2021

Structral And Biochemical Insights Into The Transition From Transcription Initiation To Elongation

Rina Fujiwara
University of Pennsylvania

Structral And Biochemical Insights Into The Transition From Transcription Initiation To Elongation

Abstract

Transcription by RNA polymerase II (Pol II) is a complex process that requires timely and coordinated regulation at multiple steps for proper gene expression. Initiation is the first step in transcription and decades of biochemical and genome-wide studies have identified proteins involved in the process and revealed their functions. Additionally, technological advancements in cryo-EM enabled researchers to visualize initiation complexes and provide mechanistic insights into initiation processes in the last several years. However, the mechanistic understanding of the transition from transcription initiation to elongation has been limited in part due to the lack of an efficient transcription initiation system in vitro. We purified yeast general transcription factors (GTFs: TFIIA, TFIIB, TBP(a component of TFIID), TFIIE, TFIIF, and TFIIH) and Pol II, all of which are necessary and sufficient for basal transcription initiation, and optimized the initiation system. Using this system, we biochemically re-examined effects of two elongation factors (Cet1-Ceg1 and Spt4/5) on promoter escape, a process in which Pol II dissociates from GTFs except TFIIF for elongation. We find that inclusion of these elongation factors has positive effects on promoter escape. Furthermore, we took advantage of our efficient system, and generated and isolated post-initiation complexes in vitro for structural characterization by cryo-EM. Our structure of the initially-transcribing complex (ITC) stalled +26 shows a large conformational change of TFIIH in the way that it is much closer to TFIIE than in the pre-initiation complex (PIC) and it loses contacts with Pol II. These changes most likely prime for Pol II to escape the promoter. In addition, the structural studies of post-initiation complex stalled +49 reveal two elongation complexes (ECs) colliding to each other as well as show the presence of EC+ITC. In the structure the colliding ECs, the trailing EC contained RNA of ~25 nt in length but has backtracked by ~10 nt upon colliding. These studies together provide a model of the process of promoter escape, where TFIIH can get kicked out by the preceding promoter-proximal EC.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Biochemistry & Molecular Biophysics

First Advisor

Kenji K. Murakami

Keywords

Initially-transcribing complex, RNA polymerase II, Transcription

Subject Categories

Biochemistry | Biophysics

STRUCTRAL AND BIOCHEMICAL INSIGHTS INTO THE TRANSITION FROM TRANSCRIPTION INITIATION TO ELONGATION

Rina Fujiwara

A DISSERTATION

in

Biochemistry and Molecular Biophysics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2021

Supervisor of Dissertation

____

Kenji Murakami, PhD
Assistant Professor of Biochemistry and Biophysics

Graduate Group Chairperson

____

Kim Sharp, PhD
Associate Professor of Biochemistry and Biophysics

Dissertation Committee

Ronen Marmorstein, Ph.D., George W. Raiziss Professor

Vera Moiseenkova-Bell, PhD., Associate Professor of Pharmacology

Benjamin Garcia, Ph.D., John McCrea Dickson M.D. Presidential Professor

ACKNOWLEDGMENTS

First, I would like to express my sincere appreciation to Dr. Kenji Murakami for letting me join his lab, and for his patience and invaluable advice he has given me throughout my PhD journey. You are very hard-working and always encourage people in the lab to work hard for our career. I look up to you for many reasons including your diligence, determination, and positiveness. I thank all the former and current members of Murakami lab, Trevor van Eeuwen, Hee Jong Kim, Jose Gorbea, Dr. Sophie Yang, and Leon Palao for helpful discussions and their help.

I am extremely grateful for having had Dr. Jeremy Wilusz as my unofficial mentor who let me come into his office to talk whenever I was seeking for advice scientifically and non-scientifically. I thank the former and current members of Wilusz lab, in particular, Drs. Deirdre Tatomer, Nebibe Mutlu, and Sarai Mendoza-Figueroa, Dongming Liang, and Yuxi Ai for having scientific discussions with me, helping me technically, and for friendship. Although I was not officially in Wilusz lab in my PhD, having many female scientists around gave me a sense of belonging. His lab meant a lot for me.

I thank Biochemistry and Biophysics Graduate Program and students in my cohort for creating an inclusive and supportive environment. I thank my committee members, Drs. Ronen Marmorstein, Vera Moiseenkova-Bell, and Benjamin Garcia for helpful discussions.

Endless thanks to all of my friends, especially my best friend, Nivedita Damodaren. I loved working together with you, our numerous spontaneous outing after the lab, and traveling together. After you left for your PhD in a different school in my 3rd year, you made sure to keep in touch with me and I looked forward to hours of weekly talking on the phone every week.

Finally, I would like to thank my family for their unconditional love and support. You let me pursue my interests wherever I want and supported all decisions I have made throughout my life. Every time I go home, you make me feel so special. I could not have completed a PhD without their support.

ABSTRACT

STRUCTURAL AND BIOCHEMICAL INSIGHTS INTO THE TRANSITION FROM TRANSCRIPTION INITIATION TO ELONGATION

Rina Fujiwara

Kenji Murakami

Transcription by RNA polymerase II (Pol II) is a complex process that requires timely and coordinated regulation at multiple steps for proper gene expression. Initiation is the first step in transcription and decades of biochemical and genome-wide studies have identified proteins involved in the process and revealed their functions. Additionally, technological advancements in cryo-EM enabled researchers to visualize initiation complexes and provide mechanistic insights into initiation processes in the last several years. However, the mechanistic understanding of the transition from transcription initiation to elongation has been limited in part due to the lack of an efficient transcription initiation system *in vitro*. We purified yeast general transcription factors (GTFs: TFIIA, TFIIB, TBP(a component of TFIID), TFIIE, TFIIF, and TFIIH) and Pol II, all of which are necessary and sufficient for basal transcription initiation, and optimized the initiation system. Using this system, we biochemically re-examined effects of two elongation factors (Cet1-Ceg1 and Spt4/5) on promoter escape, a process in which Pol II dissociates from GTFs except TFIIF for elongation. We find that inclusion of these elongation factors has positive effects on promoter escape. Furthermore, we took advantage of our efficient system, and generated and isolated post-initiation complexes in vitro for structural characterization by cryo-EM. Our structure of the initially-transcribing complex (ITC) stalled +26 shows a large conformational change of TFIIH in the way that it is much closer to TFIIE than in the pre-initiation complex (PIC) and it loses contacts with Pol II. These changes most likely prime for Pol II to escape the promoter. In addition, the structural studies of post-initiation complex stalled +49 reveal two elongation complexes (ECs) colliding to each other as well as show the presence of EC+ITC. In the structure the colliding ECs, the trailing EC contained RNA of ~25 nt in length but has backtracked by ~10 nt upon colliding. These studies together provide a model of the process of promoter escape, where TFIIH can get kicked out by the preceding promoter-proximal EC.

Table of Contents

LIST OF TABLES

LIST OF ILLUSTRATIONS

Figure **Chapter 3: Structural visualization of de novo initiation of RNA polymerase II transcription** **Page**

CHAPTER 1: INTRODUCTION

1.1 Brief history of discovery of RNA polymerases

One of the most important concepts in biology, “the central dogma of molecular biology” was introduced in 1958 by Francis Crick (Crick, 1970; Crick, 1958). The processes of flows of genetic information are replication, transcription, and translations, all of which are tightly regulated. Dysregulation of any of these processes can cause diseases. Thus, understanding mechanisms of these fundamental processes is crucial. Much of early mechanistic insights into the genetic flow of information came from studies of bacteria in 1960s (Jacob and Monod, 1961). It was only 1969 when single bacterial RNA polymerase was purified and the sigma initiation factor that stimulates RNA synthesis was identified (Burgess et al., 1969).

Despite the fast discovery of prokaryotic factor of basal transcription, there was still a little understanding of eukaryotic transcription in late 1960s with appreciation of different RNA molecules including ribosomal RNA (rRNA), messanger RNA (mRNA), and transfer RNA (tRNA). Following studies of bacterial transcription, Dr. Robert Roeder isolated three enzymes that possess catalytic activity of RNA synthesis from sea urchin and rat liver in 1969 (Roeder and Rutter, 1969). These enzymes were shown to have different preference for divalent metals, salt concentration, and DNA template for their optimal activity (Roeder and Rutter, 1969), and different sensitivity for alpha-amanitin, which is now known to be a selective inhibitor for Pol II and Pol III (Kedinger et al., 1970). Further, subfractination of rat liver nuclei revealed that Pol I resides in nucleoli and Pol II in nucleoplasm (Roeder and Rutter, 1970). These discoveries prompted many scientists to investigate sophisticated regulation of eukaryotic transcription in the future.

In the last five decades, so much research has been done on the fundamental processes of transcription.

1.2 Transcription by RNA polymerase I, II, and III

Pol I, II, and III share a conserved catalytic site and form structurally similar elongation complexes (Cramer et al., 2001; Gnatt et al., 2001; Hoffmann et al., 2015; Neyer et al., 2016). Pol I resides in nucleolus and transcribes precursors of ribosomal (r) RNAs and its activity accounts for approximately 60% of transcription activities in a eukaryotic cell (Goodfellow and Zomerdijk, 2013). Pol II and Pol III are located in nucleoplasm. Pol II is responsible for transcribing mRNAs, most of small nuclear RNAs (snRNA), microRNAs, and other non-coding RNAs. Pol III transcribes essential non-coding RNAs for cellular function such as tRNAs, the U6 snRNA, and the 5S rRNA. All the transcription by Pol I, II, and III are highly regulated at the initiation stage and misregulation can lead to diseases including cancer.

Pol I, Pol II, and Pol III require their own set of general transcription factors (GTFs) to initiate transcription. Pol III assemble at the promoter with TFIIIB that is comprised of TATA-binding protein (TBP), B-related factor (Brf1), and the B double prime (Bdp1) subunit (Abascal-Palacios et al., 2018). These factors are sufficient for Pol III transcription in vitro. Pol I requires Rrn3 and the heterotrimeric core factor (CF) for basal transcription, and additional factors, TBP and upstream activating factor (UAF) for activating transcription (Engel et al., 2017). Pol II possesses the most complex transcription machinery, requiring TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH. Additionally, Mediator is needed for stimulated transcription.

1.3 Basal transcription machinery of Pol II transcription

In cells, core promoters of actively expressed genes lie in nucleosome depleted region and flanked by -1 and +1 nucleosomes (Haberle and Stark, 2018; Jiang and Pugh, 2009; Mavrich et al., 2008; Rhee and Pugh, 2012; Weiner et al., 2010). Core promoters contain elements that allow for transcription factors to bind and facilitate formation of a pre-initiation complex (PIC). These elements include TATA box/TATA-like sequences, Initiator motif (Inr), and the downstream promoter element (DPE) (Haberle and Stark, 2018). TATA box is a promoter element conserved from yeast to human, however only ~20% of yeast genes (Basehoar et al., 2004) and ~10% of human genes (Yang et al., 2007) have this feature. The rest of the genes contains TATA-like sequences or lacks the TATA element (Vo Ngoc et al., 2017). Transcription machinery that is sufficient for recognizing promoter and basal transcription activity is composed of promoter DNA, general transcription factors (GTFs: TFIIA, TFIIB, TBP(a component of TFIID), TFIIE, TFIIF, TFIIH), and Pol II (Murakami et al., 2013b). TBP recognizes TATA in the promoter and bends DNA ~90 degrees (Geiger et al., 1996; Kim et al., 1993a; Kim et al., 1993b; Tan et al., 1996). The binding of TBP to TATA is facilitated by TFIIA and TFIIB, which increase TBP affinity for TATA box (Hieb et al., 2007; Imbalzano et al., 1994) and assist TPB binding unidirectionally (Kays and Schepartz, 2000). TFIIB binds TBP opposite side of where TFIIA is located, and also interacts with Pol II (Kostrewa et al., 2009; Liu et al., 2010), thus recruiting Pol II to the promoter. Subsequently, TFIIE and TFIIH are recruited to complete formation of the PIC that possesses basal transcription activity.

Biochemical evidence suggested cooperative functions of TFIIE and TFIIH within PIC (Goodrich and Tjian, 1994; Holstege et al., 1996; Lin and Gralla, 2005; Ohkuma and

Roeder, 1994; Watanabe et al., 2003) and the recent cryo-EM structures of yeast PICs revealed multiple interaction sites (Schilbach et al., 2021; Schilbach et al., 2017). The RING domain of Tfb3 (MAT1 in human) is located between Pol II stalk (Rpb4/7) and TFIIE zinc ribbon. The PH domain of Tfb1 (p62 in human) is apparently stabilized by Tfa1 “E-dock” which extends a “E-wing” loop that interacts with single stranded DNA upon DNA opening (Plaschka et al., 2016). Tfa1 “E-floater” interacts with Tfb1 BSD1. Additionally, Tfa1 “E-bridge” and Tfa2 “E-tether” bind C-lobe of translocase subunit Ssl2, supporting biochemical observations that THIIH activity is stimulated by TFIIE.

1.4 Mechanisms of promoter opening

Multiple cryo-EM structures of yeast (Dienemann et al., 2019a; Murakami et al., 2013c; Murakami et al., 2015b; Robinson et al., 2016a; Schilbach et al., 2021; Schilbach et al., 2017) and human PICs (Aibara et al., 2021; He et al., 2013; He et al., 2016; Rengachari et al., 2021) became available in the past several years in an effort to understand mechanisms of transcription initiation. The region downstream of TATA box of closed promoter DNA is stabilized above the Pol II cleft by the contacts with Tfg2 WH domain, Tfa1 E-wing, and Ssl2. Ssl2 binding induces ~20 degree bending of double stranded DNA (dsDNA) around 30-nt downstream of TATA box (Dienemann et al., 2019a; Murakami et al., 2015b). Further, DNA distortion in the initially melting region (~20-30 nt downstream of TATA) was observed in the absence and presence of TFIIH upon Pol II clamp closure, possibly contributing to dsDNA destabilization for promoter opening (Dienemann et al., 2019a). Very recently, a high-resolution (2.9 Å) structure of yeast PIC was determined in the presence of ADP-BeF_3 which mimics a post-hydrolysis state (Schilbach et al., 2021). This study shows an intermediate PIC structure in which 6-bp DNA is unwound, consistent with the size of the bubble observed in the single molecule study (Tomko et al., 2017). This initial 6-bp bubble was located in the upstream end (30-35 nt downstream of TATA) of the initially melting region and stabilized by the moderately charged loop protruding from Pol II clamp head and the Tfg1 charged loop. Thus these regions function in promoter melting. Furthermore, the structure suggests that the bubble propagates upstream till around 20-nt downstream of TATA box during transcription initiation (Giardina and Lis, 1993).

1.5 TSS scanning

A vast majority of eukaryotic promoters possess multiple transcription start sites (TSSs). While human PIC associates with DNA at ~30 nt upstream of TSS, in *S. cerevisiae,* PIC assembles and initiates transcription ~40-120 nt upstream of TSS (Smale and Kadonaga, 2003). As TFIIH Ssl2 (XPB in human) translocates downstream DNA, unwound DNA is fed into Pol II active sites (Fishburn et al., 2015). In TATA-containing promoters, TSS is not determined by fixed distance from TATA (Fishburn and Hahn, 2012; Murakami et al., 2015a), but rather DNA elements around TSS contribute to Pol II TSS usage (Chen and Struhl, 1985). Therefore, yeast Pol II demonstrates TSS scanning prior to recognition of TSS (Qiu et al., 2020). Although mechanistic details of TSS scanning still remains elusive, two optical tweezer based single molecule experiments using in vitro reconstituted PIC came up with conflicting models. One study showed that the distance between upstream and downstream DNA was shortened upon transcription initiation and suggested formation of on average of 85 bp open DNA prior to formation of elongation complex (Fazal et al., 2015b). The change in distance was similar in the conditions where ATP or all NTPs were present, indicating that TSS scanning by TFIIH might continue independent of RNA synthesis by Pol II. On the other hand, the other study reported TFIIH generates 6 bp open DNA in the presence of ATP, which expands to a 12-13 bp transcription bubble when all the NTPs are present. Additionally, alteration of Pol II, TFIIB, and TFIIF functions affect promoter scanning and changes TSS usage genome-wide in yeast by modulating the initiation efficiency (Qiu et al., 2020). Also, processivity of Ssl2 determines TSS scanning window (Zhao et al., 2021b). TSS selection is thus influenced by numerous factors including stability of the open complex, speed of TFIIH translocation and Pol II transcription.

1.6 Post-assembly of PIC

Expression of genes are regulated at multiple steps during transcription. Gene-specific regulation can occur during PIC assembly step that involves recruitment of DNA binding activator proteins, co-activator complexes such as Mediator, SAGA, and TFIID, enhancer-promoter interaction, and so on. Global regulation of gene expression can occur post assembly of PIC. Recent single-molecule tracking in budding yeast to investigate dynamics of initiation factors suggest that the most PICs assembled at the promoter either fail to initiate transcription or to complete initiation cycle (Nguyen et al., 2020). This suggests that transcription initiation post-assembly is one of the rate-limiting steps. Further, nearly all the genes in *S. cerevisiae* including constitutive and inducible genes possess two Pol II stalling sites near TSS under stressed conditions, suggesting that these Pol II stalling sites are checkpoint locations for proper and coordinated initiation events (Badjatia et al., 2021).

During TSS scanning, Pol II finds appropriate TSS(s) and initiates RNA synthesis, at which PIC becomes the initially transcribing complex (ITC). The ITC typically undergoes multiple rounds of abortive initiation where short RNA (2-15 nt) physically dissociate from the complex, the ITC is reverted to the open PIC, and new RNA is synthesized (Carpousis and Gralla, 1980; Goldman et al., 2009; Wade and

Struhl, 2008). During abortive initiation, Pol II remains associated with the promoter and thus the upstream part of the complex is fixed. When Pol II transitions from initiation to elongation, it must leave the promoter by achieving dissociation from GTFs except TFIIF, and at the same time, an abrupt collapse of the upstream edge of the transcription bubble occurs, a process referred as promoter escape (Luse, 2013). Early biochemical studies focusing on understanding early initiation steps are mainly done in human systems and understandings of mechanisms prior to transitioning to elongation phase is rather fragmented. Human Pol II is unstable during early transcription and gains functional stability after synthesizing 4-nt long transcript (Cai and Luse, 1987; Holstege et al., 1997; Kugel and Goodrich, 2002). Another study shows Pol II initiation complex containing 10-nt or shorter transcript is fragile during purification (Coppola and Luse, 1984). Further, Human Pol II is susceptible to transcription arrest at promoter-proximal region 9-13-nt downstream of TSS prior to promoter escape (Dvir et al., 1997b) and ATP-dependent translocase activity of TFIIH has been shown to be required for formation of the escape-competent transcription complex (Dvir et al., 1996, 1997a; Dvir et al., 1997b). Although the length of the transcript seems to be an important factor for the stability of Pol II complex, the DNA sequence also affects (Keene and Luse, 1999). Once the upstream edge of the transcription bubble reanneals, the translocase activity of TFIIH is no longer required (Pal et al., 2005). Additionally, TFIIH kinase activity that phosphorylate Ser5 position on Pol II CTD triggers dissociation of Mediator from the PIC and thus promotes promoter escape (Wong et al., 2014)

Another important factor implicated in promoter escape is TFIIB which is known to be interacting with numerous locations on Pol II near the active site, including the clamp, protrusion, wall, folk loop, and rudder (Liu et al., 2010; Sainsbury et al., 2013).

Mutations and deletion analysis of TFIIB B-reader region led to utilization of abnormal TSS and reduced transcription (Bangur et al., 1997; Chen and Hampsey, 2004; Kuehner and Brow, 2006) suggesting this region of TFIIB participates in stabilization of upstream template DNA and the DNA/RNA hybrid (Sainsbury et al., 2013).

Apart from the fact that the RNA length and sequence contribute to stability of the early initiation complex and TFIIB is involved in processes in early initiation steps, how Pol II gains escape competency and decides to escape the promoter are not completely understood. Early structural studies of yeast Pol II complexes gave some insights into these steps. The DNA/RNA hybrid in yeast Pol II active site is 8-nt based on the crystal structure (Westover et al., 2004), which explains why Pol II complexes containing shorter transcripts are unstable. As Pol II elongates RNA, the 5' end of the transcript is guided into RNA exit channel and the protein-RNA interaction within RNA exit tunnel probably contributes to stabilization of the early initiation complex. Concomitantly, the 5' end of elongating RNA clashes with the N-terminal domain of TFIIB when the length reaches ~13-nt (Sainsbury et al., 2013) and thus it has to be displaced as RNA becomes elongated. Others have shown biochemically that human TFIIB release is triggered when the transcription bubble is at least 18-nt and the transcript is 7-9 nt length, and suggested that some structural changes within the ITC that allow for TFIIB release may occur (Ly et al., 2020; Pal et al., 2005). Another study shows Human TFIIF is involved in destabilization of TFIIB during initiation (Čabart et al., 2011). Whether and how the structural changes within the ITC during initiation, if any, happen and how these would lead to ejection of GTFs are still unsolved. Currently available structures of human ITC (He et al., 2016) and yeast core ITC (lacking TFIIH) (Plaschka et al., 2015) did not explain these biochemical data. This could partially be due to that fact that these ITCs were formed without relying on the translocase activity of TFIIH (He et al., 2016; Plaschka et al., 2015). Further investigations are needed to understand what exactly happens during transcription initiation.

Soon after initiation, Pol II is subjected to a pause near the promoter before proceeding to productive elongation in higher eukaryotes (Core and Adelman, 2019). This event called "promoter-proximal pausing" is a general feature of early elongation of active genes (Muse et al., 2007; Zeitlinger et al., 2007) and many factors, including DNA/RNA sequences and proteins, are involved in establishment and release of paused Pol II (Nechaev et al., 2010). As the 5' end of RNA emerges from Pol II, Spt5, a larger subunit of DSB sensitivity-inducing factor (DSIF) associates with RNA and recruited to the elongation complex (Missra and Gilmour, 2010). Subsequently DSIF recruits negative-elongation factor (NELF). DSIF, composed of Spt4 and Spt5, binds Pol II around the RNA exit tunnel and the clamp (Bernecky et al., 2017; Vos et al., 2018a; Vos et al., 2018b), and NELF locates around the bottom of Pol II including in Pol II funnel and by the stalk (Vos et al., 2018b). Paused Pol II is in a backtracked state meaning that the 3' end of RNA is displaced from the Pol II active site and placed in the Pol II funnel (Nechaev et al., 2010). To rescue backtracked Pol II, TFIIS which stimulates endonuclease activity of Pol II is needed (Izban and Luse, 1992). Recent cryo-EM structure of Pol II-DSIF-NELF revealed how NELF establishes promoter-proximal pausing (Vos et al., 2018b). NELF interacts with and restricts the movement of trigger loop in Pol II active site, the element involved in catalysis. Binding of NELF along the Pol II funnel impedes diffusion of NTPs, making less substrates available in the active site, and it also interferes with TFIIS binding which is necessary for reactivation of backtracked Pol II. Recently, backtracking of Pol II was shown to be a general phenomenon in early elongation in vivo and rescue of the arrested Pol II by TFIIS is critical for efficient elongation (Sheridan et al., 2019). Additionally, recent studies indicate that TFIID is also involved in regulation of pausing and release of Pol II (Dollinger and Gilmour, 2021; Fant et al., 2020). Mechanisms by which TFIID regulates promoter-proximal pausing are not yet known.

Release of promoter-proximal paused Pol II into productive elongation is triggered by the kinase activity of P-TEFb which is comprised of CDK9 and Cyclin T1 (Marshall and Price, 1995). The P-TEFb target includes NELF, DSIF, and Pol II. Phosphorylation on NELF-A tentacle by P-TEFb destabilizes NELF binding to Pol II (Lu et al., 2016) and PAF can be recruited upon NELF dissociation (Vos et al., 2018a). Another elongation factor, Spt6 is recruited through RNA as well as phosphorylated of Pol II CTD linker by P-TEFb (Vos et al., 2018a). Cryo-EM structure of Pol II-DSIF-PAF-SPT6 indicates that binding of PAF and SPT6 induces conformational changes of DSIF and Pol II clamp which might allow for efficient elongation (Vos et al., 2018a). Also RTF1, a PAF subunit, was shown to allosterically stimulates elongation by interacting with the N-edge of Pol II bridge helix in the active site (Vos et al., 2020).

1.7 Pol II C-terminal domain (CTD)

As alluded to above, Pol II is targeted by kinases and phosphatases during transcription. The largest subunit of Pol II, Rpb1, which possesses the catalytic activity, contains repeats of heptapeptide with the consensus sequence $Y_1S_2P_3T_4S_5P_6S_7$ in its carboxyl-terminal domain (CTD). This is a unique feature that distinguishes Pol II from other polymerases and is conserved from fungi to human although the number of repeats varies among species (Hsin and Manley, 2012). *S. cerevisiae* possesses 26 repeats while human has 52 repeats. Pol II CTD is subjected to phosphorylation and dephosphorylation during transcription and timely modification is crucial for proper co-transcriptional processes. During initiation, hypo-phosphorylated Pol II assembles in a PIC. The kinase subunit of TFIIH, Kin28 in *S. Cerevisiae* and Cdk7 in human, phosphorylates Ser5 and Ser7 of the Pol II CTD (Akhtar et al., 2009; Buratowski, 2009; Glover-Cutter et al., 2009). Both the modification marks are enriched near the promoter region, and loss of TFIIH kinase activity using an inhibitor results in reduced level of the modification marks. The known interactor of phosphorylated Ser7 is integrator complex (Egloff et al., 2007) which has been shown to be involved in termination of snRNA (Baillat et al., 2005) as well as mRNA near the promoter region (Elrod et al., 2019; Tatomer et al., 2019; Vervoort et al., 2021). Phosphorylation at Ser5 has been shown to be important for mediator dissociation from the PIC (Søgaard and Svejstrup, 2007) and subsequently for promoter escape (Buratowski, 2009). Additionally, a number of factors are recruited to transcribing Pol II through phosphorylated Ser5, such as the capping enzyme (Cho et al., 1997; Fabrega et al., 2003; McCracken et al., 1997; Yue et al., 1997), the Set1 histone methyltransferase complex (Ng et al., 2003) and non-polyA termination factor Nrd1 (Vasiljeva et al., 2008). Phosphorylated Pol II CTD tethers the capping enzyme close to the RNA exit channel, allowing for the efficient 5'-capping of the nascent RNA as soon as it emerges from the Pol II (Martinez-Rucobo et al., 2015) and this 5'-capping is crucial for mRNA stability and recruitment of the translation machinery.

As Pol II travels further from the 5' end of the genes, Ser5 gets dephosphorylated by phosphatases such as Rtr1 (Mosley et al., 2009) and Ssu72 (Krishnamurthy et al., 2004) and the level of phosphorylated Ser2 increases. Ctk1 and Bur1 in *S. cerevisiae* and Cdk9 in human phosphorylate Ser2 and elongation factors such as Spt4/5 (*S.*

cerevisiae)/DISF (human) and NELF (metazoan specific factor). Phosphorylation of Ser2 is another checkpoint for proper transcription as Pol II enters a productive elongation phase. Phosphorylated Ser2 is important for recruiting additional elongation factors such as Spt6 (Burugula et al., 2014).

1.8 TFIID

TFIID is one of the GTFs comprised of TBP and 13-14 subunits of TBP-associated factors (TAFs) that form a trilobed complex containing lobe A, B, and C (Patel et al., 2018). TFIID functions in TBP loading for formation of PIC at the promoter (Rhee and Pugh, 2012). Depending on the promoter and conditions, the entire TFIID is not necessary for transcription in vitro, but TBP is required. In vitro study using yeast nuclear extract showed that negative effects on transcription initiation of both TATA-containing and TATA-less promoter genes seen upon depletion of Taf1 (a largest subunit of TFIID) could not be rescued by the presence of high levels of TBP (Donczew and Hahn, 2018), suggesting general requirement of TFIID. A genome-wide study suggests requirement of TFIID for all genes in vivo (Warfield et al., 2017), but the recent study showed most genes (87%) are strongly affected by TFIID depletion but not by depletion of SAGA, a co-activator complex generally important for H3 acetylation. The rest of genes (13%) are dependent on both TFIID and SAGA suggesting that they are co-activator redundant genes (Donczew et al., 2020).

1.9 Mediator

Another important factor for Pol II transcription is Mediator (Soutourina, 2018), which is the last PIC component found in Kornberg and Young labs through biochemical and genetic studies in yeast (Flanagan et al., 1991; Kelleher et al., 1990; Koleske and Young, 1994; Thompson et al., 1993). Mediator is a large protein complex comprised of

25 subunits in yeast and 30 subunits in human that form three modules called head, middle and tail, and a 4-subunit kinase module (Dotson et al., 2000; Verger et al., 2019). Mediator is required for activation of activator-dependent transcription in vivo and in vitro, and for enhancer regulated transcription (Allen and Taatjes, 2015; Flanagan et al., 1991; Koleske and Young, 1994). Though mechanistic details of how exactly Mediator stimulate Pol II transcription is still under active investigation, it is evident that numerous transcription factors interact with different Mediator subunits, which orchestrates transcriptional responses (Brzovic et al., 2011; Fondell et al., 1996; Stevens et al., 2002).

Transcription factors contain DNA-binding domains (DBDs) and activation domains (ADs). ADs are involved in recruitment of co-activators, including Mediator, SAGA, and TFIID. Unlike DBDs, ADs have poorly conserved sequences, which makes predictions of AD features challenging, but the sequence features with the activation function have been identified to be enriched for certain amino acids, such as acidic, aromatic, and hydrophobic residues. Further, most of ADs are intrinsically disordered (Staby et al., 2017). One of the most characterized kind of ADs, the acidic AD, has been shown to dynamically interact with a co-activator through “fuzzy” interactions. For instance, yeast activator Gcn4 can bind Med15 in multiple orientation through hydrophobic regions, and there is no requirement for specific sequences for activation (Tuttle et al., 2018). Additionally, ADs can phase-separate and form condensates with Mediator, which creates concentrated environment at promoter-enhancer regions, stimulating transcription of the genes (Boija et al., 2018; Chong et al., 2018; Sabari et al., 2018)

Earlier structural studies of PIC-mediator from yeast revealed that Pol II and mediator interaction is stabilized through three interfaces: Med18-Med20 binds TFIIB, Rpb1, and Rpb3-Rpb11, Med8-Med11 binds Rpb4/7 stalk, and Med9 binds Pol II foot (Plaschka et al., 2015; Robinson et al., 2016a; Schilbach et al., 2017). Due to the flexibility of the large complex especially the tail module, it had been challenging to determine high resolution structure of the entire Mediator-PIC and these studies were limited in that only the structures of head and middle modules were determined.

Very recently, medium to near atomic resolution structures of the Mediator-Pol II from thermophilic fungus, apo Mediator from mouse, and human Mediator-PIC became available. These studies provided more detailed information on how Mediator subunits within the complex interact with each other and how Mediator interacts with core PIC. They also revealed the structure of the tail module (Abdella et al., 2021; Chen et al., 2021b; El Khattabi et al., 2019; Rengachari et al., 2021; Zhang et al., 2021; Zhao et al., 2021a). All the structures show the conserved role of Med14 connecting head, middle and tail modules consistent with the previous biochemical study that defined the role of human Med14 as an architectural and a functional backbone of the complex (Cevher et al., 2014). Overall structure of the head module of human mediator is very similar to that of yeast structure except for the presence of additional subunits Med27, Med28, Med29, and Med30.

So how does Mediator and activator proteins stimulate transcription? There is no clear answer to this yet, but the structures provided some insights. Based on the structure of Mediator-Pol II from thermophilic fungus, Med15 connects with Med14 on the side of Med1 and thus activator binding with Med15 must occur opposite side of Med1 (Zhang et al., 2021). Med15 cryo-EM density revealed a bundle of seven alpha helices exposed on the surface. This bundle contains hydrophobic cleft and flanked by basic residues, both of which are critical features for activator binding. Comparison between the apo mediator and Pol II bound mediator revealed extensive conformational changes upon Pol II binding and coordinated movement of hook of the middle module and the tail module. In the structural studies of human complex, the domains involved in activator or transcription factor binding including the N-terminus of Med15, the N-terminus of Med25, and the C-terminus of Med1 were not visible, indicating these domains are flexible. Although the activation domain of VP16 was included during sample preparation, the activator domain and its binding partner, the N-terminus of Med25, were flexibly tethered leading to lack of their density in the cryo-EM map. Therefore activator binding might not induce major conformational changes in Mediator unlike previously suggested (Bernecky and Taatjes, 2012; Meyer et al., 2010).

1.10 Mediator kinase module

Mediator complex strongly associates with promoter and enhancer while the kinase module is only enriched at the enhancers (Jeronimo et al., 2016; Petrenko et al., 2016). Mediator contains a kinase module that is comprised of four subunits, which cross-links mainly with the middle module and some with the head module of Mediator based on an in vitro study (Osman et al., 2021). The majority of the crosslinks were located at the interface between Pol II and the middle module, suggesting that the kinase module inhibits mediator binding to Pol II upstream of activator sequences. Cdk8, the kinase subunit, self-phosphorylates in the presence of ATP and also phosphorylates middle and head modules at the interacting interface of the kinase module and core Mediator (Osman et al., 2021; Tsai et al., 2013). The kinase activity weakens the interaction of the kinase module and core Mediator, releasing Mediator to bind Pol II during activation. Thus the kinase module has positive and negative effects on transcription consistent with previous studies.

1.11 TFIIH with a focus on the structure of kinase module

TFIIH is a 10-subunt protein complex conserved from yeast to human. TFIIH possesses ATP-dependent translocase subunit (Ssl2 in yeast and XPB in human) and required for Pol II transcription whereas in Pol I and Pol III, DNA can be opened without a translocase activity possessing factor. TFIIH can be divided into two modules: core and kinase modules. The kinase module contains three subunits and is called TFIIK (containing Kin28, Ccl1, and Tfb3) in yeast and CAK (containing CDK7, Cyclin H, and MAT1) in human. Nogales and Murakami labs determined the structure of human CAK and yeast TFIIK, respectively (Greber et al., 2020; van Eeuwen et al., 2021a). The structure of human CAK shows MAT1 interacts with both CDK7 and Cyclin H more extensively than CDK7 and Cyclin H themselves interact, explaining its role as an assembling factor for CDK7 and Cyclin H (Greber et al., 2020). Additionally, alpha-helix in MAT1 C-terminus is in close proximity to CDK7 regulatory T-loop which shifts away from the catalytic site in the active form, consistent with the previous biochemical observation that MAT1 activates the kinase activity (Busso et al., 2000). These findings are conserved in yeast TFIIK (van Eeuwen et al., 2021a).

In the structure of PIC lacking mediator, the only visible part of the TFIIH kinase module is N-terminus of Tfb3/MAT1 that interacts with Rad3/XPD and anchors the core TFIIH on Pol II stalk (He et al., 2016; Murakami et al., 2013c; Murakami et al., 2015b). The rest of the kinase module was not seen in these structures due to its flexible tethering. This was overcome by the presence of Mediator. In the yeast mediator-bound PICs, the TFIIK density was observed, but its orientation was not clear because of the limited resolution (Robinson et al., 2016a; Schilbach et al., 2017). In the recent human structures of the complete PIC, CAK module was seen consistent with the yeast structures and was be able to be unambiguously docked to the Cryo-EM density as a rigid body (Abdella et al., 2021; Chen et al., 2021b; Rengachari et al., 2021). CAK module was stabilized by interactions of CDK7 with Med6, Med14, and Med19 in the orientation that allows for the catalytic site of CDK7 to face the hook domain of Mediator.

In the structure of the head module of yeast mediator co-crystalized with Pol II CTD peptide shows the CTD binding to the neck domain of MedHead (Robinson et al., 2012). The CTD density was seen in the same location in human PIC-mediator structure (Abdella et al., 2021) in which the CTD adopts slightly different conformation from the yeast Pol II CTD most likely due to the presence of metazoan specific subunit Med31. Another notable difference between human and yeast structures is that the Pol II CTD interacts less extensively with MedHead due to a clash with N-terminus of Med7.

1.12 Cryo-electron microscopy for structural studies

Having detailed structural information of macromolecules is essential to understand how they function. For decades, X-ray crystallography was a major technique for structure determination of proteins. Protein crystals are bombarded with X-ray and the resulting diffraction patterns can be used to find the three-dimensional positions of atoms. This technique often gives high quality information of the protein structure when protein crystals can be obtained for analysis, yet it is extremely difficult and sometimes impossible to produce crystals of large, flexible, and fragile macromolecules. Cryo-electron microscopy (cryo-EM) is a powerful tool that does not require crystallization of biomolecules for structural analysis. In cryo-EM, proteins in solution are applied onto grids and flash frozen in liquid ethane and the resulting thin layer of amorphous ice in which proteins are embedded are bombarded with electrons for data collection (Murata and Wolf, 2018). Because cryo-EM does not rely on crystallization like X-ray crystallography that captures one form of protein conformation, cryo-EM can visualize multiple dynamic states of proteins in a native-like environment, which is one of the advantages of this technique.

Technological development of both hardware and software allowed for determination of higher resolution structures by cryo-EM (Callaway, 2020; Mitra, 2019; Shen, 2018). As the technique is becoming more accessible to researchers, increasing numbers of cryo-EM structures are published in recent years. In 2017, Jacques Dubochet, Joachim Frank, and Richard Henderson were awarded a novel prize for developing cryo-EM techniques. Although it is still difficult to obtain high resolution structures of small proteins (<100 kDa), technical development enabled structural determination of 52 kDa strateptavidin at 3.2Å resolution (Fan et al., 2019) and 64 kDa hemoglobin at 3.2Å (Khoshouei et al., 2017) in recent years. Furthermore, data acquisition and analysis to obtain final structures are taking less time now than before and more complex and flexible samples can be analyzed with the advancement of technologies.

1.13 Conclusions

Transcription is a sophisticated process that involves many factors. The output of transcription can affect gene expression and dysregulation of gene expression leads to many diseases. Thus, transcription is tightly regulated at multiple places to make sure it occurs in a timely and coordinated fashion. Understanding the fundamental processes of transcription is critical and many studies have been done biochemically and on a genome-wide level since discovery of RNA polymerases about 50 years ago. Further, thanks to recent technical development of cryo-EM and data analysis tools and those who elucidated the structures of transcription complexes, our mechanistic understanding of Pol II transcription improved dramatically in the last several years.

Despite decades of studies, the mechanistic understanding of the transition from initiation to elongation is still lacking. My thesis focuses on understanding post-initiation mechanisms specifically how Pol II leaves promoter and how the initially-transcribing complex is converted to an elongation complex in *S. cerevisiae.* We reconstituted transcription initiation using purified proteins from budding yeast and isolated post-initiation complexes stalled at different distances from TSSs. In Chapter 2, we identify and characterize those post-initiation complexes and found that the yeast ITC continued to associate with the GTFs and promoter longer than expected. Addition of capping enzymes (Cet1-Ceg1 and Abd1) and an elongation factor (Spt4/5) promoted promoter escape. In Chapter 3, we show structures of post-initiation complexes, the ITC and the reinitiated, colliding Pol II-Pol II complex. These structures provide insights into promoter escape.

CHAPTER 2: THE CAPPING ENZYME FACILITATES PROMOTER ESCAPE AND ASSEMBLY OF FOLLOW-ON PREINITIATION COMPLEX FOR REINITIATION

2.1 Preface

The manuscript presented in this chapter was originally published online on November 5, 2019 (Fujiwara et al., 2019). It has been reformatted here in accordance with University of Pennsylvania dissertation formatting guideline.

Authors: Rina Fujiwara[1,2], Nivedita Damodaren[3], Jeremy E. Wilusz[1], and Kenji Murakami[1*]

Affiliations:
[1]Department of Biochemistry and Biophysics, Perelman School of Medicine, University of Pennsylvania, Philadelphia PA. 19104, USA

[2]Biochemistry and Molecular Biophysics Graduate Group, Perelman School of Medicine, University of Pennsylvania, Philadelphia PA. 19104, USA

[3]Molecular Biology Institute, University of California, Los Angeles CA, 90095

*Correspondence: kenjim@pennmedicine.upenn.edu

2.2 Respective Contributions

The majority of the experiments and analyses in this chapter are performed by me under the guidance of Dr. Kenji Murakami. Dr. Jeremy Wilusz performed northern blot in Figure 2A. The manuscript was written by me, Dr. Jeremy Wilusz and Dr. Kenji Murakami.

2.3 Abstract

After synthesis of a short nascent RNA, RNA polymerase II (pol II) dissociates general transcription factors (GTFs; TFIIA, TFIIB, TBP, TFIIE, TFIIF, TFIIH) and escapes the promoter, but many of the mechanistic details of this process remain unclear. Here, we developed an *in vitro* transcription system from the yeast *Saccharomyces cerevisiae* that allows conversion of the pre-initiation complex (PIC) to bona fide initially transcribing complex (ITC), elongation complex (EC), and re-initiation complex (EC+ITC). By biochemically isolating post-initiation complexes stalled at different template positions, we have determined the timing of promoter escape and the composition of protein complexes associated with different lengths of RNA. Almost all of the post-initiation complexes retained the GTFs when pol II was stalled at position +27 relative to the transcription start site, whereas most complexes had completed promoter escape when stalled at +49. This indicates that GTFs remain associated with pol II much longer than previously expected. Nevertheless, the long-persisting transcription complex containing RNA and all the GTFs is unstable and is susceptible to extensive backtracking of pol II. Addition of the capping enzyme and/or Spt4/5 significantly increased the frequency of promoter escape as well as assembly of a follow-on pre-initiation complex (PIC) at the promoter for re-initiation. These data indicate that elongation factors play an important role in promoter escape, and that ejection of TFIIB from the RNA exit tunnel of pol II by the growing nascent RNA is not sufficient to complete promoter escape.

2.4 Introduction

In eukaryotic transcription, RNA polymerase II (pol II) and a set of general transcription factors (GTFs), including TFIIA, TBP, TFIIB, TFIIE, TFIIF, and TFIIH, assemble in a pre-initiation complex (PIC) that is responsible for promoter opening and scanning of transcription start sites (TSSs) (Conaway and Conaway, 1993; Kornberg, 2007). Once a TSS is recognized, pol II begins to synthesize a nascent RNA, thereby converting the PIC into the initially transcribing complex (ITC), which is comprised of all the GTFs, pol II, and RNA. The transcription bubble is propagated downstream until the nascent RNA reaches a certain length. The upstream segment of the bubble then abruptly re-anneals, resulting in dissociation of all the GTFs except TFIIF from the ITC (Luse, 2013). This causes the ITC to be converted to an elongation complex (EC), which only contains pol II, TFIIF, and RNA, and this conversion step is known as promoter escape. Additional pol II and TFIIF can subsequently be recruited to the promoter to enable re-initiation of transcription using the other GTFs that remained committed to the template (Yudkovsky et al., 2000).

Once the nascent transcript reaches a length of ~20-30 nt in vivo (Rasmussen and Lis, 1993; Tome et al., 2018) and ~20 nt in vitro, 5’ capping of the nascent RNA occurs (Mandal et al., 2004; Martinez-Rucobo et al., 2015; Nilson et al., 2015). This is shortly after the 5’ end of the transcript has emerged from the pol II RNA exit tunnel. In yeast (*S. cerevisiae*), RNA 5’ capping involves three steps: (i) removal of a gamma phosphate from the 5’ end of the RNA by Cet1, (ii) transfer of guanosine monophosphate (GMP) by Ceg1, and (iii) methylation of the guanosine by Abd1 (Mao et al., 1995; Schroeder et al., 2004). Cet1 and Ceg1 form a heterodimer and are recruited to the transcription complex upon binding phosphorylated Ser5 of the pol II C-terminal domain (CTD) (Cho et al., 1997; McCracken et al., 1997; Rodriguez et al., 2000). Shortly after recruitment of Cet1-Ceg1, Spt4/5 (the yeast homolog of DSIF) also binds pol II (Lidschreiber et al., 2013) and facilitates productive elongation in vivo. It has been long assumed that promoter escape occurs after synthesis of 9-15 nt RNA (Luse, 2013), thus preceding the recruitment of Cet1-Ceg1 and Spt4/5, but this assumption has not been experimentally proven.

To better characterize the early steps in transcription, we previously developed an *in vitro* reconstituted system from yeast (*S. cerevisiae*) purified factors in which ~10-30% of the assembled PICs were capable of RNA synthesis (0.1–0.3 transcripts per template) (Murakami et al., 2013b; Murakami et al., 2015a). Single-molecule analysis using this system indicated that pol II can remain associated with promoter DNA (via interactions with GTFs) even when transcribing RNAs of ~50 nt in length (Fazal et al., 2015b). This result is inconsistent with promoter escape occurring when the RNA reaches 9-15 nt in length. Nevertheless, as only pol II and GTFs were present in these reactions, we hypothesized that additional factors, e.g. capping enzymes and/or Spt4/5, may play a role in promoter escape.

