere, F., Robert, M.C., Cusack, S*., et al.* (2013). CBC-ARS2 stimulates 3'-end maturation of multiple RNA families and favors cap-proximal processing. Nat Struct Mol Biol *20*, 1358-1366.

Hampsey, M., Singh, B.N., Ansari, A., Laine, J.P., and Krishnamurthy, S. (2011). Control of eukaryotic gene expression: gene loops and transcriptional memory. Adv Enzyme Regul *51*, 118-125.

Han, Y., Yan, C., Fishbain, S., Ivanov, I., and He, Y. (2018). Structural visualization of RNA polymerase III transcription machineries. Cell Discov *4*, 40.

Hatzoglou, M., Adamtziki, E., Margaritis, L., and Sekeris, C.E. (1985). Isolation and characterization of nuclear particles containing rapidly labelled hnRNA and snRNA in combination with a distinct set of polypeptides of Mr 74000 and 72000. Exp Cell Res *157*, 227-241.

Heidmann, S., Obermaier, B., Vogel, K., and Domdey, H. (1992). Identification of pre-mRNA polyadenylation sites in Saccharomyces cerevisiae. Mol Cell Biol *12*, 4215-4229.

Heidmann, S., Schindewolf, C., Stumpf, G., and Domdey, H. (1994). Flexibility and interchangeability of polyadenylation signals in Saccharomyces cerevisiae. Mol Cell Biol *14*, 4633-4642.

Hill, C.H., Boreikaite, V., Kumar, A., Casanal, A., Kubik, P., Degliesposti, G., Maslen, S., Mariani, A., von Loeffelholz, O., Girbig, M*., et al.* (2019). Activation of the Endonuclease that Defines mRNA 3' Ends Requires Incorporation into an 8-Subunit Core Cleavage and Polyadenylation Factor Complex. Molecular cell *73*, 1217-1231 e1211.

Hirata, A., Klein, B.J., and Murakami, K.S. (2008). The X-ray crystal structure of RNA polymerase from Archaea. Nature *451*, 851-854.

Hirose, Y., and Manley, J.L. (2000). RNA polymerase II and the integration of nuclear events. Genes Dev *14*, 1415-1429.

Hockert, J.A., Yeh, H.J., and MacDonald, C.C. (2010). The hinge domain of the cleavage stimulation factor protein CstF-64 is essential for CstF-77 interaction, nuclear localization, and polyadenylation. J Biol Chem *285*, 695-704.

Hodo, H.G., 3rd, and Blatti, S.P. (1977). Purification using polyethylenimine precipitation and low molecular weight subunit analyses of calf thymus and wheat germ DNA-dependent RNA polymerase II. Biochemistry *16*, 2334-2343.

Hofmann, I., Schnolzer, M., Kaufmann, I., and Franke, W.W. (2002). Symplekin, a constitutive protein of karyo- and cytoplasmic particles involved in mRNA biogenesis in Xenopus laevis oocytes. Mol Biol Cell *13*, 1665-1676.

Hoiby, T., Zhou, H., Mitsiou, D.J., and Stunnenberg, H.G. (2007). A facelift for the general transcription factor TFIIA. Biochim Biophys Acta *1769*, 429-436.

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

Hu, X., Malik, S., Negroiu, C.C., Hubbard, K., Velalar, C.N., Hampton, B., Grosu, D., Catalano, J., Roeder, R.G., and Gnatt, A. (2006). A Mediator-responsive form of metazoan RNA polymerase II. Proceedings of the National Academy of Sciences *103*, 9506-9511.

Hubbard, R.E., and Kamran Haider, M. (2001). Hydrogen Bonds in Proteins: Role and Strength. In eLS (John Wiley & Sons, Ltd).

Irniger, S., and Braus, G.H. (1994). Saturation mutagenesis of a polyadenylation signal reveals a hexanucleotide element essential for mRNA 3' end formation in Saccharomyces cerevisiae. Proc Natl Acad Sci U S A *91*, 257-261.

Jamonnak, N., Creamer, T.J., Darby, M.M., Schaughency, P., Wheelan, S.J., and Corden, J.L. (2011). Yeast Nrd1, Nab3, and Sen1 transcriptome-wide binding maps suggest multiple roles in post-transcriptional RNA processing. RNA *17*, 2011-2025.