In this study, we first improved our *in vitro* transcription system and have now achieved at least 90% efficiency (by determining the extent of template usage). We then used this system to re-examine the mechanism of transcription initiation, specifically the characteristics of the ITC in the absence and presence of capping enzymes (Cet1-Ceg1 and Abd1) and Spt4/5. Transcription complexes were stalled at various positions on a G-less *SNR20* promoter and isolated by glycerol gradient sedimentation. This revealed the composition of transcription complexes associated with different lengths of RNA. Our data indicate that the ITC, which contains all the GTFs, pol II, and RNA, persists at least until Cet1-Ceg1 and Spt4/5 are recruited. Cet1-Ceg1 and Spt4/5 then facilitate the transition from initiation to elongation by promoting promoter escape.

2.5 Results

In vitro reconstituted transcription system with ~90% efficiency

We previously developed a transcription initiation reconstituted system using yeast proteins that had been purified from *E. coli* or yeast (Fig. S1) and showed that 10-30% of the assembled PICs were active (Murakami et al., 2013b; Murakami et al., 2015a). To reveal additional insights into the transition from initiation to elongation, we set out to further optimize this system and achieve higher transcription efficiency. We generated a set of U2 snRNA promoter (*SNR20*) variants that have a G-free region between the TSS (+1) and a G-stop at +27, +39, +49, or +85 (named G-less 27, G-less 39, G-less 49, and G-less 85, respectively) (Fig. 1A). Inclusion of chain-terminating 3'-O-methyl GTP instead of GTP in the reactions enabled pol II to be efficiently stalled at the end of the G-free region. We thus reasoned that this approach should allow accurate quantification of the efficiency of initiation, including independent measurements of the efficiencies of the first round of transcription vs. re-initiation. It should nevertheless be noted that the *SNR20* promoters yielded various lengths of transcripts due to multiple TSSs at positions +1 to +7 (Fig. 1B).

By adjusting the concentrations of factors added to the reactions (Fig. S2A-C), we identified conditions that allowed ~90% efficiency (by determining the extent of template usage, defined as the percentage of DNA templates that were transcribed) from the G-less *SNR20* templates as measured by incorporation of [α-^{32}P] UTP into nascent transcripts (Table 1). As in our previous studies (Murakami et al., 2013b; Murakami et al., 2015a), we use almost equimolar amounts of the GTFs (TFIIA, TFIIB, TBP, TFIIE, TFIIH and TFIIK) and promoter DNA. Here, however, we found that the efficiency increased by ~2-3 fold upon addition of 4-fold molar excess pol II and TFIIF relative to DNA (Fig. S2A) and by ~1.4 fold in the presence of Sub1 (Fig. S2B), the yeast homolog of PC4 (Ge and Roeder, 1994; Henry et al., 1996). No increase in efficiency was observed upon titrating other GTFs (TFIIA, TFIIB, TFIIE, TFIIH, or TFIIK) (Fig. S2C), and thus all subsequent analyses were done using reactions that contain Sub1 along with excess TFIIF and pol II. As expected, the level of nascent transcripts produced increased over time, but peaked once the reactions had been incubated for 20 min (Fig. S2D and S2E). At this time point, the efficiency of the first round of transcription was very high, ranging from 0.9 to 0.98 depending on the promoter variant examined (Table 1).

Reconstituted transcription initiation system supports 5' capping and re-initiation.

To further define the capabilities of the optimized transcription initiation system, we tested whether transcripts generated from the G-less templates become capped by 3'-O-methyl GTP and recombinant Cet1-Ceg1 and Abd1, which were added to the initiation reactions at the time of NTP addition. Given that Abd1 might have roles in transcription initiation independent of its methylation activities (Schroeder et al., 2000; Schroeder et al., 2004), S-adenosyl methionine (SAM), which is required for methylation of the cap, was not included unless otherwise noted. A shift in RNA mobility by ~1 nt was observed upon addition of Cet1-Ceg1, consistent with addition of a 5'-cap, and this effect was enhanced by addition of Abd1, Spt4/5, and/or SAM (Fig. 1B and Fig. S3A-B). To quantitate the extent of 5'-capping, RNAs were isolated from the transcription reactions and then treated with CIP/PNK followed by digestion with a 5'-3' exonuclease (Fig. S4A). In the absence of Cet1-Ceg1, RNAs had triphosphorylated 5' ends and were susceptible to digestion by the exonuclease, as expected (Fig. S4B-C). In contrast, when a 4-fold molar excess of Cet1-Ceg1 relative to DNA was added to the reactions, ~41% of the G-less 27 (Fig. S4B) and ~80% of the G-less 49 (Fig. S4C) transcripts from the first round of initiation became capped, consistent with the efficiencies estimated from the ~1 nt shifts in RNA mobility (Fig. S3A). The difference in capping efficiency between templates is likely due to the 5’ ends of the shorter transcripts being less accessible for capping (discussed further below). It should be noted that a similar ~1 nt shift in RNA mobility was observed when capping enzymes were added after transcription had been completed, confirming that the mobility shift was indeed due to capping and not a shift in the TSS (Fig. S4D). Based on these results, we conclude that our transcription system supports 5’ capping.

Upon further examining the transcripts generated from the different promoter templates, we noted that the G-less 49 (Fig. 1B, lanes 17-24) and G-less 85 templates (Fig. S5) produced transcripts with ~25 nt stepwise decrease in length. These results are analogous to a previous study that showed successive pol II stacking when pol II was stalled in an in vitro human transcription system (Szentirmay and Sawadogo, 1994). This ~25 nt decrease in length is also in good agreement with previous in vitro footprinting analysis of two colliding pol II elongation complexes (Hobson et al., 2012). We thus reasoned that the shorter transcripts may represent products of transcription re-initiation. To address this hypothesis, we focused on the G-less 49 template that yielded ~49 nt and ~25 nt transcripts (Fig. 1B, lanes 17-24). Consistent with re-initiation from the same TSS as the 49 nt transcripts, Northern blot analysis revealed that the ~25 nt transcripts could be detected with a probe antisense to nt 1-25, but not with a probe to nt 26-49 (Fig. 2A). These data strongly suggest that transcription from the G-less 49 template results in a pol II elongation complex that has escaped the promoter and is stalled at +49 and that another pol II has initiated transcription by re-utilizing the promoter to generate the ~25 nt RNAs.

To further support re-initiation, we mapped the location of pol II molecules on the G-less templates by performing potassium permanganate ($KMnO_4$) footprinting. G-less 27 and G-less 49 DNA with the 5' ends of template (Fig. 2B) or non-template (Fig. 2C) strands were radiolabeled, and the DNA bound by stalled transcription complexes were reacted with $KMnO_4$. The DNA templates were then cleaved at reactive residues by treatment with piperidine and analyzed by denaturing PAGE gel electrophoresis (Fig. 2B-C). The increase in $KMnO_4$ reactivity in the presence of NTPs (i.e. post-transcription complexes) compared to in the absence of NTPs (i.e. PIC) was observed at residues downstream of TSS where pol II is stalled (Fig. 2D). On the G-less 27 template, a ~17 bp $KMnO_4$ hyperreactive region from residues +10 to +27 was observed (Fig. 2B-C bottom), which is consistent with the pol II active center being localized at the stall position. A more extended $KMnO_4$ hyperreactive region (residues -1 to +38) was observed on the G-less 49 template (Fig. 2B-C top). Although the signals are faint, residues ~20-40 bp downstream of the TATA box (residues -66 to -49) were also slightly reactive to $KMnO_4$ on both the G-less 27 and G-less 49 templates (Fig. 2B-D). This reactive region is consistent with the location where initial melting occurs through the translocase activity of TFIIH (Fazal et al., 2015b; Fishburn et al., 2015; Murakami et al., 2015a; Pal et al., 2005) and could be indicative of the presence of ITCs (Choi et al., 2004; Holstege et al., 1997; Pal et al., 2005). Taken together, we conclude that the improved transcription initiation system can support 5' capping as well as re-initiation (and validation by gradient sedimentation is described below that further supports that re-initiation is indeed occurring). Of note, the efficiency of the first round of initiation from the G-less 49 template is ~90%, but re-initiation is only ~28% efficient (Table 1).

Promoter escape is nearly completed when pol II is stalled at +49

To define the timing of promoter escape as well as how quickly the subsequent pol II can associate with the template, we sought to use glycerol gradient sedimentation to isolate complexes stalled at different template positions. After completion of the transcription reactions but before gradient sedimentation, non-hydrolyzable ATP (AMP-PNP) was added to potentially inhibit the translocase activity of TFIIH (Holstege et al., 1997). This was done to minimize structural changes during glycerol gradient sedimentation while retaining the structural assembly (Dvir et al., 1997a). Sedimentation gradients were then fractionated and protein stoichiometry in each fraction analyzed by SDS-PAGE. As a control, we first analyzed the sedimentation of the PIC and capping enzymes on the G-less 49 template under conditions where there should be only abortive RNA (~2 nt) synthesis (Fig. 3A). This is because only ATP and 3'-O-methyl GTP (but not CTP or UTP) were added to these reactions. Note that ATP was added because the PIC may behave as a slightly larger complex in the presence of ATP due to the binding of the capping enzymes through the phosphorylated pol II CTD (Cho et al., 1997; McCracken et al., 1997; Suh et al., 2010). In a glycerol gradient, we found that the capping enzymes and pol II interacted at nearly a 1:1 molar ratio irrespective of presence of transcripts that extend outside the RNA exit tunnel of pol II (Fig. 3A-C). In contrast, when TFIIK activity was inhibited, no CTD phosphorylation was observed (Fig. S6A-C) and the capping enzymes no longer co-sedimented with the PIC (Fig. S6D-E). It should be noted that excess pol II and TFIIF present in the reactions had no effect on sedimentation of the PIC (compare the sedimentation profiles in Fig. 3A to Fig. 4A, where roughly stoichiometric amounts of DNA, pol II, and TFIIF were present) and instead accumulated in fractions 14-17 (Fig. 3A and 3C).

We next compared this sedimentation profile to that obtained from G-less 49 transcription reactions that had post-initiation complexes stalled at +49 (Fig. 3D). Analysis of the RNAs present in each fraction revealed two major (fractions 5-8 and 14-17) and one minor (fractions 11-12) post-initiation complex (Fig. 3E). Notably, the complex that migrated fastest (fractions 5-8) contained all the GTFs, pol II, and capping enzymes (Fig. 3F), but with pol II and the TFIIF subunits present at ~1.6- and 1.5-fold molar excess, respectively, compared to their levels in the PIC from Fig. 3A (quantification in Fig. S7). This suggests that this post-initiation complex underwent the transition from initiation to promoter escape, thereby allowing binding of a second pol II. Indeed, both ~49 nt and ~25 nt transcripts, representing the first and second rounds of transcription, respectively, were observed in fractions 5-8 (Fig. 3E and Fig. S8). The molar ratio of the first and second rounds of transcripts was estimated as ~5:1.8 based on the band intensities, indicating that the complex is a mixture of EC+PIC (~64%) and EC+ITC (~36%) (Fig. S8). The second major post-initiation complex (fractions 14-17 in Fig. 3D-E) is EC containing pol II, TFIIF, Cet1-Ceg1, Abd1, and ~49 nt transcripts (but not TFIIE, TFIIH, TFIIA, and TBP) (Fig. 3G). Note that TFIIB, which is not compatible with EC, appeared in these fractions, due to co-migration of free pol II-TFIIF-TFIIB that did not engage in transcription (compare Fig. 3C and 3G). The third post-initiation complex (fractions 11-12 in Fig. 3D-E) migrated as fast as the control PIC (Fig. 3A-B) and contained the entire complement of the PIC polypeptides, capping enzymes, and ~49 nt RNAs (Fig. 3H). This complex is thus likely to be the ITC, which has not undergone promoter escape, as evidenced by the presence of all PIC components, in contrast to the other two post-initiation complexes. However, the third complex is much less populated than the other two post-initiation complexes, and thus we conclude that promoter escape usually occurs before +49.

Promoter escape is often completed when the RNA length is longer than 22 nt

To then further clarify the timing of promoter escape, we sedimented G-less 27 PIC (Fig. 4A) and post-initiation complexes (Fig. 4B-C) and analyzed whether promoter escape had been completed when pol II was stalled at +27. Similar to the results obtained with the G-less 49 template, a fast migrating post-initiation complex was observed (fractions 6-9 in Fig. 4B) that contained mainly ~24-27 nt transcripts (Fig. 4C) as well as all the GTFs, pol II, and capping enzymes (Fig. 4D). Subunits of pol II and TFIIF were present at 1.9- and 1.6- fold molar excess compared to those in PIC from Fig. 4A (quantification in Fig. S9), suggesting that the first pol II underwent promoter escape to allow a second pol II to bind. Given the ~25-nt spacing between pol II molecules that was observed on the G-less 49 template (Fig. 1-2), any re-initiated products from the G-less 27 template should be ~3 nt or shorter. This length is too short to be retained in a transcription complex (Luse, 2013) and thus we assigned this complex exclusively to EC+PIC (hereafter referred to as the re-initiation complex). Besides this major complex present in fractions 6-9, we observed a smaller number of G-less 27 transcription complexes that converted to ECs (Fig. 4B-C fractions 15-17) and ITC (Fig. 4B-C fraction 11).

24-27 nt transcripts were predominately retained in the re-initiation complex (Fig. 4C fractions 6-9), but 21-22 nt transcripts (derived from initiation from TSSs downstream of +1) largely remained on the top of the gradient after sedimentation (Fig. 4C, fractions 25-29). These RNAs may have been released from either ITCs or ECs, but we thought that ITCs seemed more likely as it has been previously suggested that the EC is very stable (Komissarova et al., 2003). To confirm these prior observations about the EC, we formed artificial ECs on the G-less 27 and G-less 49 templates by pre-annealing a 9-mer RNA at the TSS and then adding TFIIF, TFIIB, ATP, CTP, UTP, and 3'-O-methyl GTP to allow the 9 nt RNA to be elongated in the same conditions as the transcription initiation assay (Fig. S10A-C). Glycerol gradient sedimentation revealed that the 22, 27 and 49 nt RNAs were all predominantly present in the center of the gradient (Fig. S10C), confirming that elongated RNAs are stably retained in ECs. Based on these data, it is highly likely that the 21-22 nt RNAs that did not migrate into the gradient (Fig. 4C, fractions 25-29) were associated with ITCs but were then released from the complex during their isolation. These data suggest that promoter escape usually happens after the nascent RNA is longer than 22 nt in length.

Pol II in the ITC is susceptible to extensive backtracking

A reconstituted human system previously demonstrated that (i) short (<9 nt) RNAs can be released from the ITC when it reconverts to PIC upon addition of non-hydrolyzable ATP (Holstege et al., 1997) and that (ii) complexes stalled at promoter proximal positions (up to ~32 nt) are susceptible to extensive backtracking (Pal et al., 2001; Ujvari et al., 2002). We, therefore, reasoned that RNA release from the yeast ITC may similarly be due to extensive pol II backtracking that is caused by the addition of non-hydrolyzable ATP at the end of the transcription reaction and/or removal of ATP during glycerol gradient sedimentation. To explore this idea, we first used TFIIS cleavage assays to directly examine whether pol II backtracking occurs in the stalled transcription complexes (Fig. 4E). TFIIS stimulates the intrinsic activity of pol II to cleave the 3' end of the transcript when pol II backtracks, allowing for replacement of the new 3' end at the pol II active site and transcription restart (Cheung and Cramer, 2011; Sigurdsson et al., 2010). Transcription complexes that had been stalled on the G-less 27 or G-less 49 template were combined with TFIIS, incubated for an additional 6 min, and the resulting transcripts analyzed on a denaturing RNA gel (Fig. 4E). The level of ~49 nt transcripts derived from the G-less 49 template was largely insensitive to addition of TFIIS, whereas about 24% of the transcripts derived from the G-less 27 template were degraded upon addition of TFIIS (Fig. 4E, Lanes 4-6). The short (~25 nt) transcripts from the second round of transcription on the G-less 49 template were also highly sensitive to TFIIS (Fig. 4E, Lanes 1-3). These results indicate that transcription complexes stalled at promoter proximal positions (up to ~+27) are prone to extensive backtracking, whereas little backtracking is observed with the EC transcribing ~49-nt RNA.

Next, to address whether the observed extensive backtracking is an inherent feature of the ITC but not the EC, the artificial EC on the G-less 27 template was subjected to TFIIS cleavage assays (Fig. S10D-E). ~27 nt transcripts in the artificial EC were, as expected, cleaved by ~1-nt at their 3' termini, but were otherwise insensitive to addition of TFIIS (Fig. S10E, lanes 1-3). Note that substantial amounts of cleaved RNA were observed in the artificial EC (Fig. 10E) as well as from G-less 49 (Fig. 4E, left) which are probably due to repetitive partial backtracking followed by TFIIS reactivation of transcription, and not complete backtracking that would result in transcript release. Lastly, we tested whether more extensive pol II backtracking occurs in the ITC when the translocase activity of TFIIH, which exerts a forward force on pol II (Fazal et al., 2015b; Fishburn et al., 2015; Murakami et al., 2015a; Pal et al., 2005), is impeded by addition of a non-hydrolyzable ATP analogue. Unlike in the human system (Holstege et al., 1997), addition of 2 mM AMP-PNP (in the presence of 800 µM ATP) had no effect on pol II backtracking (lanes 7-9 vs 10-12 in Fig. S10E). This suggests that the continuous translocase activity of TFIIH may be maintained after addition of 2 mM AMP-PNP and that the large amounts of RNA that we observed to be released from the ITC with the G-less 27 template (Fig. 4C) was likely caused by removal of ATP during gradient sedimentation.

Cet1-Ceg1 and Spt4/5 facilitate the transition from initiation to elongation

Considering that the ITC persists until a nascent transcript reaches a length of ~22-23 nt, which is roughly when the capping enzymes bind (Martinez-Rucobo et al., 2015), we hypothesized that the capping enzyme may play a role in promoter escape and recruitment of a new incoming pol II for re-initiation. To test this model, we used gradient sedimentation to compare transcription complexes stalled at +49 or +27 in the presence (Fig. 5A and 5C) or absence of the capping enzymes (Fig. 5B and 5D). Omission of Cet1-Ceg1 and Abd1 from the reactions reduced the population of G-less 49 re-initiation complex (EC+PIC) (fractions 2-6 in Fig. 5A-B), while increasing the population of the ECs that failed to assemble a follow-on PIC (fractions 10-12 in Fig. 5A and 5B). Analogous experiments revealed that addition of Cet1-Ceg1 alone without Abd1 (Fig. S11A) or SAM (Fig. S11B) was sufficient to promote formation of the re-initiation complex. Omission of Cet1-Ceg1 and Abd1 likewise reduced the population of the G-less 27 re-initiation complex (EC+PIC) (fractions 3-5 in Fig. 5C-D), while increasing the population of ITCs as indicated by increased amounts of released RNAs (fractions 17-20 in Fig. 5C-D). Notably, omission of the capping enzymes resulted in release of almost all lengths of the nascent transcripts from ITCs (Fig. 5D), whereas only shorter transcripts (21–23 nt in length) were released in the presence of the capping enzymes (Fig. 5C). This suggests that RNA length may play a critical role in promoter escape, likely by contributing to recruitment of Cet1-Ceg1 to the transcription complex. Taken together, our results suggest that Cet1-Ceg1 facilitates promoter escape and the assembly of a follow-on PIC for re-initiation.

Spt4/5, the yeast homologue of DSIF, is recruited soon after recruitment of Cet1-Ceg1 (Lidschreiber et al., 2013). We thus asked whether Spt4/5 also promotes promoter escape and the association of a new incoming pol II for re-initiation. Indeed, glycerol gradient sedimentation of G-less 27 transcription complexes showed that formation of the re-initiation complex (EC+PIC) was enhanced upon addition of Spt4/5 in a similar manner to addition of Cet1-Ceg1 (Fig. S12). We thus conclude that both Cet1-Ceg1 and Spt4/5 act to promote the transition from initiation to elongation.

Arresting pol II in the ITC facilitates the conversion to the EC, but not the assembly of a follow-on PIC for re-initiation

Pol II in the ITC stalled at +27 is prone to extensive backtracking, but not arrested, as evidenced by RNA release upon gradient sedimentation (Fig. 4C and 5C-D). We thus sought a way to isolate the ITC with the G-less 27 template by inducing pol II arrest and thereby preventing RNA release. Previous studies of bacterial RNA polymerase (RNAP) demonstrated that nucleotide analogues incorporated at the 3'-terminus of RNA generally induce RNAP backtracking followed by stable arrest via destabilizing the 3'-proximal RNA-DNA hybrid (Shaevitz et al., 2003). By taking advantage of successive U residues clustered at +17, +18, +19, +20, +23, +24, +25, and +26 of the G-less 27 template, we screened UTP analogues and found that 4'-thio UTP can be incorporated by pol II without reducing initiation activity and that the resulting 4'-

thio RNAs in the post-initiation complexes are less sensitive to addition of TFIIS than RNAs without 4’-thio UTP (Fig. S13A-B). Using G-less 27 template DNA, we then stalled post-initiation complexes at +27, sedimented them on a glycerol gradient and analyzed 4’-thio RNAs by denaturing PAGE (Fig. S13C top). Whereas ~21-27-nt RNAs containing standard uridine were largely released from the ITC and were present on the top of the gradient (Fig. 5D), 4’-thio RNAs were near completely present in fractions corresponding to ECs (Fig. S13C top). No released transcripts were observed, as indicated by the absence of RNA on the top of the gradient (Fig. S13C top). Incorporation of 4’-thio UTP thus strongly prevented release of nascent transcripts during gradient sedimentation presumably by inducing pol II arrest before +17. Notably, almost no re-initiation complexes were observed with 4’-thio UTP (Fig. S13C top), although promoter escape occurred as shown by the presence of ECs. These data suggest that promoter escape is not necessarily followed by the assembly of a follow-on PIC even in the presence of excess pol II and TFIIF, and further highlight the critical roles of the capping enzyme and Spt4/5 in both promoter escape and the assembly of a follow-on PIC.

Given that 4’-thio UTP can prevent RNA release from the ITC, we attempted to isolate the ITC by stalling pol II upstream of +27. When pol II was stalled at +26 in the presence of 4’-thio UTP, we for the first time observed a major peak of RNA at the position where the PIC migrates (Fig. 6A-B fractions 9-10; Fig. S13C bottom) along with a peak for the EC (Fig. 6A-B fractions 14-16). All the GTFs, pol II and RNA are present in fraction 10 (Fig. 6B-C) indicating the presence of the ITC with 26 nt RNA. Taken together, these results definitively confirm that ITC can persist longer than previously expected.

2.6 Discussion

The transition from transcription initiation to elongation is a major rate-limiting step at many mRNA genes (Wade and Struhl, 2008), but key details of how the PIC transitions through the ITC to the EC have remained unclear. Here, we gained important insights into this transition by using an improved in vitro transcription system and yeast *SNR20* promoter DNA. Compared to our prior in vitro system (Murakami et al., 2013b; Murakami et al., 2015a), we were able to increase the transcription efficiency by ~3-4 fold. By then isolating and characterizing naturally generated post-initiation complexes, we found that the ITC, which contains pol II, GTFs, and a nascent transcript, often persists much longer than previously expected. In particular, we find that promoter escape and assembly of a follow-on PIC are facilitated by the capping enzyme and Spt4/5.

Previous crystal structures have shown that the RNA 5' end begins to clash with the N-terminal region of TFIIB (TFIIB$_N$) when the nascent RNA reaches a length of 9-15 nt (Bushnell et al., 2004; Kostrewa et al., 2009; Liu et al., 2010; Sainsbury et al., 2013). It has thus been thought that TFIIB and then other GTFs (TFIIA, TBP, TFIIE, and TFIIH) dissociate from the ITC while the upstream end of the initial bubble collapses (Čabart et al., 2011; Luse, 2013; Pal et al., 2005). In contrast to this model, our results indicate that all the GTFs, including TFIIB, can be associated with complexes transcribing 26 nt (Fig. 6A-C) or even 49 nt long RNAs (Fig. 3D-E). Nevertheless, the ITC is generally unstable and susceptible to long-range backtracking (Fig. 4E), and it undergoes the transition to the EC in an RNA-length dependent manner (Fig. 3-4). Human pol II complexes associated with short RNAs up to ~50 nt in length are also known to be prone to backtracking (Pal et al., 2001; Ujvari et al., 2002), suggesting an evolutionarily conserved feature of the early stages of eukaryotic transcription. There are important differences between human and yeast ITC, however. Backtracked human transcription complexes can be arrested and restart transcription with assistance from TFIIS (Pal et al., 2001). In contrast, we found that backtracking of yeast ITCs led to RNA release from the complex and termination of transcription unless pol II arrest is induced, e.g. by the use of 4'-thio UTP (Fig. 5C-D). It remains unclear why the RNA is released from the ITC only in the yeast system, but it may be due to fundamental differences in promoter architecture between human and yeast genes, for example, the spacing between the TATA box and TSS (Yang et al., 2007).

In this study, the ITC (Fig. 5C-D), but not the EC (Fig. S11), dissociated the RNA from the complex during glycerol gradient sedimentation. This difference allowed us to determine the timing of promoter escape as a function of position downstream of the TSS (Fig. 3-4). When transcription complexes were stalled at +49, glycerol gradient sedimentation revealed that a majority of the pol II escaped the promoter, as indicated by the presence of EC, EC+PIC, and EC+ITC (Fig. 3D-G). Interestingly, when pol II was stalled at +27, a larger proportion of transcription complexes transcribing 24-27 nt RNA escaped the promoter than those transcribing shorter (21-23 nt) RNAs (Fig. 4B-D). This indicates that, at least with *SNR20* promoter DNA we tested in this study, a major structural change may occur when the RNA length reaches ~23 nt. The timing of promoter escape we observed differs from the prevailing assumption, which was based on in vitro $KMnO_4$ footprinting (Holstege et al., 1997; Pal et al., 2005). We found that $KMnO_4$ footprinting was not sensitive enough to assess the timing of promoter escape (at least by bulk measurements; Fig. 2B-D) as $KMnO_4$ reactivity was observed mainly within the ~15 bp transcription bubble surrounding the pol II active center (Fig. 2D). Addition of $KMnO_4$, a strong oxidizing agent, may collapse the extended bubble in the ITC (Fazal et al., 2015b), whereas the extremely stable ~15 bp transcription bubble

(Gnatt et al., 2001) remains unwound. We instead found that glycerol gradient sedimentation gave clear indications of what factors are present in transcription complexes transcribing various lengths of RNA and thus a clearer view of when promoter escape occurs.

Upon addition of capping enzymes to the in vitro system, we noticed that the capping efficiency of ~27 nt transcripts was ~2-fold lower than that of ~49 nt transcripts (Fig. S4B-C). The differences in capping efficiency may be due to the fact ~79% of G-less 49 (Fig. 3E) transcription complexes escaped the promoter, compared to only ~43% of G-less 27 (Fig. 4C) complexes. This is in contrast to a previous study that used a reconstituted mammalian system and showed that the capping efficiency of short (23 nt) and long (223 nt) transcripts was indistinguishable (Noe Gonzalez et al., 2018). In our study, the capping enzyme was added at the same time as addition of NTPs (i.e. transcription initiation), whereas the capping enzyme was added in the previous study to stalled transcription complexes that had been washed with high salt. High salt should result in dissociation of some GTFs from the ITC, essentially forming ECs (Nilson et al., 2015; Noe Gonzalez et al., 2018) Thus, the differences in capping efficiency between the G-less 27 and G-less 49 transcription complexes may suggest that recruitment of the capping enzymes to the pol II surface is restricted due to the presence of GTFs. Nilson et al. indeed observed that the capping efficiency is much lower on human transcription complexes containing 21 nt RNAs after a low salt wash (which should not remove GTFs) compared to after a high salt wash (Nilson et al., 2015). Alternatively, backtracking of the ITC (Fig. 4E) may cause the 5' ends of transcripts to be pulled back inside the RNA exit tunnel of pol II, limiting the access of the capping enzyme on RNAs and resulting in less efficiently capped ~27 nt RNAs than ~49 nt RNAs.

Our results further indicate that the conversion of the ITC to EC is assisted by

Cet1-Ceg1 and Spt4/5 (Fig. 5 and Fig. S12). In light of recent studies that mapped binding sites for Cet1-Ceg1 immediately adjacent to Rpb4/7 of pol II, which overlap with those for TFIIE in the PIC (perhaps also in ITC) (Martinez-Rucobo et al., 2015) (Noe Gonzalez et al., 2018), we propose that the recruitment of Cet1-Ceg1 to pol II with an emerging nascent RNA may lead to ejection of TFIIE from the ITC and promoter escape. Similarly, interactions between Spt4/5 and an emerging nascent RNA (Bernecky et al., 2017; Ehara et al., 2017) in the ITC may likewise aid promoter escape. In addition to Cet1-Ceg1 and Spt4/5, some stress-responsive transcription factors that stimulate transcription restart from pol II backtracking in early transcription may also play a role in the transition from initiation to elongation (Damodaren et al., 2017). Moreover, the transition could be regulated by the +1 nucleosome (Nagai et al., 2017). It will be of considerable interest to explore in the future how a variety of transcription factors and chromatin factors positively or negatively regulate the transition from transcription initiation to elongation.

Finally, our improved system represents a significant technical advance that will be highly useful for biochemical and structural studies of transcription initiation as well as the transition to elongation. By taking advantage of this highly efficient transcription initiation system, structural determination of the naturally formed ITC can be now pursued to provide further molecular insights into the mechanism of promoter escape.

2.7 Figures

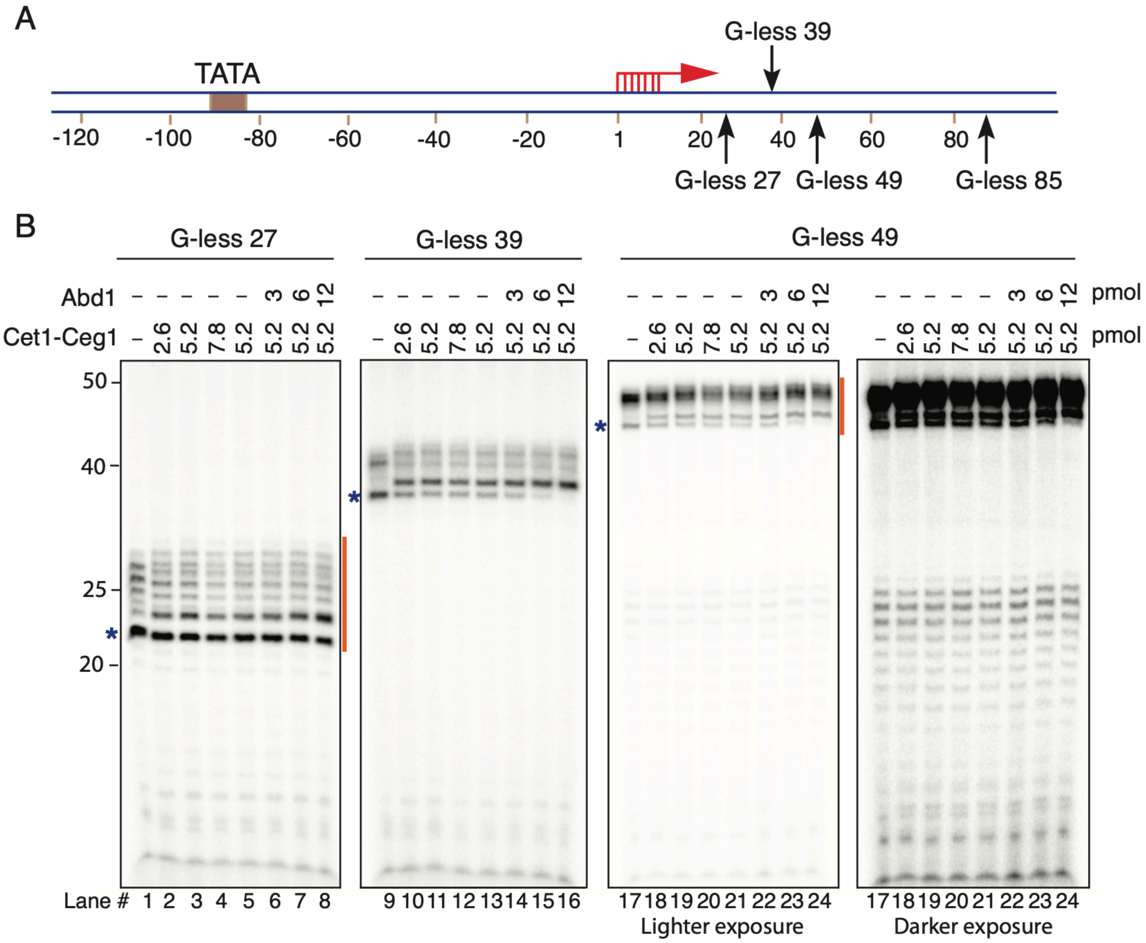


Figure 1. Transcription initiation assay with a series of G-less *SNR20* promoter variants. (A) Schematic diagram of the *SNR20* promoter variants. The transcription start sites (red arrows) and G-stops (black arrows) are indicated. **(B)** PIC was formed on the indicated *SNR20* promoter DNA variants with TFIIA, TFIIB, TBP, TFIIE, TFIIF, TFIIH, Sub1, and pol II. Transcription was initiated by addition of ATP, CTP, 3'-O-methyl GTP, UTP, [α-^{32}P]UTP and then incubated for 20 min at 30°C. Varying concentrations of Cet1-Ceg1 and Abd1 were added at the time of addition of NTPs. Orange lines indicate the sets of transcripts that were used to calculate transcription efficiency in Table 1. The bands indicated by blue asterisks are used for quantification of the shift in RNA mobility in Fig. S3A.

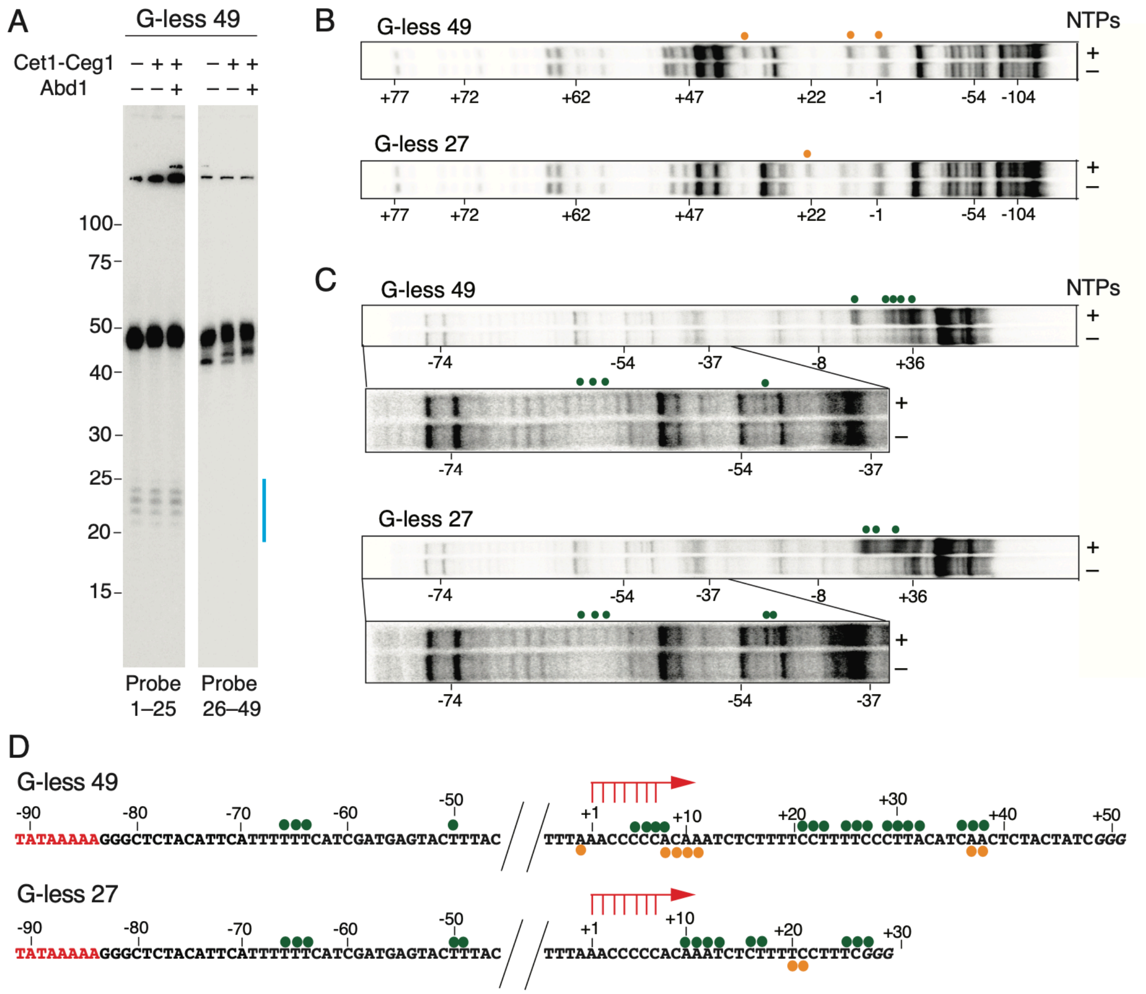


Figure 2. Reconstituted transcription initiation system supports 5' capping and re-initiation. (A) To confirm re-initiation from the G-less 49 promoter, Northern blot analysis was performed on G-less 49 transcripts that were generated in the presence or absence of Cet1-Ceg1 and Abd1. Probes antisense to nt 1-25 or nt 26-49 were used. Transcripts generated by the second round of transcription are indicated by a blue line. **(B-C)** Potassium permanganate ($KMnO_4$) footprinting assays with the G-less 49 (upper) and the G-less 27 (lower) templates (-122/+97) to detect single-stranded regions. The 5' ends of the template (B) or non-template (C) strands were labeled with ^{32}P. After incubation of the initiation reactions for 20 min, 18 mM $KMnO_4$ was added. In C, a ~50 bp region downstream of the TATA box is enlarged and shown with darker exposure (lower panels). $KMnO_4$ reactive positions on template and non-template strands are indicated by orange and green dots, respectively. **(D)** G-less 49 (top) and G-less 27 (bottom) DNA sequences showing $KMnO_4$ reactive residues. The TATA box is shown in red and transcription start sites are indicated by red arrows.

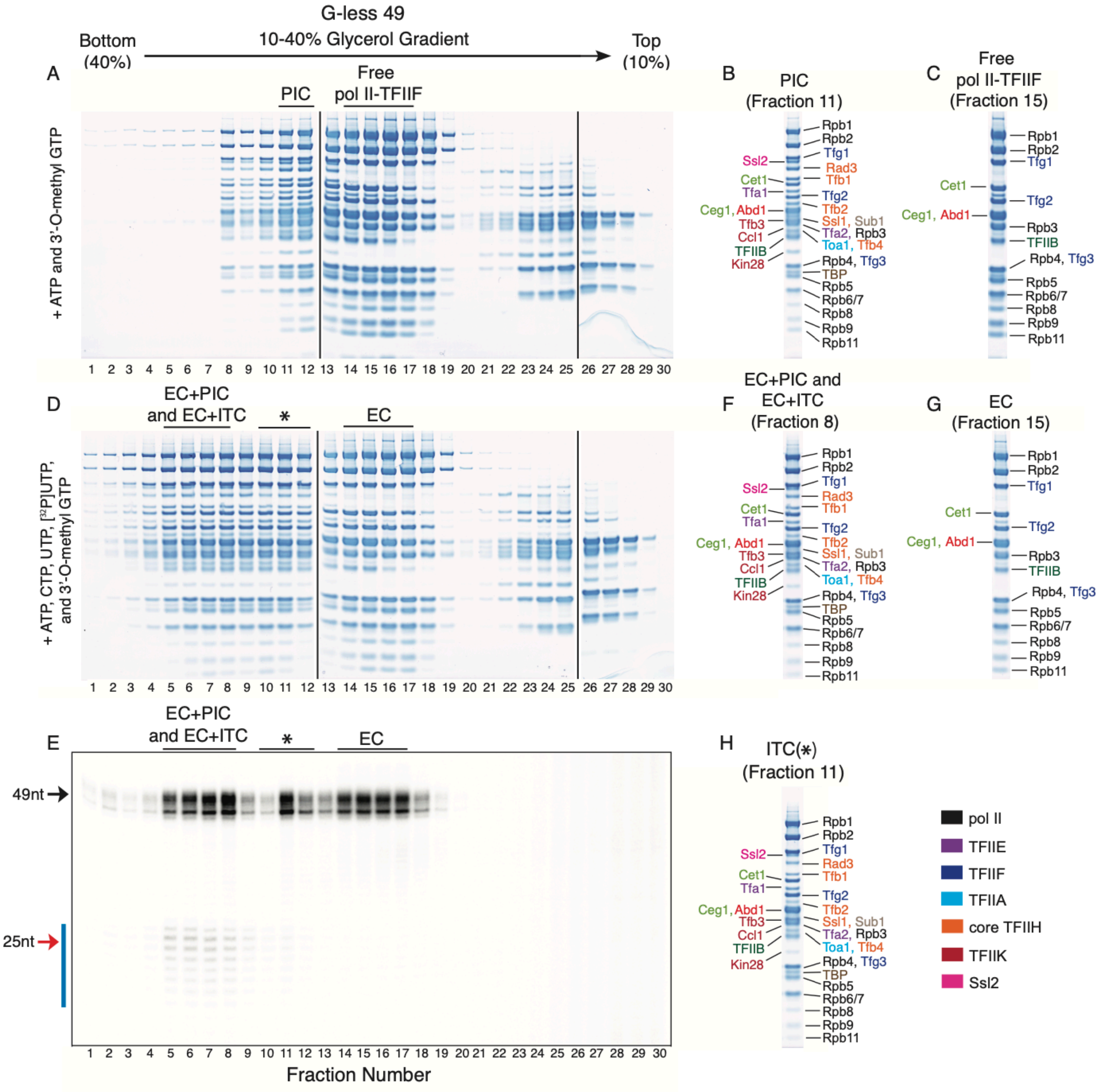


Figure 3. Separation of the post-initiation complexes stalled at +49. (A) PIC was assembled with 4-fold excess pol II and TFIIF relative to G-less 49 template DNA and then combined with ATP, 3'-O-methyl GTP, 4-fold excess Cet1-Ceg1, and 8-fold excess Abd1 (relative to the DNA). Reactions were incubated for 20 min at 30°C, combined with 2 mM AMP-PNP, and sedimented on a 10-40% glycerol gradient. ~130 μL per fraction were isolated and then analyzed by SDS-PAGE. **(B-C)** Protein identification of the PIC in fraction 11 (B) and free pol II-TFIIF in fraction 15 (C). **(D)** PIC was assembled in the same manner as in A. Transcription was initiated by addition of ATP, CTP, 3'-O-methyl GTP, UTP, [α-^{32}P]UTP, Cet1-Ceg1 and Abd1. The post-initiation complexes were combined with 2 mM AMP-PNP, sedimented on a gradient, and analyzed as in A. **(E)** RNA analysis of fractions isolated in D by denaturing Urea-PAGE. The transcripts (~25 nt) from the second round of initiation are indicated by a blue line. **(F-H)** Protein identification of the EC+PIC and EC+ITC in fraction 8 (F), EC in fraction 15 (G), and ITC in fraction 11 (H). Asterisk in D, E, and H indicates the presence of the ITC+PIC.