Jang, S., Cook, N.J., Pye, V.E., Bedwell, G.J., Dudek, A.M., Singh, P.K., Cherepanov, P., and Engelman, A.N. (2019). Differential role for phosphorylation in alternative polyadenylation function versus nuclear import of SR-like protein CPSF6. Nucleic acids research *47*, 4663-4683.

Ji, Z., Lee, J.Y., Pan, Z., Jiang, B., and Tian, B. (2009). Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. Proc Natl Acad Sci U S A *106*, 7028-7033.

Jishage, M., Malik, S., Wagner, U., Uberheide, B., Ishihama, Y., Hu, X., Chait, B.T., Gnatt, A., Ren, B., and Roeder, R.G. (2012). Transcriptional regulation by Pol II(G) involving mediator and competitive interactions of Gdown1 and TFIIF with Pol II. Molecular cell *45*, 51-63.

Kamieniarz-Gdula, K., Gdula, M.R., Panser, K., Nojima, T., Monks, J., Wisniewski, J.R., Riepsaame, J., Brockdorff, N., Pauli, A., and Proudfoot, N.J. (2019). Selective Roles of Vertebrate PCF11 in Premature and Full-Length Transcript Termination. Molecular cell *74*, 158-172 e159.

Kang, J.G., Hahn, M.Y., Ishihama, A., and Roe, J.H. (1997). Identification of sigma factors for growth phase-related promoter selectivity of RNA polymerases from Streptomyces coelicolor A3(2). Nucleic acids research *25*, 2566-2573.

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

Keon, B.H., Schafer, S., Kuhn, C., Grund, C., and Franke, W.W. (1996). Symplekin, a novel type of tight junction plaque protein. J Cell Biol *134*, 1003-1018.

Kettenberger, H., Armache, K.J., and Cramer, P. (2004). Complete RNA polymerase II elongation complex structure and its interactions with NTP and TFIIS. Molecular cell *16*, 955-965.

Kim, H., Erickson, B., Luo, W., Seward, D., Graber, J.H., Pollock, D.D., Megee, P.C., and Bentley, D.L. (2010). Gene-specific RNA polymerase II phosphorylation and the CTD code. Nat Struct Mol Biol *17*, 1279-1286.

Kim, K.Y., and Levin, D.E. (2011). Mpk1 MAPK association with the Paf1 complex blocks Sen1-mediated premature transcription termination. Cell *144*, 745-756.

Kim, M., Ahn, S.H., Krogan, N.J., Greenblatt, J.F., and Buratowski, S. (2004a). Transitions in RNA polymerase II elongation complexes at the 3' ends of genes. EMBO J *23*, 354-364.

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

Kim, M., Vasiljeva, L., Rando, O.J., Zhelkovsky, A., Moore, C., and Buratowski, S. (2006). Distinct pathways for snoRNA and mRNA termination. Molecular cell *24*, 723-734.

Kinghorn, A.D. (1987). Peptides of Poisonous Amanita Mushrooms. Theodor Wieland , Alexander Rich. The Quarterly Review of Biology *62*, 308-309.

Kireeva, M.L., Komissarova, N., Waugh, D.S., and Kashlev, M. (2000). The 8-nucleotide-long RNA:DNA hybrid is a primary stability determinant of the RNA polymerase II elongation complex. J Biol Chem *275*, 6530-6536.

Klebe, G. (2015). Applying thermodynamic profiling in lead finding and optimization. Nat Rev Drug Discov *14*, 95-110.

Komissarova, N., Becker, J., Solter, S., Kireeva, M., and Kashlev, M. (2002). Shortening of RNA:DNA hybrid in the elongation complex of RNA polymerase is a prerequisite for transcription termination. Molecular cell *10*, 1151-1162.

Kuehner, J.N., Pearson, E.L., and Moore, C. (2011). Unravelling the means to an end: RNA polymerase II transcription termination. Nat Rev Mol Cell Biol *12*, 283-294.