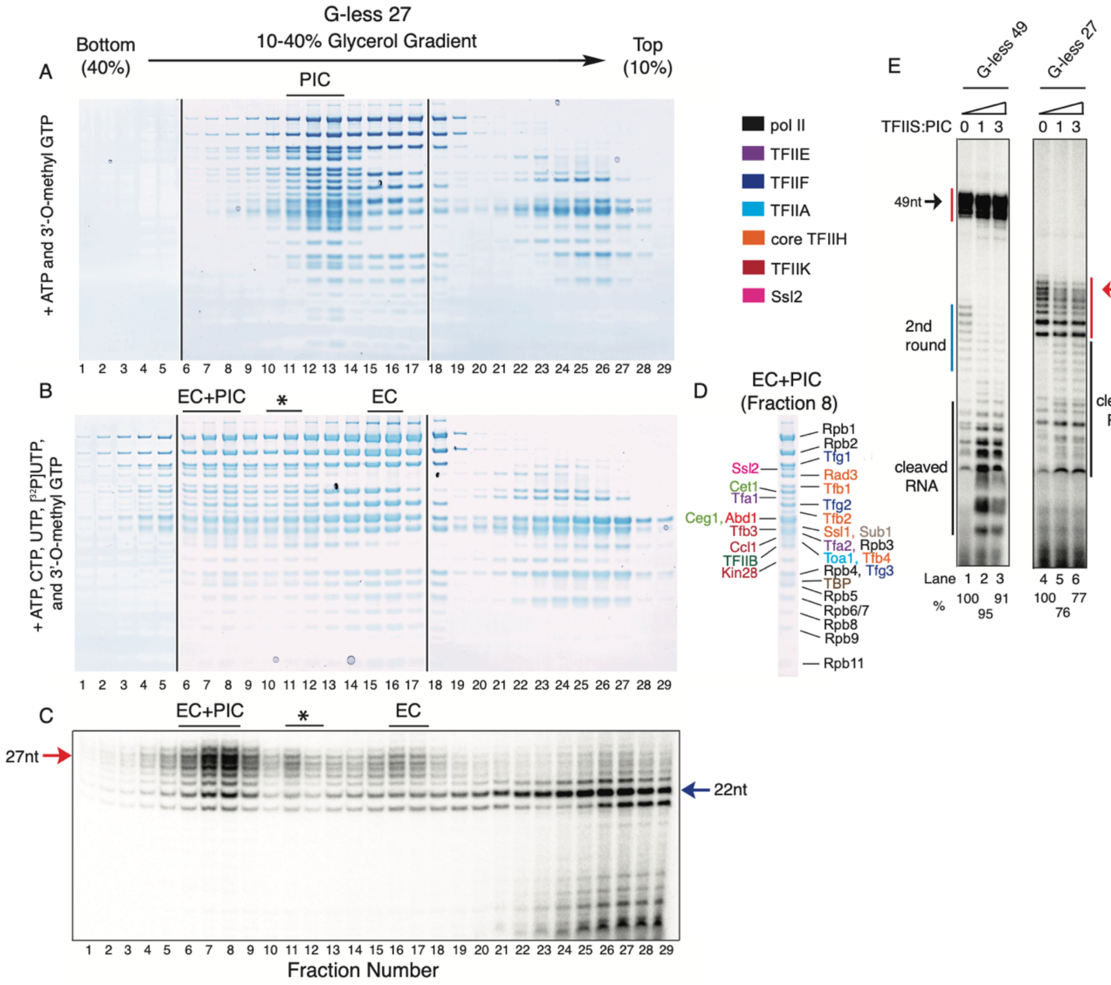


Figure 4. Separation of the post-initiation complexes stalled at +27 that are prone to extensive backtracking. (A-B) Separation and SDS-PAGE analysis of PICs (A) and post-initiation complexes (B) with G-less 27 DNA template were performed in the same manner as in Fig. 3 except that the control PIC in A was assembled with only 1.4-fold excess pol II and TFIIF relative to G-less 27 DNA template DNA, and received 1.5-fold excess Cet1-Ceg1 and Abd1 (relative to the DNA). **(C)** RNA analysis of fractions from B by denaturing Urea-PAGE. 27-nt and 22-nt transcripts are indicated by red and blue arrows, respectively. Asterisk in B and C indicates the presence of the ITC+PIC. **(D)** Protein identification of EC+PIC in fraction 8. **(E)** TFIIS cleavage assay of complexes stalled at promoter proximal positions. Transcription initiation assays with G-less 49 (left) or G-less 27 (right) DNA templates were performed as described in Fig. 1. After 20 min, reactions were combined with the indicated concentrations of TFIIS, incubated for another 6 min, and cleaved transcripts were analyzed by denaturing PAGE. Note that the cleaved RNAs indicated by black lines were not used for determining how far pol II backtracks as many of these RNAs are 3' fragments that were generated by partial backtracking followed by TFIIS-induced cleavage and subsequent resumption of transcription. Thus full-length transcripts indicated by red lines were quantified and the amount compared to the control (0 pmol TFIIS) is shown below the gel.

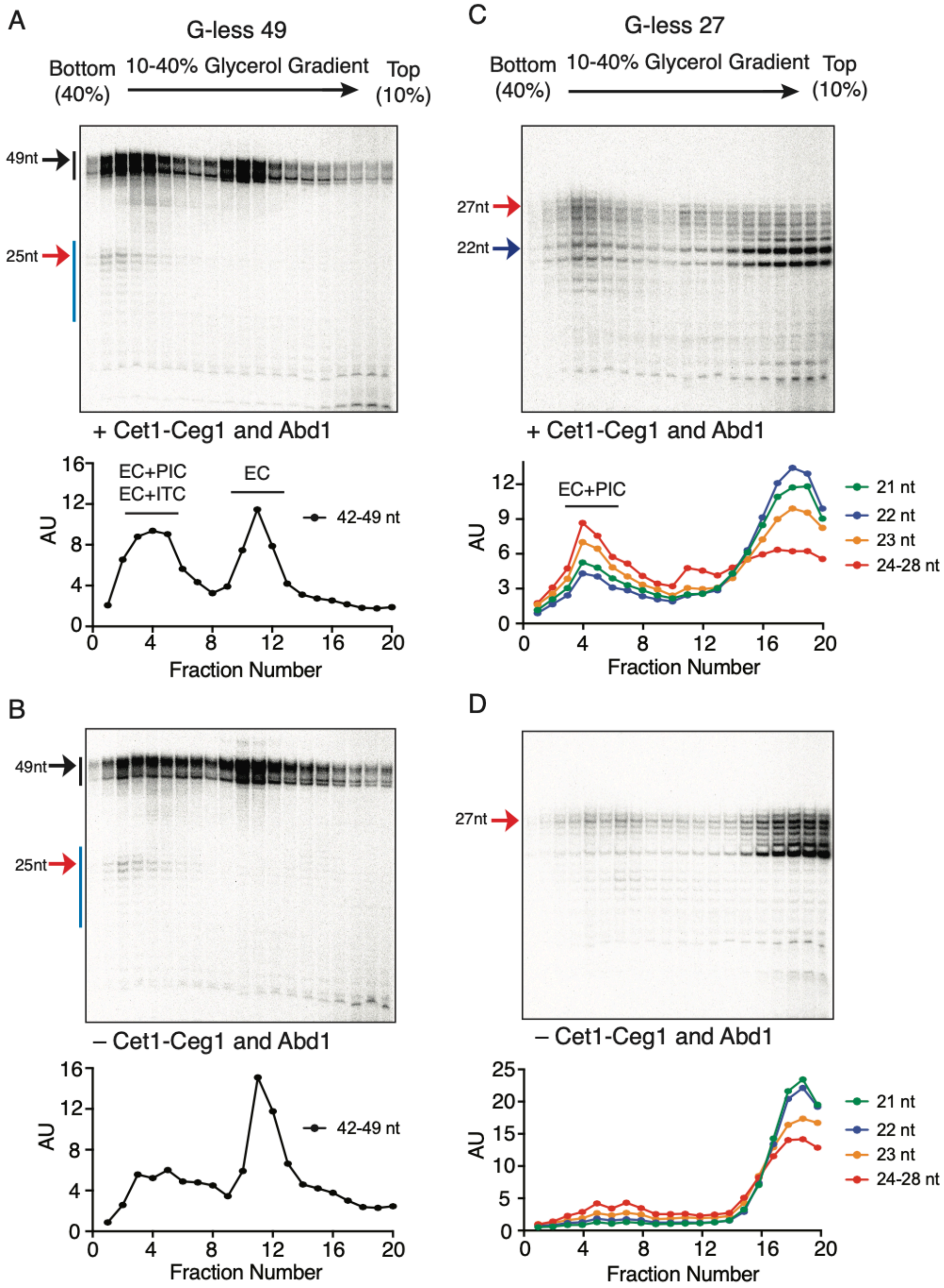


Figure 5. Cet1-Ceg1 facilitates promoter escape and assembly of a follow-on PIC for re-initiation. Transcription was initiated with the G-less 49 (A, B) or G-less 27 (C, D) template DNA in the presence (A, C) or absence (B, D) of Cet1-Ceg1 and Abd1. All reactions received 2 mM AMP-PNP prior to 10-40% glycerol gradient sedimentation. The gradients were fractionated and ~180 µL per fraction was isolated prior to analyzing RNA by denaturing PAGE. The levels of RNA in each fraction was quantified by ImageJ, normalized to fraction 1, and the relative RNA levels were plotted.

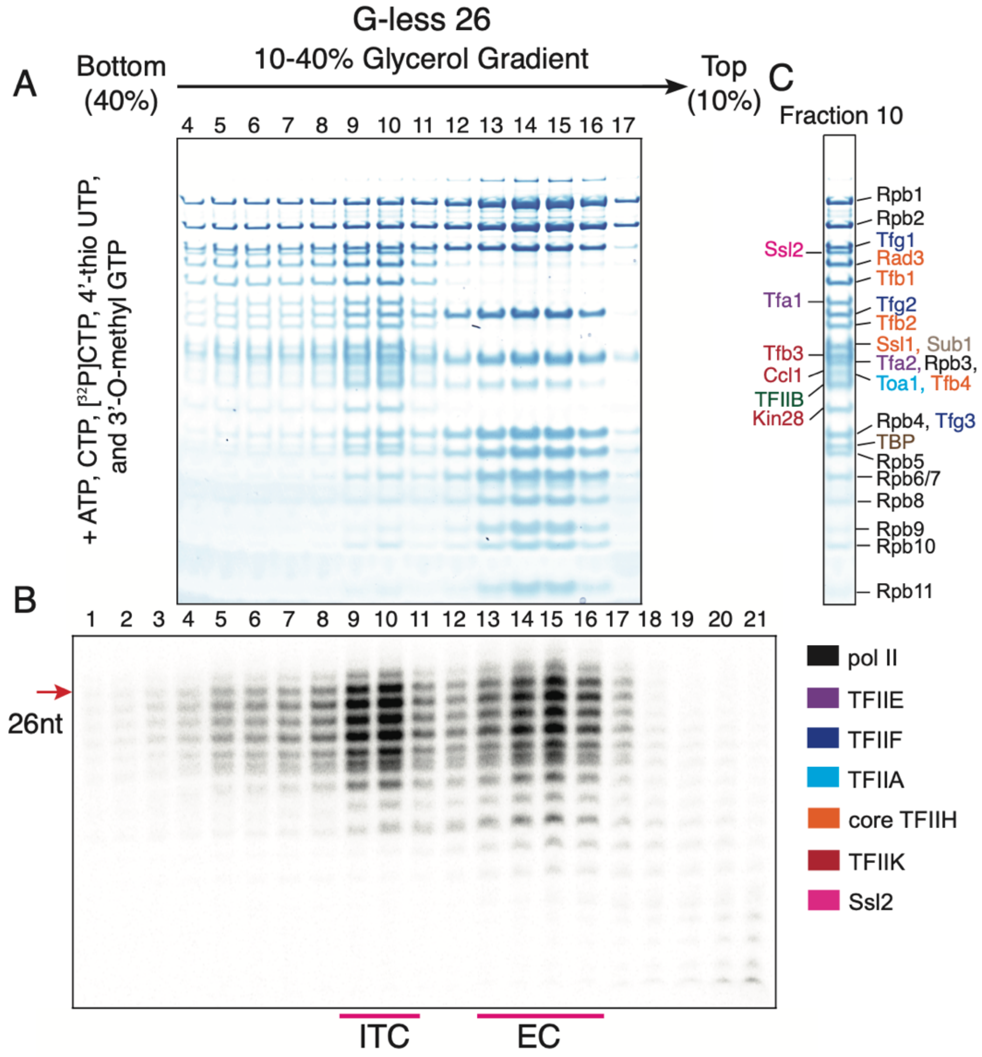


Figure 6. 4'-Thio UTP suppresses RNA release and allows transcripts to be trapped in the ITC. PIC assembly and initiation reactions were performed in the same manner as in Fig. 3D-E except that pol II was stalled at +26 using G-less 26 template DNA, transcription was initiated in the presence of [^{32}P]CTP and 630 µM 4'-thio UTP instead of [^{32}P]UTP and 500 µM UTP, and the concentration of CTP in the reaction was lowered to 630 µM from 800 µM. Centrifugation was performed in the same manner as before. The gradient was fractionated and ~130 uL per fraction was isolated. **(A)** Proteins in fraction 4-17 were analyzed by SDS-PAGE. **(B)** RNA in fractions 1-21 was analyzed by denaturing PAGE. **(C)** Protein identification of fraction 10.

2.8 Main Table

Table 1. Transcription efficiency (efficiency of template usage) with *SNR20* promoter. Transcription was initiated in the presence of Sub1 (3 pmol) and 4-fold molar excess pol II and TFIIF (5.2 pmol) relative to DNA (1.3 pmol). The efficiency of template usage, defined as the percentage of templates that were transcribed, was calculated based on the incorporation of radiolabeled UTP (see Methods). G-less 27, n = 3; G-less 49, n = 5.

DNA template	Efficiency from 1st round of initiation (%) ± S.E.M	Efficiency from 2nd round of initiation (%) ± S.E.M
G-less 27	98.9 ± 6.6	NA
G-less 49	90.8 ± 10.0	28.4 ± 1.4

2.9 Supplemental Figures

Figure S1

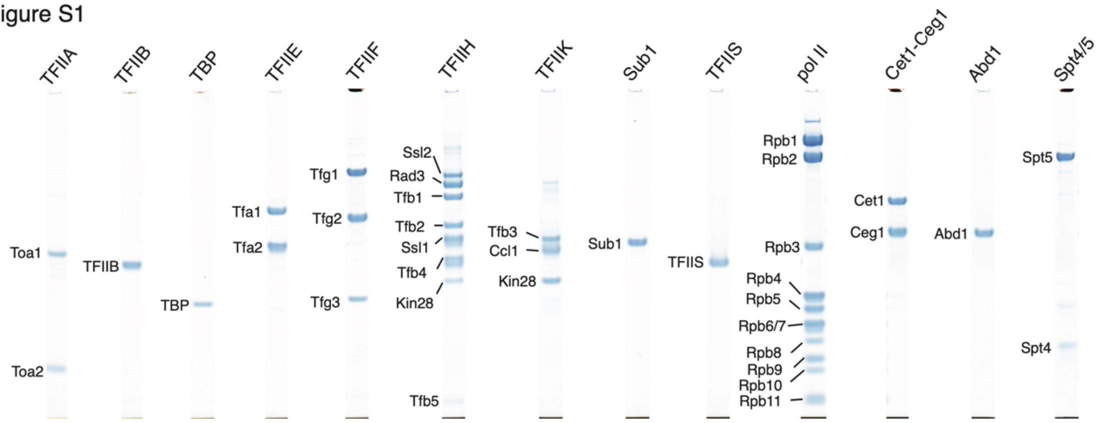


Fig. S1. Purified proteins used in transcription initiation assays. TFIIA, TFIIB, TBP, Sub1, TFIIS, Cet1-Ceg1, Abd1, and Spt4/5 were recombinantly expressed and purified. TFIIE, TFIIF, TFIIH, TFIIK, and pol II were purified from yeast.

Figure S2

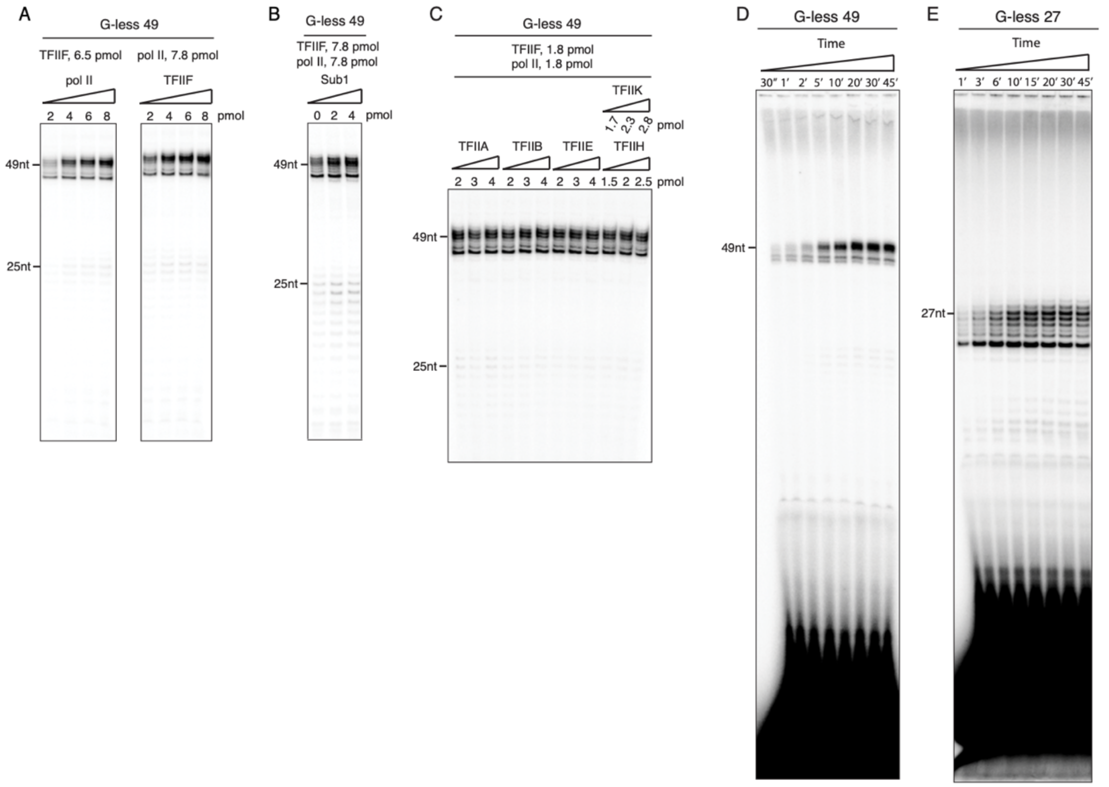


Fig. S2. Optimization of transcription initiation assays. (A-C) Varying concentrations of pol II and TFIIF (A), Sub1 (B), and GTFs (TFIIA, TFIIB, TFIIE, and TFIIH + TFIIK) (C) were added for PIC assembly. Transcription initiation reactions were performed as described in Fig. 1. (D-E) Time course of transcription initiation reaction with G-less 49 (D) and G-less 27 (E) templates.

Figure S3

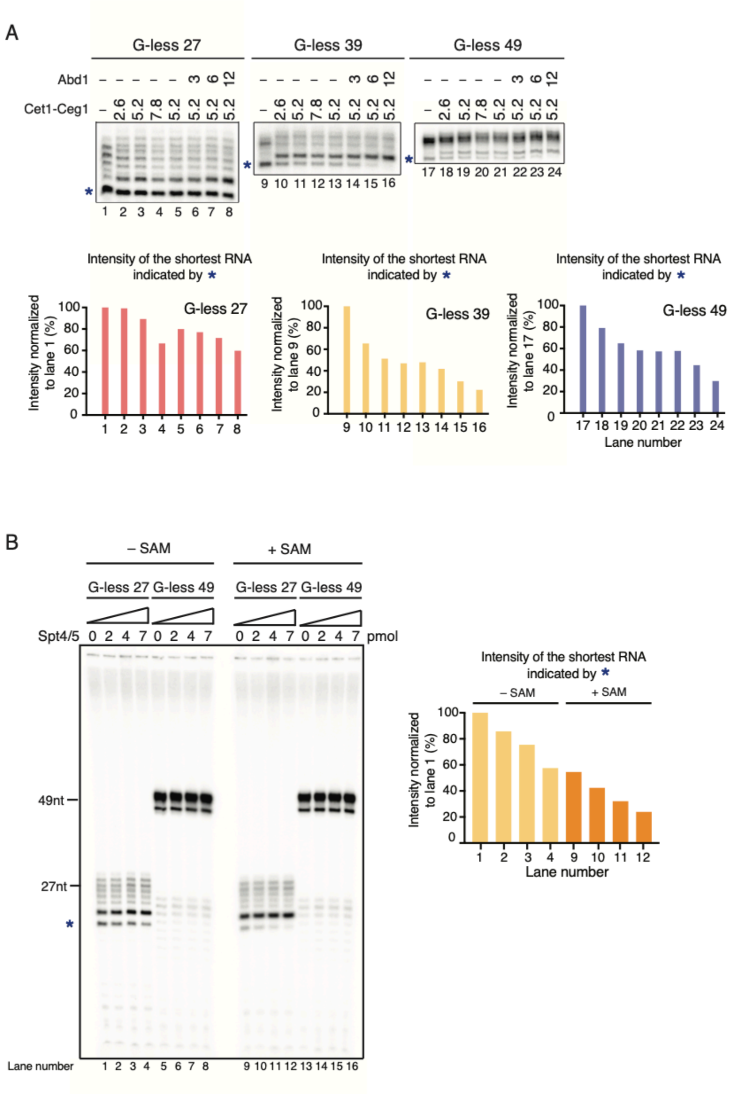


Fig. S3. 1 nt shift in RNA mobility was observed upon addition of Cet1-Ceg1, and the shift was enhanced by the presence of Abd1, Spt4/5, or SAM. (A) Image J was used to quantify the intensities of the shortest transcript generated in the first round of transcription of each G-less template (indicated by blue asterisk) in Fig. 1B. Note that the gels shown in A are the same gels shown in Fig. 1B. The intensities were normalized, and plotted in the graphs below the gels. Loss of intensity of the shortest transcript indicates enhanced 1 nt shift. (B) Varying concentrations of Spt4/5 were added to the transcription initiation reactions at the time of NTP addition. These reactions also contained 5.2 pmol of Cet1-Ceg1 and 10.4 pmol of Abd1. Lanes 1-8 were performed in the absence of SAM; lanes 9-16 were performed in the presence of 100 µM SAM. The intensities of bands corresponding to the shortest G-less 27 transcript (indicated by blue asterisk) were quantified, normalized to control (0 pmol Spt4/5, Lane 1), and plotted in the graph.

Figure S4

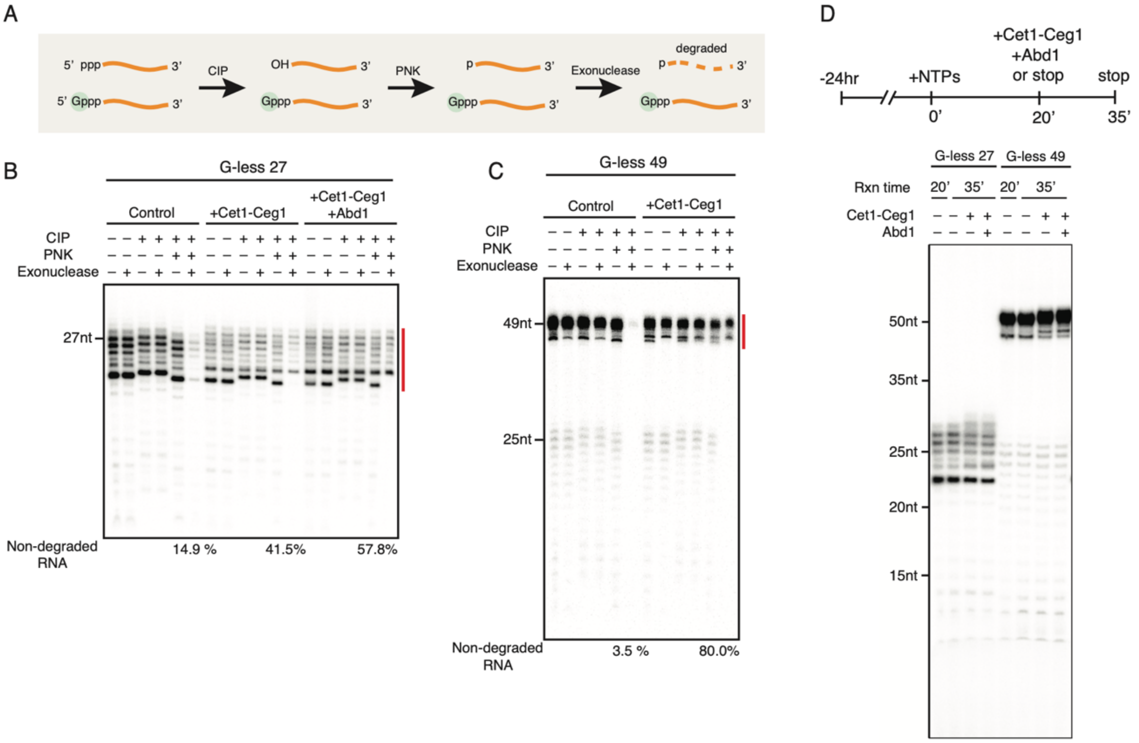


Fig. S4. The observed ~1 nt shift in transcript mobility is due to Cet1-Ceg1 guanylylating the 5' end of RNA, not a shift in TSS. (A) Schematic of exonuclease assay to determine the 5' modification status of transcribed RNAs. (B-C) After transcription initiation reactions with G-less 27 (B) and G-less 49 (C) template DNAs were performed in the presence or absence of capping enzymes, RNA was isolated and treated with CIP to dephosphorylate triphosphates followed by PNK to mono-phosphorylate the 5' end of the uncapped RNA. Only uncapped RNA, but not 5'-capped RNA (guanylylated RNA), was converted to the mono-phosphorylated form by CIP and PNK treatments, and subsequently degraded by the 5'-3' exonuclease. The percentage of non-degraded RNA is indicated. Red lines indicate the transcripts that were used to calculate the % of non-degraded RNA. (D) Transcription was initiated at time 0. After 20 min, 5.2 pmol of Cet1-Ceg1 and 10.4 pmol of Abd1 were added and reactions was subsequently stopped (time = 35 min). Transcripts were then analyzed by denaturing Urea-PAGE.

Figure S5

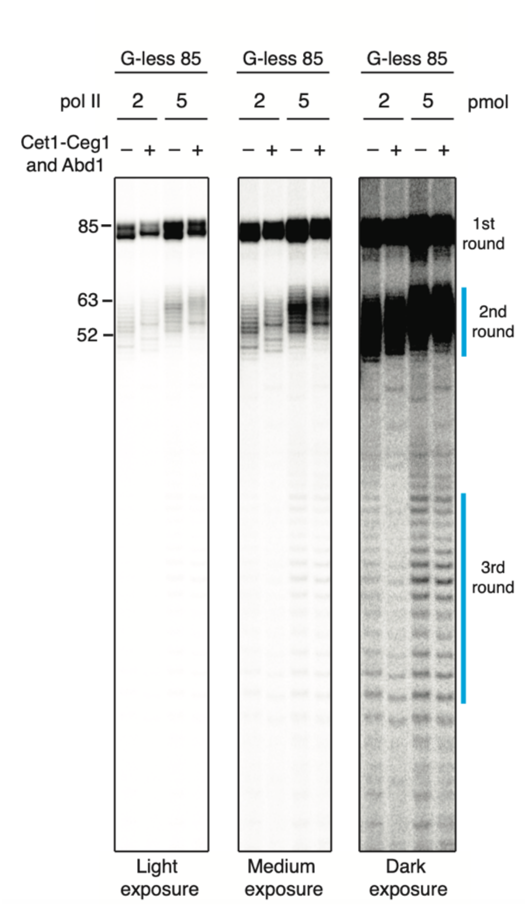


Fig. S5. RNA signals are ~25 nt apart, supporting pol II stacking as a result of re-initiation. PIC was assembled on G-less 85 DNA template with varying concentrations of pol II. As indicated, the reactions were performed in the presence or absence of Cet1-Ceg1 and Abd1. Transcripts were then analyzed by denaturing Urea-PAGE. Transcripts generated by 2nd and 3rd rounds of initiation are indicated by blue lines.

Figure S6

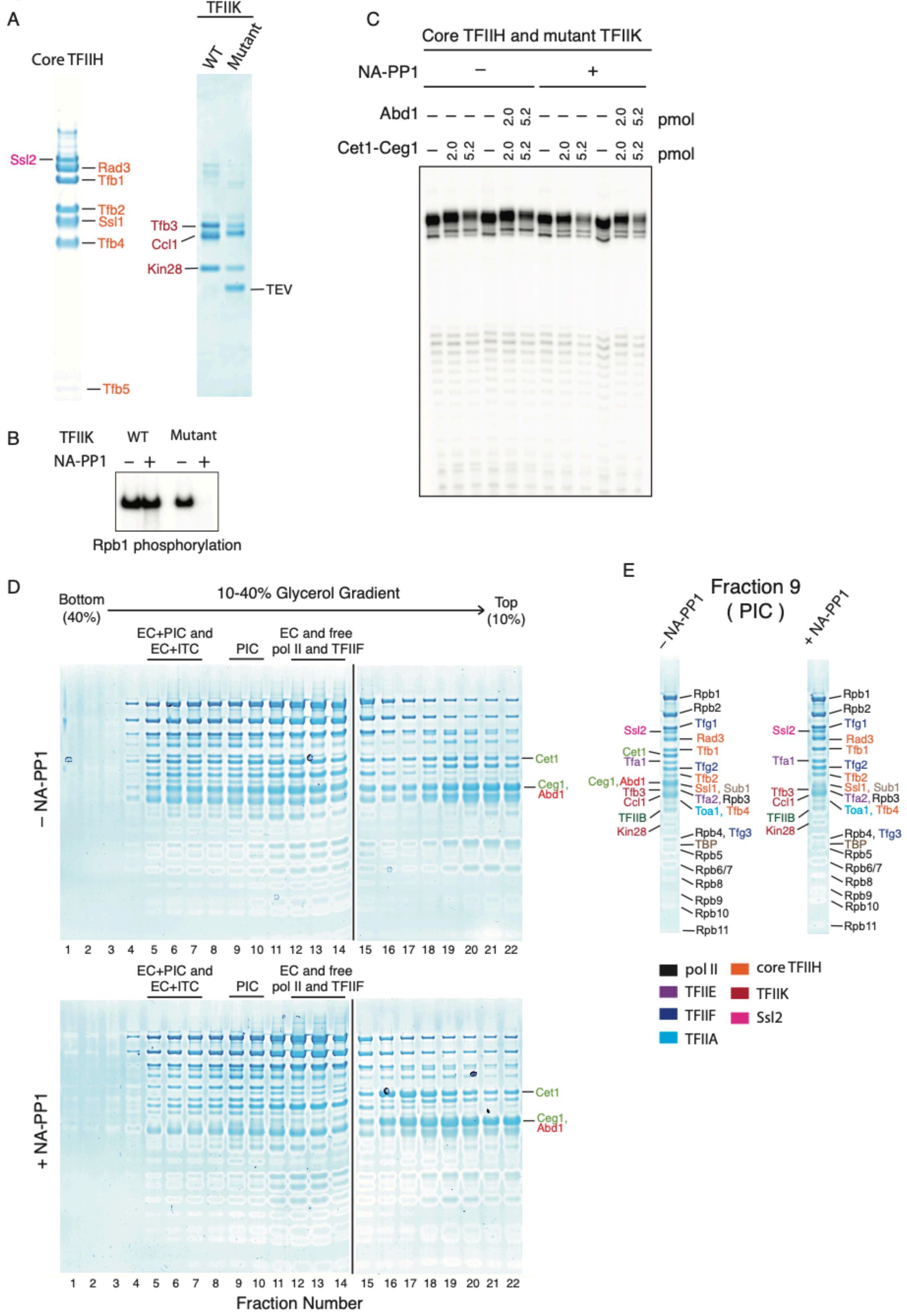
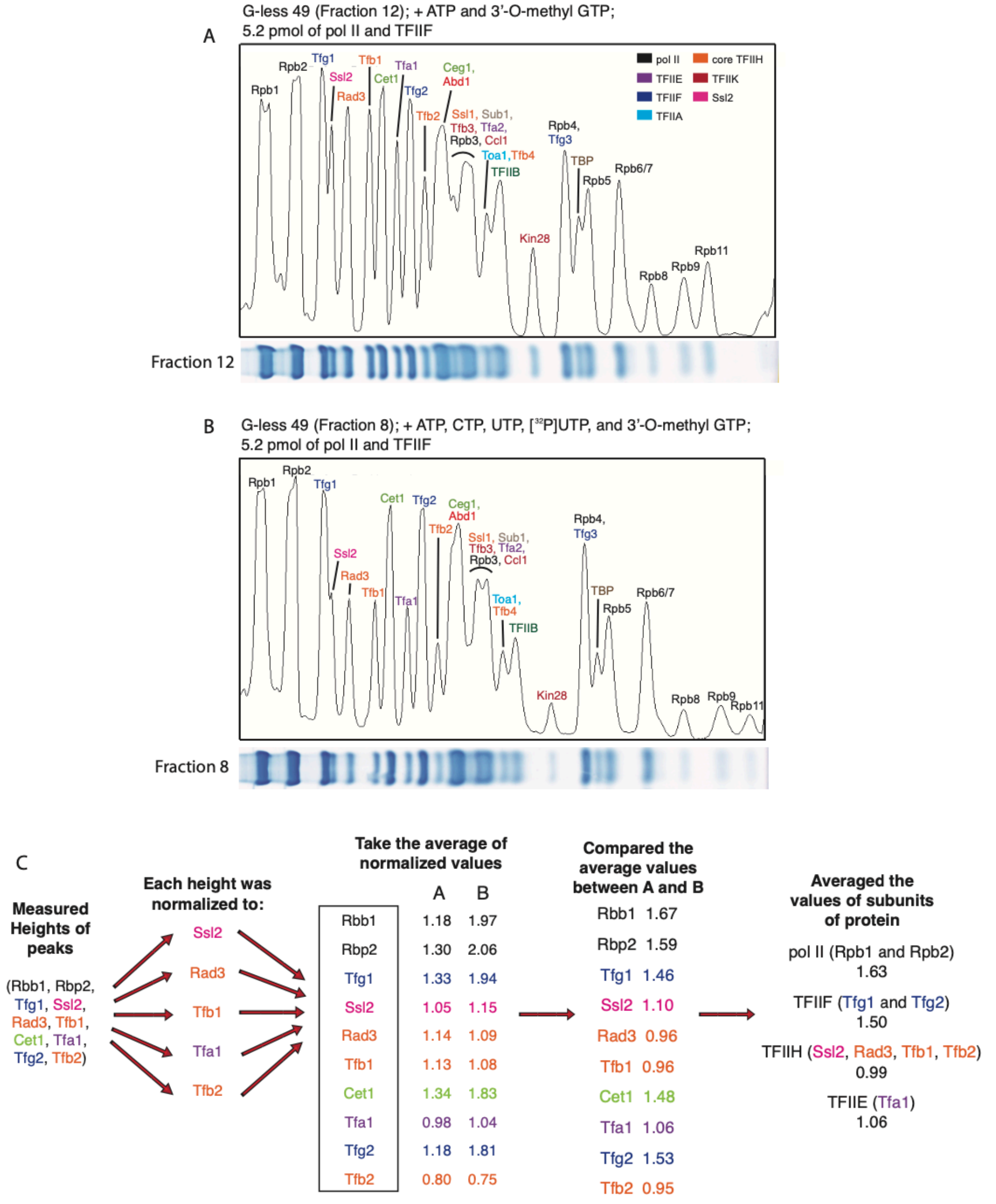

A
Core TFIIH
Ssl2
Rad3
Tfb1
Tfb2
Ssl1
Tfb4
Tfb5
TFIIK
WT
Mutant
Tfb3
Ccl1
Kin28
TEV
B
TFIIK
WT
Mutant
NA-PP1
– + – +
Rpb1 phosphorylation
C
Core TFIIH and mutant TFIIK
NA-PP1
–
+
Abd1 – – – – 2.0 5.2 – – – – 2.0 5.2 pmol
Cet1-Ceg1 – 2.0 5.2 – 2.0 5.2 – 2.0 5.2 – 2.0 5.2 pmol
D
Bottom (40%)
10-40% Glycerol Gradient
Top (10%)
EC+PIC and EC+ITC
PIC
EC and free pol II and TFIIF
– NA-PP1
+ NA-PP1
Cet1
Ceg1, Abd1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Fraction Number
E
Fraction 9 (PIC)
– NA-PP1
+ NA-PP1
Rpb1
Rpb2
Tfg1
Ssl2
Rad3
Tfb1
Cet1
Tfa1
Tfg2
Tfb2
Ceg1, Abd1
Ssl1, Sub1
Tfb3
Ccl1
Tfa2, Rpb3
Toa1, Tfb4
TFIIB
Kin28
Rpb4, Tfg3
TBP
Rpb5
Rpb6/7
Rpb8
Rpb9
Rpb10
Rpb11
pol II
core TFIIH
TFIIE
TFIIK
TFIIF
Ssl2
TFIIA

Fig. S6. Phosphorylation of the CTD of pol II is sufficient for Cet1-Ceg1 binding in vitro. (A) Purification of core TFIIH (7 subunits), wild type TFIIK, and inhibitor sensitive TFIIK. The inhibitor sensitive TFIIK has a mutation in Kin28 (L83G) (Murakami et al., 2015a). **(B)** PICs were assembled on the G-less 49 template using core TFIIH and wild type or mutant TFIIK. Transcription reactions were initiated as described in the methods section except that they were initiated in the presence of [γ-^{32}P] ATP and 750 µM NA-PP1, an inhibitor of TFIIK kinase activity. The control reactions (–NA-PP1) received an equivalent volume of DMSO. Phosphorylation of Rpb1 was analyzed by SDS-PAGE followed by radioautography. **(C)** PICs were assembled on the G-less 49 template in the same manner as in B and the reactions were performed in the absence or presence of NA-PP1. Varying concentrations of Cet1-Ceg1 and Abd1 were added to the reactions at the time of transcription initiation, as indicated. **(D)** Transcription initiation reactions were performed in the presence of 4-fold excess Cet1-Ceg1 and 8-fold excess Abd1 (relative to DNA). G-less 49 post-initiation complexes that had been generated in the absence (top) or presence (bottom) of NA-PP1 received 2 mM AMP-PNP and then were subjected to glycerol gradient sedimentation. The gradients were fractionated (~180 µL per fraction), and proteins were analyzed by SDS-PAGE. **(E)** Protein identification of PIC in fractions 9 (left –NA-PP1; right +NA-PP1) indicated a lack of Cet1-Ceg1 binding to pol II in the presence of NA-PP1.

Figure S7



Measured Heights of peaks

(Rbb1, Rbp2, Tfg1, Ssl2, Rad3, Tfb1, Cet1, Tfa1, Tfg2, Tfb2)

Each height was normalized to:

Ssl2
Rad3
Tfb1
Tfa1
Tfb2

Take the average of normalized values

	A	B
Rbb1	1.18	1.97
Rbp2	1.30	2.06
Tfg1	1.33	1.94
Ssl2	1.05	1.15
Rad3	1.14	1.09
Tfb1	1.13	1.08
Cet1	1.34	1.83
Tfa1	0.98	1.04
Tfg2	1.18	1.81
Tfb2	0.80	0.75

Compared the average values between A and B

Rbb1	1.67
Rbp2	1.59
Tfg1	1.46
Ssl2	1.10
Rad3	0.96
Tfb1	0.96
Cet1	1.48
Tfa1	1.06
Tfg2	1.53
Tfb2	0.95

Averaged the values of subunits of protein

pol II (Rpb1 and Rpb2)
1.63

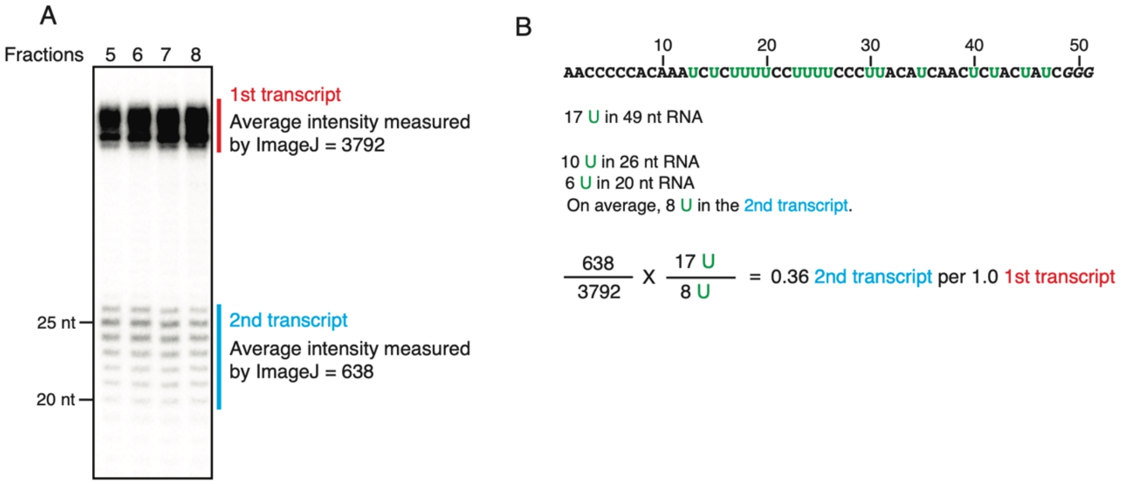
TFIIF (Tfg1 and Tfg2)
1.50

TFIIH (Ssl2, Rad3, Tfb1, Tfb2)
0.99

TFIIE (Tfa1)
1.06

Fig. S7. Protein stoichiometry of the G-less 49 re-initiation complex. (A-B) To determine the stoichiometry of pol II, TFIIF, TFIIE, and TFIIH of the G-less 49 re-initiation complex in fraction 8 in Fig 3D, SDS-PAGE of fraction 12 in Fig. 3A (PIC, as a control) and fraction 8 in Fig. 3D was scanned and plotted using ImageJ. **(C)** The amounts of Rpb1, Rpb2, Tfg1, Ssl2, Rad3, Tfb1, Cet1, Tfa1, Tfg2, and Tfb2 in the PIC and re-initiation complex were determined by measuring the heights of the peaks and then normalizing to TFIIE and TFIIH (Ssl2, Rad3, Tfb1, Tfa1, and Tfb2), based on the fact that one molecule each of TFIIE and TFIIH is present in one PIC. Stoichiometry of the remaining subunits (Rpb1, Rpb2, Tfg1, Cet1, Tfg2) in the re-initiation complex was then determined by comparing the normalized values of each respective subunit in the re-initiation complex with those in the PIC control.

Figure S8


A
Fractions 5 6 7 8
1st transcript
Average intensity measured by ImageJ = 3792
25 nt
20 nt
2nd transcript
Average intensity measured by ImageJ = 638
B
10
20
30
40
50
AACCCCCACAAAUCUCUUUUCCUUUUCCCUUACAUCAACUCUACUAUCGGG

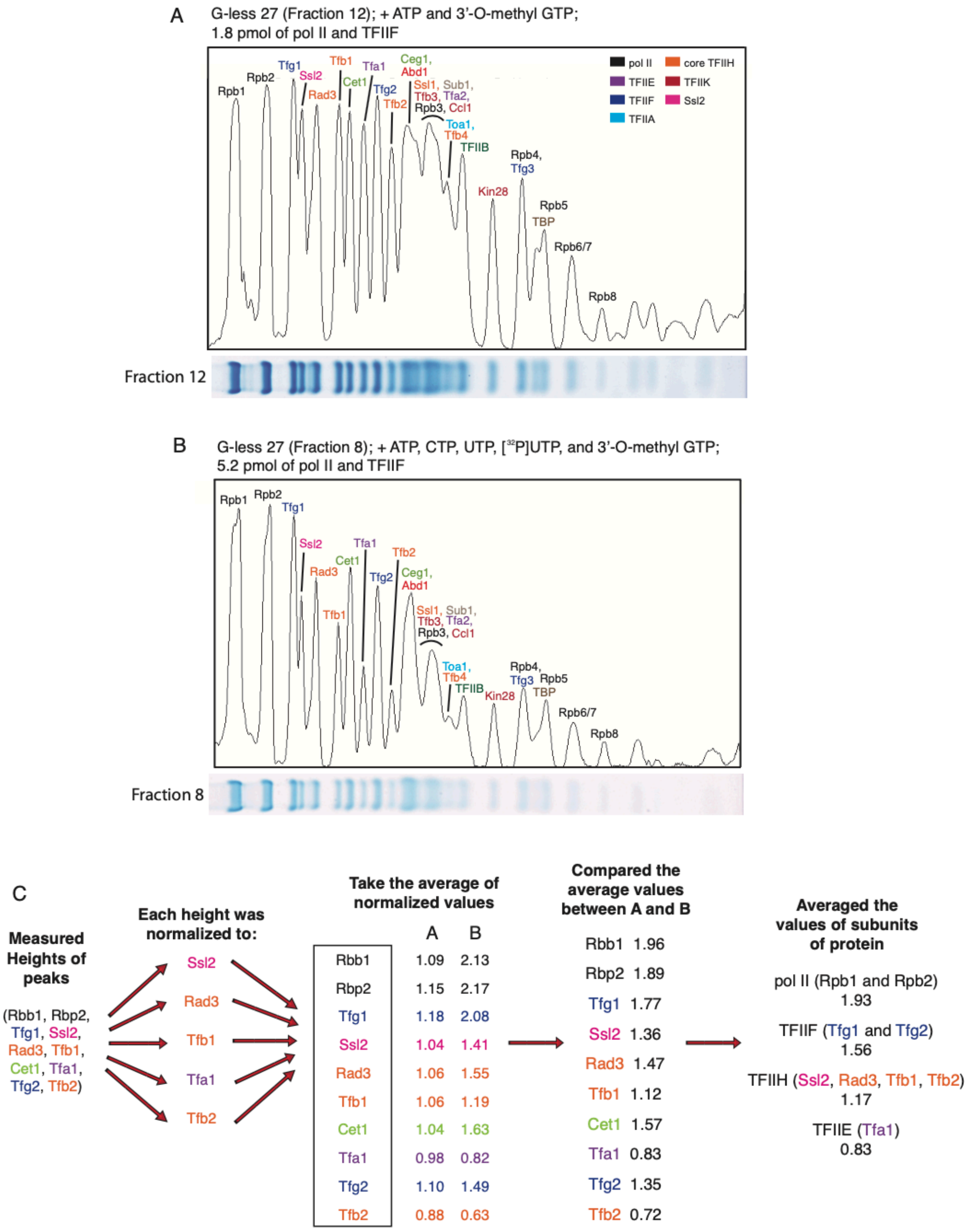
17 U in 49 nt RNA

10 U in 26 nt RNA
6 U in 20 nt RNA
On average, 8 U in the 2nd transcript.

$$\frac{638}{3792} \times \frac{17\ \mathrm{U}}{8\ \mathrm{U}} = 0.36\ \text{2nd transcript per 1.0 1st transcript}$$

Fig. S8. Quantification of transcripts generated from 1st and 2nd rounds in the G-less 49 re-initiation complex reveals presence of EC+PIC and EC+ITC. (A) To determine the ratio of EC+PIC and EC+ITC in fraction 5-8 in Fig. 3E, intensities of the 1st and 2nd transcripts indicated by red and blue lines, respectively, were measured using ImageJ. The average of the intensity values is indicated. **(B)** Sequence of the G-less 49 transcript (top) and calculation for determining the number of 2nd transcripts per 1st transcript (bottom) are shown.

Figure S9


A G-less 27 (Fraction 12); + ATP and 3'-O-methyl GTP; 1.8 pmol of pol II and TFIIF

B G-less 27 (Fraction 8); + ATP, CTP, UTP, [^{32}P]UTP, and 3'-O-methyl GTP; 5.2 pmol of pol II and TFIIF

C

Measured Heights of peaks

(Rbb1, Rbp2, Tfg1, Ssl2, Rad3, Tfb1, Cet1, Tfa1, Tfg2, Tfb2)

Each height was normalized to:

Ssl2
Rad3
Tfb1
Tfa1
Tfb2

Take the average of normalized values

	A	B
Rbb1	1.09	2.13
Rbp2	1.15	2.17
Tfg1	1.18	2.08
Ssl2	1.04	1.41
Rad3	1.06	1.55
Tfb1	1.06	1.19
Cet1	1.04	1.63
Tfa1	0.98	0.82
Tfg2	1.10	1.49
Tfb2	0.88	0.63

Compared the average values between A and B

Rbb1	1.96
Rbp2	1.89
Tfg1	1.77
Ssl2	1.36
Rad3	1.47
Tfb1	1.12
Cet1	1.57
Tfa1	0.83
Tfg2	1.35
Tfb2	0.72

Averaged the values of subunits of protein

pol II (Rpb1 and Rpb2)
1.93

TFIIF (Tfg1 and Tfg2)
1.56

TFIIH (Ssl2, Rad3, Tfb1, Tfb2)
1.17

TFIIE (Tfa1)
0.83

Fig. S9. Protein stoichiometry of the G-less 27 re-initiation complex. Stoichiometry of pol II, TFIIF, TFIIE, and TFIIH of the G-less 27 re-initiation complex (fraction 8 in Fig 4D) was determined in the same manner as in Fig S7. **(A-B)** SDS-PAGE of fraction 12 in Fig. 4A (PIC, as a control) and fraction 8 in Fig. 4D (EC+PIC) was scanned and plotted using ImageJ. **(C)** Quantification of the protein subunits was performed as described in Fig. S7C.

Figure S10

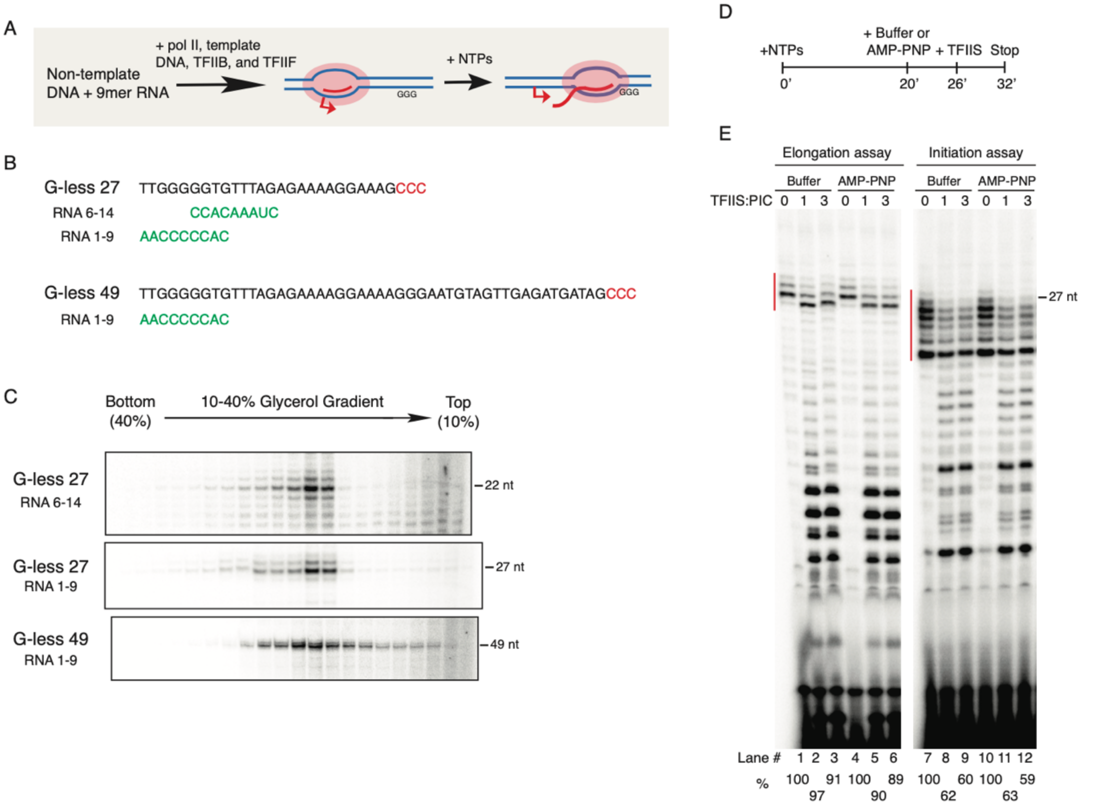


Fig. S10. RNA is retained in artificial elongation complexes during gradient sedimentation and is insensitive to addition of TFIIS in TFIIS cleavage assay. (A) Schematic of artificial elongation assay procedure. Artificial elongation complexes were formed by combining template DNA, a complementary 9-mer RNA, pol II, non-template DNA, TFIIF, and TFIIB. NTPs (ATP, CTP, UTP, [^{32}P]UTP, and 3'-O-methyl GTP) were added and the reactions were incubated for 20 min. TSS is indicated by a red arrow. **(B)** Sequences of 9-mer RNAs (in green) and template strand DNAs that were transcribed (in black). To generate G-less 27 elongation complex with 22 nt transcript, RNA 6-14 complementary to template strand at positions +6 to +14 was annealed and then elongated to +27. To generate G-less 27 and 49 elongation complexes, RNA 1-9 was annealed to each of template strands at positions +1 to +9 and elongated to +27, and +49, respectively. **(C)** The artificial elongation complexes that were stalled after synthesis of 22, 27 or 49 nt RNAs were combined with 2 mM AMP-PNP and subjected to 10-40% glycerol gradient sedimentation. The gradients were fractionated (~180 µL per fraction) and RNA was analyzed by Urea-PAGE. **(D)** Schematic of TFIIS cleavage assays. Transcription initiation and elongation reactions with the G-less 27 template were performed as described in Methods. No capping enzymes were added in the reactions. Transcription was initiated at time 0. After 20 min, 2 mM AMP-PNP or buffer for control was added to the reactions, incubated for 6 min, combined with various amounts of TFIIS, and incubated for additional 6 min before adding stop buffer. The cleaved RNA was analyzed on a 18% acrylamide denaturing gel. **(E)** The level of full-length transcripts was insensitive to TFIIS (indicated by red lines) in the artificial G-less 27 elongation complex, unlike the G-less 27 initiation complex. Also, the level of full-length transcripts was insensitive to addition of 2 mM AMP-PNP in either elongation (lanes 1-3 vs 4-6) or initiation (lanes 7-9 vs 10-12) complexes. Full-length transcripts indicated by red lines were quantified and the amount compared to the control (0 pmol TFIIS) is shown below the gels.

Figure S11

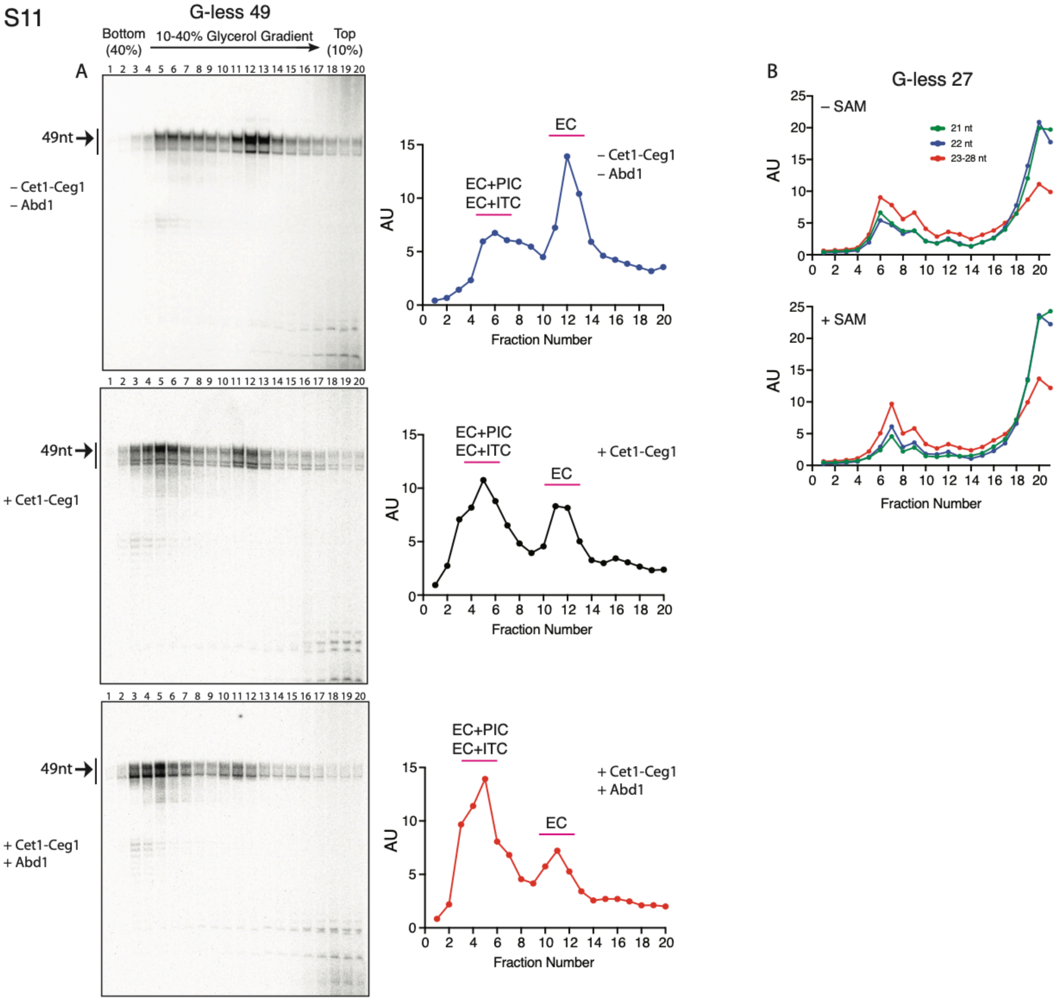


Fig. S11. Abd1 has no effect on assembly of a follow-on PIC and the presence of SAM does not affect promoter escape. (A) Transcription initiation assays and glycerol gradient sedimentation were performed with G-less 49 template DNA in the same manner as in Fig. 5A-B. Assays were performed in in the absence of Cet1-Ceg1 and Abd1 (top), in the presence of Cet1-Ceg1 (middle), or in the presence of both Cet1-Ceg1 and Abd1 (bottom). All the reactions received 2 mM AMP-PNP prior to glycerol gradient sedimentation. RNAs from each ~180 uL fraction were analyzed by denaturing PAGE and the levels of ~49 nt RNAs (indicated by black lines) were quantified (right panels) as in Fig. 5A-B. **(B)** Transcription initiation assay with G-less 27 in the absence (top) or presence (bottom) of 100 μM SAM. Glycerol gradient sedimentation and RNA analysis were performed as described in A-C.

Figure S12

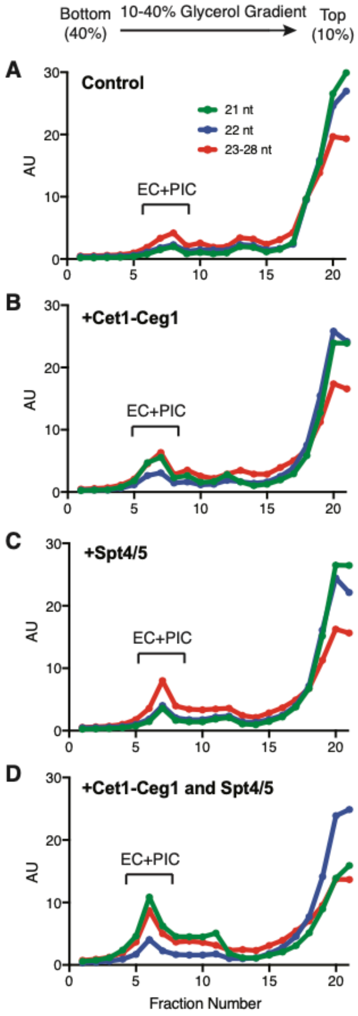


Fig. S12. Spt4/5 facilitates promoter escape. (A-D) Transcription initiation reactions with the G-less 49 template were performed in the absence of Cet1-Ceg1 and Spt4/5 (A), in the presence of Cet1-Ceg1 (B), the presence of Spt4/5 (C), or in the presence of both Cet1-Ceg1 and Spt4/5 (D). Cet1-Ceg1 and/or Spt4/5 were added to the reactions at the time of addition of NTPs. All the reactions received 2 mM AMP-PNP and were subjected to glycerol gradient sedimentation. RNA levels of each ~180 uL fraction were determined by ImageJ and plotted as in Fig. 5. Peaks corresponding to EC+PIC are indicated.

Figure S13

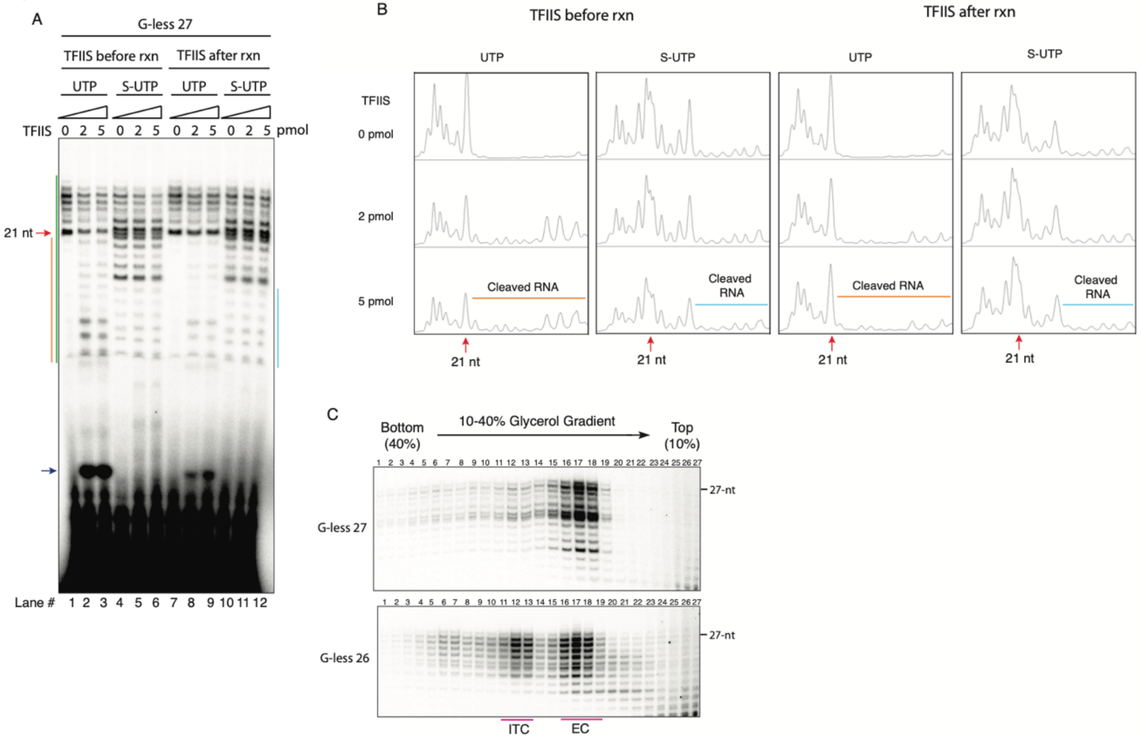


Fig. S13. 4'-Thio UTP reduces backtracking of pol II. (A) PIC was assembled as described in Fig. 1 except that lanes 1-6 received varying concentrations of TFIIS at the time of PIC assembly. Transcription was initiated in the presence of 800 μM ATP, 630 μM CTP, 49.5 nM [^{32}P]CTP, 250 μM 3'-O-methyl GTP, and 630 μM UTP or 4'-thio UTP. Reactions in lanes 1-6 were terminated after 20 min incubation. Reactions in lanes 7-12 received varying concentrations of TFIIS after the 20 min initiation reaction and were incubated an additional 5 min. ~2 nt cleaved products are indicated by a blue arrow. **(B)** The intensity of the bands in the region indicated by the green line in A was blotted using ImageJ. Red arrows indicate 21 nt RNA. Cleaved RNAs are indicated by orange and blue lines. **(C)** Transcription initiation reactions with G-less 27 (top) and G-less 26 (bottom) were performed in the presence of 4'-thio UTP. Reactions received 2 mM AMP-PNP prior to glycerol gradient sedimentation. The gradients were fractionated (~130 μL per fraction) into a total of 27 fractions. RNA in all fractions was analyzed on denaturing gels.

2.10 Materials and Methods

Expression and purification of proteins

TFIIA, TFIIB, TBP, and Sub1 were purified from bacteria and TFIIE, TFIIF, and TFIIH were purified from yeast as previously described (Gibbons et al., 2012; Murakami et al., 2015a). pSBET-His7-ABD1 and pSBET-His7-CET1-CEG1 plasmids were provided by Dr. Buratowski. Abd1 and Cet1-Ceg1 were separately expressed from bacteria and purified as previously described with minor modifications (Cho et al., 1998; Takase et al., 2000). pST69-His6-Spt5-Strep-Spt4 plasmid was a gift from Dr. Reese. Recombinant Spt4/5 was prepared as described (Crickard et al., 2016) with some modifications. Detailed purification methods for Cet1-Ceg1, Abd1, and Spt4/5 are in the supplemental material.

In vitro transcription initiation assays

SNR20 promoter DNA was obtained as described (Murakami et al., 2015a). A series of point mutations were performed using QuikChange Site-Directed Mutagenesis (NEB). SNR20 (–122/+97) was amplified by PCR and purified as previously published (Murakami et al., 2015a). All transcription assays were performed as previously described (Murakami et al., 2013b; Murakami et al., 2015a) with modifications. PIC was formed on 1.3 pmol of DNA fragment with 2 pmol of TFIIA, 3 pmol of TFIIB, 1.5 pmol of TBP, 3 pmol of TFIIE, 5.2 pmol of TFIIF, 1.5 pmol of TFIIH, 1.8 pmol of TFIIK, 3 pmol of Sub1, and 5.2 pmol of pol II in 5 μl of buffer 300 (50 mM HEPES [pH 7.6], 300 mM potassium acetate, 5 mM DTT, and 5% glycerol). The mixture was diluted with 5 μl of buffer 10 (20 mM HEPES [pH 7.6], 10 mM potassium acetate, 5 mM magnesium sulfate, 5 mM DTT) and incubated on ice for 24 hours. After 20 min of preincubation at 30 °C, the reaction was initiated by addition of 10 μl of 2X NTPs containing 1.6 mM ATP, 1.6

mM CTP, 1 mM UTP, 0.5 mM 3'-O-methyl GTP, 1 unit of RNaseOUT, 66-132 nM [α-^{32}P] UTP (2-4 μCi), 5.2 pmol Ceg1-Ceg1, and 10.4 pmol of Abd1 in buffer 10. Transcription initiation samples without capping enzymes received an equal volume of the capping enzyme buffer. The reaction was carried out for 20 min, and quenched by addition of 190 μl of stop buffer (300 mM sodium acetate [pH 5.5], 5 mM EDTA, 0.7% SDS, 0.1 mg/ml glycogen, 0.013 mg/ml of proteinase K [Sigma]) followed by 15 min incubation at 41°C. RNAs were recovered by ethanol precipitation, dried, and dissolved in formamide before running a urea acrylamide denaturing gel. In TFIIS-induced cleavage assays, transcription reactions were performed as described above and were followed by addition of 1 μl of TFIIS (1.5 μM or 4.5 μM) and 6-min incubation at 30°C. Methods for calculation of the transcription efficiency (template usage) is in the supplemental material.

Northern Blotting

RNAs were separated by 15% denaturing polyacrylamide gel electrophoresis (National Diagnostics) and electroblotted/UV crosslinked to Hybond N+ membrane (GE Healthcare RPN303B). ULTRAhyb-oligo Hybridization Buffer (Thermo Fisher Scientific AM8663) was used as per the manufacturer's instructions. Oligonucleotide probes were designed to anneal to the 5' end (nt 1-25; 5'-AAAGGAAAAGAGATTTGTGGGGGTT-3') or the 3' end (nt 26-49; 5'- CGATAGTAGAGTTGATGTAAGGGA-3') of the G-less 49 transcripts. Blots were viewed with the Typhoon 9500 scanner (GE Healthcare).

Separation of post-initiation complexes

45.5 pmol of PICs were assembled as described above. The reaction was initiated by adding 2X NTPs containing 4-fold excess Cet1-Ceg1 and 8-fold excess Abd1 relative to DNA. The reaction was stopped by addition of 2 mM AMP-PNP and loaded on a 10-40% glycerol gradient (v/v) containing 80 mM potassium acetate, 20 mM HEPES (pH7.6), 5 mM DTT, and 2 mM MgOAc. All the gradients in this study were prepared at 82.6 degree tilt except the gradient in Fig 6, which was prepared at 74 degree tilt. All centrifugation was performed for 14 hours at 30,000 rpm in a Beckman SW60 Ti rotor. After the gradient was fractionated (~130 µL per fraction) using a PGF Piston Gradient Fractionator (BioComp Instruments, Inc.), 90 µL was subjected to TCA precipitation for protein analysis by SDS-PAGE and 20 µL was subjected to ethanol precipitation for RNA analysis as described above. When only RNA analysis was performed without protein analysis, 5.2 pmol of PICs were assembled, initiated, and sedimented as described above. After the gradient was fractionated (~180 µL per fraction), 150 µL was subjected to ethanol precipitation for RNA analysis. Note that the gradients in Fig. 5, S6, S11A are fractionated using a peristaltic pump and the rest of the gradients in this study are fractionated using a fractionator.

$KMnO_4$ footprinting assays

The 5'-end of the template or non-template strand of SNR20 (–122/+97) DNA was ^{32}P-labeled by T4 PNK. Transcription initiation assay was performed in the presence of Abd1 and Cet1-Ceg1 as described above except that 20 mM HEPES pH 7.6 in buffer 10 and buffer 300 was substituted with 20 mM potassium/sodium phosphate pH 7.5 and that DTT was removed. After 20 min reaction at 30°C, the reaction mixture received 18 mM $KMnO_4$ (as indicated) and was incubated for 2 min 30 sec at 30°C. $KMnO_4$ was quenched by addition of 1.2 M 2-mercaptoethanol. RNA was extracted by ethanol precipitation and the dried pellet was incubated with 150 µL of 0.1 M piperidine for 30 min at 90°C. The sample was mixed with 1 µL of 12.5 mg/ml yeast tRNA, 15 µL of

3 M sodium acetate pH 5.5, and 600 µL of 100% ethanol and incubated at –80°C for 1.5 hr before centrifugation at 15,000 rpm for 30 min. The pellet was washed with 200 µL of 80% ethanol and dried. RNA was analyzed on a denaturing 8% polyacrylamide gel.

Expression and purification of the capping enzymes

All protein purifications were performed at 4℃. pSBET-His7-ABD1 plasmid was a gift from Dr. Buratowski. The plasmid was transformed into Escherichia coli Rosetta 2(DE3) competent cells. Abd1 was expressed and purified as previously described with minor modifications (Takase et al., 2000). Briefly, the expression of Abd1 was induced at OD600 ~ 0.6 by addition of 1 mM IPTG at 37℃. Cells were harvested by centrifugation after 18 hr of induction. The cells were resuspended in lysis buffer (0.2% Triton, 50 mM Tris-HCl pH 7.5, 150 mM NaCl, 10% sucrose, 1 mM DTT, 1 mM benzamidine, 100 µM leupeptin, 10 µM pepstatin A, 1 mM PMSF) followed by sonication. The supernatant was loaded onto a 5 mL Ni column and the protein was eluted using a gradient of buffer A (20 mM Tris-HCl pH 8.0, 500 mM NaCl, 15 mM imidazole) and buffer B (20 mM Tris-HCl pH 8.0, 500 mM NaCl, 400 mM imidazole). The eluted sample was dialyzed against buffer C (25 mM Tris-HCl pH 8.0, 50 mM NaCl, 0.5 mM EDTA, 1 mM DTT, 0.05% Triton, 5% glycerol) before it was loaded onto a 2 mL phosphocellulose column. The protein was eluted using buffer containing 25 mM Tris-HCl, 1 mM DTT, 10% glycerol, and 1M NaCl. The sample was further purified using a Superdex 200 (GE Healthcare), concentrated, and stored in buffer containing 100 mM potassium acetate, 5 mM HEPES pH 7.5, and 3 mM DTT.

pSBET-His7-CET1-CEG1 plasmid was a gift from Dr. Buratowski. The plasmid was transformed into Escherichia coli Rosetta 2 (DE3) competent cells. Cet1-Ceg1 was expressed in a slightly different manner from Abd1 in that the cell culture was incubated at 21℃ for 16 hours after induction. The cells were lysed by sonication in buffer containing 0.2% Triton, 20 mM Tris-HCl pH 8.0, 500 mM NaCl, 1 mM DTT, 1 mM benzamidine, 100 μM leupeptin, 10 μM pepstatin A, 1 mM PMSF. The debris was removed by centrifugation and the supernatant was loaded onto a 5 mL Ni column. The sample was eluted by the gradient of buffer A (20 mM Tris-HCl pH 8.0, 100 mM NaCl, and 15 mM imidazole) and buffer B (20 mM Tris-HCl pH 8.0, 100 mM NaCl, and 300 mM imidazole). The final purification was performed using a Superdex 200 (GE Healthcare) in the same manner as Abd1 above.

Expression and purification of recombinant Spt4/5

pST69-His6-Spt5-Strep-Spt4 plasmid was a gift from Dr. Reese. Purification of Spt4/5 was prepared as described (Crickard et al., 2016) with some modifications. BL21-Codon Plus RIPL was transformed with the plasmid and was grown to an OD of 0.4-0.5. The expression of Spt4/5 was induced by addition of 0.4 mM IPTG and 10 μM Zn acetate for 18 hours at 18℃. The harvested cells were rinsed with wash buffer (30 mM Tris-HCl pH 7.5, 300 mM ammonium sulfate, and 10 μM Zn acetate), resuspended in wash buffer supplemented with 4 mM 2-mercaptoethanol, 1 mM benzamidine, 100 μM leupeptin, 10 μM pepstatin A, 1 mM PMSF, 0.1% Triton, and 20mM imidazole, and lysed by sonication. The lysate was cleared by centrifugation and the supernatant was filtered using 0.45 μm PES syringe filter (Sartorius) before it was loaded ono a 1 mL Ni column. The column was washed with wash buffer supplemented with 50 mM imidazole. The proteins were eluted by gradient using wash buffer supplemented with 500 mM imidazole. The peak fractions were pooled and loaded onto a 5 mL Q column equilibrated in Q buffer (30 mM Tris-HCl pH 7.5, 100 mM ammonium sulfate, and 10 µM Zn acetate, 3 mM DTT, and 5 % glycerol). Spt4/5 was eluted with a gradient of ammonium sulfate in Q buffer from 100 mM to 800 mM and concentrated with 100k MWCO Vivaspin 6 concentrator (GE Healthcare).

Calculation of transcription initiation efficiency

To calculate the transcription initiation efficiency, a dilution series of the stock radiolabeled UTP was blotted on a filter paper and exposed to a phosphorimager screen with the assay gel. Band intensities and standard blots were measured using Image J. A linear regression was fitted through the standard points and its equation was used to calculate moles of ^{32}P UTP in the transcripts. Transcription efficiencies were then determined as follows: transcription efficiency = moles of ^{32}P UTP ÷ number of U in a transcript × ([cold UTP])/([hot UTP]) ÷ moles of preinitiation complex × 100.

Reconstitution of artificial elongation complexes

Artificial elongation complexes were assembled as previously described with modifications (Cabart et al., 2014). Briefly, RNA [5'-ACCCCCACA-3'] or [5'-CCACAAAUC-3'] was annealed with DNA templates ([5'-ACTACACTTGATCCACCCGAAAGGAAAAGAGATTTGTGGGGGTTTAAAAAAAAAACAAGTAGA -3'] for G-less 27 and [5'-GTTACACTGAAAAGACCCGATAGTAGAGTTGATGTAAGGGAAAAGGAAAAGAGATTTGTGGGGGTTTAAAAAAAAAACAAGTAGA -3'] for G-less 49 over 1 min gradient from 60 to 4°C. Non template DNA ([5'-TCTACTTGTTTTTTTTTTAAACCCCCACAAATCTCTTTTCCTTTCGGGTGGATCAAGT

GTAGT-3’] for G-less 27 and [5’-TCTACTTGTTTTTTTTTTAAACCCCCACAAATCTCTTTTCCTTTTCCCTTACATCAACTCTACTATCGGGTCTTTTCAGTGTAAC-3’] for G-less 49, and the RNA-DNA mixture was combined with equimolar amounts of pol II on ice. The assembled elongation scaffold was purified using an illustra MicroSpin G-50 column (GE Healthcare life Science). We combined 4 pmol of the purified scaffold with 4.2 pmol of TFIIF and 4.2 pmol of TFIIB in 2 µl of buffer 300 and 2 µl of buffer 10, and incubated on ice for 1 hour. RNA was extended by addition of 2X NTPs as in in vitro initiation assays.

Exonuclease experiments

Transcription initiation assays were performed as described above. RNA was isolated using acid phenol chloroform (Thermo Fisher Scientific) followed by ethanol precipitation. Isolated RNA was treated with CIP (NEB) to remove phosphate groups, then T4 polynucleotide kinase (NEB) to add monophosphates to the 5’ ends of the RNAs, and finally Terminator 5’-phosphate-dependent exonuclease (Epicentre) for RNA degradation. RNA was purified by acid phenol chloroform extraction with ethanol precipitation after each step of the enzyme treatments. Negative controls received an equal volume of water instead of enzymes. RNA was analyzed on a urea acrylamide gel.

CHAPTER 3: STRUCTRAL VISUALIZATION OF DENOVO INITIATION OF RNA POLYMERASE II TRANSCRIPTION

2.1 Preface

The manuscript presented in this chapter was originally published on BioRxiv on May 4, 2021. It has been submitted to a peer-reviewed journal and was under review while the thesis was being prepared. It has been reformatted here in accordance with University of Pennsylvania dissertation formatting guideline.

Authors: Chun Yang[1†], Rina Fujiwara[1,2†], Hee Jong Kim[1,2], Jose J. Gorbea Colón[1,2], Stefan Steimle[3], Benjamin A. Garcia[1,4], and Kenji Murakami[1*]

Affiliations:
[1]Department of Biochemistry and Biophysics, Perelman School of Medicine, University of Pennsylvania, Philadelphia PA. 19104, USA

[2]Biochemistry and Molecular Biophysics Graduate Group, Perelman School of Medicine, University of Pennsylvania, Philadelphia PA. 19104, USA

[3]Epigenetics Institute, Department of Biochemistry and Biophysics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA

[4]Penn Center for Genome Integrity, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

[†]Contributed equally to this work.

*Correspondence: kenjim@pennmedicine.upenn.edu

2.2 Respective Contributions

Cryo-EM sample preparation and analyses in Figure 1-3 were performed by me under the guidance of Dr. Kenji Murakami. Dr. Kenji Murakami and Dr. Stefan Steimle and I prepared the models in Figure 1-3. Dr. Chun Yang performed Cryo-EM sample preparation and analyses in Figure 4-6. Hee Jong Kim performed cross-linking MS under the guidance of Dr. Benjamin A. Garcia. Jose J. Gorbea Colón performed integrative modeling and prepared Figure S5. Figure 7 was prepared by Dr. Chun Yang. Figure S1-3 were prepared by me, and Figure S6 was prepared by Dr. Chun Yang. Kenji Murakami wrote introduction and discussion, Dr. Kenji Murakami and I wrote the sections related to Figure 1-3. Dr. Kenji Murakami and Dr. Chun Yang wrote the sections related to Figure 4-7.

2.3 Abstract

Structural studies of the initiation-elongation transition of RNA polymerase II (pol II) transcription were previously facilitated by the use of synthetic oligonucleotides. Here we report structures of initiation complexes de novo converted from pre-initiation complex (PIC) through catalytic activities and stalled at different template positions. Contrary to previous models, the closed-to-open promoter transition was accompanied by a large positional change of the general transcription factor TFIIH which became in closer proximity to TFIIE for the active delivery of the downstream DNA to the pol II active center. The initially-transcribing complex (ITC) reeled over 80 base pairs of the downstream DNA by scrunching, while retaining the fixed upstream contact, and underwent the transition to elongation when it encountered promoter-proximal pol II from a preceding round of transcription. TFIIH is therefore conducive to promoter melting, TSS scanning, and promoter escape, extending far beyond synthesis of a short transcript.

2.4 Introduction

RNA polymerase II (pol II) and the six general transcription factors (GTFs) assemble in a transcription pre-initiation complex (PIC), which recognizes promoter DNA before every round of transcription, and opens the double-stranded DNA to expose and select a transcription start site (TSS) (Conaway and Conaway, 1993; Kornberg, 2007). Following the TSS recognition, the PIC transitions to the initially-transcribing complex (ITC), which is responsible for synthesizing a nascent transcript and subsequently transitions to an elongation complex (EC), followed by re-initiation. This set of transitions is universal across all eukaryotes and overlaid by many additional regulatory steps involving elongation factors such as DSIF (Spt4/5 in yeast) and initiation factors such as Mediator (Adelman and Lis, 2012; Conaway and Conaway, 2012; Wade and Struhl, 2008). It is also commonly thought that polymerases from successive rounds of transcription are located immediately adjacent to each other at promoter proximal regions of actively transcribed genes, yet the precise nature and the significance of the interaction remain unknown (Ehrensberger et al., 2013).

The largest GTF, TFIIH comprising 10 subunits, is an integral component of the PIC: the translocase subunit Ssl2 (XPB in humans) acts as a molecular motor during promoter opening, TSS scanning, and initial RNA chain elongation (Bradsher et al., 2000; Dvir et al., 1997a; Fazal et al., 2015a; Fishburn et al., 2015; Qiu et al., 2020; Spangler et al., 2001). The other subunits comprise the six-subunit structural core (Greber et al., 2019) and the three-subunit kinase termed TFIIK for pol II CTD phosphorylation (van Eeuwen et al., 2021a). Previous structural studies of open promoter complexes provided information about locations of GTFs and the DNA path (He et al., 2013; He et al., 2016; Schilbach et al., 2017). However the open promoter template was not obtained by the catalytic activity of TFIIH. Thus it remains to be determined how TFIIH is responsible for the initiation process.

To elucidate the mechanisms of the initiation, we have recently developed an in vitro transcription system, in which pol II and six GTFs (TFIIA, TBP, TFIIB, TFIIE, TFIIF, and TFIIH) isolated from the yeast *Saccharomyces cerevisiae*, melt double-stranded promoter DNA, and initiate RNA synthesis de novo with high efficiency (Fujiwara and Murakami, 2019; Murakami et al., 2013a; Murakami et al., 2015a). Resulting post-initiation complexes could be stalled at different template positions on a series of G-less promoter mutants and isolated in abundant homogeneous form by glycerol gradient sedimentation (Fujiwara et al., 2019).

Here we report cryo-EM structures of such de novo initiation complexes stalled at two different template positions (Figure 1A). The first cryo-EM structure with the G-less 26 template revealed an ITC containing all the GTFs, pol II, and a nascent RNA on a bona fide open promoter DNA. Compared to previous structures of PICs (Dienemann et al., 2019b; Murakami et al., 2015b; Schilbach et al., 2017), the ITC underwent a large positional change in TFIIH for the active delivery of the downstream DNA to the pol II active center. By contrast, the second structure with the G-less 49 template revealed successive elongation complexes (EC+EC), in which two polymerases that completed promoter escape were in close contact with each other. From the combination of these structures with previous biochemical and biophysical studies (Fazal et al., 2015a; Fujiwara et al., 2019), we arrive at a picture of the initiation process driven by TFIIH, in which the preceding pol II stalled at promoter-proximal regions blocks TFIIH translocation of the trailing ITC and ultimately occludes TFIIH on a promoter.

2.5 Results

Isolation of bona fide ITCs and Cryo-EM analysis

ITCs were obtained by transcription reaction with the G-less 26 SNR20 promoter fragment, in the procedure we have previously established (Fujiwara et al., 2019) (Figure 1A). Briefly the protocol entails combining a 33-subunit PIC with a G-less SNR20 promoter fragment, supplemented with ~4-8 fold molar excess pol II and GTFs relative to PIC, followed by addition of NTPs for transcription reaction. ITCs were stalled at position +26 relative to the TSS (+1) by use of chain-terminating 3'-O-methyl GTP instead of normal GTP, while inclusion of 4'-thio UTP instead of normal UTP induced pol II arrest and thereby prevented the extensive backtracking, which would otherwise have completely collapsed back to the closed complex (PIC) with concomitant RNA release during subsequent gradient sedimentation (Fujiwara et al., 2019). The reaction mixture was sedimented on a 10-40% of glycerol gradient to remove free nucleotides and excess GTFs and pol II. The resulting ITC contained equimolar amounts of GTFs and pol II, and transcripts of ~20-26 nucleotides initiating from positions +1 to +7 (Figures S1A-B). Due to the similarity in size, ITC were not separable on the gradient from residual PICs that did not engage in transcription and/or those that collapsed back from ITC.

Knowing the heterogeneity of the specimen, aliquots of peak fractions were embedded in ice, disclosing fields of monodispersed particles (Figure S1C). We imaged ~4 million particles using Titan Krios electron microscopes equipped with a K3 direct electron detector. 2D class averaging of particles yielded a set of homogeneous classes, which showed clear division in two parts: a well-ordered pol II and a disordered TFIIH (Figure S1D). For some classes, DNA was identifiable on TFIIH. We selected a subset of particles (~0.8 million) through 2D class averaging and subjected them to *ab initio* calculation of an initial map (Figure S1E). To sort out variability in positions of TFIIH and DNA, the ~1.8 million particles were subjected to iterative global 3D classifications, which revealed three forms of PICs (hereinafter PIC1, PIC2, PIC3) and one form of the ITC, accounting for 137K, 117K, 69K, and 120K particles, respectively. In each form of PICs and ITC, TFIIH and DNA were poorly ordered due to their flexibility compared to pol II. For reconstruction of PIC1-3, TFIIH was subjected to focused classification and refinement, and composited back to the entire map (Figure S1F-N). For reconstruction of the ITC, three masks were created, the first containing the active center of pol II, the second containing TFIIH, and the third containing DNA-TFIIA-TBP-TFIIE (Tfa1 and Tfa2 WH domains)-TFIIF (Tfg2 WH domains)-TFIIB (cyclin domains). Three segments were subtracted from images with respective masks, subjected to local 3D classifications and refinement, and then composited back to the entire complex (Figure S1O-Q). Focused classification of pol II active center in ITC map enabled removal of particles that had only poor density of the DNA-RNA hybrid.

Three forms of PICs and the ITC differed from each other in locations and conformations of TFIIH and DNA path (Figures 1B-D). PIC1 was a good match to previous structures of yeast 31-subunit PIC (EMDB 3114 and EMDB0092) (Dienemann et al., 2019b; Murakami et al., 2015b), and was resolved at higher resolution (3.2-4.6 Å) than before, attesting to our sample preparation and data analysis strategy. In PIC1, TFIIH was resolved at near atomic resolution (4.6 Å), allowing us to define two different DNA-binding modes for DNA translocation, as described in detail below. PIC2 and PIC3 were refined to 3.2-7.3 Å and 3.4-11.8 Å, respectively, and were distinct from PIC1 in the position and the conformation of TFIIH and DNA. The ITC was refined to 3.2-9.9 Å resolution, revealing a bona fide open promoter DNA and a short 6-bp DNA-RNA hybrid in the pol II active center, which differs from previous open promoter complexes with artificial templates (He et al., 2016), by translations of more than ~30 Å (over 50 Å for some TFIIH subunits (Tfb1, Tfb2, Tfb4, Tfb5)) in the location of TFIIH.

Two DNA-binding modes of TFIIH in PIC1

The previous cryo-EM model of PIC (EMDB 3114 and EMDB0092) was well fitted into density of PIC1. The fit showed some differences in degree and position of the DNA bend ~20-30 bp downstream of the TATA box, which may relate to different promoter sequences (SNR20 in this study vs HIS4 in previous studies). The relationship between the DNA bend/distortion and promoter melting was previously characterized (Dienemann et al., 2019b). The model of TFIIH was built using the previous 3.9 Å-resolution cryo-EM structure of yeast TFIIH bound to Rad3-Rad23-Rad33 as an initial model (van Eeuwen et al., 2021b). The model of PIC1 was subjected to iterative refinements with Coot (Emsley et al., 2010) and Phenix (Liebschner et al., 2019) (Supplemental Table 1), and then used as an initial template for model building of PIC2, PIC3 and ITC.

Focused 3D classification of TFIIH in PIC1 revealed two forms of TFIIH at 4.6 Å and 7.6 Å resolution (orange vs steel blue in Figures 1E-F, Figure S1F). One form had a good match to the previous structure of the pre-translocation state of TFIIH in the PIC (Dienemann et al., 2019b; Murakami et al., 2015b; Schilbach et al., 2017) (orange, upper panel of Figure 1E), while the other form revealed a ~60° rotation of the domain that consists of Tfb5 and the C-terminal region of Tfb2, accompanied by a rotation of the C-terminal ATPase domain of Ssl2 (Ssl2C) relative to the rest of TFIIH (steel blue, lower panel of Figure 1E) as previously observed in the structure of TFIIH-Rad4-Rad23-Rad33 (van Eeuwen et al., 2021b). In the former (orange in Figures 1E-F), a ~13-bp segment of

DNA double helix was bent, deep within the DNA-binding groove between the two ATPase domains, in close contact with the five DNA binding motifs (Ic, IVa, IV, V, Vb, as previously defined (Fairman-Williams et al., 2010)) (referred to as strong-binding state), whereas, in the latter (steel blue in Figures 1E-F), the DNA was relatively straight, only in contact with two DNA binding motifs (IVa, Ic) (referred to as weak-binding state). In the weak-binding state, the detachment of the DNA from motifs IV, V, and Vb was accompanied by the rotation of Ssl2C along with Tfb5-Tfb2C (Figure 1F), consistent with the previously suggested role of Tfb5 (p8 in humans) in stimulating Ssl2's catalytic activity (Coin et al., 2006; Ranish et al., 2004). The remaining DNA-Ssl2 interactions by the two motifs IVa and Ic were altered, enabling a slight rotation of the DNA along its axis, likely coupled with DNA translocation (Figure S1R). Thus the weak-binding state may represent the post-translocation state, although nucleotides were not directly resolved.

Distinct forms of PICs represent the path to the open promoter complex.

PIC2 and PIC3 differed from PIC1 in locations and conformations of TFIIH and DNA path, as readily apparent in initial rounds of 3D classification (Figure S1E), and there are several notable differences between three forms of the PIC (Figure 2). First, PIC2 and PIC3 differed from PIC1 by ~20 Å and ~30 Å shifts in the location of TFIIH (Figure 2A), and by repositioning of Ssl2 on DNA by one turn of dsDNA (~10 bp), accompanied by greater degrees of DNA distortion ~20-30 bp downstream of the TATA box (Figures 2B-C). Second, PIC2 and PIC3 revealed TFIIH in the weak-binding state, while PIC1 primarily revealed the strong-binding state, suggesting that locations of TFIIH in PICs would shift the conformational equilibrium among coexisting translocation states. Third, in PIC3, Tfa1 (TFIIE) and the RING domain of Tfb3 (one of three TFIIK subunits) were dissociated from the pol II clamp and Rpb4/7, resulting in a shift in their positions by ~10 Å relative to those in PIC1/PIC2, such that TFIIH and TFIIE less closely contacted pol II (lower panels of Figures 1B-D). Our previous exonuclease footprinting demonstrated that omission of TFIIK, while retaining a high-level of TFIIK-independent transcription, causes upstream shift of the downstream boundary of the PIC (by ~5 residues) (Murakami et al., 2015a), suggesting removal of TFIIK may be able to mimic the transition to PIC3. Irrespective of these significant differences between three forms, promoter DNA was nevertheless associated only with GTFs and not with pol II in all forms, requiring for the translocase activity of TFIIH for promoter melting.