Kuhn, U., Gundel, M., Knoth, A., Kerwitz, Y., Rudel, S., and Wahle, E. (2009). Poly(A) tail length is controlled by the nuclear poly(A)-binding protein regulating the interaction between poly(A) polymerase and the cleavage and polyadenylation specificity factor. J Biol Chem *284*, 22803-22814.

Lafarga, M., Casafont, I., Bengoechea, R., Tapia, O., and Berciano, M.T. (2009). Cajal's contribution to the knowledge of the neuronal cell nucleus. Chromosoma *118*, 437-443.

Landick, R. (2004). Active-site dynamics in RNA polymerases. Cell *116*, 351-353.

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

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

Liu, X., Bushnell, D.A., and Kornberg, R.D. (2013). RNA polymerase II transcription: structure and mechanism. Biochim Biophys Acta *1829*, 2-8.

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

Long, J.C., and Caceres, J.F. (2009). The SR protein family of splicing factors: master regulators of gene expression. Biochem J *417*, 15-27.

Louder, R.K., He, Y., Lopez-Blanco, J.R., Fang, J., Chacon, P., and Nogales, E. (2016). Structure of promoter-bound TFIID and model of human pre-initiation complex assembly. Nature *531*, 604-609.

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

Luo, Z., Lin, C., and Shilatifard, A. (2012). The super elongation complex (SEC) family in transcriptional control. Nat Rev Mol Cell Biol *13*, 543-547.

MacDonald, C.C., Wilusz, J., and Shenk, T. (1994). The 64-kilodalton subunit of the CstF polyadenylation factor binds to pre-mRNAs downstream of the cleavage site and influences cleavage site location. Mol Cell Biol *14*, 6647-6654.

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

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

Mapendano, C.K., Lykke-Andersen, S., Kjems, J., Bertrand, E., and Jensen, T.H. (2010). Crosstalk between mRNA 3' end processing and transcription initiation. Molecular cell *40*, 410-422.

Martin, G., Gruber, A.R., Keller, W., and Zavolan, M. (2012). Genome-wide analysis of pre-mRNA 3' end processing reveals a decisive role of human cleavage factor I in the regulation of 3' UTR length. Cell Rep *1*, 753-763.

Martin, G., and Keller, W. (1996). Mutational analysis of mammalian poly(A) polymerase identifies a region for primer binding and catalytic domain, homologous to the family X polymerases, and to other nucleotidyltransferases. EMBO J *15*, 2593-2603.

Martin, G., Keller, W., and Doublie, S. (2000). Crystal structure of mammalian poly(A) polymerase in complex with an analog of ATP. EMBO J *19*, 4193-4203.

Martin, G., Moglich, A., Keller, W., and Doublie, S. (2004). Biochemical and structural insights into substrate binding and catalytic mechanism of mammalian poly(A) polymerase. J Mol Biol *341*, 911-925.

Marzluff, W.F., Wagner, E.J., and Duronio, R.J. (2008). Metabolism and regulation of canonical histone mRNAs: life without a poly(A) tail. Nat Rev Genet *9*, 843-854.

Mas, A. (2005). Mushrooms, amatoxins and the liver. J Hepatol *42*, 166-169.

Masamha, C.P., and Wagner, E.J. (2018). The contribution of alternative polyadenylation to the cancer phenotype. Carcinogenesis *39*, 2-10.

Masamha, C.P., Xia, Z., Yang, J., Albrecht, T.R., Li, M., Shyu, A.B., Li, W., and Wagner, E.J. (2014). CFIm25 links alternative polyadenylation to glioblastoma tumour suppression. Nature *510*, 412-416.

Mastronarde, D.N. (2003). SerialEM: A program for automated tilt series acquisition on Tecnai microscopes using prediction of specimen position. Microscopy and Microanalysis *9*.

Mathew, R., and Chatterji, D. (2006). The evolving story of the omega subunit of bacterial RNA polymerase. Trends Microbiol *14*, 450-455.

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

McLauchlan, J., Gaffney, D., Whitton, J.L., and Clements, J.B. (1985). The consensus sequence YGTGTTYY located downstream from the AATAAA signal is required for efficient formation of mRNA 3' termini. Nucleic acids research *13*, 1347-1368.

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

Merino E, Y.C. (2005). Transcription attenuation: a highly conserved regulatory strategy used by bacteria. Trends in Genetics *21*, 260-264.

Meselson, M., and Stahl, F.W. (1958). The Replication of DNA in Escherichia Coli. Proc Natl Acad Sci U S A *44*, 671-682.

Michelot, D., and Labia, R. (1988). alpha-Amanitin: a possible suicide substrate-like toxin involving the sulphoxide moiety of the bridged cyclopeptide. Drug Metabol Drug Interact *6*, 265-274.

Milo, R.P., Rob. (2017). Cell Biology by the Numbers: What are the concentrations of different ions in cells? book.bionumbers.org.

Miwa, K., Kojima, R., Obita, T., Ohkuma, Y., Tamura, Y., and Mizuguchi, M. (2016). Crystal Structure of Human General Transcription Factor TFIIE at Atomic Resolution. J Mol Biol *428*, 4258-4266.

Moreira, M.C., Klur, S., Watanabe, M., Nemeth, A.H., Le Ber, I., Moniz, J.C., Tranchant, C., Aubourg, P., Tazir, M., Schols, L.*, et al.* (2004). Senataxin, the ortholog of a yeast RNA helicase, is mutant in ataxia-ocular apraxia 2. Nat Genet *36*, 225-227.

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

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

Nag, A., Narsinh, K., and Martinson, H.G. (2007). The poly(A)-dependent transcriptional pause is mediated by CPSF acting on the body of the polymerase. Nat Struct Mol Biol *14*, 662-669.

Naji, S., Bertero, M.G., Spitalny, P., Cramer, P., and Thomm, M. (2008). Structure-function analysis of the RNA polymerase cleft loops elucidates initial transcription, DNA unwinding and RNA displacement. Nucleic acids research *36*, 676-687.

Narita, T., Yung, T.M., Yamamoto, J., Tsuboi, Y., Tanabe, H., Tanaka, K., Yamaguchi, Y., and Handa, H. (2007). NELF interacts with CBC and participates in 3' end processing of replication-dependent histone mRNAs. Molecular cell *26*, 349-365.

Naville, M., and Gautheret, D. (2010). Transcription attenuation in bacteria: theme and variations. Brief Funct Genomics *9*, 178-189.

Neuwald, A.F., and Poleksic, A. (2000). PSI-BLAST searches using hidden markov models of structural repeats: prediction of an unusual sliding DNA clamp and of beta-propellers in UV-damaged DNA-binding protein. Nucleic acids research *28*, 3570-3580.

O'Sullivan, J.M., Tan-Wong, S.M., Morillon, A., Lee, B., Coles, J., Mellor, J., and Proudfoot, N.J. (2004). Gene loops juxtapose promoters and terminators in yeast. Nat Genet *36*, 1014-1018.

Osborne, C.S., Chakalova, L., Brown, K.E., Carter, D., Horton, A., Debrand, E., Goyenechea, B., Mitchell, J.A., Lopes, S., Reik, W.*, et al.* (2004). Active genes dynamically colocalize to shared sites of ongoing transcription. Nat Genet *36*, 1065-1071.

Pauws, E., van Kampen, A.H., van de Graaf, S.A., de Vijlder, J.J., and Ris-Stalpers, C. (2001). Heterogeneity in polyadenylation cleavage sites in mammalian mRNA sequences: implications for SAGE analysis. Nucleic acids research *29*, 1690-1694.

Pearson, E.L., and Moore, C.L. (2013). Dismantling promoter-driven RNA polymerase II transcription complexes in vitro by the termination factor Rat1. J Biol Chem *288*, 19750-19759.

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

Plant, K.E., Dye, M.J., Lafaille, C., and Proudfoot, N.J. (2005). Strong polyadenylation and weak pausing combine to cause efficient termination of transcription in the human Ggamma-globin gene. Mol Cell Biol *25*, 3276-3285.

Porrua, O., Boudvillain, M., and Libri, D. (2016). Transcription Termination: Variations on Common Themes. Trends Genet *32*, 508-522.