Bona fide ITC structure

Locations of GTFs in the ITC largely correspond to those in the PIC3 except some differences in orientations of TFIIH and TFIIE (Figure 3A, Movie S1); as in PIC3, the Tfb3 RING domain was absent on Rpb4/7 (not visualized in the map), so that TFIIH and TFIIE less closely contacted pol II (Figure 3A). The promoter DNA of the ITC was suspended above the pol II cleft, bound by TFIIH at the downstream end, and by the remaining general transcription factors (GTFs) at upstream end. A ~36 bp segment (positions –116 to –79) of the upstream DNA bound to TFIIA, TBP, TFIIB (cyclin domains), TFIIE (Tfa1 and Tfa2 WH1 and WH2 domains), and TFIIF (Tfg2 WH domains) was clearly discerned (Figure 3B). The upstream edge of the transcription bubble in the ITC (position –79), corresponding to the position of the 25° bend in the PIC3, was stabilized by the WH domain of Tfa1 (the large subunit of TFIIE) (Figure 3B). Although the downstream DNA bound to Ssl2 was poorly ordered, there was a discernable density attributable to a ~9-bp short segment of DNA double helix bound to Ssl2 in the focused classification of TFIIH (Figures S3A-B). In between, the DNA of over ~100 bp was missing except the region of the DNA (–2 to +9) that was accommodated in the pol II active center (Figure 3C): the region of the DNA between positions –79 and –2 presumably looped out or scrunched (Fazal et al., 2015a; Kapanidis et al., 2006; Liu et al., 2010), while the downstream DNA that bridges between Ssl2 and the pol II active site, likely a straight DNA double helix, was disordered (schematically illustrated in Figure 3E). It is important to note that the downstream DNA was not observed deep in pol II downstream cleft, which markedly contrasts to the transcribing complex (EC) (Gnatt et al., 2001; Kettenberger et al., 2004) (inset of Figure 3E).

The short DNA-RNA hybrid observed in the active center of the ITC was a good match to the 6-bp DNA-RNA hybrid previously observed by X-ray crystallography in a complex with TFIIB (Sainsbury et al., 2013) (Figure 3C): eleven nucleotides of the template strand at positions –2 to +9 were identifiable with discernible backbone phosphate positions. The six ribonucleotides of RNA in the hybrid were identifiable at positions from i–1 to i–6 relative to the nucleotide addition site, i+1. The 5'-terminal nucleotide (position i–6) was in direct contact with the finger domain of TFIIB (Figure 3C). Despite of inclusion of 4'-thio UTP, pol II was evidently subjected to extensive backtracking from +26 to +6, so that the RNA was stabilized in the ITC, that would otherwise have been incompatible with TFIIB (Bushnell et al., 2004) (Figure 3E). Consistent with this model, there was a density attributable to the backtracked RNA (at positions i+3 to i+5) in the pol II funnel. The observed path of the backtracked RNA coincides with that of the backtracked EC (see below).

As a key feature of the ITC, the positional constraint of TFIIH imposed by the rigid straight double-stranded DNA (as in PIC1) as well as by the contact between the Tfb3 RING domain and Rpb4/7/Tfa1 was relieved due to the promoter melting, such that

TFIIH was stabilized through protein-protein interactions that were absent in the canonical PIC (PIC1) (Figure 3D, Figure S3, Movie S1): the primary contact was made between the Tfb1 BSD2 domain of TFIIH and the Tfa1 WH2 domain of TFIIE. The second contact was made between the Rad3 Arch domain of TFIIH and the Tfa2 WH2 domain of TFIIE (Movie S1). Although not modeled, there was a density adjacent to the Tfb1 BSD1 domain of TFIIH, which may be attributed to the C-terminal region of Tfa1. These interactions are in good agreement with a number of cross-links, most of which were obtained in the PIC lacking TFIIK (Murakami et al., 2013c) (e.g., a cross-link between K268 of Tfb1 and K194 of Tfa2, Figure 3D and S3C-F).

The TFIIH-TFIIE interactions described above (Figure 3D) apparently serve as a critical point of contact between TFIIH and the remaining GTFs, such that TFIIH rotates the downstream DNA for unwinding, while retaining fixed upstream contact (Fishburn et al., 2015; Kim et al., 1997). Without TFIIH being held by this anchor point, TFIIH itself may freely rotate around the DNA axis. Based on real-time observations of single PICs (Fazal et al., 2015a), this translocation reels dozens of base pairs of the downstream DNA independently of pol II transcription (*i.e.,* only with dATP that allows for DNA translocation by Ssl2), and continues even after the point (~+7–+12) at which TFIIB is displaced from the RNA exit tunnel, in good agreement with biochemical isolation of stable ITCs stalled at ~+26–27 (Fujiwara et al., 2019).

The structure of ECs colliding head-to-end (EC+EC)

In contrast to the G-less 26 complex that formed such long-persisting ITC, our previous biochemical studies demonstrated that the G-less 49 complex contained a pol II that escaped the promoter (+49), and another pol II that initiated transcription by re-utilizing the promoter to generate the ~25 nt RNAs (thus referred to as re-initiation complex) (Fujiwara et al., 2019). Upon removal of ATP during gradient sedimentation, the ~25 nt transcripts from the second round of transcription were retained in the G-less 49 complex, in contrast to the G-less 26 complex that released transcripts of similar lengths by extensive backtracking of pol II (Fujiwara et al., 2019). Inclusion of 4'-thio-UTP, instead of normal UTPs, was needed to induce pol II arrest and prevent RNA release for the structure determination of the ITC with the G-less 26 template, as described above. This indicates that an EC stalled at promoter proximal regions (~+49) serves to play a positive role in the trailing ITC, as previously suggested (Ehrensberger et al., 2013).

To isolate G-less 49 complexes for structural study, transcription reaction with the G-less 49 template was initiated by adding NTPs (ATP, CTP, and UTP) with chain-terminating 3'-O-methyl GTP, and following gradient sedimentation revealed two major peaks of the re-initiation complex (fractions 17-18 and 22-24 in Figure 4A). Aliquots of each peak were subjected to cryo-EM analysis in a similar manner to the G-less 26 complex (Figure S4). Consistent with protein analysis by SDS-PAGE (Figure 4B), initial two rounds of 2D classification of the slower sedimenting fractions yielded a set of well-ordered homogeneous classes of two colliding pol II molecules (referred to as EC+EC) (Figures 4C and 4E), while the faster sedimenting fractions yielded similar classes of two colliding pol II molecules associated with a set of GTFs (referred to as EC+ITC) (Figures 4D and 4F). After interactive 3D classifications to remove some residual PICs (Figure S4F, see also Figure 4B), the structure of the EC+EC was refined to 3.5 Å resolution (Figure 5A), while the considerable variability in the distance between two pol II molecules prevented refinement of EC+ITC past about 15-Å resolution.

In the structure of EC+EC, two colliding ECs span over ~74 bp of DNA (from positions –8 to +66 relative to TSS) (Figure 5). There was a well-ordered density corresponding to TFIIF only on the trailing EC, but not the leading EC (Figure 5A). A previous crystallographic model of an EC complex with a 9-bp DNA-RNA hybrid (PDB ID: 5C4J) (Barnes et al., 2015) was fitted into two corresponding densities with some deviations in the non-template strand of the transcription bubble. Also a previous model of pol II-TFIIF (PDB ID: 5FYW) was fitted without any deviations except a ~10°-rotation of Rpb4/7 subunits of the leading EC, that enabled a direct contact with TFIIF of the trailing EC (Figures 5B-C).

The active site (the nucleotide addition site) of the leading EC was located at the G-stop (+49), while that of the trailing EC was located at +14 (Figures 6A-B, and 6G). This suggests that the trailing pol II that had reached ~+25 to transcribe a ~25-nt RNA, was subjected to extensive (~11bp) backtracking, and arrested at +14. Without this backtracking, two ECs require substantial structural changes in the protein component or/and the DNA component to avoid steric clash at the interface. Consistent with the pol II backtracking, there was density attributed to this backtracked RNA in the funnel of the trailing EC, but not the leading EC (Figures 6C-E). Two-body refinement revealed a ~6°-rotational motion relative to each other, while maintaining the 35-bp spacing between two nucleotide addition sites (Movie S2).

Following the substantial backtracking of the trailing EC, specific protein-protein interactions were established at the interface between two ECs (Figure 5B). There were two major points of contact: the first point of contact involved two loops (residues 148-168 and residues 185-197) protruding from Rpb1 clamp of the trailing EC, and the loop of Rpb2 (residues 97-113) and the Rpb12 zinc ribbon (residues 35-50) of the leading

EC. The second point of contact involved the tip (residues 134-137) of the dimerization domain of Tfg1 (TFIIF) of the trailing EC, and the tip (residues 125-127) of Rpb7 of the leading EC. Also there were some weak EM densities that were not modeled at the interface. These densities may be attributed to YEATS (residues 1-137) and ET domains (residues 174-244) of Tfg3 as well as the C-terminal WH domain (residues 671-735) of Tfg1 based on an integrative modeling derived from XL-MS (Figure S5).

The template DNA was overall Z-shaped with two kinks at the two active centers of pol II (Figures 6A-B). The 24-bp DNA (from +15 to +38) bridging between two active centers was clearly discerned and modeled with a straight B-form DNA (Figure 6B). The density of the DNA-RNA hybrid in each active center was traceable (Figures 6B-D): in the leading EC, 16 ribonucleotides of the 49-nt transcript were visualized: 9 ribonucleotides from the 3' end formed a hybrid with the template DNA (positions from +41 to +49), with the 3' end of the transcript (3'-O-methyl guanosine 5′-monophosphate) being aligned at the nucleotide addition site (designated i+1 position) in the active center, while a stretch of adjacent seven ribonucleotides was in the RNA exit tunnel (Figures 6A-B and 6G). In the trailing pol II, 15 ribonucleotides of the ~25-nt transcript were discernible (Figures 6C and 6G). 9 ribonucleotides formed a hybrid with the template DNA (positions from +6 to +14) in the active site, and adjacent five ribonucleotides of the backtracked RNA were observed in the pol II pore and funnel (Figure 6E): two ribonucleotides of the backtracked RNA at positions i+2 and i+3 were in the pore as previously observed by X-ray crystallography(Wang et al., 2009), while three ribonucleotides at positions i+4, i+5, and i+6 lie on a positively charged patch composed of Lys619 and Lys620 of Rpb1 in the funnel, not observed in any previous transcribing complex structures (Figure 6F). The path of the backtracked RNA differed from that of the backtrack site previously observed by X-ray crystallography (Figure 6F) (Cheung and Cramer, 2011). The backtracked RNA is nonetheless incompatible with TFIIS, and must be displaced from these sites for TFIIS-induced transcription resumption from the backtracked state (Cheung and Cramer, 2011).

Two lines of evidence support the specificity and functional significance of the EC+EC complex. First, previous exonuclease footprinting of elongation complexes without TFIIF, exhibited greater variability in the distance between two ECs upon head-to-end collision as well as much more extensive backtracking of the trailing EC (~50 bp backtracking upon encountering a leading EC)(Saeki and Svejstrup, 2009), supporting the specificity of the EC+EC conferred by TFIIF. Second, the trailing EC stalled at +14 completed promoter escape, whereas initiation complexes stalled at any positions before +27 failed to escape promoter in our single round transcription system (Fujiwara et al., 2019). The leading EC stalled at +49 from a preceding round of transcription likely plays a positive role in promoter escape of a trailing ITC, rather than simply acting as a roadblock of TFIIH translocation.

Promoter escape of the ITC is facilitated by a transcribing pol II at promoter proximal regions

Direct support for the role of the leading EC in promoter escape of the trailing ITC came from cryo-EM analysis of the EC+ITC (Figures 4D and 4F). All 2D class averages showed a large (~500 kDa) density attributable to TFIIH in a space between two polymerases (indicated by orange arrow heads in Figure 4D). The assignment of TFIIH was further validated by a comparison with a 2D projection from a 3D model of EC+core ITC (ITC lacking TFIIH) (Figure S7B). Of the eight class averages we obtained, the top four populated classes maintained a similar spacing between the EC and the ITC as in the EC+EC, through the direct protein-protein interactions described above (upper row in Figure 4D). In these class averages, the DNA double helix was accommodated in the pol II downstream cleft of the trailing ITC, while TFIIH was dissociated from the DNA and displaced from its position in the ITC of the G-less26 complex (schematically illustrated in Figure 4F, see also Figure S7A). This conformational change of the ITC, as an irreversible critical transition from initiation to elongation (see Discussion), was evidently facilitated by the presence of the leading EC. In the other classes, two ECs were apparently separated from each other, suggesting that the trailing ITC was subjected to more extensive backtracking (lower row in Figure 4D), which may further require TFIIS for transcription resumption from the backtracked state.

2.6 Discussion

Structural and mechanistic studies of transcription initiation involving TFIIH have been hampered by poor efficiency of initiation reaction in vitro (commonly ~0.01-0.1 transcripts per PIC). Previous structural models of transition from initiation to elongation were derived from complexes with artificially open templates, and not obtained by the catalytic activity of TFIIH. Thus how TFIH directs promoter melting, TSS scanning, and promoter escape (Dvir et al., 1997a; Fishburn et al., 2016; Luse, 2013; Qiu et al., 2020; Spangler et al., 2001) remains to be resolved. To dispel this long-standing mystery of the transcription initiation process, we have developed a highly efficient in vitro reconstitution from the yeast at quality and quantity amenable to structure determination (Fujiwara and Murakami, 2019; Murakami et al., 2013a). As a notable achievement reported here, we have arrived at a complete description of pol II transcription from initiation by the 33-subunit PIC through promoter escape, to finally reach elongation. Our structural data provide a direct evidence that the ITC retaining all GTFs continues until it encounters an

EC at promoter proximal regions from a preceding round of transcription, and that two polymerases occlude TFIIH binding to facilitate promoter escape. Promoter escape, viewed in the past as no more than dissociation of pol II from promoter, now appears mechanistically varied, with important regulatory consequences.

Three distinct forms of the PIC were defined in this study: relative to PIC1 in a form similar to previous structures (Dienemann et al., 2019b; Murakami et al., 2015b), PIC2 and PIC3 exhibited ~20 Å and ~30 Å shifts in the location of TFIIH, and repositioning of Ssl2 on DNA by one turn of dsDNA, along with greater degrees of DNA distortion. In PIC1, the location of TFIIH is constrained primarily by the contact with the relatively straight and rigid double-stranded DNA. Upon the DNA distortion in PIC2/PIC3, the positional constraint of TFIIH imposed by the double-stranded DNA is relieved, such that TFIIH is rather stabilized by direct protein-protein contacts with TFIIE. Locations of GTFs of the PIC3 closely correspond to those in the ITC, indicating the functional significance of PIC3, as well as PIC2, as intermediates on a path to the open complex formation. However, apparently a transition from PIC1 to PIC2/PIC3 requires rebinding of TFIIH on DNA, due to the upstream shift in the location of Ssl2 on DNA. Energy barriers required for this rebinding may indicate some functional differences between PIC2/PIC3 and PIC1.

The possible functional differences between distinct forms of the PIC may relate to two forms of TFIIH: only the weak-binding state, which most likely represent a post-translocation state of TFIIH, was exclusively identified in PIC2/PIC3, whereas the strong-binding state (pre-translocation state) was apparently favored in PIC1. This suggests that locations of TFIIH in PICs would shift the conformational equilibrium among coexisting translocation states to regulate translocase activity of TFIIH. Although previous and current structures of TFIIH did not directly resolve nucleotide states, there is a consensus observation that the strong-binding state (pre-translocation state) was exclusively identified from specimens with a non-hydrolysable ATP analogue or without ATP, while the weak-binding state (post-translocation state) was identified only when ATP was provided (this study and (van Eeuwen et al., 2021b)).

In previous structures of the open PIC (He et al., 2016; Schilbach et al., 2017), downstream dsDNA was stably accommodated in the pol II downstream cleft and the further downstream end was simultaneously bound by TFIIH. Considering that these models resemble those from other transcription systems devoid of equivalent translocases and that a similar open complex could be formed in the absence of TFIIH (Plaschka et al., 2016), they may represent the pathway to TFIIH-independent transcription (Alekseev et al., 2017; Holstege et al., 1995; Parvin and Sharp, 1993). By contrast, in our bona fide ITC, TFIIH precluded such stable accommodation of the downstream DNA in the pol II downstream cleft, and directed open complex formation that were essentially maintained by GTFs, but not pol II. Due to the lack of the direct contact with the downstream DNA, pol II may have a degree of freedom of lateral movement along the template, and thus confer TFIIH-dependent properties in TSS utilization, initial RNA synthesis, and promoter escape (Bradsher et al., 2000; Dvir et al., 1997a; Fishburn et al., 2016; Fujiwara et al., 2019; Murakami et al., 2015a; Spangler et al., 2001).

Previous biochemical and biophysical data suggest that a bona fide ITC is long-persisting and that promoter escape occurs after synthesis of dozens of nucleotides (Fazal et al., 2015a; Fujiwara et al., 2019; Luse, 2019). However this is unlikely to occur in cells as the initially transcribing pol II is thought to encounter another pol II or a nucleosome at promoter proximal regions shortly after the initiation of transcription (Ehrensberger et al., 2013). Thus our G-less 49 complex may represent a more complete picture of promoter escape occurring in vivo, as an ITC is formed in the presence of EC stalled at +49 from a preceding round of transcription (Figure 7). Contrary to expectation, an EC at promoter proximal regions supported a positive role in transcription rather than acting as a transcriptional roadblock; TFIIH of the ITC was occluded between two transcribing polymerases, followed by partial dissociation of TFIIH (resulting in EC+ITC) or complete dissociation of TFIIH (resulting in EC+EC) (stage 3 or 4 in Figure 7). In the EC+EC, the trailing pol II completed promoter escape after transcribing ~25 nt RNA, while in the EC+ITC, the trailing ITC apparently failed to escape promoter, but successfully accommodated the downstream DNA in the pol II downstream cleft (upper row of Figure 4D). Both structures markedly contrast to the G-less 26 ITC that failed to escape promoter in the absence of such EC in front of it (Figure 3). Although previous biochemical data suggest that the 8-9-bp DNA-RNA hybrid is the minimum requirement for the formation of pol II-DNA-RNA complex (Kireeva et al., 2000), we posit the interaction between pol II and the downstream DNA confers additional stability to keep the polymerase in register at the 3’-end of RNA. Before this transition, TFIIH continuously draws dozens of DNA base pairs by scrunching (Fazal et al., 2015a; Tomko et al., 2017), on which initially transcribing pol II (as well as TSS scanning pol II) has a degree of freedom of lateral movement along the template. Therefore the stable accommodation of the downstream DNA in the pol II downstream cleft, which marks the end of the requirement for TFIIH, represents an irreversible critical transition from initiation to elongation. This explains why the G-less 27 complex stalled at +27 released transcripts by extensive backtracking of pol II upon removal of ATP during gradient sedimentation, whereas the G-less 49 complex completely retained transcripts of similar length (~25 nt or shorter) in the trailing ITC (Fujiwara et al., 2019). It should be noted that even after the entry of downstream DNA into the pol II cleft, in some ITCs, GTFs remained bound to pol II (Figure 4D), which may further require additional elongation factors such as the capping enzyme and/or Spt4/5 to displace TFIIE from pol II (Fujiwara et al., 2019).

Lastly, when the structures of pol II (EC)-DSIF-NELF complex (Vos et al., 2018b), in a canonical form of prompter-proximal paused pol II in mammalian systems, is aligned with the leading EC of the EC+EC complex, there is no steric clash of the trailing EC with NELF, but partial clash with DSIF. Also Mediator, which serves a critical role in promoter escape (Jeronimo and Robert, 2014; Takahashi et al., 2011; Wong et al., 2014), has no steric clash with the leading EC when the trailing EC is aligned with the PIC-Mediator complex (Robinson et al., 2016b). It will be of great interest to pursue possible positive/negative regulations of promoter escape by such general factors and determine the underling structural basis.

2.7 Main Figures

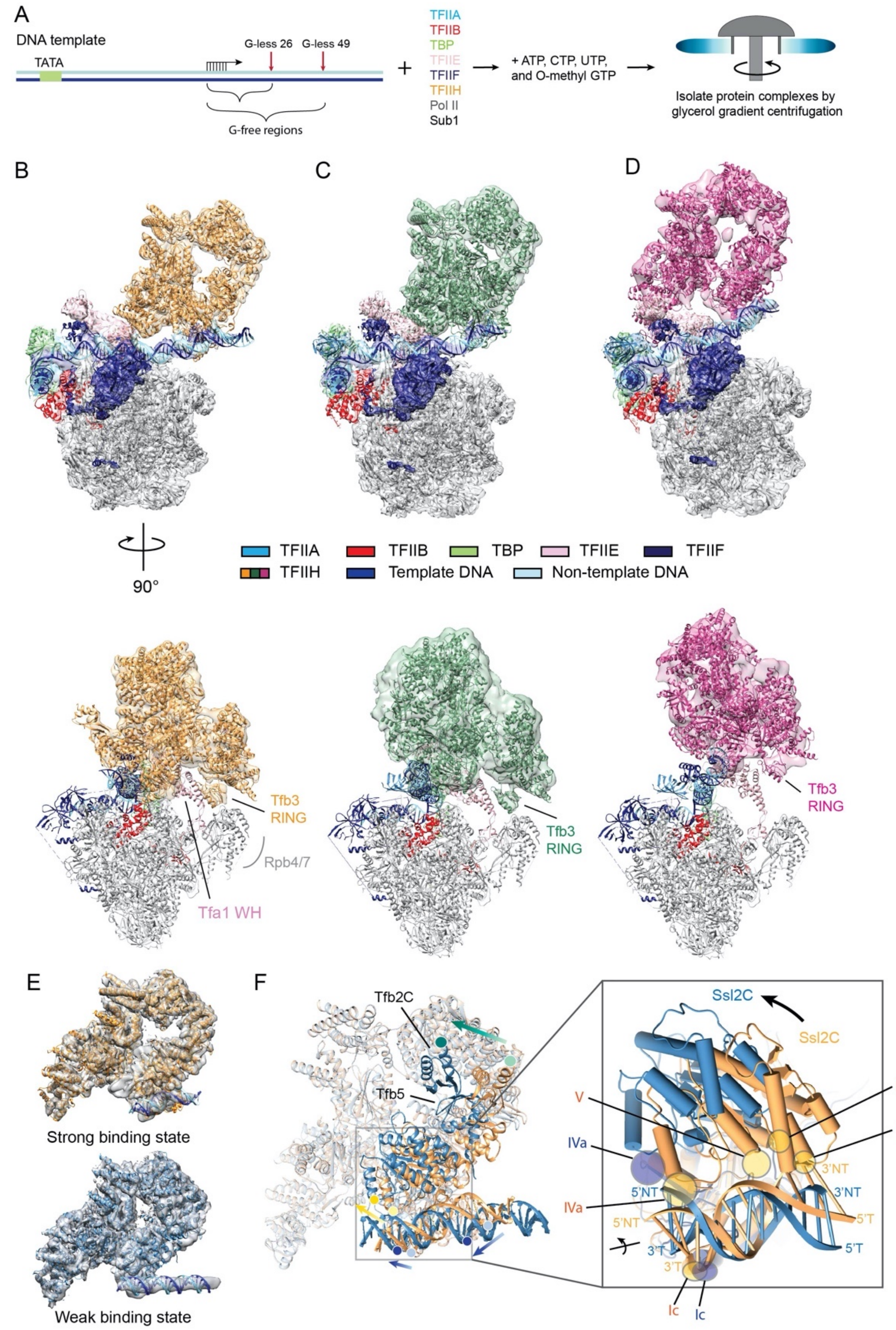


Figure 1. Structures of three forms of pre-initiation complexes. **(A)** Schematic representation for isolation of the transcription complexes analyzed in this study. **(B-D)** Composite density maps and the models of three forms of PIC on G-less 26 DNA template; PIC1 (B), PIC2 (C), and PIC3 (D). Side view (top) and front view (bottom) are shown. Same colors are used throughout the manuscript unless otherwise noted: TFIIA (steel blue), TFIIB (red), TBP (light green), TFIIE (pink), TFIIF (dark blue), TFIIH (orange, green, or magenta), template DNA (blue), non-template DNA (sky blue). Tfb3 interacts with Rpb4/7 in PIC1 and PIC2, but dissociates in PIC3. **(E)** Composite density maps and models of strong (top) and weak (bottom) binding states of TFIIH, with corresponding models in orange and steel blue, respectively. **(F)** Comparison of TFIIH and downstream DNA in the strong and weak binding states. Same residues of Tfb2C, Ssl2C, and DNA in the two states are marked by circular dots and the directions of movements are indicated by green, yellow, and blue arrows, respectively. Inset, Ssl2–DNA interactions suggested by the model. The five DNA binding motifs (Ic, IVa, IV, V, Vb) are indicated.

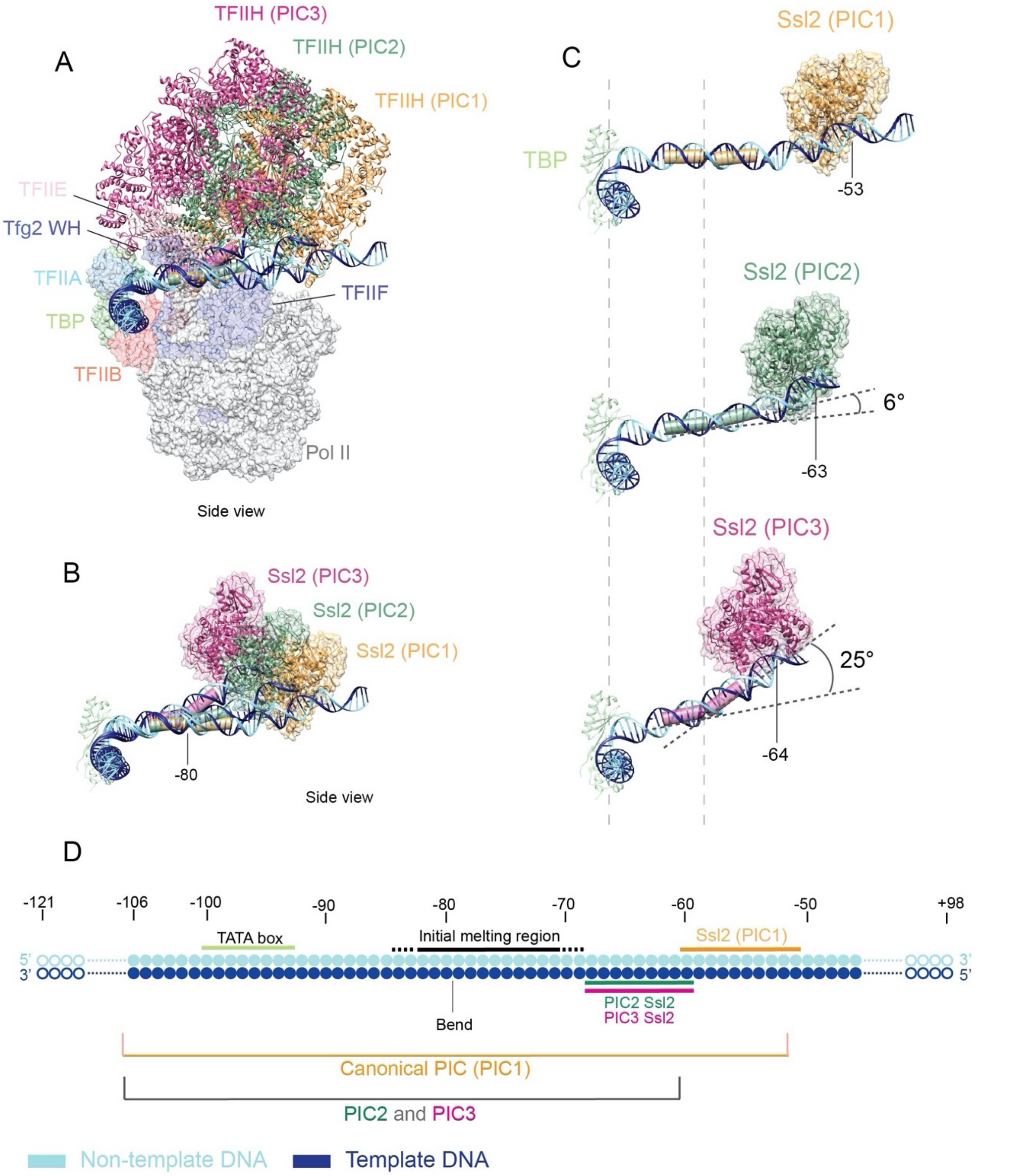


Figure 2. Distortion of promoter DNA in PIC1-3. Coloring as in Figure 1. **(A)** Comparison of locations of TFIIH and DNA in PIC1-3 relative to Pol II. TFIIH shifts ~ 20 Å and ~30 Å in PIC2 and PIC3, respectively, relative to that in PIC1. **(B)** Paths of promoter DNA. Ssl2 contacting DNA is shown. **(C)** Ssl2 binds ~47 bp and ~37 bp downstream of the TATA box in PIC1 and PIC2/PIC3, respectively. DNA bends by ~6° and ~25° in PIC2 and PIC3, respectively. Dashed lines indicate positions of the bend at ~–80 and the TATA box at –100. The numbering is relative to the TSS. **(D)** A schematic showing PIC1-3 occupancy on the promoter DNA.

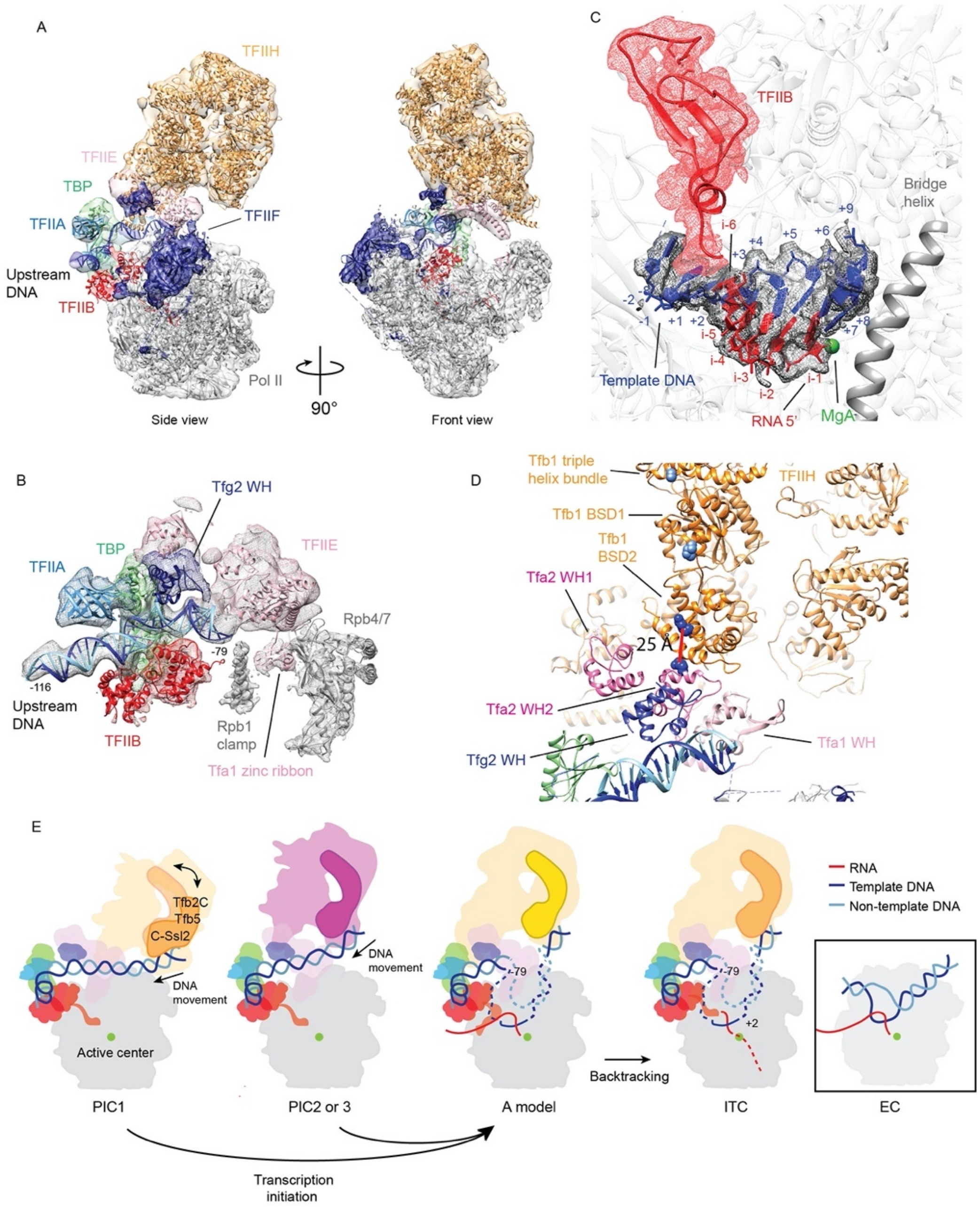


Figure 3. Structure of ITC with the G-less 26 template. (A) Cryo-EM map and a corresponding model of the ITC. **(B)** EM density map shows the DNA-RNA hybrid in the pol II active center in contact with TFIIB. **(C)** EM density shows the upstream DNA bound to TFIIA, TBP, TFIIB (cyclin domains), TFIIE (Tfa1 and Tfa2 WH1 and WH2 domains), and TFIIF (Tfg2 WH domains). **(D)** TFIIE-TFIIH interactions in the ITC. Tfb1 BSD2 domain (orange) is in contact with Tfa2 WH2 (hot pink). Red line indicates a cross-link between K268 of Tfb1 and K194 of Tfa2. K581 of the triple helix bundle and K179 of the BSD1 (sky blue) forms cross-links with the C-terminal region of Tfa1. **(E)** Schematic of transition from PICs to ITC. Inset, schematic of EC viewed from the same orientation as PIC1-3 and ITC.

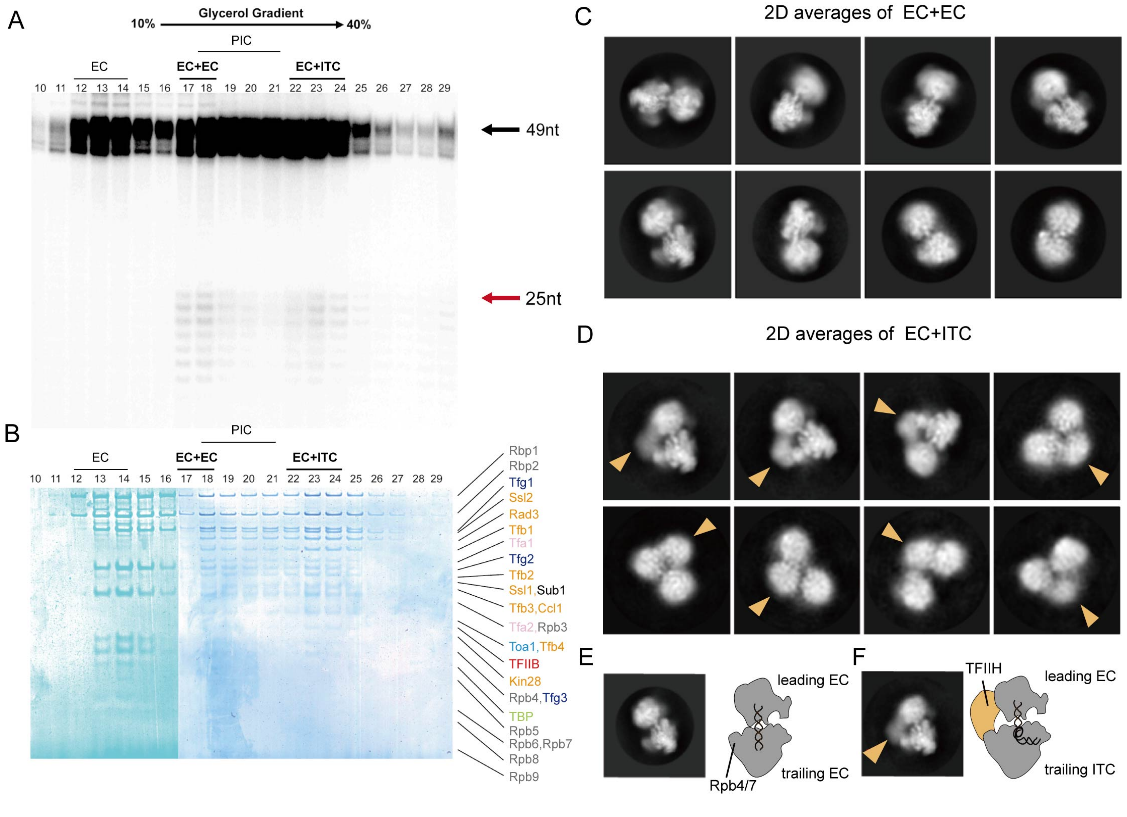


Figure 4. G-less 49 complexes contain two major post-initiation complexes, EC+EC and EC+ITC. (A-B) Transcription complexes with the G-less 49 template were subjected to glycerol gradient sedimentation. RNA analysis of the fractions by denaturing Urea-PAGE gel (A) and protein analysis of the fractions by SDS-PAGE gel (B) revealed EC+EC (fractions 17-18) and EC+ITC (fractions 22-24). 49-nt and 25-nt transcripts from the first and the second rounds of transcription are indicated by black and red arrows. Note that both complexes have some contamination of PICs (fractions 18-21). **(C)** Eight representative reference-free 2D class averages of EC+EC. **(D)** Same as (C) but for EC+ITC. A large density attributable to TFIIH (indicated by orange arrow heads) is located between EC and ITC. **(E)** A representative 2D class average of EC+EC, with a schematic model, showing the two well-featured densities corresponding to the leading EC and trailing EC, respectively. The density of DNA bridging two polymerases is clearly discernable. **(F)** Same as (E) but for EC+ITC, showing a large density attributable to TFIIH (orange) compared with 2D class averages of EC + EC. The density of DNA is clearly discernable as in (E).

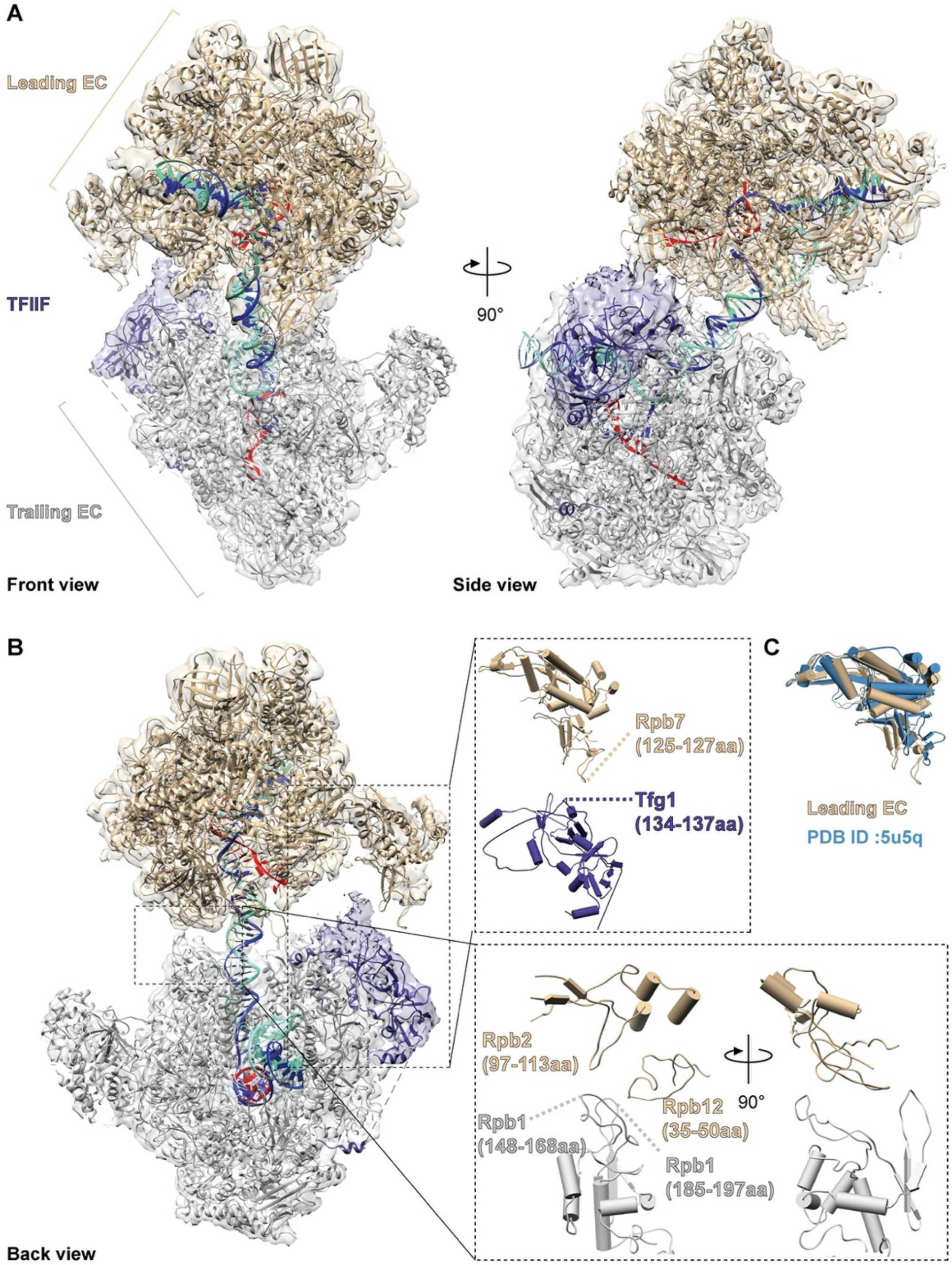


Figure 5. The structure of ECs colliding head-to-end (EC+EC). (A) Front (left) and side views (right) of the cryo-EM reconstruction with the model. **(B)** Interactions between two ECs viewed from the back. The Rpb4/7 of the leading EC contacts TFIIF associated with the trailing EC, while Rpb2 and Rpb12 of the leading EC contacts the Rpb1 of the trailing EC. The leading EC, the trailing EC, TFIIF, template DNA, non-template DNA and RNA are colored in tan, gray, navy blue, blue, aquamarine and red, respectively.

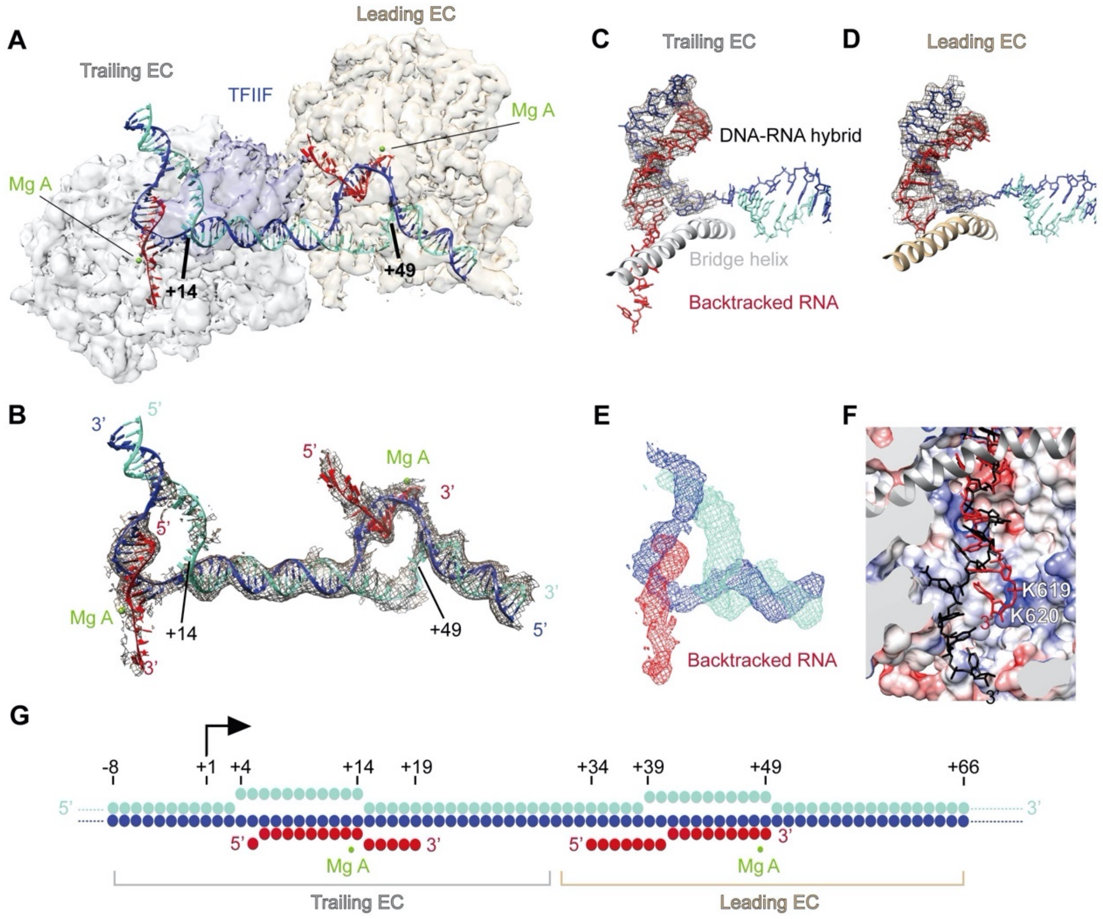


Figure 6. Nucleic acids of ECs colliding head-to-end (EC+EC). (A) Unsharpened cryo-EM densities of the trailing EC (gray), the leading EC (tan) and TFIIF (navy blue) are shown as surface, contoured at 2.07 sigma. Template DNA, non-template DNA and RNA are colored in blue, aquamarine and red throughout. Mg A in active center is shown as sphere and colored in green. **(B)** Composite cryo-EM density of nucleic acids. Unsharpened cryo-EM densities are shown as mesh (space gray), contoured at level 2.07 sigma. **(C)** The DNA-RNA hybrid of the trailing EC. Sharpened cryo-EM densities of the DNA-RNA hybrid of the trailing EC are shown as mesh (space gray), contoured at level 4.1 sigma. The bridge helix is shown in gray. **(D)** Same as (C) but for the leading EC, contoured at 3.86 sigma. **(E)** Cryo-EM density of nucleic acids of the trailing EC showing the backtracked RNA (red). Unsharpened cryo-EM densities of nucleic acids are shown as mesh (RNA, red; template DNA, blue; Non-template DNA, aquamarine), contoured at level 1.6 sigma. **(F)** Superposition of the backtracked RNA of the trailing EC (red) with the backtrack site of the arrested Pol II previously determined by crystallography (PDB:3PO2, black). Backtracked RNA in trailing EC shifts towards a positive charged patch that consists of K619 and K620 of Rpb1 in funnel. **(G)** Schematic of nucleic acids. Modelled nucleotides of promoter template are shown with filled circles. TSS (+1) is indicated by black arrow.