Porrua, O., Hobor, F., Boulay, J., Kubicek, K., D'Aubenton-Carafa, Y., Gudipati, R.K., Stefl, R., and Libri, D. (2012). In vivo SELEX reveals novel sequence and structural determinants of Nrd1-Nab3-Sen1-dependent transcription termination. EMBO J *31*, 3935-3948.

Porrua, O., and Libri, D. (2013). A bacterial-like mechanism for transcription termination by the Sen1p helicase in budding yeast. Nat Struct Mol Biol *20*, 884-891.

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

Prusiner, S.B. (1991). Molecular biology of prion diseases. Science *252*, 1515-1522.

Punjani, A., Rubinstein, J.L., Fleet, D.J., and Brubaker, M.A. (2017). cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. Nat Methods *14*, 290-296.

Raabe, T., Bollum, F.J., and Manley, J.L. (1991). Primary structure and expression of bovine poly(A) polymerase. Nature *353*, 229-234.

Ream, T.S., Haag, J.R., Wierzbicki, A.T., Nicora, C.D., Norbeck, A.D., Zhu, J.K., Hagen, G., Guilfoyle, T.J., Pasa-Tolic, L., and Pikaard, C.S. (2009). Subunit compositions of the RNA-silencing enzymes Pol IV and Pol V reveal their origins as specialized forms of RNA polymerase II. Molecular cell *33*, 192-203.

Richard, P., and Manley, J.L. (2009). Transcription termination by nuclear RNA polymerases. Genes Dev *23*, 1247-1269.

Richardson, J.P. (1993). Transcription termination. Crit Rev Biochem Mol Biol *28*, 1-30.

Robert, F., Douziech, M., Forget, D., Egly, J.M., Greenblatt, J., Burton, Z.F., and Coulombe, B. (1998). Wrapping of promoter DNA around the RNA polymerase II initiation complex induced by TFIIF. Molecular cell *2*, 341-351.

Roberts, J.W. (1969). Termination factor for RNA synthesis. Nature *224*, 1168-1174.

Roeder, R.G., and Rutter, W.J. (1969). Multiple forms of DNA-dependent RNA polymerase in eukaryotic organisms. Nature *224*, 234-237.

Roth, M.B., Murphy, C., and Gall, J.G. (1990). A monoclonal antibody that recognizes a phosphorylated epitope stains lampbrush chromosome loops and small granules in the amphibian germinal vesicle. J Cell Biol *111*, 2217-2223.

Rougvie, A.E., and Lis, J.T. (1990). Postinitiation transcriptional control in Drosophila melanogaster. Mol Cell Biol *10*, 6041-6045.

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

Ruegsegger, U., Blank, D., and Keller, W. (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*, 243-253.

Ruepp, M.D., Schumperli, D., and Barabino, S.M. (2011). mRNA 3' end processing and more--multiple functions of mammalian cleavage factor I-68. Wiley Interdiscip Rev RNA *2*, 79-91.

Ryan, K., Calvo, O., and Manley, J.L. (2004). Evidence that polyadenylation factor CPSF-73 is the mRNA 3' processing endonuclease. RNA *10*, 565-573.

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

Sahebi, M., Hanafi, M.M., van Wijnen, A.J., Azizi, P., Abiri, R., Ashkani, S., and Taheri, S. (2016). Towards understanding pre-mRNA splicing mechanisms and the role of SR proteins. Gene *587*, 107-119.

Sainsbury, S., Bernecky, C., and Cramer, P. (2015). Structural basis of transcription initiation by RNA polymerase II. Nat Rev Mol Cell Biol *16*, 129-143.

Sainsbury, S., Niesser, J., and Cramer, P. (2013). Structure and function of the initially transcribing RNA polymerase II-TFIIB complex. Nature *493*, 437-440.

Sanford, J.R., Gray, N.K., Beckmann, K., and Caceres, J.F. (2004). A novel role for shuttling SR proteins in mRNA translation. Genes Dev *18*, 755-768.

Sariki, S.K., Sahu, P.K., Golla, U., Singh, V., Azad, G.K., and Tomar, R.S. (2016). Sen1, the homolog of human Senataxin, is critical for cell survival through regulation of redox homeostasis, mitochondrial function, and the TOR pathway in Saccharomyces cerevisiae. FEBS J *283*, 4056-4083.

Schilbach, S., Hantsche, M., Tegunov, D., Dienemann, C., Wigge, C., Urlaub, H., and Cramer, P. (2017). Structures of transcription pre-initiation complex with TFIIH and Mediator. Nature *551*, 204-209.

Schul, W., van Driel, R., and de Jong, L. (1998). A subset of poly(A) polymerase is concentrated at sites of RNA synthesis and is associated with domains enriched in splicing factors and poly(A) RNA. Exp Cell Res *238*, 1-12.

Sentenac, A. (1985). Eukaryotic RNA polymerases. CRC Crit Rev Biochem *18*, 31-90.

Shepard, P.J., and Hertel, K.J. (2009). The SR protein family. Genome Biol *10*, 242.

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

Skourti-Stathaki, K., Proudfoot, N.J., and Gromak, N. (2011). Human senataxin resolves RNA/DNA hybrids formed at transcriptional pause sites to promote Xrn2-dependent termination. Molecular cell *42*, 794-805.

Spector, D.L. (1993). Macromolecular domains within the cell nucleus. Annu Rev Cell Biol *9*, 265-315.

Steinmetz, E.J., and Brow, D.A. (1996). Repression of gene expression by an exogenous sequence element acting in concert with a heterogeneous nuclear ribonucleoprotein-like protein, Nrd1, and the putative helicase Sen1. Mol Cell Biol *16*, 6993-7003.

Sullivan, K.D., Steiniger, M., and Marzluff, W.F. (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*, 322-332.

Sun, Y., Zhang, Y., Hamilton, K., Manley, J.L., Shi, Y., Walz, T., and Tong, L. (2018). Molecular basis for the recognition of the human AAUAAA polyadenylation signal. Proc Natl Acad Sci U S A *115*, E1419-E1428.

Takagaki, Y., MacDonald, C.C., Shenk, T., and Manley, J.L. (1992). The human 64-kDa polyadenylylation factor contains a ribonucleoprotein-type RNA binding domain and unusual auxiliary motifs. Proc Natl Acad Sci U S A *89*, 1403-1407.

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

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

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

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

Temin, H.M., and Mizutani, S. (1970). RNA-dependent DNA polymerase in virions of Rous sarcoma virus. Nature *226*, 1211-1213.

Thiebaut, M., Kisseleva-Romanova, E., Rougemaille, M., Boulay, J., and Libri, D. (2006). Transcription termination and nuclear degradation of cryptic unstable transcripts: a role for the nrd1-nab3 pathway in genome surveillance. Molecular cell *23*, 853-864.

Thompson, N.E., Aronson, D.B., and Burgess, R.R. (1990). Purification of eukaryotic RNA polymerase II by immunoaffinity chromatography. Elution of active enzyme with protein stabilizing agents from a polyol-responsive monoclonal antibody. J Biol Chem *265*, 7069-7077.

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

Tolhuis, B., Palstra, R.J., Splinter, E., Grosveld, F., and de Laat, W. (2002). Looping and interaction between hypersensitive sites in the active beta-globin locus. Molecular cell *10*, 1453-1465.

Torices, R., and Muñoz-Pajares, A.J. (2015). PHENIX: An R package to estimate a size-controlled phenotypic integration index. Applications in Plant Sciences *3*.

Tze-Bin Chou, Z.Z.a.P.M.B. (1987). Developmental expression of a regulatory gene is programmed at the level of splicing. The EMBO Journal *6*, 4095-4104,.

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

Vos, S.M., Farnung, L., Boehning, M., Wigge, C., Linden, A., Urlaub, H., and Cramer, P. (2018a). Structure of activated transcription complex Pol II-DSIF-PAF-SPT6. Nature *560*, 607-612.

Vos, S.M., Farnung, L., Urlaub, H., and Cramer, P. (2018b). Structure of paused transcription complex Pol II-DSIF-NELF. Nature *560*, 601-606.

Wagschal, A., Rousset, E., Basavarajaiah, P., Contreras, X., Harwig, A., Laurent-Chabalier, S., Nakamura, M., Chen, X., Zhang, K., Meziane, O.*, et al.* (2012). Microprocessor, Setx, Xrn2, and Rrp6 co-operate to induce premature termination of transcription by RNAPII. Cell *150*, 1147-1157.