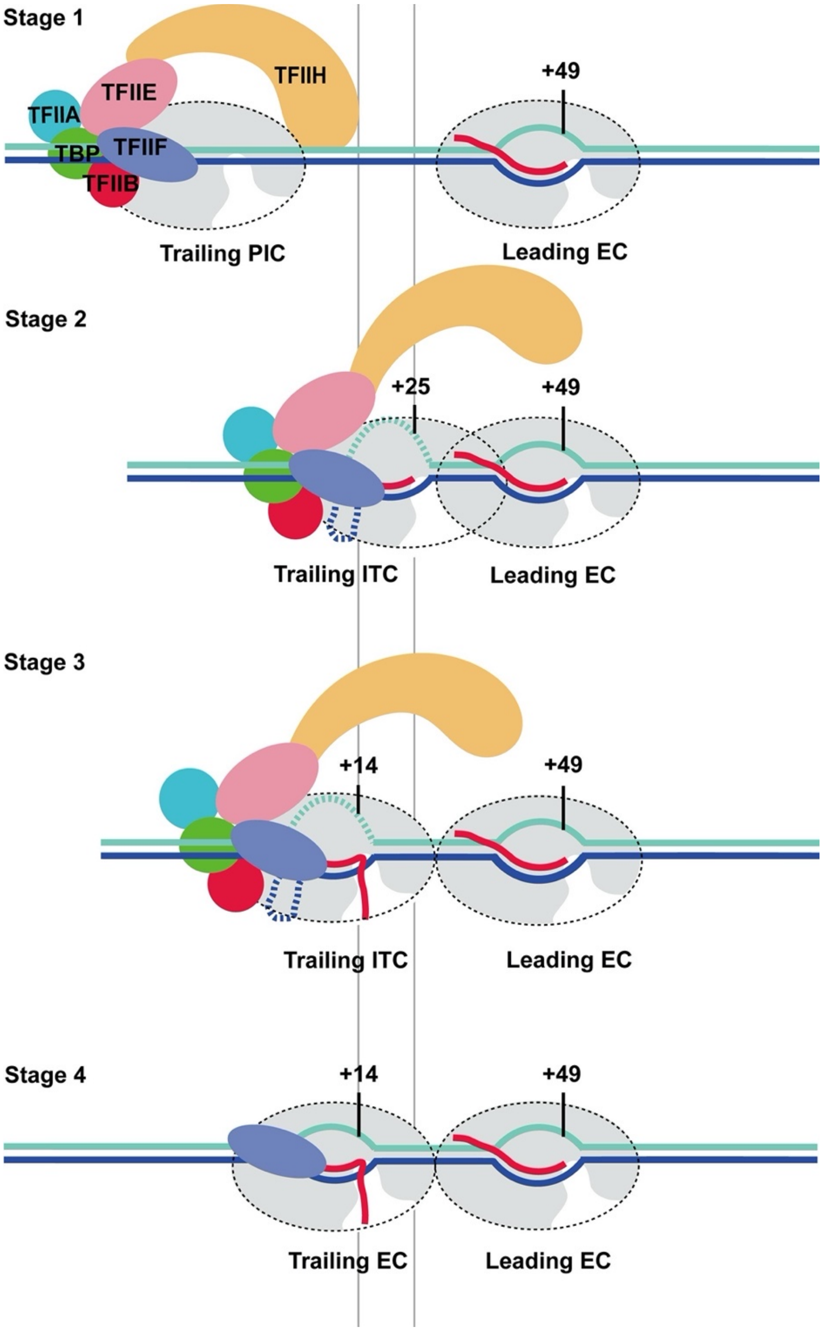
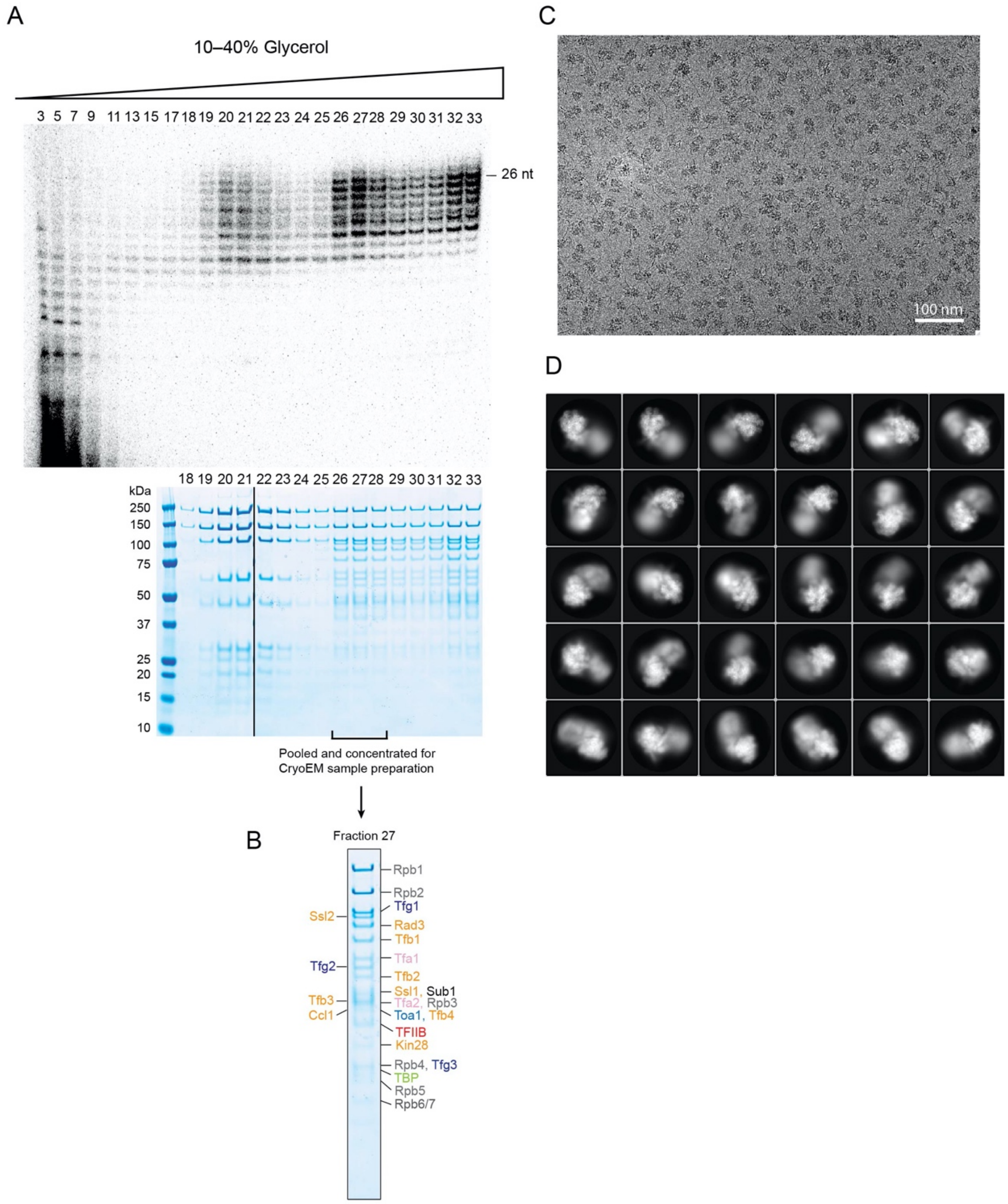
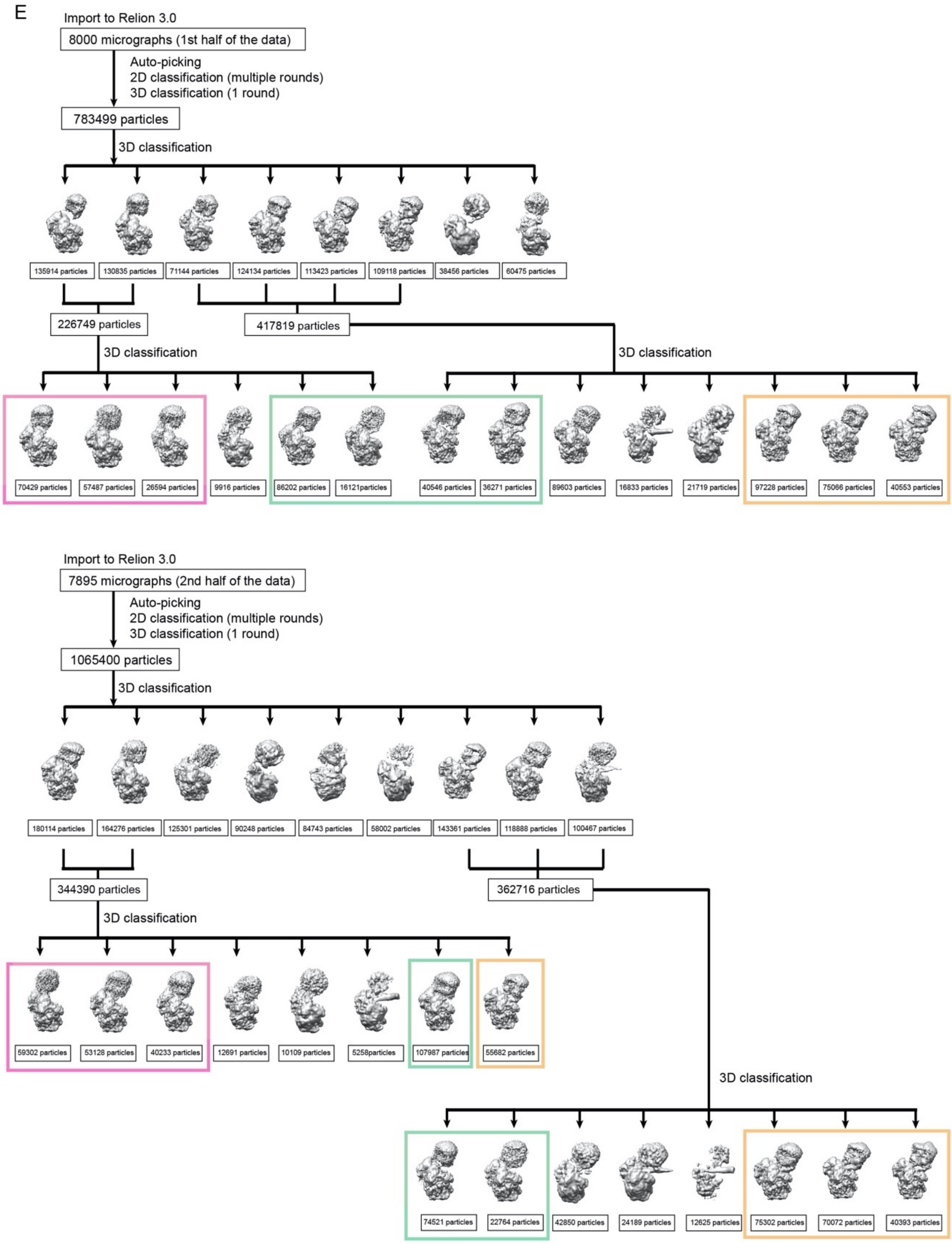
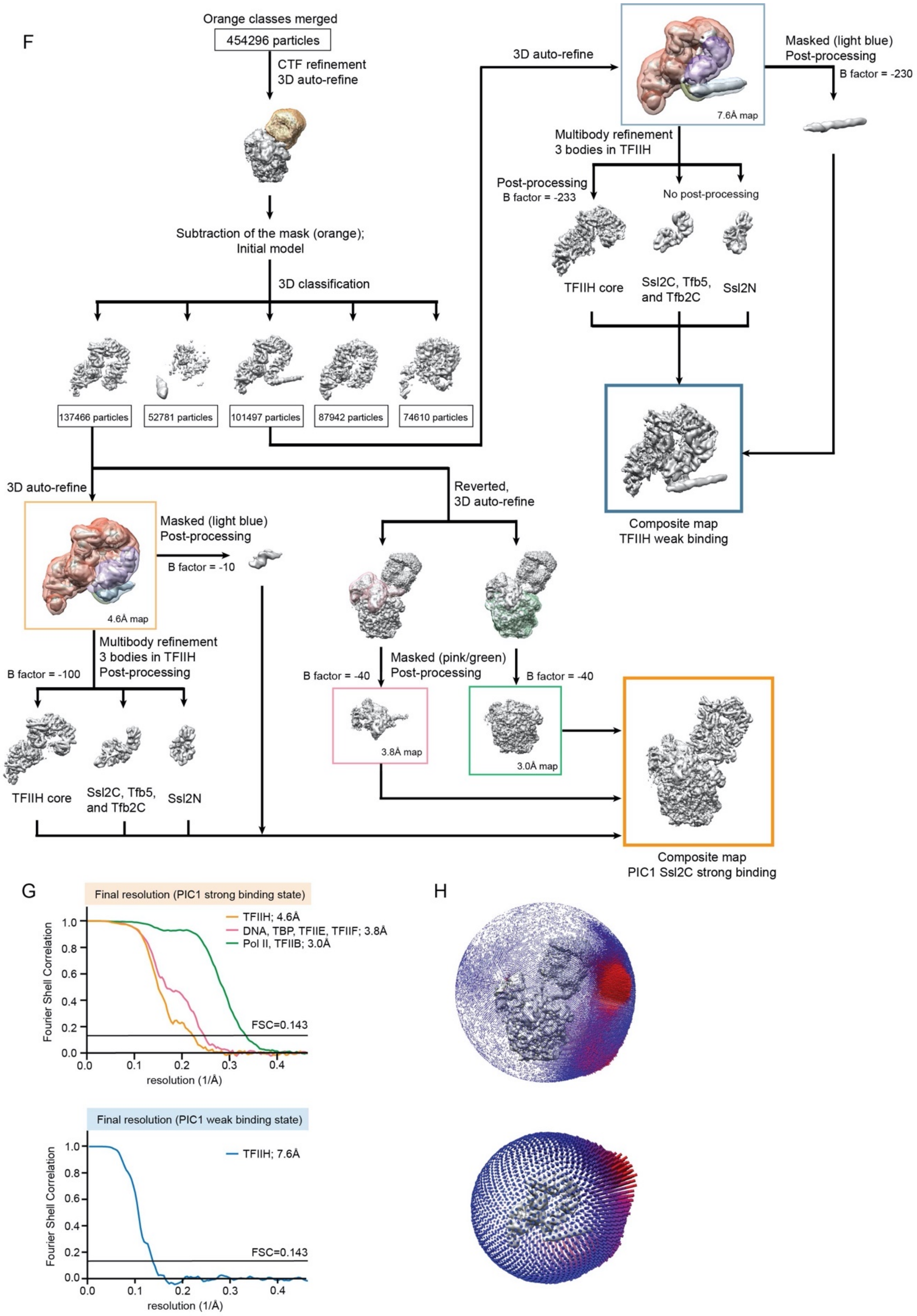


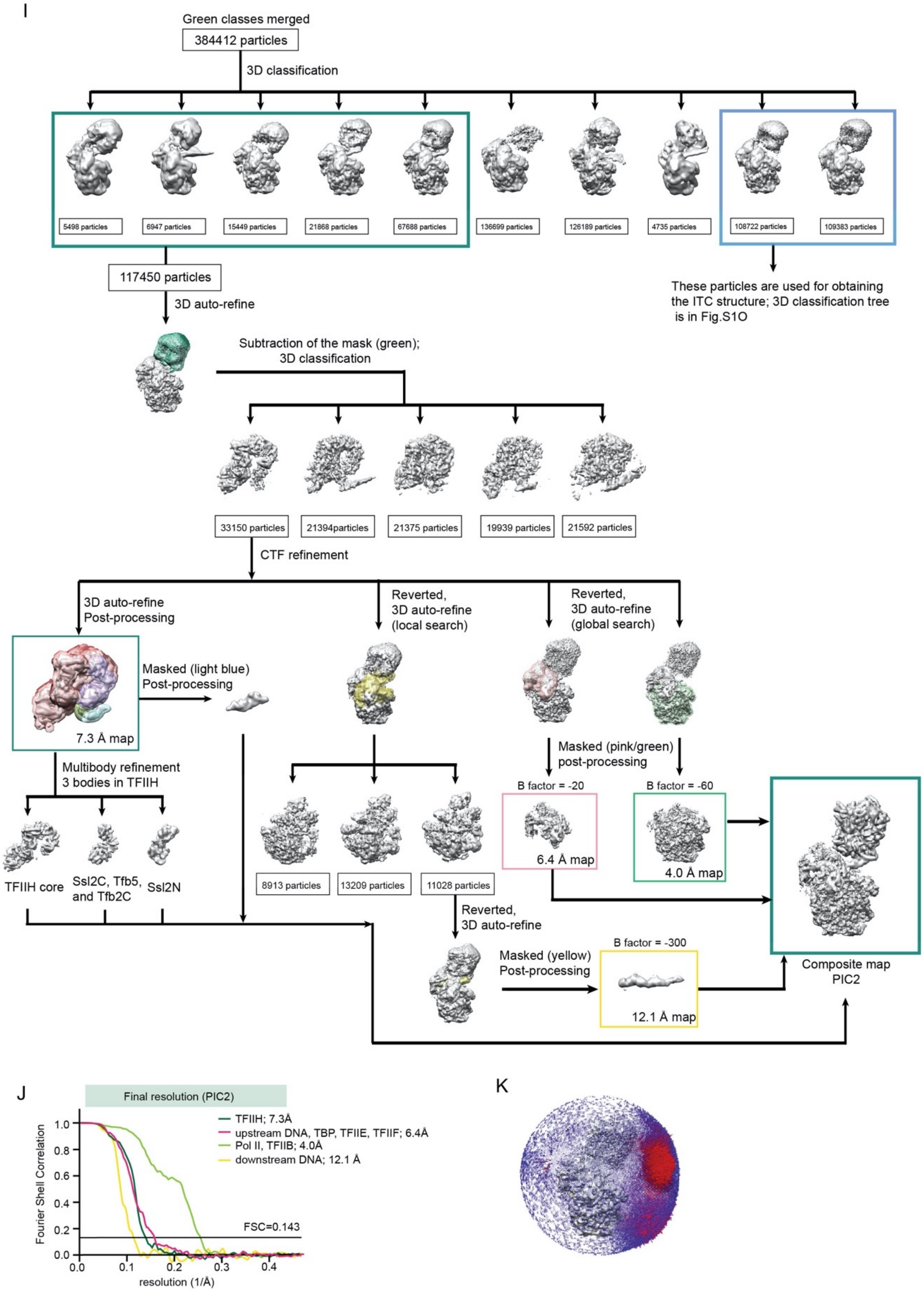
Figure 7. Schematic of promoter escape facilitated by the leading EC. A PIC is assembled on promoter, while an EC is stalled at promoter proximal regions on the G-less 49 template (Stage 1). PIC reels ~80 bp of downstream DNA by scrunching and initiates transcription, while retaining fixed upstream contact within the complex. Initially-transcribing pol II in the trailing ITC encounters the leading EC stalled at +49 and occludes TFIIH binding (stage 2). The trailing EC is backtracked by ~11bp and arrested at +14 (EC+ITC, Stage 3), followed by dissociation of GTFs and bubble collapse (EC+EC, Stage 4).

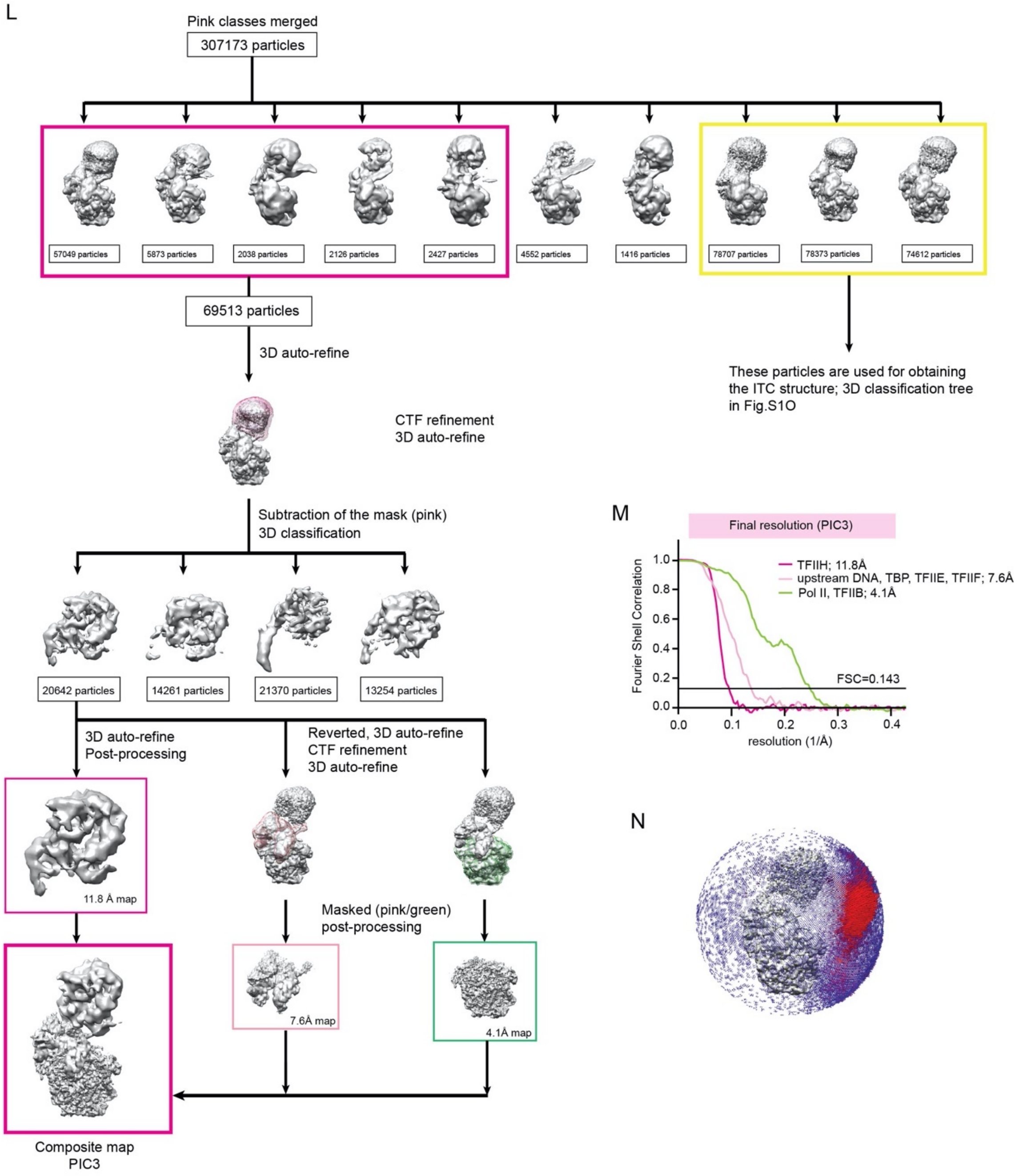
2.8 Supplemental Figures

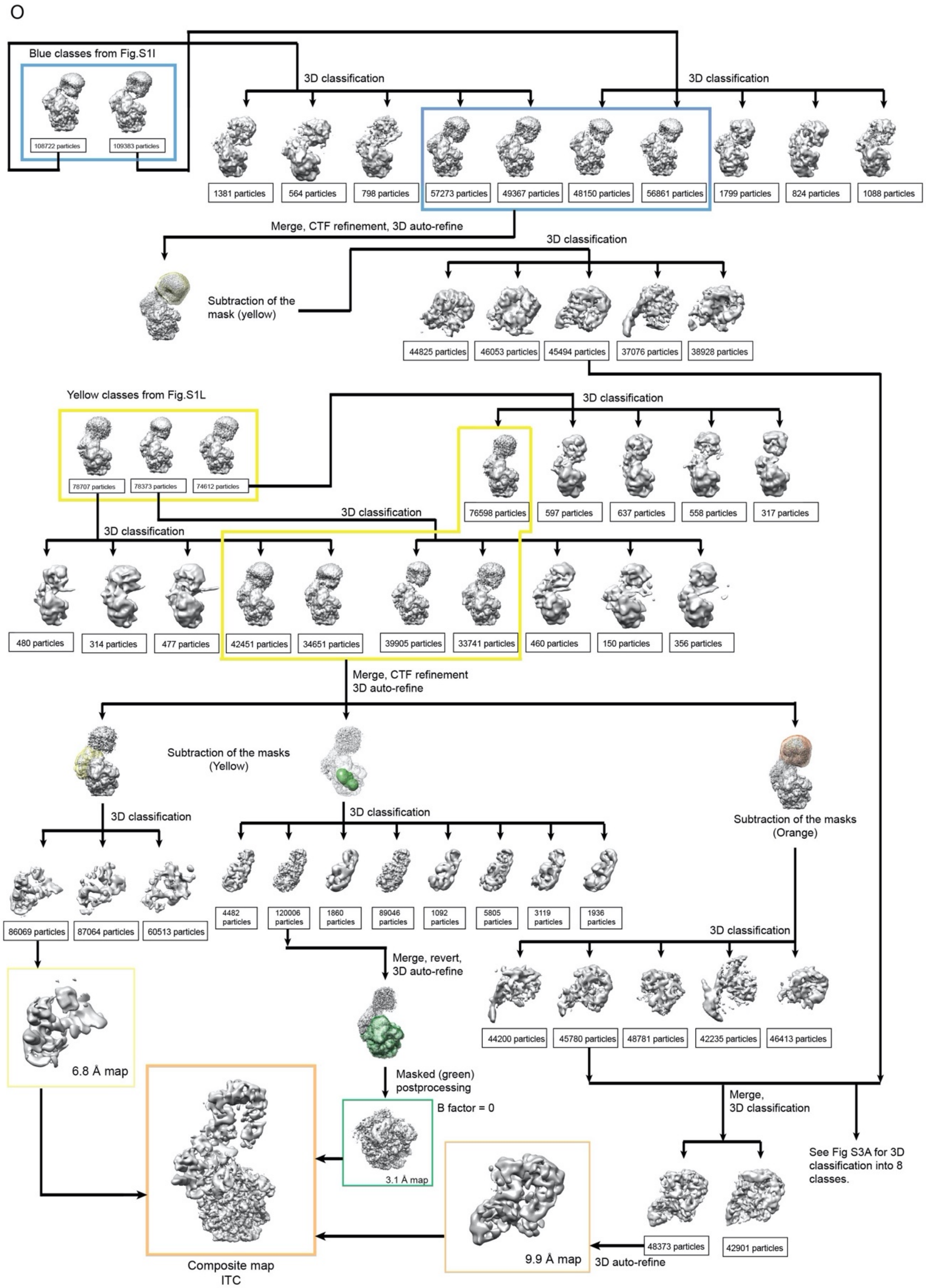


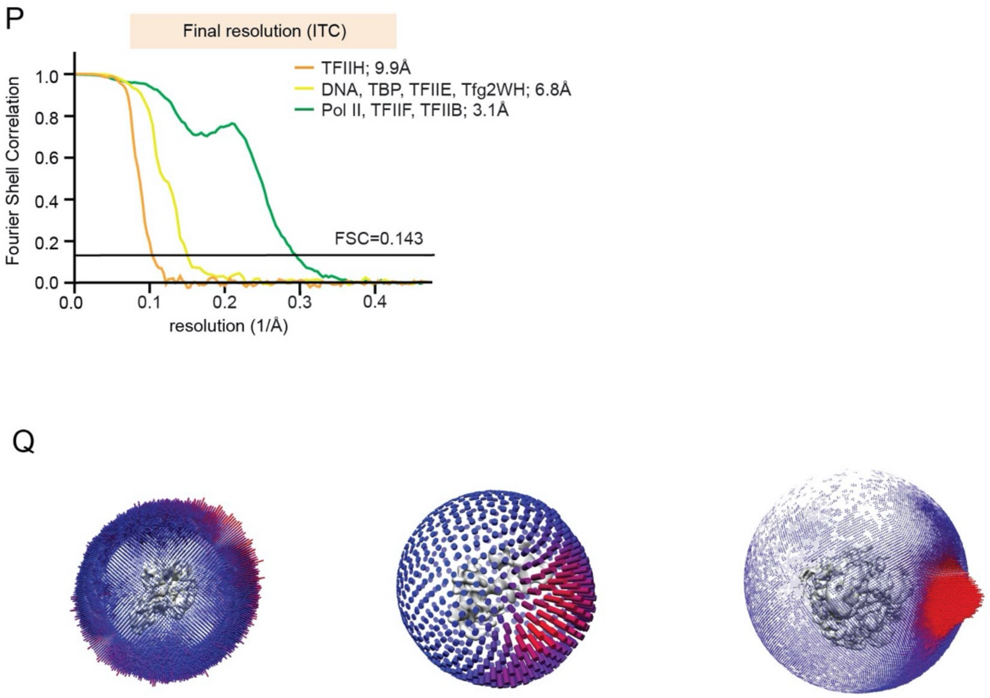

E
Import to Relion 3.0
8000 micrographs (1st half of the data)
Auto-picking
2D classification (multiple rounds)
3D classification (1 round)
783499 particles
3D classification
135914 particles
130835 particles
71144 particles
124134 particles
113423 particles
109118 particles
38456 particles
60475 particles
226749 particles
417819 particles
3D classification
3D classification
70429 particles
57487 particles
26594 particles
9916 particles
86202 particles
16121particles
40546 particles
36271 particles
89603 particles
16833 particles
21719 particles
97228 particles
75066 particles
40553 particles
Import to Relion 3.0
7895 micrographs (2nd half of the data)
Auto-picking
2D classification (multiple rounds)
3D classification (1 round)
1065400 particles
3D classification
180114 particles
164276 particles
125301 particles
90248 particles
84743 particles
58002 particles
143361 particles
118888 particles
100467 particles
344390 particles
362716 particles
3D classification
59302 particles
53128 particles
40233 particles
12691 particles
10109 particles
5258particles
107987 particles
55682 particles
3D classification
74521 particles
22764 particles
42850 particles
24189 particles
12625 particles
75302 particles
70072 particles
40393 particles


F
Orange classes merged
454296 particles
CTF refinement
3D auto-refine
Subtraction of the mask (orange);
Initial model
3D classification
137466 particles
52781 particles
101497 particles
87942 particles
74610 particles
3D auto-refine
4.6Å map
Masked (light blue)
Post-processing
B factor = -10
Multibody refinement
3 bodies in TFIIH
Post-processing
B factor = -100
TFIIH core
Ssl2C, Tfb5, and Tfb2C
Ssl2N
Reverted, 3D auto-refine
Masked (pink/green)
Post-processing
B factor = -40
B factor = -40
3.8Å map
3.0Å map
Composite map
PIC1 Ssl2C strong binding
3D auto-refine
7.6Å map
Masked (light blue)
Post-processing
B factor = -230
Multibody refinement
3 bodies in TFIIH
Post-processing
B factor = -233
No post-processing
TFIIH core
Ssl2C, Tfb5, and Tfb2C
Ssl2N
Composite map
TFIIH weak binding
G
Final resolution (PIC1 strong binding state)
TFIIH; 4.6Å
DNA, TBP, TFIIE, TFIIF; 3.8Å
Pol II, TFIIB; 3.0Å
FSC=0.143
Fourier Shell Correlation
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.1
0.2
0.3
0.4
resolution (1/Å)
Final resolution (PIC1 weak binding state)
TFIIH; 7.6Å
FSC=0.143
Fourier Shell Correlation
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.1
0.2
0.3
0.4
resolution (1/Å)
H


I
Green classes merged
384412 particles
3D classification
5498 particles
6947 particles
15449 particles
21868 particles
67688 particles
136699 particles
126189 particles
4735 particles
108722 particles
109383 particles
117450 particles
3D auto-refine
These particles are used for obtaining the ITC structure; 3D classification tree is in Fig.S1O
Subtraction of the mask (green); 3D classification
33150 particles
21394particles
21375 particles
19939 particles
21592 particles
CTF refinement
3D auto-refine
Post-processing
Reverted, 3D auto-refine (local search)
Reverted, 3D auto-refine (global search)
7.3 Å map
Masked (light blue) Post-processing
Masked (pink/green) post-processing
B factor = -20
B factor = -60
6.4 Å map
4.0 Å map
Multibody refinement
3 bodies in TFIIH
TFIIH core
Ssl2C, Tfb5, and Tfb2C
Ssl2N
8913 particles
13209 particles
11028 particles
Reverted, 3D auto-refine
Masked (yellow) Post-processing
B factor = -300
12.1 Å map
Composite map PIC2
J
Final resolution (PIC2)
Fourier Shell Correlation
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.1
0.2
0.3
0.4
resolution (1/Å)
TFIIH; 7.3Å
upstream DNA, TBP, TFIIE, TFIIF; 6.4Å
Pol II, TFIIB; 4.0Å
downstream DNA; 12.1 Å
FSC=0.143
K


L
Pink classes merged
307173 particles
57049 particles
5873 particles
2038 particles
2126 particles
2427 particles
4552 particles
1416 particles
78707 particles
78373 particles
74612 particles
69513 particles
3D auto-refine
These particles are used for obtaining the ITC structure; 3D classification tree in Fig.S1O
CTF refinement
3D auto-refine
Subtraction of the mask (pink)
3D classification
20642 particles
14261 particles
21370 particles
13254 particles
3D auto-refine
Post-processing
Reverted, 3D auto-refine
CTF refinement
3D auto-refine
11.8 Å map
Masked (pink/green)
post-processing
7.6Å map
4.1Å map
Composite map
PIC3
M
Final resolution (PIC3)
Fourier Shell Correlation
1.0
0.8
0.6
0.4
0.2
0.0
TFIIH; 11.8Å
upstream DNA, TBP, TFIIE, TFIIF; 7.6Å
Pol II, TFIIB; 4.1Å
FSC=0.143
0.0
0.1
0.2
0.3
0.4
resolution (1/Å)
N


O
Blue classes from Fig.S1I
108722 particles
109383 particles
3D classification
1381 particles
564 particles
798 particles
57273 particles
49367 particles
48150 particles
56861 particles
3D classification
1799 particles
824 particles
1088 particles
Merge, CTF refinement, 3D auto-refine
3D classification
Subtraction of the mask (yellow)
44825 particles
46053 particles
45494 particles
37076 particles
38928 particles
Yellow classes from Fig.S1L
78707 particles
78373 particles
74612 particles
3D classification
76598 particles
597 particles
637 particles
558 particles
317 particles
3D classification
3D classification
480 particles
314 particles
477 particles
42451 particles
34651 particles
39905 particles
33741 particles
460 particles
150 particles
356 particles
Merge, CTF refinement
3D auto-refine
Subtraction of the masks (Yellow)
Subtraction of the masks (Orange)
3D classification
3D classification
86069 particles
87064 particles
60513 particles
4482 particles
120006 particles
1860 particles
89046 particles
1092 particles
5805 particles
3119 particles
1936 particles
Merge, revert, 3D auto-refine
3D classification
44200 particles
45780 particles
48781 particles
42235 particles
46413 particles
6.8 Å map
Masked (green) postprocessing
B factor = 0
3.1 Å map
Merge, 3D classification
See Fig S3A for 3D classification into 8 classes.
48373 particles
42901 particles
9.9 Å map
3D auto-refine
Composite map ITC


P
Final resolution (ITC)
TFIIH; 9.9Å
DNA, TBP, TFIIE, Tfg2WH; 6.8Å
Pol II, TFIIF, TFIIB; 3.1Å
Fourier Shell Correlation
1.0
0.8
0.6
0.4
0.2
0.0
FSC=0.143
0.0
0.1
0.2
0.3
0.4
resolution (1/Å)
Q

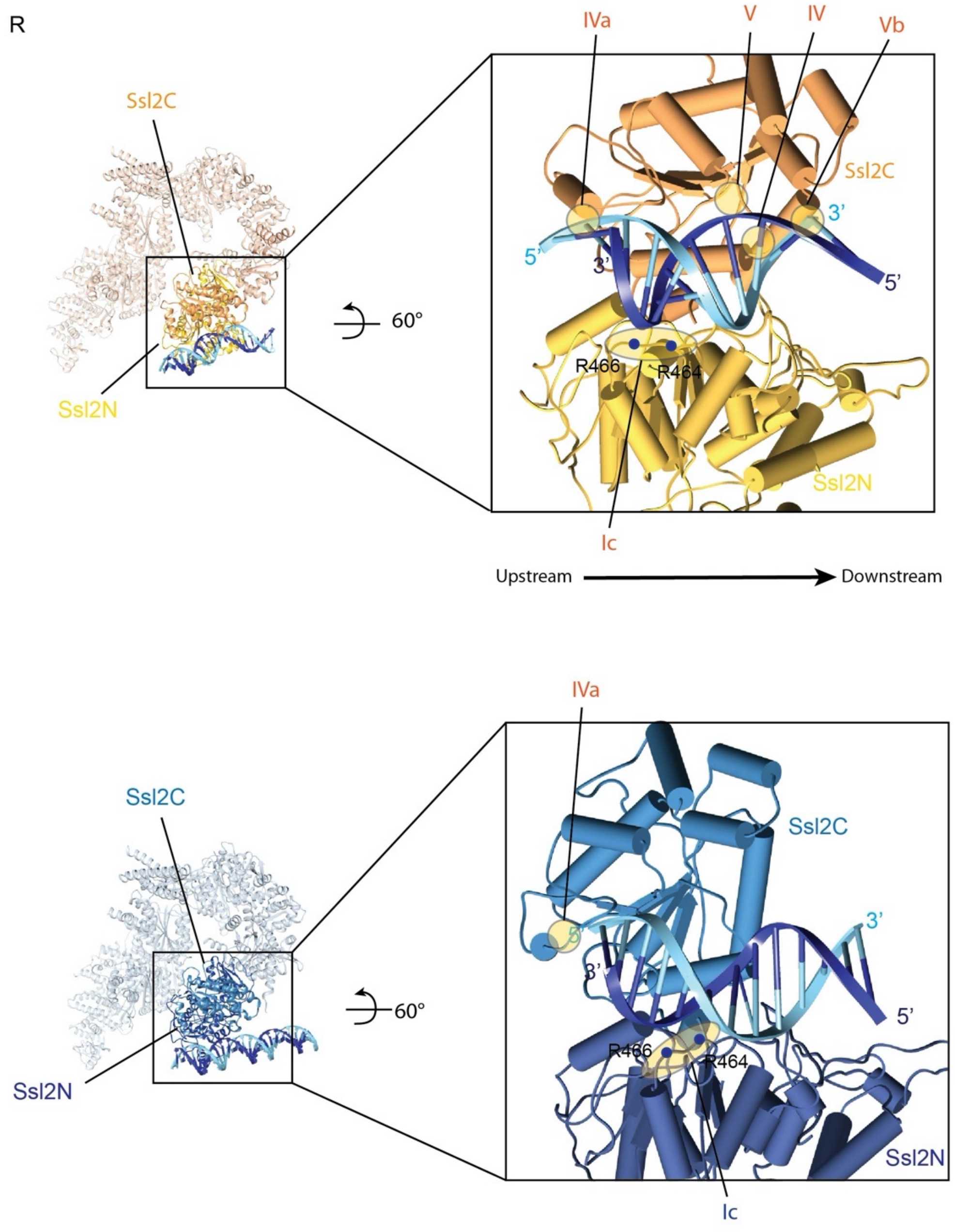


Figure S1. Preparation of *S. cerevisiae* ITC, related to Figure 1. **(A)** Transcription complexes formed on the G-less 26 template were separated by 10-40% glycerol gradient sedimentation. The gradient was fractionated into 40 fractions, each of which contained ~100 µL. The presence of RNA (top) and proteins (bottom) in the fractions was confirmed by urea denaturing gel and by SDS-PAGE, respectively. **(B)** SDS-PAGE analysis of *S. cerevisiae* ITC sample after isolation. **(C)** A representative cryo-EM image. **(D)** Representative 2D class averages of PIC1-3 and ITC from RELION 3.1. **(E)** Early

cryo-electron microscopy processing pipeline. A total of 15895 images was collected. The data were divided into two sets and processed with RELION 3.0 and 3.1. After a few rounds of 3D classification, maps were manually inspected and grouped into three based on the location of TFIIH relative to pol II. Particles that are used for further analysis and to reconstruct PIC1, PIC2, and PIC3 maps are indicated with orange, green, and pink squares, respectively. The ITC map was reconstructed from a subset of particles in green and pink squares (see also Fig. S1I, L, and O). **(F-H)** Cryo-electron microscopy processing pipeline (F), map resolution (G), and angular distribution (H) for the PIC1 strong and weak binding states. **(I-K)** Cryo-electron microscopy processing pipeline (I), map resolution (J), and angular distribution (K) for the PIC2. **(L-N)** Cryo-electron microscopy processing pipeline (L), map resolution (M), and angular distribution (N) for the PIC3. **(O)** Cryo-electron microscopy processing pipeline for the ITC. **(P-Q)** Map resolution (P) and angular distribution (Q) for the ITC. **(R)** Comparison of interactions between Ssl2 and downstream DNA in PIC1 strong (left) and weak (right) binding states. The C- and N-terminal domains of Ssl2 in strong binding states is colored in orange and yellow, respectively. The C- and N-terminal domains of Ssl2 in weak binding states is colored in blue and light blue, respectively. Light yellow circles indicate interaction sites between Ssl2 and downstream DNA. The motifs (IVa, V, IV, Vb, Ic) are based on the previous studies (Fairman-Williams et al., 2010; Schilbach et al., 2017). The N-terminal domain of Ssl2 interacts with DNA strand in strong binding state whereas it relocates near the minor grove in weak binding state. Additionally, three (V, IV, Vb) out of four interaction sites between the C-terminal domain of Ssl2 and DNA dissociate in the weak binding state.

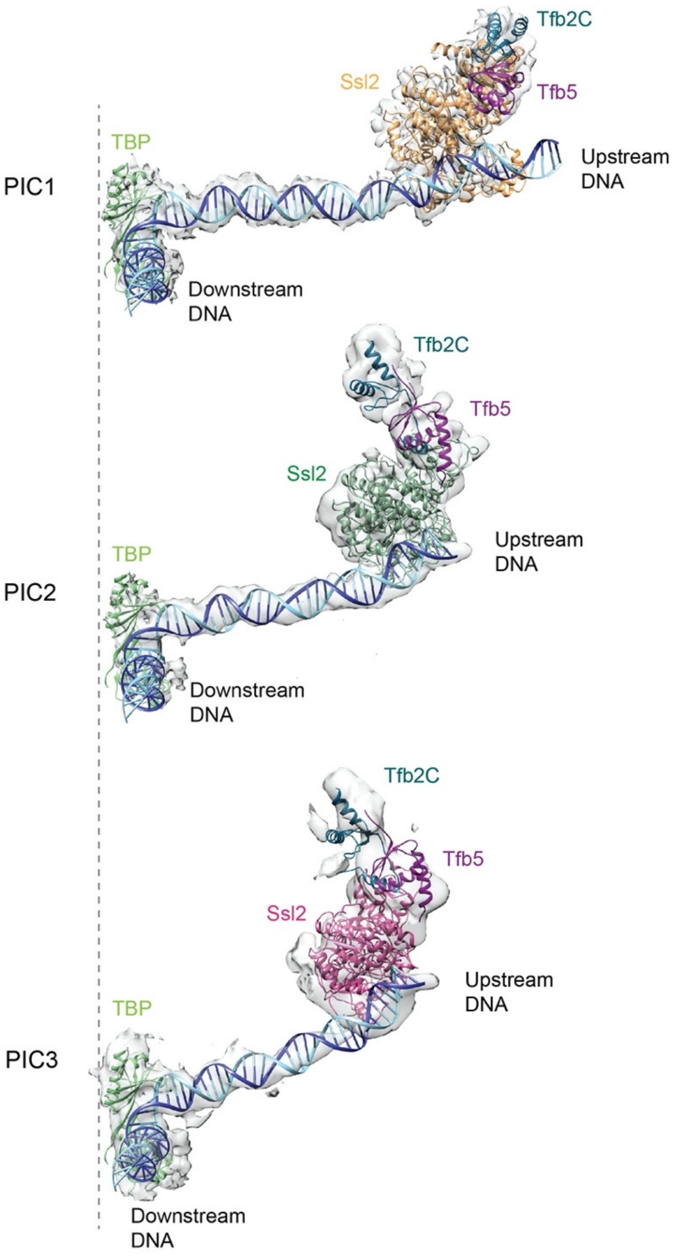


Figure S2. The EM density and the fitted model of DNA, Ssl2, Tfb5, Tfb2C, and TBP in PIC1 (top), PIC2 (middle), and PIC3 (bottom), related to Figure 2. TBP is aligned for depiction. Coloring as in Figure 1 except for Tfb5 in purple and Tfb2C in dark cyan.