Wahle, E. (1991). Purification and characterization of a mammalian polyadenylate polymerase involved in the 3' end processing of messenger RNA precursors. J Biol Chem *266*, 3131-3139.

Wahle, E. (1995). Poly(A) tail length control is caused by termination of processive synthesis. J Biol Chem *270*, 2800-2808.

Wang, D., Bushnell, D.A., Westover, K.D., Kaplan, C.D., and Kornberg, R.D. (2006). Structural basis of transcription: role of the trigger loop in substrate specificity and catalysis. Cell *127*, 941-954.

Wang, Y., Fairley, J.A., and Roberts, S.G. (2010). Phosphorylation of TFIIB links transcription initiation and termination. Curr Biol *20*, 548-553.

Weitzer, S., and Martinez, J. (2007). The human RNA kinase hClp1 is active on 3' transfer RNA exons and short interfering RNAs. Nature *447*, 222-226.

Werner, F., and Grohmann, D. (2011). Evolution of multisubunit RNA polymerases in the three domains of life. Nat Rev Microbiol *9*, 85-98.

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

Whitelaw, E., and Proudfoot, N. (1986). Alpha-thalassaemia caused by a poly(A) site mutation reveals that transcriptional termination is linked to 3' end processing in the human alpha 2 globin gene. EMBO J *5*, 2915-2922.

Wieland, H., and Hallermayer, R. (1941). Ober die Giftstoffe des Knollen- blS_tterpilzes. VI. Amanitin, das Hauptgift des Knollenbl/itterpilzes. Liebigs Ann. Chem *548*, 1-18.

Wieland, T. (1986). Peptides of poisonous amanita mushrooms. (Springer Series in Molecular Biology Springer).

Wieland, T., and Fischer, E. (1948). Ober Elektrophorese auf Filtrierpapier. Naturwissenschaften *35*.

Wieland, T., Gotzendorfer, C., Dabrowski, J., Lipscomb, W.N., and Shoham, G. (1983). Unexpected similarity of the structures of the weakly toxic amanitin (S)-sulfoxide and the highly toxic (R)-sulfoxide and sulfone as revealed by proton nuclear magnetic resonance and X-ray analysis. Biochemistry *22*, 1264-1271.

Wienland, T., and Faulstich, H. (1991). Fifty years of amanitin. Experientia *47*, 1186-1193.

Wong, W., Bai, X.-C., Sleebs, B.E., Triglia, T., Brown, A., Thompson, J.K., Jackson, K.E., Hanssen, E., Marapana, D.S., Fernandez, I.S.*, et al.* (2017). Mefloquine targets the Plasmodium falciparum 80S ribosome to inhibit protein synthesis. Nature Microbiology *2*, 17031.

Wu, J.Y., and Maniatis, T. (1993). Specific interactions between proteins implicated in splice site selection and regulated alternative splicing. Cell *75*, 1061-1070.

Wu, Y.M., Chang, J.W., Wang, C.H., Lin, Y.C., Wu, P.L., Huang, S.H., Chang, C.C., Hu, X., Gnatt, A., and Chang, W.H. (2012). Regulation of mammalian transcription by Gdown1 through a novel steric crosstalk revealed by cryo-EM. EMBO J *31*, 3575-3587.

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

Xiang, K., Tong, L., and Manley, J.L. (2014). Delineating the structural blueprint of the pre-mRNA 3'-end processing machinery. Mol Cell Biol *34*, 1894-1910.

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

Yang, Q., Coseno, M., Gilmartin, G.M., and Doublie, S. (2011). Crystal structure of a human cleavage factor CFI(m)25/CFI(m)68/RNA complex provides an insight into poly(A) site recognition and RNA looping. Structure *19*, 368-377.

Yao, J., Ardehali, M.B., Fecko, C.J., Webb, W.W., and Lis, J.T. (2007). Intranuclear distribution and local dynamics of RNA polymerase II during transcription activation. Molecular cell *28*, 978-990.