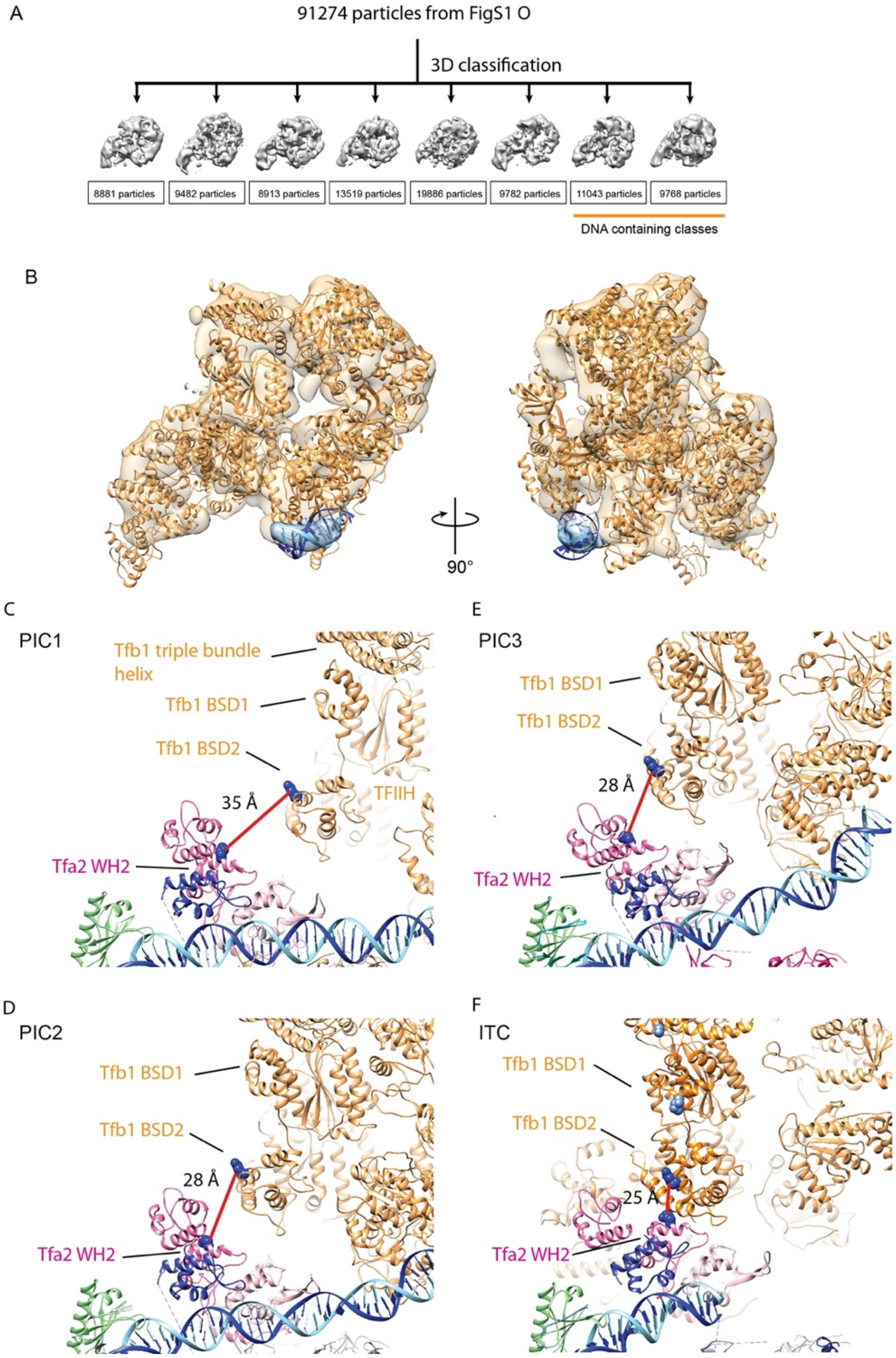
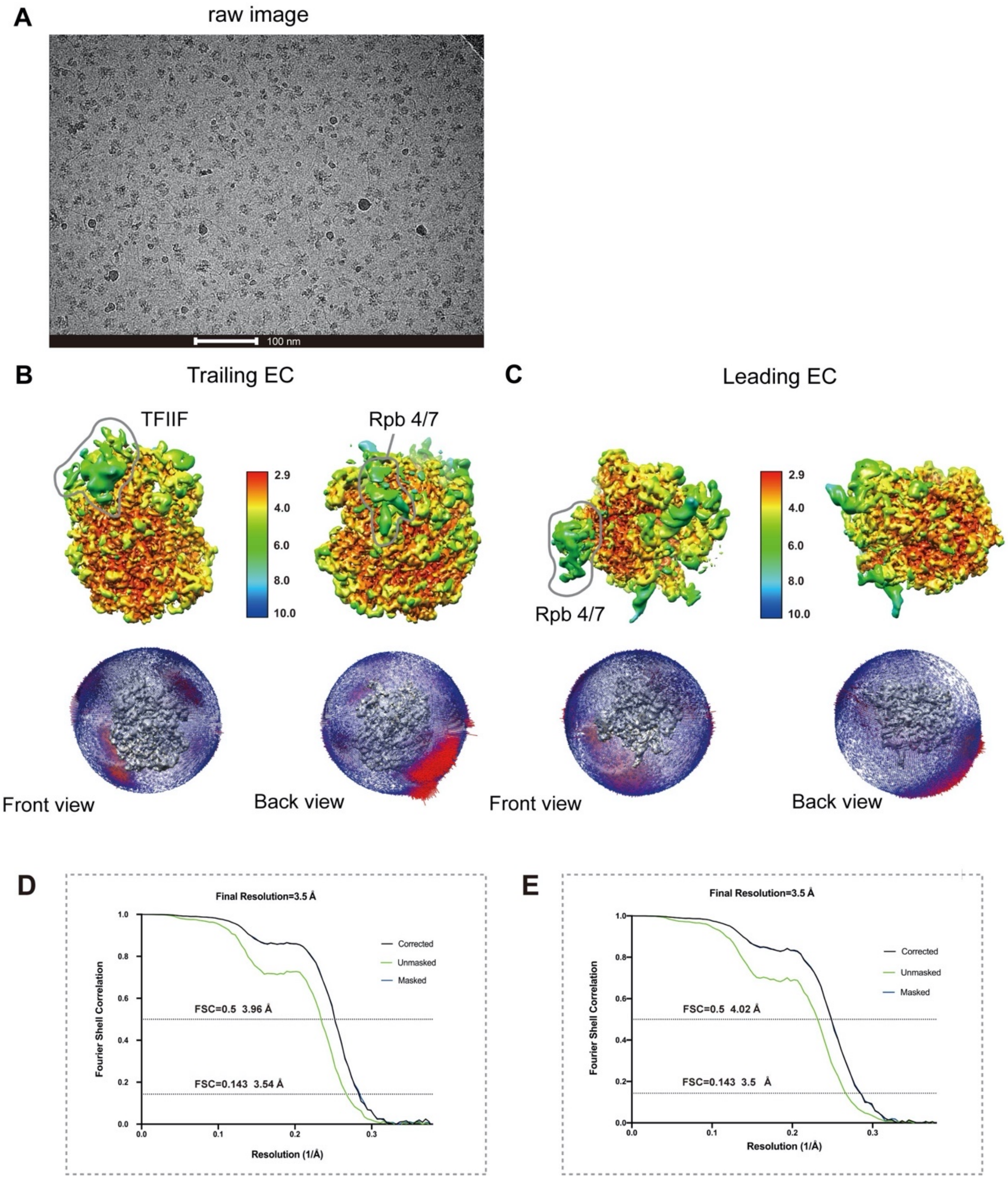


Figure S3. Focused classification of TFIIH in the ITC and a comparison of distances between TFIIH and TFIIE in PIC1-3 and ITC, related to Figure 3. **(A)** Focused classification of TFIIH into 8 classes shows two TFIIH density maps containing dsDNA along Ssl2 C- and N terminal domains. **(B)** Representative TFIIH map containing the dsDNA density. **(C-F)** Distances between Tfb1 (TFIIH)-Tfa2 WH2 (TFIIE) in PIC1(C), PIC2 (D), PIC3 (E) and ITC (F). Red line indicates a cross-link between K268 of Tfb1 and K194 of Tfa2, observed in the PIC lacking TFIIK.


A
raw image
100 nm
B
Trailing EC
TFIIF
Rpb 4/7
2.9
4.0
6.0
8.0
10.0
Front view
Back view
C
Leading EC
Rpb 4/7
2.9
4.0
6.0
8.0
10.0
Front view
Back view
D
Final Resolution=3.5 Å
Fourier Shell Correlation
Resolution (1/Å)
Corrected
Unmasked
Masked
FSC=0.5 3.96 Å
FSC=0.143 3.54 Å
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.1
0.2
0.3
E
Final Resolution=3.5 Å
Fourier Shell Correlation
Resolution (1/Å)
Corrected
Unmasked
Masked
FSC=0.5 4.02 Å
FSC=0.143 3.5 Å
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.1
0.2
0.3

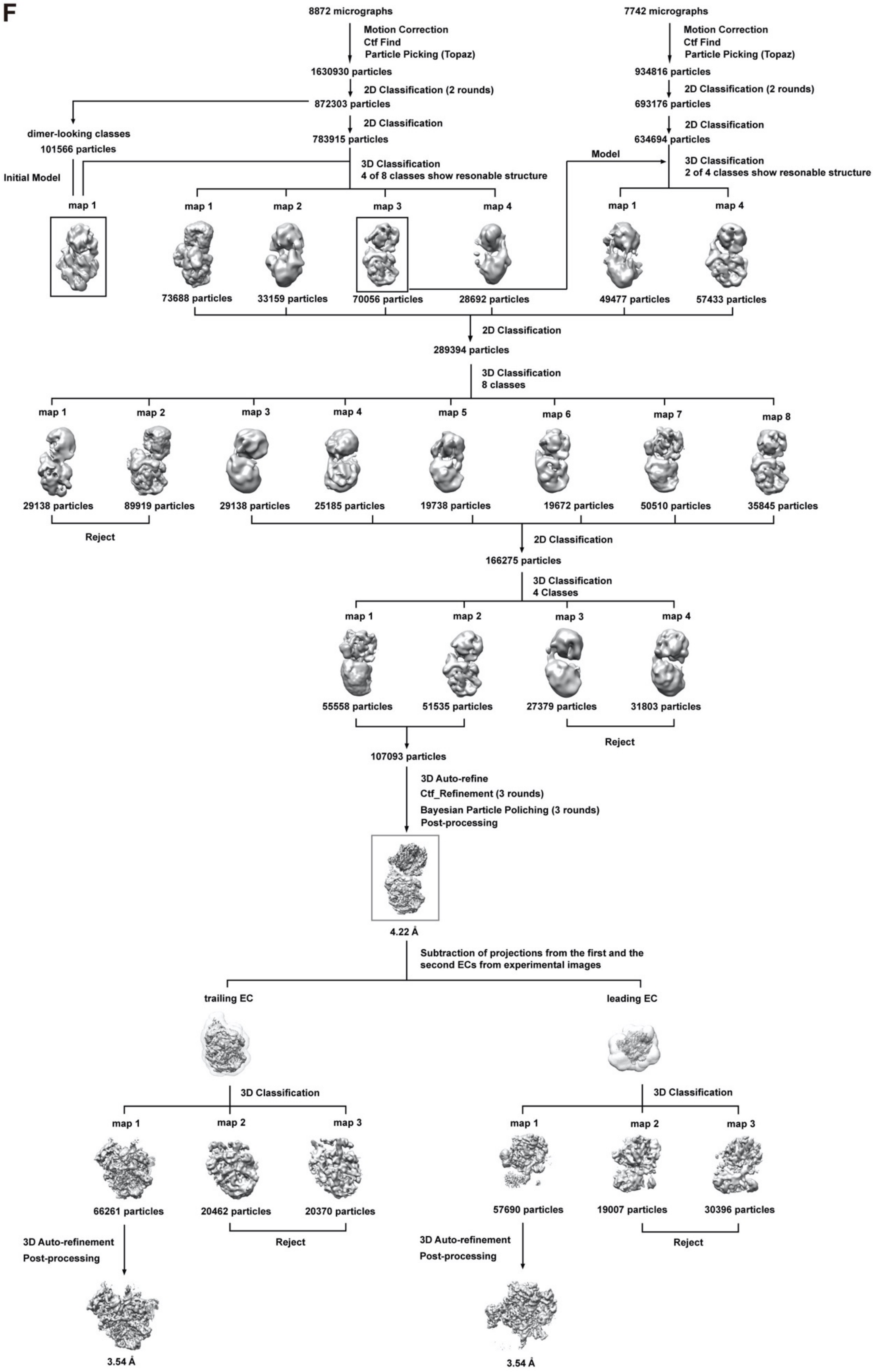

F
8872 micrographs
Motion Correction
Ctf Find
Particle Picking (Topaz)
1630930 particles
2D Classification (2 rounds)
872303 particles
2D Classification
783915 particles
dimer-looking classes
101566 particles
Initial Model
3D Classification
4 of 8 classes show resonable structure
7742 micrographs
Motion Correction
Ctf Find
Particle Picking (Topaz)
934816 particles
2D Classification (2 rounds)
693176 particles
2D Classification
634694 particles
Model
3D Classification
2 of 4 classes show resonable structure
map 1
map 1
map 2
map 3
map 4
map 1
map 4
73688 particles
33159 particles
70056 particles
28692 particles
49477 particles
57433 particles
2D Classification
289394 particles
3D Classification
8 classes
map 1
map 2
map 3
map 4
map 5
map 6
map 7
map 8
29138 particles
89919 particles
29138 particles
25185 particles
19738 particles
19672 particles
50510 particles
35845 particles
Reject
2D Classification
166275 particles
3D Classification
4 Classes
map 1
map 2
map 3
map 4
55558 particles
51535 particles
27379 particles
31803 particles
Reject
107093 particles
3D Auto-refine
Ctf_Refinement (3 rounds)
Bayesian Particle Poliching (3 rounds)
Post-processing
4.22 Å
Subtraction of projections from the first and the
second ECs from experimental images
trailing EC
leading EC
3D Classification
map 1
map 2
map 3
66261 particles
20462 particles
20370 particles
Reject
3D Auto-refinement
Post-processing
3.54 Å
3D Classification
map 1
map 2
map 3
57690 particles
19007 particles
30396 particles
Reject
3D Auto-refinement
Post-processing
3.54 Å

Figure S4. Summary of Cryo-EM data analysis of the EC+EC complex, related to Figure 4. **(A)** A representative raw micrograph of the EC+EC. **(B)** Local resolution (top) and corresponding angular distribution of particles (bottom) of the trailing EC from front view (left) and back view (right). **(C)** Same as (B) but for the leading EC. **(D)** Estimation of average resolution showing focused refinement of trailing EC. **(E)** Estimation of average resolution showing focused refinement of leading EC. The lines indicate the FSC between the half maps of the reconstruction. **(F)** Cryo-EM data processing and workflow of the EC+EC complex. Particle numbers used to identify each map are shown below the corresponding map. The number of the map is corresponding to the number of class in 3D Classification.

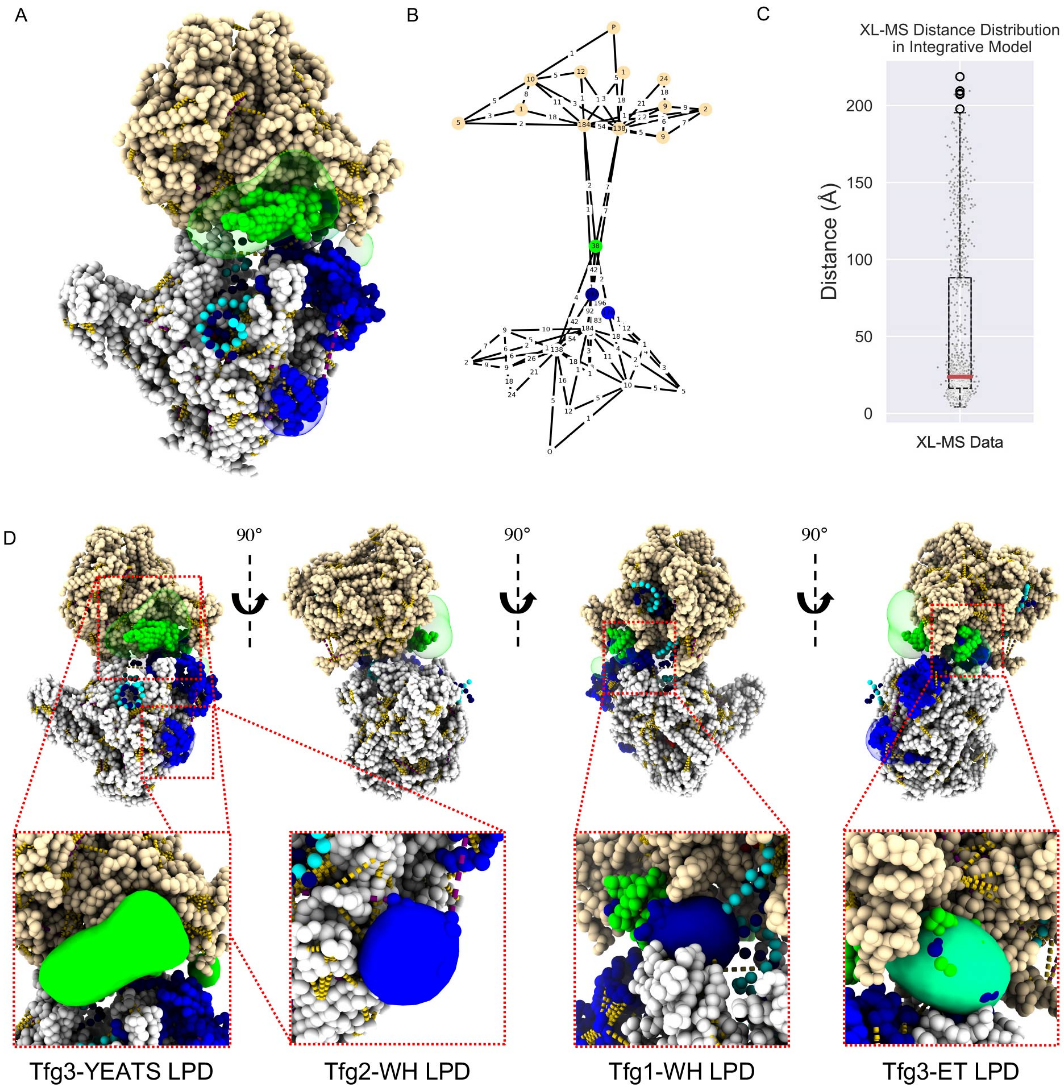


Figure S5. Integrative Modeling results for EC+EC with TFIIF, related to Figure 5. An integrative modeling approach implemented on the Integrative Modeling Platform (IMP) was used to model the EC+EC with TFIIF based on XL-MS data along with high-resolution models or homology models comprising most of the complex's mass. Integrative modeling resulted in one main structural cluster satisfying >%80 of the crosslinking data at 35Å at a sampling precision of 12.8Å. **(A)** Top-view of the EC+EC scaffold with TFIIF. Leading EC shown in tan, trailing EC shown in white. Tfg1 shown in navy, Tfg2 shown in blue, and Tfg3 shown in green. Crosslinks satisfied at 35Å distance displayed in dashed yellow lines, crosslinks satisfied at 35-50Å in dashed purple lines, and localization probability density envelopes for Tfg1-WH, Tfg2-WH, and Tfg3-ET and Tfg3-YEATS domains shown in transparent surface at Chimera standard deviation level 6. **(B)** UCSF Chimera X Crosslink Network diagram for crosslinks satisfied at 50Å. Nodes colored according to scheme in A, and numbers correspond to satisfied intrasubunit crosslinks within nodes: 184 (Rpb1) 138 (Rpb2), 10 (Rpb3), 9 (Rpb4), 2 (Rpb7), 192 (Tgf1), 176 (Tfg2), 38 (Tfg3). Lines between nodes indicate inter-subunit cross-links with numbers of satisfied XLs. **(C)** Boxplot distance distribution of crosslinks, median in red, box encompasses second and third quartiles, whiskers encompass 95% of datapoints. Individual datapoints overlayed onto the boxplot for clarity. **(D)** Rotated views of integrative model showcasing localization probability density envelopes for each TFIIF component queried.

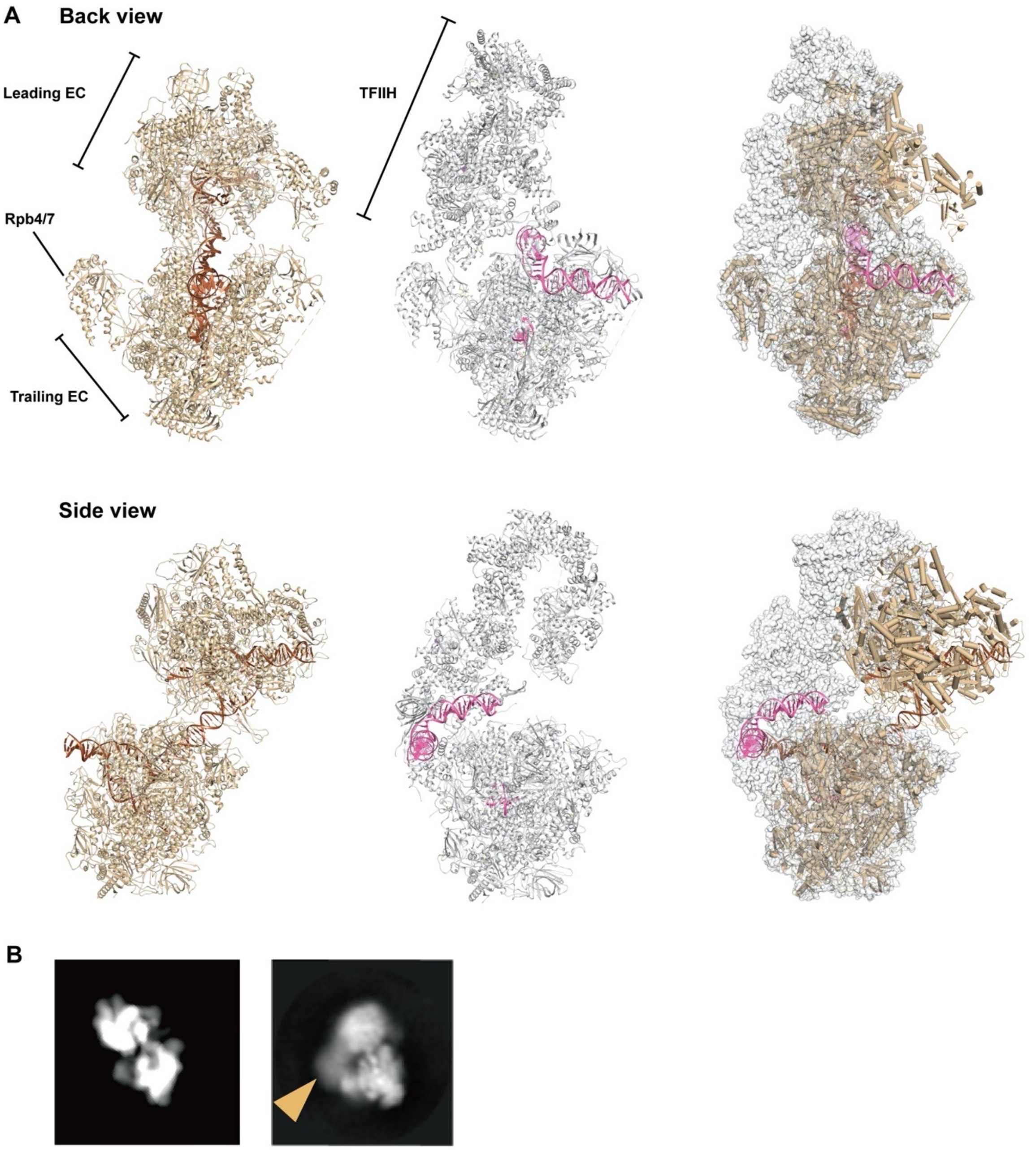


Figure S7. A comparison between the EC+EC stalled at +49 and the ITC stalled at +26, related to Figure 7. **(A)** The EC+EC stalled at +49 (left) and the ITC stalled at +26 (middle). Right, when the trailing EC of the EC+EC and the pol II in the ITC are aligned, there is a steric clash between the leading EC and TFIIH of the trailing ITC. The EC+EC stalled at +49 is colored in tan and shown as pipes and planks. The ITC stalled at +26 is colored in gray and shown as surface. **(B)** Left, 2D projection of 3D model of EC+core ITC (the ITC lacking TFIIH). Right, 2D class average of EC+ITC. Same as Figure 4F. The difference density indicated by orange arrow is attributed to TFIIH.

2.9 Supplemental Table

Table S1. Cryo EM image collection and processing statistics.

	PIC-1	TFIIH-2	PIC-2	PIC-3	ITC	EC+EC
Data collection and processing						
Magnification	81,000	81,000	81,000	81,000	81,000	81000
Voltage (kV)	300	300	300	300	300	300
Electron exposure ($e^-/Å^2$)	45	45	45	45	45	50
Defocus range (μm)	-0.5 to -2.5	-0.5 to -2.5	-0.5 to -2.5	-0.5 to -2.5	-0.5 to -2.5	-1 to -2.5
Pixel size (Å)	0.53	0.53	0.53	0.53	0.53	0.54
Map sharpening B factor (Å)	-10 to -100*	0 to -233*	0 to -300*	0 to -40*	0*	-34.8[e] / -19.9[f]
Map resolution (Å)	3.0[a]/3.8[b]/4.6[c]	7.6	4.0[a]/6.4[b]/ 7.3[c]/12.1[d]	4.1[a]/7.6[b]/11.8[c]	3.1[a]/6.8[b]/9.9[c]	3.5[e] / 3.5[f]
FSC threshold	0.143	0.143	0.143	0.143	0.143	0.143
EMDB entry	23904	23907	23905	23906	23908	23789
Model Refinement						
Model resolution (Å)	3.46 (2.7)	8.44 (4.17)	4.20 (3.56)	7.25 (3.65)	4.08 (3.21)	4.87 (4.02)
FSC threshold	0.5 (0.143)	0.5 (0.143)	0.5 (0.143)	0.5 (0.143)	0.5 (0.143)	0.5 (0.143)
PDB entry	7ML0	7ML3	7ML1	7ML2	7ML4	7MEI
Model composition						
Non-hydrogen atoms	64,255	22,609	64,603	64,571	62,839	68,421
Protein residues	7,996	2,847	8,085	8,086	7,925	8176
Nucleotides	132	58	114	113	88	179
Ligands	140	129	142	140	138	20
R.m.s deviations						
Bond lengths (Å)	0.025	0.011	0.010	0.008	0.009	0.011
Bond angles (°)	1.188	0.935	0.850	0.880	0.706	1.911

Validation						
MolProbity score	3.23	3.68	3.45	3.12	3.15	2.75
Clashscore	22.28	37.60	33.61	25.55	24.67	37.50
Poor rotamer (%)	13.09	18.50	13.79	7.80	9.06	0.9
Ramachandran plot						
Favored (%)	90.27	84.68	89.22	90.25	90.37	82.97
Allowed (%)	9.52	15.21	10.65	9.57	9.49	16.72
Disallowed (%)	0.21	0.11	0.13	0.18	0.14	0.48
Model vs. Data						
CC (mask)	0.58	0.60	0.61	0.37	0.48	0.70
CC (box)	0.70	0.75	0.77	0.73	0.57	0.87
CC (volume)	0.63	0.59	0.61	0.34	0.57	0.70
CC (peaks)	0.53	0.48	0.53	0.23	0.45	0.67
CC (main chain)	0.62	0.69	0.70	0.60	0.50	0.78
CC (side chain)	0.62	0.69	0.70	0.61	0.50	0.74

[a]Pol II, TFIIB; [b]DNA, TBP, TFIIE, TFIIF; [c]TFIIH; [d]downstream DNA; [e]Leading EC; [f]Trailing EC; *see FigS1 F, I, L, O for the B factor values applied.

2.10 Materials and Methods

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and virus strains		
BL21(DE3) Competent Cells	Thermo Fisher Scientific	Cat#EC0114
Chemicals, peptides, and recombinant proteins		
25% Glutaraldehyde Solution in H_2O, EM grade	Sigma Aldrich	Cat#111-30-8
4-thiouridine-5'-triphosphate	TriLink	Cat#N-1025-5
3'-O-methylguanosine-5'triphosphate	TriLink	Cat#N-1058-5
100 mM ATP, CTP, UTP	Thermo Fisher Scientific	Cat#R0481
Salmon sperm DNA	Thermo Fisher Scientific	Cat#15632011
Glycogen	Roche	Cat#10901393001
Proteinase K	Sigma Aldrich	Cat#P4850
RNaseOUT Recombinant Ribonuclease Inhibitor	Thermo Fisher Scientific	Cat#10777019
UTP [α-^{32}P]	PerkinElmer	Cat#NEG507H250UC
CTP [α-^{32}P]	PerkinElmer	Cat#BLU008H250UC
Disuccinimidyl Dibutyric Urea	Thermo Fisher Scientific	Cat#A35459
SDS, 10% solution	Thermo Fisher Scientific	Cat#AM9822
EDTA	Thermo Fisher Scientific	Cat#15575020
Deposited data		
PIC1 structure	This paper	PDB: 7ML0
PIC2 structure	This paper	PDB: 7ML1
PIC3 structure	This paper	PDB: 7ML2
ITC structure	This paper	PDB: 7ML4

TFIIH weak binding state structure	This paper	PDB: 7ML3
Composite structure of EC+EC	This paper	PDB: 7MEI
Leading EC structure (focused-refinement)	This paper	PDB: 7MKA
Trailing EC structure (focused-refinement)	This paper	PDB: 7MK9
PIC1 cryo-EM composite map	This paper	EMD: 23904
PIC2 cryo-EM composite map	This paper	EMD: 23905
PIC3 cryo-EM composite map	This paper	EMD: 23906
ITC cryo-EM composite map	This paper	EMD: 23908
TFIIH weak binding state cryo-EM map (Focused refinement)	This paper	EMD: 23907
Composite cryo-EM Map of EC+EC	This paper	EMD: 23789
Leading EC cryo-EM Map	This paper	EMD: 23888
Trailing EC cryo-EM Map	This paper	EMD: 23887
Raw MS-MS data		
Experimental models: organisms/strains		
S. cerevisiae: C-TAP TFB3 and ∆TFB6 CB010 (Matα pep4::HIS3 prb1::LEU2 prc1::HISG can1 ade2 trp1 ura3 his3 leu2–3,112 cir-o GAL+ RAF+ SUC+ tfb6::kanMX6 TFB3::TAP::Kl.TRP1)	Murakami et al. 2012	N/A
S. cerevisiae: C-TAP TFG2 CB010 (Matα pep4::HIS3 prb1::LEU2 prc1::HISG can1 ade2 trp1 ura3 his3 leu2–3,112 cir-o GAL+ RAF+ SUC+ TFG2::TAP::Kl.TRP1)	Murakami et al. 2012	N/A
S. cerevisiae: C-TAP TFA2 CB010 (Matα pep4::HIS3 prb1::LEU2 prc1::HISG can1 ade2 trp1 ura3 his3 leu2–3,112 cir-o GAL+ RAF+ SUC+ TFA2::TAP::Kl.TRP1)	Murakami et al. 2012	N/A
S. cerevisiae: C-TAP Rpb3 CB010 (Matα pep4::HIS3 prb1::LEU2 prc1::HISG can1 ade2 trp1 ura3 his3 leu2–3,112 cir-o GAL+ RAF+ SUC+ Rpb3::TAP::Kl.TRP1)	Murakami et al. 2012	N/A

S. cerevisiae: C-TAP TFB4 and △TFB6 CB010 (Matα pep4::HIS3 prb1::LEU2 prc1::HISG can1 ade2 trp1 ura3 his3 leu2–3,112 cir-o GAL+ RAF+ SUC+ tfb6::kanMX6 TFB4::TAP::Kl.TRP1)	Murakami et al. 2012	N/A
Recombinant DNA		
Plasmid: Full-length *S.c* TBP (pRSFDuet)	Murakami et al. 2012	N/A
Plasmid: Full-length *S.c* Toa1 (pRSFDuet)	Murakami et al. 2012	N/A
Plasmid: Full-length *S.c* Toa2 (pRSFDuet)	Murakami et al. 2012	N/A
Plasmid: Full-length *S.c* Toa1-Toa2 (pET47b)	Adachi et al. 2017	N/A
Plasmid: Full-length *S.c* TFIIB (pET28)	Bratkowski et al. 2017	N/A
Plasmid: Full-length *S.c* Sub1 (pCold II)	Fazal et al., 2015	N/A
Plasmid: SNR20 G-less 26 promoter fragment [-133/+86]	Fujiwara et al. 2019	N/A
Plasmid: SNR20 G-less 49 promoter fragment [-133/+86]	Fujiwara et al. 2019	N/A
Sequence-Based Reagents		
DNA Oligos: template ssDNA : 5'-AGG TCA TTT CAG TTG TTA CAC TGA AAA GAC CCC TCT CGA TCC GCA TAC GCA GGT AAA AGG AAA AGA TGT GGG GGT GGG TTT AAA AAA AAA ACA AG-3'	This paper	N/A
DNA Oligos: non-template ssDNA: 5'-CTT GTT TTT TTT TTA AAC GGC AAA AAC ACA GAA TTC CTT TTA CCT GCG TAT GCC TCG GTT CCT TCC AGT TTT CAG TGT AAC AAC TGA AAT GAC CT-3'	This paper	N/A
RNA Oligos: RNA1: 5'-ACCCCCACA-3'	This paper	N/A
RNA Oligos: RNA2: 5'-AAAACAAAUAUGCAUAUUAUCGAGAGG-3'	This paper	N/A
Primer: SNR20 G-less 49 and G-less 26 Forward:5'-GCCGTTTCCGATGGG CCACTCGGTGAAAA-3'	This paper	N/A
Primer: SNR20 G-less 49 and G-less 26 Reverse:5'-GGTAATGAGCCTCAT TGAGGTCATTTCAGTTGTTACA-3'	This paper	N/A

Software and algorithms		
Relion version 3.0/3.1	Zivanov et al. 2018	https://www3.mrc-lmb.cam.ac.uk/relion//index.php/Main_Page/; RRID:SCR_016274
cryoSPARC version 3.1	Punjani et al. 2017	https://cryosparc.com/; RRID:SCR_016501
UCSF Chimera	Pettersen et al. 2004	https://www.cgl.ucsf.edu/chimera/; RRID:SCR_004097
Phenix version 1.18.2	Adams et al. 2010	https://www.phenix-online.org/; RRID:SCR_014224
Coot version 0.8.9.2	Emsley et al., 2010	https://www2.mrc-lmb.cam.ac.uk/personal/pemsley/coot/; RRID:SCR_014222
Topaz	Bepler et al., 2019	http://cb.csail.mit.edu/cb/topaz/

Integrative Modeling Platform version 2.13.0	Andre Sali Lab	https://integrativemodeling.org/; RRID:SCR_002982

Protein purification

Recombinant TFIIA,TFIIB ,TBP, and Sub1 were overexpressed and purified from bacteria. TFIIE, TFIIF, TFIIH, and pol II were isolated from yeast as previously described (Fujiwara and Murakami, 2019).

Cryo-EM sample preparation of the ITC with the G-less 26 template

G-less 26 DNA templates derived from the SNR20 promoter were obtained by PCR as previously described (Fujiwara and Murakami, 2019) and purified using Superose 6 10/300 (GE Healthcare) in buffer 300 (20 mM Hepes (pH 7.6), 300 mM potassium acetate, 5 mM DTT, and 2 mM magnesium acetate). To assemble PIC on the G-less 26 DNA template, the following were mixed in 515 μL of buffer 300 containing additional 5% glycerol: 0.26 μM DNA template, 0.5 μM TFIIA, 0.7 μM TFIIB, 1.2 μM TBP, 0.6 μM TFIIE, 1.04 μM TFIIF, 0.44 μM TFIIH, 0.44 μM TFIIK, 1.04 μM pol II, and 0.6 μM Sub1. The mixture was then diluted by adding an equal volume of buffer 10 (20 mM Hepes (pH 7.6), 10 mM potassium acetate, 5 mM magnesium sulfate, 5 mM DTT) and incubated on ice for 24 hours. After pre-incubation for 20 min at 30°C, 3/4th of the PIC mixture received 2x NTP solution consisting of 1.6 mM ATP, 1.34 mM CTP, 2 mM 4’-thio UTP, 0.5 μM 3’-O-methyl GTP, 10 mM magnesium acetate, and 0.5 U/μL RNaseOUT in buffer 10, and 1/4th of the mixture received 2x NTP solution containing 44 nM [α-^{32}P] CTP (33 μCi). Transcription initiation was carried out for 20 min at 30°C and a total of 1.5

mL of the cold sample was immediately loaded onto three pre-cooled glycerol gradients (500 µL per gradient) prepared with buffer A (20 mM Hepes (pH7.6), 50 mM potassium acetate, 5 mM DTT, and 2 mM magnesium acetate, and 10% glycerol (v/v)) and buffer B (20 mM Hepes (pH7.6), 50 mM potassium acetate, 5 mM DTT, and 2 mM magnesium acetate, 0.125% glutaraldehyde, and 40% glycerol (v/v)). The 1/4th of the mixture (500 µL) that was incubated with hot NTP solution was loaded onto a glycerol gradient without glutaraldehyde. After centrifugation for 14 h at 30,000 rpm in a Beckman SW60 Ti rotor, the gradients were fractionated using a PGF Piston Gradient Fractionator (BioComp Instruments, Inc.) into ~100 µL per fraction and crosslinking reaction was quenched by addition of 50 mM glycine (pH 7.6). To perform RNA analysis of the fractions, 70 µL from non-crosslinked sample was incubated for 15 min at 42°C with 160 µL of stop buffer containing 390 mM sodium acetate (pH 5.5), 8 mM EDTA, 0.6% SDS, 0.06 mg/mL glycogen, 0.03 mg/mL proteinase K, and 0.03 mg/mL salmon sperm DNA, and subjected to ethanol precipitation. RNA was then analyzed by urea denaturing gel. For protein analysis, 20 µL per fraction was analyzed by SDS-PAGE (Figure S1A).

Cryo-EM grid preparation and data collection of the G-less 26 sample

The appropriate fractions from crosslinked sample were pooled and concentrated ~8 fold with a 100k MWCO spin concentrator and dialyzed against buffer 50 (20 mM Hepes (pH 7.6), 50 mM potassium acetate, 5 mM DTT, and 4 mM magnesium acetate) for 45 min. For cryo-EM grid preparation, 2.7 µL of the G-less 26 sample were applied onto glow-discharged Quantifoil R0.6/1 200-mesh holey carbon grids (Electron Microscopy Sciences), blotted for ~1.7 second, and plunge frozen in liquid ethane with a Leica EM CPC manual plunger (Leica Microsystems). The grids were loaded onto a Titan Krios electron microscope operating at 300kV equipped with Gatan K3 Summit direct electron detector with Gatan quantum energy filter (slit width of 20 eV) at CryoEM core facility at University of Massachusetts. The data were collected automatically using SerialEM at a nominal magnification of 81,000x, with a defocus range of 0.5 μm to 2.5 μm, and with a 30 frame exposure taken over ~2.4 sec with a total electron dose of ~45 e^{-}/Å^{2}. A total of 15895 images was collected.

Cryo-EM data processing of the G-less 26 sample

All image processing of G-less 26 sample was performed using RELION 3.0 and 3.1 (Scheres, 2012). A total of 15,895 images was processed in two sets (8,000 and 7,895 images) in the same manner. The movie frames were aligned using RELION's own implementation with a binning factor of 2 and the CTF was determined using CTFFIND-4.1(Mindell and Grigorieff, 2003). At this point, a total of 56 images were excluded for the further analysis. Initially, particles were picked automatically from 700 images and subjected to a few rounds of reference-free 2D classification. Some of the resulting 2D classes were low-pass filtered to 20Å and then used to pick particles from the two sets of the data, resulting in 1,940,218 and 1,872,221 particles from the first and second sets of images, respectively. These particles are separately subjected to three rounds of reference-free 2D classification. The 2D classes containing detailed features were selected from the first set and used to generate an initial model. We then performed 3D classification, using the initial model with a low-pass filter of 60Å, from 1,637,582 particles and 1,705,180 particles separately. After three rounds of global 3D classification, the resulting EM maps were aligned on pol II, and maps that had TFIIH at similar location relative to pol II were combined, resulting into three groups. The first group (454,296 particles) that was similar to the canonical PIC (Murakami et al., 2015b) was subjected to per-particle CTF refinement by first estimating beam shift, trefoil, and

4th order aberrations, then magnification anisotropy, and finally per-particle defocus, and per-micrograph astigmatism. To improve the map quality of TFIIH, which was poorly ordered in the entire map, a soft mask was created around TFIIH of the 3D refined map, subtracted, and the subtracted images were used to generate an initial model. Then 3D classification was performed with image alignment. The resulting classes revealed two interpretable maps: strong (137,466 particles) and weak (101,497 particles) binding states of TFIIH, which were 3D-refined and post-processed to a resolution of 4.6Å and 7.6Å, respectively. All the reported resolutions are based on the gold-standard Fourier shell correlation (FSC) using 0.143 criterion. To further improve map quality of TFIIH, the maps were segmented into 3 bodies for multibody refinement: Tfb3, Rad3, Tfb1, Ssl1, Tfb4, and Tfb2N in body 1, Ssl2N in body 2, and Ssl2C, Tfb2C, and Tfb5 in body 3. The core PIC (cPIC) maps for both strong and weak binding states were generated by reverting the TFIIH maps to obtain entire maps, then 3D refinement using global search, and finally postprocessing pol II and TFIIB, and the rest of cPIC separately by applying appropriate masks. Resolution for the cPIC parts ranged between 3.0Å and 3.9Å.

The maps in the second group (384,412 particles) were combined and subjected to one more round of global 3D classification. For structural determination of PIC2, the resulting maps that showed the density for downstream dsDNA (117,450 particles) were combined and 3D refined. As in PIC1, focused refinement of TFIIH was performed, resulting in 33,159 particles in the best class. The per-particle CTF was determined as described above, and then TFIIH was 3D refined. This resulted TFIIH map with a resolution of 7.3Å. Multibody refinement (Nakane et al., 2018) was performed as described above to improve the map quality of TFIIH. To improve the map quality corresponding to downstream dsDNA, a soft mask around the DNA was created from the entire map obtained by reverting TFIIH and 3D refining using only local search, subtracted, and 3D classified. The map containing the best DNA density (11,028 particles) was reverted and 3D refined using only local search. This yielded the DNA density at 12.1Å resolution. To obtain cPIC2, the particles in the best TFIIH class was reverted and 3D refined using global search. Post-processing of the cPIC was performed as in PIC1. The resolution ranged between 4.0Å and 6.4Å.

To reconstruct PIC3, map in the third group (307,173 particles) were combined and subjected to another round of 3D classification. Similar to PIC2 analysis above, classes containing downstream dsDNA (69,513 particles) were merged and 3D refined, and then TFIIH was subjected to a focused refinement. The best class was 3D refined and post-processed to a resolution of 11.8Å. The cPIC for PIC3 was obtained in the same manner as PIC1 and the resolution ranged between 4.1 and 7.6Å.