Yilmaz, I., Ermis, F., Akata, I., and Kaya, E. (2015). A Case Study: What Doses of Amanita phalloides and Amatoxins Are Lethal to Humans? Wilderness & Environmental Medicine *26*, 491-496.

Zanotti, G., Wieland, T., Benedetti, E., Di Blasio, B., Pavone, V., and Pedone, C. (1989). Structure-toxicity relationships in the amatoxin series. Synthesis of S-deoxy[gamma(R)-hydroxy-Ile3]-amaninamide, its crystal and molecular structure and inhibitory efficiency. Int J Pept Protein Res *34*, 222-228.

Zeng, C., Kim, E., Warren, S.L., and Berget, S.M. (1997). Dynamic relocation of transcription and splicing factors dependent upon transcriptional activity. EMBO J *16*, 1401-1412.

Zhang, K. (2016). Gctf: Real-time CTF determination and correction. Journal of Structural Biology *193*, 1-12.

Zhang, X., Henderson, I.R., Lu, C., Green, P.J., and Jacobsen, S.E. (2007). Role of RNA polymerase IV in plant small RNA metabolism. Proc Natl Acad Sci U S A *104*, 4536-4541.

Zhang, Z., Fu, J., and Gilmour, D.S. (2005). CTD-dependent dismantling of the RNA polymerase II elongation complex by the pre-mRNA 3'-end processing factor, Pcf11. Genes Dev *19*, 1572-1580.

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

Zhao, W., and Manley, J.L. (1996). Complex alternative RNA processing generates an unexpected diversity of poly(A) polymerase isoforms. Mol Cell Biol *16*, 2378-2386.

Zheng, S.Q., Palovcak, E., Armache, J.P., Verba, K.A., Cheng, Y., and Agard, D.A. (2017). MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. Nat Methods *14*, 331-332.

Zhu, Y., Wang, X., Forouzmand, E., Jeong, J., Qiao, F., Sowd, G.A., Engelman, A.N., Xie, X., Hertel, K.J., and Shi, Y. (2018). Molecular Mechanisms for CFIm-Mediated Regulation of mRNA Alternative Polyadenylation. Molecular cell *69*, 62-74 e64.

Zivanov, J., Nakane, T., Forsberg, B.O., Kimanius, D., Hagen, W.J., Lindahl, E., and Scheres, S.H. (2018). New tools for automated high-resolution cryo-EM structure determination in RELION-3. eLife *7*.

Curriculum Vitae

Personal details

Name:	Xiangyang Liu
Date of birth:	April 28, 1989
Nationality:	Chinese

Education

Sept. 2004 – July. 2008	Quwo Middle School, Linfen, Shanxi, China
	High school diploma
Sept. 2008 – July. 2012	China Pharmaceutical University, Nanking, China
	Subject: Marine pharmacy
	Degree: Bachelor
Sept. 2012 - July. 2015	Shanghai institute for Biological Sciences, Chinese
	Academy of Sciences, Shanghai, China
	Subject: Bioengineering
	Degree: Master

Research activities

2010.6-2011.9	Laboratory for microbiology,
	China Pharmaceutical University, Nanking, China
2012.2-2012.5	Institute for Nutritional Science, Shanghai Institute for
	Biological sciences, Chinese Academy of Sciences,
	Shanghai, China
	Bachelor thesis
Sept. 2012 - July. 2015	Shanghai Institute of Biochemistry and cell Biology
	Chinese Academy of Sciences, Shanghai, China

	Master thesis
Since Sept. 2015	Cramer Laboratory
	Max-Planck-Institute for biophysical chemistry,
	Goettingen, Germany
	PhD Student
Focus areas:	Structural Biology

Publications

Liu, X.; Farnung, L.; Wigge, C.; Cramer, P.: Cryo-EM structure of a mammalian RNA polymerase II elongation complex inhibited by α-amanitin. Journal of Biological Chemistry 293 (19), pp. 7189 - 7194 (2018)

Wang Y, Ding Z, **Liu X**, Bao Y, Huang M, Wong CCL, Hong X, Cong Y*. Architecture and subunit arrangement of the complete Saccharomyces cerevisiae COMPASS complex. Scientific Reports, 2018,8:17405.