To reconstruct ITC, three maps that revealed the upstream edge of the bubble after one round of 3D classification of the third group (Figure S1L) were subjected to another round of 3D classification (Figure S1O). The resulting maps (227,346 particles) containing no downstream dsDNA were combined and then per-particle CTF was determined in the same manner as PIC particles above. The 3D auto-refined map did not show clear density for the DNA-RNA hybrid although some density was apparent in the active site of pol II, indicating variability of the DNA/RNA hybrid in position and length. Thus, to improve the quality of the hybrid density, a soft mask around the active site was generated, subtracted, and the resulting images were subjected to focused 3D classification without image alignment with higher regularization parameter (T=30). Classes that contained strong density for the hybrid (120,006 particles) were combined, reverted and 3D refined to obtain the entire map. cITC was postprocessed with a mask to a resolution of 3.1Å. The quality of the density for upstream DNA, TFIIE, TBP, Tfg2WH was also improved by a similar manner as the hybrid but with alignment during focused 3D classification, which yielded a map (86,069 particles) with a resolution of 6.8Å. To improve the map quality of TFIIH in the ITC, focused 3D classification with alignment was performed after per-particle CTF refinement. The best class (45,780 particles) was combined with focused refined map resulting from a subset of particles in the second group that did not show density for downstream dsDNA, then the particles were subjected to one round of 3D classification. This yielded a 9.9Å TFIIH map for the ITC.

Model building of PIC1-3 and ITC

Maps with and without B-factor sharpening were used to build models of PIC1-3. For cPIC, the previous model from the yeast PIC (PDB ID: 5OQJ) was used as an initial template. For TFIIH, the previous model from the 3.9 Å resolution cryo-EM structure in a form of DNA repair (PDB ID: 7K01) was used as an initial model. Promoter DNA was manually built by combining short (~10bp) B-form DNA segments. A combined model containing cPIC, TFIIH and promoter DNA was iteratively subjected to manual refinement with Coot (Emsley et al., 2010) and rigid body refinement with Phenix1.16 (Liebschner et al., 2019). Each subunit of pol II and GTFs was constrained as a rigid body, while base pairs of DNA double helix were maintained throughout refinement with Phenix. The ITC was modeled essentially as in PIC1-3. The model of the 6-nt DNA-RNA hybrid with TFIIB was built using the previous X-ray crystallographic model (PDB ID: 4BBS) as a template, and refined using Phenix. Figures were prepared using UCSF Chimera (Pettersen et al., 2004).

Cryo-EM sample preparation with the G-less 49 template

The G-less 49 template derived from the SNR20 promoter was obtained by PCR as previously described (Fujiwara and Murakami, 2019) and purified using Superose 6 10/300 (GE Healthcare) in buffer 300. To assemble PIC on the G-less 49 DNA template, the following were mixed in 240 µL of buffer 300: 0.26 µM DNA template, 0.4 µM TFIIA, 1.2 µM TFIIB, 2.4 µM TBP, 0.6 µM TFIIE, 1 µM TFIIF, 0.6 µM holoTFIIH, 0.36 µM TFIIK, 1.04 µM pol II, and 0.4 µM Sub1. The mixture was then diluted by adding an equal volume of buffer 10 (20 mM Hepes (pH 7.6), 10 mM potassium acetate, 5 mM magnesium sulfate, 5 mM DTT) and incubated on ice for 24 hours. After pre-incubation for 20 min at 30°C, 3/4th of the PIC mixture received 2x NTP solution consisting of 1.6 mM ATP, 1.6 mM CTP, 1 mM UTP, 0.5 µM 3'-O-methyl GTP, 10 mM magnesium acetate, and 0.5 U/µL RNaseOUT in buffer 10, for cryo-EM analysis, while 1/4th of the mixture received 2x NTP solution containing 44 nM [α-^{32}P] UTP (33 µCi) for characterization of proteins and RNA (Figures 4A-B). Transcription initiation was carried out for 20 min at 30°C and the sample was immediately loaded onto a gradient prepared with buffer A and buffer B as for the G-less 26 complex. 240 µl (with [α-^{32}P] UTP) and 720 µl (with cold UTP) were sedimented without and with glutaraldehyde, respectively. After centrifugation for 13 h at 30,000 rpm in a Beckman SW60 Ti rotor, the gradients were fractionated using a PGF Piston Gradient Fractionator (BioComp Instruments, Inc.) into ~130 µL per fraction and crosslinking reaction was quenched by addition of 40 mM glycine (pH 7.6). To perform RNA analysis of the fractions, 100 µL from non-crosslinked sample was incubated for 15 min at 42°C with 110 µL of stop buffer containing 390 mM sodium acetate (pH 5.5), 8 mM EDTA, 0.6% SDS, 0.06 mg/mL glycogen, 0.03 mg/mL proteinase K, and 0.03 mg/mL salmon sperm DNA, and subjected to ethanol precipitation, followed by RNA analysis by urea denaturing gel (Figure 4A). For protein analysis, 20 µL per fraction was analyzed by SDS-PAGE (Figure 4B).

To prepare cryo-EM grids, samples were dialyzed into EM buffer (20 mM HEPES (pH 7.6), 50 mM potassium acetate, 5 mM DTT, 2 mM magnesium acetate) for 30 minutes prior to making grids. EC+EC samples were applied to R1.2/1.3 400 mesh quantifoil holey carbon grids (Electron Microscopy Sciences), and EC+ITC samples were applied to R2/2 300 mesh quantifoil holey carbon grids (Electron Microscopy Sciences). All grids were glow-discharged (easiGlow, Pelco) for 2 min before deposition of 2uL of dialyzed sample, and subsequently blotted for 1.5 (EC+EC samples) or 2 seconds (EC +ITC samples) using Whatman Grade 41 filter paper (Sigma-Aldrich) and flash-frozen in liquid ethane with a Leica EM CPC manual plunger (Leica Microsystems). EM grids were prepared in batches and the freezing conditions were optimized by screening on a FEI TF20 microscope operating at 200 kV and equipped with a FEI Falcon III direct electron detection camera at the Electron Microscopy Research Lab (the University of Pennsylvania).

For EC+EC, two datasets (8872 and 7742 micrographs) were collected at Frederick National Laboratory (sponsored by the National Cancer Institute) using a NCEF Titan Krios transmission electron microscope operating at 300 kV, equipped with a K3 Bioquantum detector and a Bioquantum energy quantum filter. Images were collected by image shift and at a nominal magnification of 81,000x in super-resolution mode (pixel size of 0.54 Å) at a defocus range between 1 and 2.5 μm. The exposure time was 3.2 s at a nominal dose of 50 e^-/Å2, movies were divided into 40 frames.

Cryo-EM images of EC+ITC were collected at the Pacific Northwest Cryo-EM center using a Titan Krios transmission electron microscope operating at 300 kV, equipped with a K3 direct detection camera (Gatan) and a Bioquantum energy quantum filter. Data was collected by image shift and at a nominal magnification of 105,000x in super-resolution mode (pixel size of 0.415 Å) at a defocus range between 0.9 and 2.2 μm. A total of 29,626 images were collected over 5 days. The exposure time was 2.1 s at a nominal dose of 45 e^-/Å2, movies were divided into 66 frames.

Image processing and 3D reconstruction of the EC+EC

Cryo-EM images of EC+EC were processed by a combination of cryoSPARC v3.1 (Punjani et al., 2017), Relion 3.1 (Scheres, 2012), and Topaz (Bepler et al., 2019). The two datasets were motion-corrected with MotionCorr2 (Zheng et al., 2017) , and then CTF corrected with CTFFIND4 (Mindell and Grigorieff, 2003). A total of 1,630,930 particles were extracted with 340 pixel box after particle-picking using Topaz from the first dataset, and then the resultant particles were screened by two rounds of reference-free 2D classification, from which classes containing two ECs, accounting for 101,566 particles, were selected to calculate initial model. Subsequently one round of 3D classification was carried out, yielding four reasonable 3D classes. From the second dataset, 934,816 particles were extracted with 340 pixel box with Topaz, and then was subjected to three rounds of reference-free 2D classification followed by one round of 3D classification using the map obtained from the first dataset as a reference, yielding two reasonable 3D classes. The four 3D classes from the first dataset and the two classes from the second dataset were combined to perform further iterative rounds of 2D and 3D classifications, resulting in two maps showing the leading EC and the trailing EC. The resulting two 3D classes, accounting for a total of 107,093 particles, were subjected to 3D auto-refinement with a soft-edged mask, CTF-refinement, Bayesian polishing, and Post-processing, leading to a reconstruction of the entire structure at 4.22 angstrom resolution. To push resolution, each elongation complex was subtracted using soft-edged masks encompassing the leading EC and the trailing EC respectively with 220 pixel box for each, and subjected to focused 3D classification followed by 3D auto-refinement. Lastly, a 3.5 angstrom map of the leading EC containing 57,690 particles and a 3.5 angstrom map of the trailing EC containing 66,261 particles were combined to generate a composite map using the vop maximum command in UCSF chimera (Pettersen et al., 2004).

All 2D classification, 3D classification, 3D refinement, Bayesian polishing, CTF-refinement, mask creation and post-process procedures described above were employed with Relion 3.1.0. Local resolution estimation for each elongation complex was performed with cryoSPARC v3.1. Resolution was reported on the basis of the gold-standard Fourier shell correlation (FSC) (0.143 criterion).

Cryo-EM images of EC +ITC were processed using a combination of Relion 3.1.1 (Scheres, 2012), and sphier-crYOLO (Wagner et al., 2019). Datasets were motion-corrected with MotionCorr2 (Zheng et al., 2017) then CTF corrected with CTFFIND4 (Mindell and Grigorieff, 2003). Particles were picked with sphier-crYOLO and then extracted with 530 pixel box. A total of 988,153 particles were subjected to three rounds of 2D classification with Relion, yielding eight 2D classes accounting for ~9,707 particles (Figure 4D).

<u>Model building and refinement of the EC+EC</u>

Structural models were built using COOT (Emsley et al., 2010) and Phenix (Liebschner et al., 2019), the process was described as follows.

For the leading EC, structural models of pol II, the upstream DNA, the downstream DNA, the transcription bubble, and the DNA-RNA hybrid (PDB:5C4J) were placed into the map and subjected to rigid-body refinement with Phenix. The DNA-RNA hybrid was then manually refined using COOT against auto-sharpened map generated by Phenix. The 7 nt extended ssRNA (PDB : 6gml) in the exit tunnel was fitted into density, and then subjected to refinement with Phenix and COOT. The trailing EC was modeled as for the leading EC. The 5 nt backtracked ssRNA (15-19 nt, PDB:3PO2) was rigid body fitted into the density and then subjected to iterative refinement with Phenix and COOT. The TFIIF except Tfg2 WH domain (PDB:5FYW) was fitted into density using UCSF Chimera. The dsDNA bridging two elongation complexes was built with a B form DNA and then manually adjusting using COOT. Lastly, the entire model was interactively subjected to rigid-body refinement with Phenix and manual refinement with COOT. The final refinement was done with Phenix, with validation report (Table S1). All figures were generated using UCSF Chimera (Pettersen et al., 2004).

XL-MS of the EC+EC

For XL-MS, the EC+EC was assembled with an 95-bp artificial template containing two 15-bp mismatch bubbles with 35-bp spacing between the two nucleotide addition sites: 151 pmol of template DNA, 300 pmol of RNA1, 300 pmol of RNA2, and 154 pmol non-template DNA were combined in buffer10 (20 mM Hepes pH 7.6, 10 mM potassium acetate, 10 mM DTT , 5 mM magnesium sulfate), incubated at 95 ℃ for 5 min, and then annealed by slowly cooling down to room temperature. Then 368 pmol of pol II and 368 pmol of TFIIF were added to the template in buffer 150 (50 mM Hepes pH 7.6, 150 mM potassium acetate, 5 mM DTT, 3mM magnesium sulfate, and 5% glycerol) on ice overnight , and then dialyzed into buffer 100 (20 mM Hepes pH 7.6, 100 mM potassium acetate, 2 mM DTT , 2 mM magnesium acetate, and 10% glycerol) for 4 hours to remove primary amines. 6 mM (final concentration) of disuccinimidyl dibutyric urea (DSBU, Thermo Fisher Scientific) was added to the 0.755 mg /mL EC+EC sample, and incubated on ice for 2 hours, and then quenched by adding 50 mM (final concentration) ammonium bicarbonate. Crosslinked proteins were precipitated with 20% (w/v) trichloroacetic acid (TCA, Sigma) on ice for 60 minutes. Proteins were pelleted by centrifugation at 15000 rpm for 15 min, and then washed with 10% TCA in 100mM Tris-HCl and then with acetone (Fisher Scientific). After supernatant was decanted, the pellet was air-dried at room temperature, and then stored at –80°C for analysis by mass spectrometry.

Crosslinked sample was prepared, acquired and analyzed as previously described (van Eeuwen et al., 2021a) with minor modifications: 1) digested crosslinked peptides were fractionated by high pH reverse phase fractionation kit (ThermoFisher Scientific Cat. 84868) and each fractionation was acquired separately. 2) Acquired data were first searched with SequestHTTM and Percolator (Kall et al., 2007) in Proteome DiscovererTM 2.4 to confirm all subunit existence from the samples. The curated FASTA database, containing only subunits of interest, was fed to the crosslinking identification pipeline described previously (van Eeuwen et al., 2021a). A total of 1085 cross-links, comprising 512 within pol II, 330 within TFIIF, and 243 between pol II and TFIIF, were identified. False positive discovery rate (FDR) was estimated based on a target-decoy analysis (Rinner et al., 2008), where decoy was generated by shuffled sequences but with protease sites retained. The FDR was set to 1% to filter each acquisition.

Integrative model building (IMP) of TFIIF on EC+EC

Possible locations of TFIIF domains (the Tfg1 and Tfg2 C-terminal WH domains, and the Tfg3 ET and Tfg3 YEATS domains (PDB IDs 1I27, 1BBY, 6LQZ, 5D7E) on the EC+EC were simulated by the Integrative Modeling Platform (IMP) as described in previous work (van Eeuwen et al., 2021a) with a few modifications. A yeast homology model of the Tfg1 C-terminal WH domain was generated from the human homolog (PDB ID 1I27) using Modeller (Webb and Sali, 2016). A homology model of the Tfg3 ET domain bound to Tfg1 EB (residues 615-623) was generated using the Tfg3 ET domain bound to Sth1 EB using Modeller as well. The homology models with lowest DOPE score were selected for IMP. All models of the WH domains and Tfg3 ET and Tfg3 YEATS domains were subjected to a short energy minimization with UCSF Chimera. Integrative Modeling was then carried out to satisfy XL-MS data. The EC-EC and the four TFIIF domains Tfg1 WH, Tfg2 WH, Tfg3 ET (bound to Tfg1 residues 615-623), and Tfg3 YEATS were treated as five independent rigid bodies during IMP. Regions with high-resolution description were represented as beads representing 1 residue each, whereas remaining regions were instead represented by flexible coarse-grained spheres encompassing 10-40 residues.

The model was then subjected to extensive stochastic sampling within IMP while being subjected to a scoring function enforcing basic model parameters such as backbone connectivity and volume non-overlap, as well as integrating data obtained from XL-MS. The resulting model pool was filtered for good-scoring models based on satisfaction criteria of at least 80% of XLs at a distance of 35Å. The total score distribution of the resulting model pool could be described as a Gaussian distribution. Good-scoring models with scores lower than one standard deviation from the mean were selected for further analysis. Implementing IMP’s sampling convergence module on this model pool, resulted in a single structural cluster at a sampling precision of 12.8Å in accordance with standard sampling convergence criteria (Viswanath et al., 2017). Results were displayed in UCSF Chimera X (Pettersen et al., 2021) and are available in GitHub.

CHAPTER 4: PERSPECTIVES AND FUTURE DIRECTIONS

4.1 Summary of Major Conclusions

Expression of genes is a highly regulated process and decades of studies have identified proteins needed for assembly of a pre-initiation complex and an elongation complex as Pol II moves into gene bodies, and for activation of transcription initiation, as well as many key steps that regulate the transcription output. Several rate limiting steps exist post-initiation and one of them is the transition from the initially-transcribing complex to an elongation complex (Badjatia et al., 2021; Nguyen et al., 2020; Rosen et al., 2020). Early biochemical studies in human systems have provided numerous insights into the process, including requirement of the TFIIH translocase activity for promoter escape (Dvir et al., 1996, 1997a; Dvir et al., 1997b) and the timing of bubble collapse and the stability of the early initiation complex as a function of the lengths of RNA synthesized and bubble size (Cai and Luse, 1987; Holstege et al., 1997; Kugel and Goodrich, 2002).

In addition, structural characterization of yeast Pol II-TFIIB complexes combined with biochemical studies have explained how TFIIB is critical for TSS selection and stabilization of the early transcription initiation complex (Liu et al., 2010; Sainsbury et al., 2013; Bangur et al., 1997; Chen and Hampsey, 2004; Kuehner and Brow, 2006). These findings suggest that TFIIB ejection is a key event for the upstream DNA template to reaneal and hence for promoter escape. Based on the structural studies, the N-terminal domain of TFIIB needs to be displaced from the RNA exit tunnel when the length of the elongating transcript reaches ~13 nt (Sainsbury et al., 2013). Consistent with this, the timing of TFIIB release has been shown to be somewhere between 7-16 nt downstream of TSS in human (Ly et al., 2020; Tran and Gralla, 2008). Based on these previous studies, it has been assumed that the process of promoter escape would be completed before elongation factors are recruited. Nevertheless, an optical tweezer-based study of transcription initiation using yeast PIC indicated that Pol II remained associated with promoter through GTFs even after transcribing transcripts of ~50 nt in length (Fazal et al., 2015b). The idea that elongation factors (capping enzymes and Spt4/5) may be involved in promoter escape had never been tested. Additionally, structure of the initially-transcribing complex that initiates transcription de novo was lacking. The structure of the initially-transcribing complex would provide mechanistic insight into promoter escape.

In Chapter 2, transcription initiation was reconstituted from yeast proteins (TFIIA, TFIIB, TBP, TFIIE, TFIIF, TFIIH, Sub1, and Pol II) and a fragment of SNR20 promoter DNA, a canonical TATA-containing promoter DNA, as a template. By creating G-less region and using O-methyl GTP, Pol II was stalled at varying positions from the TSSs. Stalling of Pol II on the DNA template allowed for calculation of the template usage including independent measurements of the first round of initiation vs. reinitiation. After optimization of the reconstituted transcription initiation system, that supports 5’ capping and re-initiation, post-initiation complexes stalled at +27 and +49 were analyzed. Separation of post-initiation complex by glycerol gradient sedimentation followed by protein and RNA analysis revealed that promoter escape is often completed when ~22-nt of RNA is synthesized. Further, TFIIS cleavage assay indicated that Pol II in the ITC stalled at +27 was susceptible to extensive backtracking. The ITC was unstable relative to EC containing a transcript with same length and sequence. Notably, omission of capping enzymes dramatically reduced fraction of complexes that can undergo promoter escape, indicating that capping enzymes facilitate promoter escape. Similarly, Spt4/5, another elongation factor that is recruited to transcription complex soon after recruitment of capping enzymes, had a positive effect on promoter escape. Finally, by incorporating thio-UTP into the transcript, which reduced Pol II backtracking, we showed the presence of ITC containing 26-nt RNA.

Taking advantage of the stabilized ITC on G-less 26 DNA, we isolated and analyzed the ITC sample by cryo-EM to investigate the features of the ITC, which is included in Chapter 3. Cryo-EM analysis of the sample revealed three distinct forms of PICs as well as the ITC. Three forms of the PICs had different paths of downstream DNA, degree of DNA distortion in the initially melting region, and positions of TFIIH, likely representing conformational changes that occur during transcription initiation. Notably, the position of TFIIH in the ITC was similar to that in one of the three forms of the PICs. In the ITC, Tfb3 was no longer seen on between Pol II stalk and TFIIE. TFIIH shifted in the way that the distance between Tfb1 and TFIIE is now shorter than that in the canonical form of PIC, allowing for maximum interaction between them. Concomitantly, TFIIE shifts away from the Pol II clamp region leaving the zinc ribbon the only direct connection between TFIIE and Pol II. These interactions lost between Pol II and GTFs might contribute to preparation of the ITC to escape the promoter. Additionally, we determined the structure of the reinitiated G-less 49 transcription complexes, which were biochemically characterized in Chapter 2. The structure revealed two colliding elongation complexes. The trailing Pol II which contained RNA of ~25 nt in length backtracked upon colliding with the preceding Pol II stalled at +49. The ITC structure on G-less 26 and G-less 49 re-initiation complexes together suggest that promoter proximal paused Pol II had a positive effect on promoter escape.

4.2 Perspectives and future directions

Transcription initiation is the first step for gene regulation and one way to directly increase transcription output is to enhance initiation. To understand this fundamental process of gene expression, decades of studies on transcription initiation have been done. Yet, mechanistic understanding of transcription initiation and the transition from initiation to elongation are still limited. Recent breakthrough of cryo-EM has enabled many structural studies that have provided insights into assembly of the PIC and initiation process at near atomic resolution. For instance, complete structures of Mediator-PIC were determined this year (Abdella et al., 2021; Chen et al., 2021b; Rengachari et al., 2021) and it was unknown until very recently that formation of the initial transcription bubble initiates at the upstream edge of the initially melting region (Schilbach et al., 2021).

In the studies described in Chapter 2 and 3, we reconstituted transcription initiation using purified factors that are necessary and sufficient for basal transcription. Mediator and activators, which are required for stimulated transcription, were not included. The kinase module of TFIIH is known to interact with Mediator based on the structural studies and Kin28 (the kinase subunit of TFIIH) mediated phosphorylation of Pol II CTD is known to facilitate promoter escape via Mediator dissociation from Pol II (Wong et al., 2014). Additionally, biochemical study using yeast nuclear extract suggested upon promoter escape, a reinitiation scaffold containing Mediator and GTFs, except TFIIB and TFIIF, is left at the promoter and is stabilized by the presence of the activator for recruitment of Pol II, TFIIF, and TFIIB for efficient reinitiation (Yudkovsky et al., 2000). These studies could suggest that promoter escape and Mediator dissociation may be mechanistically related. Whether Mediator dissociation event itself would have any effects on promoter escape has not been investigated. Additionally, the genome-wide study of protein architecture of yeast genome by ChIP-exo suggests that PIC assembly might be mechanistically tied to PIC assembly of adjacent genes unless divided by insulators (Rossi et al., 2021). It may be possible that transcription initiation nearby the gene is mechanistically linked to its promoter escape. Also, phase separation plays a role in transcription especially during initiation and the transition from initiation to elongation (Hnisz et al., 2017; Rawat et al., 2021). Although this would be complicated to reconstitute and there is no such system currently, investigating if and how this would drive initiation and the transition would be of interest.

The promoter DNA used in our studies is a canonical TATA-containing promoter which represents only a small fraction of promoters in the genome. A vast majority of promoters contains TATA-less sequences, and transcription of both kinds of promoters depends on TFIID (Donczew et al., 2020; Warfield et al., 2017), a GTF involved in promoter recognition and TBP loading (Chen et al., 2021a; Louder et al., 2016; Patel et al., 2018). Biochemical characterization of transcription on TATA-less and TATA-containing promoters using nuclear extract showed that a high concentration of TBP can restore transcription upon Tfa1 (TFIID subunit) depletion on TATA-containing promoters but not on TATA-less promoters (Donczew and Hahn, 2018). Based on a recent structural study, TFIID can assemble in a PIC and interacts with the core module of TFIIH. Although the significance of this interaction between TFIID and TFIIH is still poorly understood, TFIID might function during early transcription initiation process in addition to TBP loading.

Building up the previous finding from an optical tweezer based study (Fazal et al., 2015b), it would also be interesting to perform mechanistic studies of initiation and the transition using optical tweezers and single-molecule FRET. Most of understanding of eukaryotic transcription come from ensemble measurements. Single-molecule approaches would be able separate major and minor processes during transcription and thus would be able to provide more detailed mechanisms depending on the design of the experiment. For example, there may be productive and unproductive pathways that are distinct to each other during TSS scanning or early synthesis of RNA. If so, when the checkpoints for these pathways lie would be interesting to investigate. Further, initiation mechanisms in the presence of co-activators should still be studied.

Our structure of the ITC on the G-less26 DNA template has revealed conformational changes of TFIIH and TFIIE as the PIC initiate transcription, despite the limited resolution of these regions due to its unstable and flexible nature. We were unable to determine interacting sites that could be mutated and functionally be tested. In addition, DNA between ~20-nt downstream of TATA and Ssl2 (TFIIH translocase subunit) was not seen. Further studies need to be done to achieve a higher resolution structure of the ITC and provide mechanistic insights into the transition from initiation to elongation. In a structural study of transcription initiation by bacterial RNA polymerase (RNAP) utilized the transcription factor TraR, which had been shown to allosterically inhibit transcription initiation, in order to stabilize intermediate complexes en route to forming an open complex (Chen et al., 2020). Although many of the factors that are involved in initiation and post-initiation, pooled CRISPR screening using reporter genes may be able to identify additional factors that could modulate initiation processes. If one could identify such factor(s), similar approach can be taken for structural studies of eukaryotic transcription.

Finally, structural studies of Pol II transcription post-initiation in the past used pre-synthesized RNA and mismatched transcription bubble in DNA to facilitate formation of DNA/RNA hybrid that can be accommodated in the Pol II active site. Our structural study shows the first Pol II transcription complexes formed as a result of efficient PIC assembly and initiation by the TFIIH translocase activity. Thus it underscores the development of transcription initiation system in vitro and provides a foundation for the further mechanistic studies of initiation and post-initiation processes using biochemical and biophysical approaches.

References

Abascal-Palacios, G., Ramsay, E.P., Beuron, F., Morris, E., and Vannini, A. (2018). Structural basis of RNA polymerase III transcription initiation. Nature *553*, 301-306.

Abdella, R., Talyzina, A., Chen, S., Inouye, C.J., Tjian, R., and He, Y. (2021). Structure of the human Mediator-bound transcription preinitiation complex. Science *372*, 52-56.

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

Aibara, S., Schilbach, S., and Cramer, P. (2021). Structures of mammalian RNA polymerase II pre-initiation complexes. Nature.

Akhtar, M.S., Heidemann, M., Tietjen, J.R., Zhang, D.W., Chapman, R.D., Eick, D., and Ansari, A.Z. (2009). TFIIH kinase places bivalent marks on the carboxy-terminal domain of RNA polymerase II. Mol Cell *34*, 387-393.

Alekseev, S., Nagy, Z., Sandoz, J., Weiss, A., Egly, J.M., Le May, N., and Coin, F. (2017). Transcription without XPB Establishes a Unified Helicase-Independent Mechanism of Promoter Opening in Eukaryotic Gene Expression. Mol Cell *65*, 504-514 e504.

Allen, B.L., and Taatjes, D.J. (2015). The Mediator complex: a central integrator of transcription. Nat Rev Mol Cell Biol *16*, 155-166.

Badjatia, N., Rossi, M.J., Bataille, A.R., Mittal, C., Lai, W.K.M., and Pugh, B.F. (2021). Acute stress drives global repression through two independent RNA polymerase II stalling events in Saccharomyces. Cell Rep *34*, 108640.

Baillat, D., Hakimi, M.A., Näär, A.M., Shilatifard, A., Cooch, N., and Shiekhattar, R. (2005). Integrator, a multiprotein mediator of small nuclear RNA processing, associates with the C-terminal repeat of RNA polymerase II. Cell *123*, 265-276.

Bangur, C.S., Pardee, T.S., and Ponticelli, A.S. (1997). Mutational analysis of the D1/E1 core helices and the conserved N-terminal region of yeast transcription factor IIB (TFIIB): identification of an N-terminal mutant that stabilizes TATA-binding protein-TFIIB-DNA complexes. Mol Cell Biol *17*, 6784-6793.

Barnes, C.O., Calero, M., Malik, I., Graham, B.W., Spahr, H., Lin, G., Cohen, A.E., Brown, I.S., Zhang, Q., Pullara, F., *et al.* (2015). Crystal Structure of a Transcribing RNA Polymerase II Complex Reveals a Complete Transcription Bubble. Mol Cell *59*, 258-269.

Basehoar, A.D., Zanton, S.J., and Pugh, B.F. (2004). Identification and distinct regulation of yeast TATA box-containing genes. Cell *116*, 699-709.

Bepler, T., Morin, A., Rapp, M., Brasch, J., Shapiro, L., Noble, A.J., and Berger, B. (2019). Positive-unlabeled convolutional neural networks for particle picking in cryo-electron micrographs. Nat Methods *16*, 1153-1160.

Bernecky, C., Plitzko, J.M., and Cramer, P. (2017). Structure of a transcribing RNA polymerase II-DSIF complex reveals a multidentate DNA-RNA clamp. Nat Struct Mol Biol *24*, 809-815.

Bernecky, C., and Taatjes, D.J. (2012). Activator-mediator binding stabilizes RNA polymerase II orientation within the human mediator-RNA polymerase II-TFIIF assembly. J Mol Biol *417*, 387-394.

Boija, A., Klein, I.A., Sabari, B.R., Dall'Agnese, A., Coffey, E.L., Zamudio, A.V., Li, C.H., Shrinivas, K., Manteiga, J.C., Hannett, N.M.*, et al.* (2018). Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains. Cell *175*, 1842-1855.e1816.

Bradsher, J., Coin, F., and Egly, J.M. (2000). Distinct roles for the helicases of TFIIH in transcript initiation and promoter escape. J Biol Chem *275*, 2532-2538.

Brzovic, P.S., Heikaus, C.C., Kisselev, L., Vernon, R., Herbig, E., Pacheco, D., Warfield, L., Littlefield, P., Baker, D., Klevit, R.E.*, et al.* (2011). The acidic transcription activator Gcn4 binds the mediator subunit Gal11/Med15 using a simple protein interface forming a fuzzy complex. Mol Cell *44*, 942-953.

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

Burgess, R.R., Travers, A.A., Dunn, J.J., and Bautz, E.K. (1969). Factor stimulating transcription by RNA polymerase. Nature *221*, 43-46.

Burugula, B.B., Jeronimo, C., Pathak, R., Jones, J.W., Robert, F., and Govind, C.K. (2014). Histone deacetylases and phosphorylated polymerase II C-terminal domain recruit Spt6 for cotranscriptional histone reassembly. Mol Cell Biol *34*, 4115-4129.

Bushnell, D.A., Westover, K.D., Davis, R.E., and Kornberg, R.D. (2004). Structural basis of transcription: an RNA polymerase II-TFIIB cocrystal at 4.5 Angstroms. Science *303*, 983-988.

Busso, D., Keriel, A., Sandrock, B., Poterszman, A., Gileadi, O., and Egly, J.M. (2000). Distinct regions of MAT1 regulate cdk7 kinase and TFIIH transcription activities. J Biol Chem *275*, 22815-22823.

Cabart, P., Jin, H., Li, L., and Kaplan, C.D. (2014). Activation and reactivation of the RNA polymerase II trigger loop for intrinsic RNA cleavage and catalysis. Transcription *5*, e28869.

Čabart, P., Újvári, A., Pal, M., and Luse, D.S. (2011). Transcription factor TFIIF is not required for initiation by RNA polymerase II, but it is essential to stabilize transcription factor TFIIB in early elongation complexes. Proc Natl Acad Sci U S A *108*, 15786-15791.

Cai, H., and Luse, D.S. (1987). Transcription initiation by RNA polymerase II in vitro. Properties of preinitiation, initiation, and elongation complexes. J Biol Chem *262*, 298-304.

Callaway, E. (2020). Revolutionary cryo-EM is taking over structural biology. Nature *578*, 201.

Carpousis, A.J., and Gralla, J.D. (1980). Cycling of ribonucleic acid polymerase to produce oligonucleotides during initiation in vitro at the lac UV5 promoter. Biochemistry *19*, 3245-3253.

Cevher, M.A., Shi, Y., Li, D., Chait, B.T., Malik, S., and Roeder, R.G. (2014). Reconstitution of active human core Mediator complex reveals a critical role of the MED14 subunit. Nat Struct Mol Biol *21*, 1028-1034.

Chen, B.S., and Hampsey, M. (2004). Functional interaction between TFIIB and the Rpb2 subunit of RNA polymerase II: implications for the mechanism of transcription initiation. Mol Cell Biol *24*, 3983-3991.

Chen, J., Chiu, C., Gopalkrishnan, S., Chen, A.Y., Olinares, P.D.B., Saecker, R.M., Winkelman, J.T., Maloney, M.F., Chait, B.T., Ross, W.*, et al.* (2020). Stepwise Promoter Melting by Bacterial RNA Polymerase. Mol Cell *78*, 275-288.e276.

Chen, W., and Struhl, K. (1985). Yeast mRNA initiation sites are determined primarily by specific sequences, not by the distance from the TATA element. Embo j *4*, 3273-3280.

Chen, X., Qi, Y., Wu, Z., Wang, X., Li, J., Zhao, D., Hou, H., Li, Y., Yu, Z., Liu, W.*, et al.* (2021a). Structural insights into preinitiation complex assembly on core promoters. Science *372*.

Chen, X., Yin, X., Li, J., Wu, Z., Qi, Y., Wang, X., Liu, W., and Xu, Y. (2021b). Structures of the human Mediator and Mediator-bound preinitiation complex. Science.

Cheung, A.C., and Cramer, P. (2011). Structural basis of RNA polymerase II backtracking, arrest and reactivation. Nature *471*, 249-253.

Cho, E.J., Rodriguez, C.R., Takagi, T., and Buratowski, S. (1998). Allosteric interactions between capping enzyme subunits and the RNA polymerase II carboxy-terminal domain. Genes Dev *12*, 3482-3487.

Cho, E.J., Takagi, T., Moore, C.R., and Buratowski, S. (1997). mRNA capping enzyme is recruited to the transcription complex by phosphorylation of the RNA polymerase II carboxy-terminal domain. Genes Dev *11*, 3319-3326.

Choi, W.S., Lin, Y.C., and Gralla, J.D. (2004). The Schizosaccharomyces pombe open promoter bubble: mammalian-like arrangement and properties. J Mol Biol *340*, 981-989.

Chong, S., Dugast-Darzacq, C., Liu, Z., Dong, P., Dailey, G.M., Cattoglio, C., Heckert, A., Banala, S., Lavis, L., Darzacq, X.*, et al.* (2018). Imaging dynamic and selective low-complexity domain interactions that control gene transcription. Science *361*.

Coin, F., Proietti De Santis, L., Nardo, T., Zlobinskaya, O., Stefanini, M., and Egly, J.M. (2006). p8/TTD-A as a repair-specific TFIIH subunit. Mol Cell *21*, 215-226.

Conaway, R.C., and Conaway, J.W. (1993). General initiation factors for RNA polymerase II. Annu Rev Biochem *62*, 161-190.

Conaway, R.C., and Conaway, J.W. (2012). The Mediator complex and transcription elongation. Biochim Biophys Acta.

Coppola, J.A., and Luse, D.S. (1984). Purification and characterization of ternary complexes containing accurately initiated RNA polymerase II and less than 20 nucleotides of RNA. J Mol Biol *178*, 415-437.

Core, L., and Adelman, K. (2019). Promoter-proximal pausing of RNA polymerase II: a nexus of gene regulation. Genes Dev *33*, 960-982.

Cramer, P., Bushnell, D.A., and Kornberg, R.D. (2001). Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution. Science *292*, 1863-1876.

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

Crick, F.H. (1958). On protein synthesis. Symp Soc Exp Biol *12*, 138-163.

Crickard, J.B., Fu, J., and Reese, J.C. (2016). Biochemical Analysis of Yeast Suppressor of Ty 4/5 (Spt4/5) Reveals the Importance of Nucleic Acid Interactions in the Prevention of RNA Polymerase II Arrest. J Biol Chem *291*, 9853-9870.

Damodaren, N., Van Eeuwen, T., Zamel, J., Lin-Shiao, E., Kalisman, N., and Murakami, K. (2017). Def1 interacts with TFIIH and modulates RNA polymerase II transcription. Proc Natl Acad Sci U S A *114*, 13230-13235.

Dienemann, C., Schwalb, B., Schilbach, S., and Cramer, P. (2019a). Promoter Distortion and Opening in the RNA Polymerase II Cleft. Mol Cell *73*, 97-106.e104.

Dienemann, C., Schwalb, B., Schilbach, S., and Cramer, P. (2019b). Promoter Distortion and Opening in the RNA Polymerase II Cleft. Mol Cell *73*, 97-106 e104.

Dollinger, R., and Gilmour, D.S. (2021). Regulation of Promoter Proximal Pausing of RNA Polymerase II in Metazoans. Journal of Molecular Biology *433*, 166897.

Donczew, R., and Hahn, S. (2018). Mechanistic Differences in Transcription Initiation at TATA-Less and TATA-Containing Promoters. Mol Cell Biol *38*.

Donczew, R., Warfield, L., Pacheco, D., Erijman, A., and Hahn, S. (2020). Two roles for the yeast transcription coactivator SAGA and a set of genes redundantly regulated by TFIID and SAGA. Elife *9*.

Dotson, M.R., Yuan, C.X., Roeder, R.G., Myers, L.C., Gustafsson, C.M., Jiang, Y.W., Li, Y., Kornberg, R.D., and Asturias, F.J. (2000). Structural organization of yeast and mammalian mediator complexes. Proc Natl Acad Sci U S A *97*, 14307-14310.

Dvir, A., Conaway, R.C., and Conaway, J.W. (1996). Promoter escape by RNA polymerase II. A role for an ATP cofactor in suppression of arrest by polymerase at promoter-proximal sites. J Biol Chem *271*, 23352-23356.

Dvir, A., Conaway, R.C., and Conaway, J.W. (1997a). A role for TFIIH in controlling the activity of early RNA polymerase II elongation complexes. Proc Natl Acad Sci U S A *94*, 9006-9010.

Dvir, A., Tan, S., Conaway, J.W., and Conaway, R.C. (1997b). Promoter escape by RNA polymerase II. Formation of an escape-competent transcriptional intermediate is a prerequisite for exit of polymerase from the promoter. J Biol Chem *272*, 28175-28178.

Egloff, S., O'Reilly, D., Chapman, R.D., Taylor, A., Tanzhaus, K., Pitts, L., Eick, D., and Murphy, S. (2007). Serine-7 of the RNA polymerase II CTD is specifically required for snRNA gene expression. Science *318*, 1777-1779.

Ehara, H., Yokoyama, T., Shigematsu, H., Yokoyama, S., Shirouzu, M., and Sekine, S.I. (2017). Structure of the complete elongation complex of RNA polymerase II with basal factors. Science *357*, 921-924.

Ehrensberger, A.H., Kelly, G.P., and Svejstrup, J.Q. (2013). Mechanistic interpretation of promoter-proximal peaks and RNAPII density maps. Cell *154*, 713-715.

El Khattabi, L., Zhao, H., Kalchschmidt, J., Young, N., Jung, S., Van Blerkom, P., Kieffer-Kwon, P., Kieffer-Kwon, K.R., Park, S., Wang, X*., et al.* (2019). A Pliable Mediator Acts as a Functional Rather Than an Architectural Bridge between Promoters and Enhancers. Cell *178*, 1145-1158.e1120.

Elrod, N.D., Henriques, T., Huang, K.L., Tatomer, D.C., Wilusz, J.E., Wagner, E.J., and Adelman, K. (2019). The Integrator Complex Attenuates Promoter-Proximal Transcription at Protein-Coding Genes. Mol Cell *76*, 738-752.e737.

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

Engel, C., Gubbey, T., Neyer, S., Sainsbury, S., Oberthuer, C., Baejen, C., Bernecky, C., and Cramer, P. (2017). Structural Basis of RNA Polymerase I Transcription Initiation. Cell *169*, 120-131.e122.

Fabrega, C., Shen, V., Shuman, S., and Lima, C.D. (2003). Structure of an mRNA capping enzyme bound to the phosphorylated carboxy-terminal domain of RNA polymerase II. Mol Cell *11*, 1549-1561.

Fairman-Williams, M.E., Guenther, U.P., and Jankowsky, E. (2010). SF1 and SF2 helicases: family matters. Curr Opin Struct Biol *20*, 313-324.

Fan, X., Wang, J., Zhang, X., Yang, Z., Zhang, J.C., Zhao, L., Peng, H.L., Lei, J., and Wang, H.W. (2019). Single particle cryo-EM reconstruction of 52 kDa streptavidin at 3.2 Angstrom resolution. Nat Commun *10*, 2386.

Fant, C.B., Levandowski, C.B., Gupta, K., Maas, Z.L., Moir, J., Rubin, J.D., Sawyer, A., Esbin, M.N., Rimel, J.K., Luyties, O.*, et al.* (2020). TFIID Enables RNA Polymerase II Promoter-Proximal Pausing. Mol Cell *78*, 785-793.e788.

Fazal, F.M., Meng, C.A., Murakami, K., Kornberg, R.D., and Block, S.M. (2015a). Real-time observation of the initiation of RNA polymerase II transcription. Nature.

Fazal, F.M., Meng, C.A., Murakami, K., Kornberg, R.D., and Block, S.M. (2015b). Real-time observation of the initiation of RNA polymerase II transcription. Nature *525*, 274-277.

Fishburn, J., Galburt, E., and Hahn, S. (2016). Transcription Start Site Scanning and the Requirement for ATP during Transcription Initiation by RNA Polymerase II. J Biol Chem *291*, 13040-13047.

Fishburn, J., and Hahn, S. (2012). Architecture of the yeast RNA polymerase II open complex and regulation of activity by TFIIF. Mol Cell Biol *32*, 12-25.

Fishburn, J., Tomko, E., Galburt, E., and Hahn, S. (2015). Double-stranded DNA translocase activity of transcription factor TFIIH and the mechanism of RNA polymerase II open complex formation. Proc Natl Acad Sci U S A *112*, 3961-3966.

Flanagan, P.M., Kelleher, R.J., 3rd, Sayre, M.H., Tschochner, H., and Kornberg, R.D. (1991). A mediator required for activation of RNA polymerase II transcription in vitro. Nature *350*, 436-438.

Fondell, J.D., Ge, H., and Roeder, R.G. (1996). Ligand induction of a transcriptionally active thyroid hormone receptor coactivator complex. Proc Natl Acad Sci U S A *93*, 8329-8333.

Fujiwara, R., Damodaren, N., Wilusz, J.E., and Murakami, K. (2019). The capping enzyme facilitates promoter escape and assembly of a follow-on preinitiation complex for reinitiation. Proc Natl Acad Sci U S A *116*, 22573-22582.

Fujiwara, R., and Murakami, K. (2019). In vitro reconstitution of yeast RNA polymerase II transcription initiation with high efficiency. Methods *159-160*, 82-89.

Ge, H., and Roeder, R.G. (1994). Purification, cloning, and characterization of a human coactivator, PC4, that mediates transcriptional activation of class II genes. Cell *78*, 513-523.

Geiger, J.H., Hahn, S., Lee, S., and Sigler, P.B. (1996). Crystal structure of the yeast TFIIA/TBP/DNA complex. Science *272*, 830-836.

Giardina, C., and Lis, J.T. (1993). DNA melting on yeast RNA polymerase II promoters. Science *261*, 759-762.

Gibbons, B.J., Brignole, E.J., Azubel, M., Murakami, K., Voss, N.R., Bushnell, D.A., Asturias, F.J., and Kornberg, R.D. (2012). Subunit architecture of general transcription factor TFIIH. Proc Natl Acad Sci U S A *109*, 1949-1954.

Glover-Cutter, K., Larochelle, S., Erickson, B., Zhang, C., Shokat, K., Fisher, R.P., and Bentley, D.L. (2009). TFIIH-associated Cdk7 kinase functions in phosphorylation of C-terminal domain Ser7 residues, promoter-proximal pausing, and termination by RNA polymerase II. Mol Cell Biol *29*, 5455-5464.

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.

Goldman, S.R., Ebright, R.H., and Nickels, B.E. (2009). Direct detection of abortive RNA transcripts in vivo. Science *324*, 927-928.

Goodfellow, S.J., and Zomerdijk, J.C. (2013). Basic mechanisms in RNA polymerase I transcription of the ribosomal RNA genes. Subcell Biochem *61*, 211-236.

Goodrich, J.A., and Tjian, R. (1994). Transcription factors IIE and IIH and ATP hydrolysis direct promoter clearance by RNA polymerase II. Cell *77*, 145-156.

Greber, B.J., Perez-Bertoldi, J.M., Lim, K., Iavarone, A.T., Toso, D.B., and Nogales, E. (2020). The cryoelectron microscopy structure of the human CDK-activating kinase. Proc Natl Acad Sci U S A *117*, 22849-22857.

Greber, B.J., Toso, D.B., Fang, J., and Nogales, E. (2019). The complete structure of the human TFIIH core complex. Elife *8*.

Haberle, V., and Stark, A. (2018). Eukaryotic core promoters and the functional basis of transcription initiation. Nat Rev Mol Cell Biol *19*, 621-637.

He, Y., Fang, J., Taatjes, D.J., and Nogales, E. (2013). Structural visualization of key steps in human transcription initiation. Nature *495*, 481-486.

He, Y., Yan, C., Fang, J., Inouye, C., Tjian, R., Ivanov, I., and Nogales, E. (2016). Near-atomic resolution visualization of human transcription promoter opening. Nature *533*, 359-365.

Henry, N.L., Bushnell, D.A., and Kornberg, R.D. (1996). A yeast transcriptional stimulatory protein similar to human PC4. J Biol Chem *271*, 21842-21847.

Hieb, A.R., Halsey, W.A., Betterton, M.D., Perkins, T.T., Kugel, J.F., and Goodrich, J.A. (2007). TFIIA changes the conformation of the DNA in TBP/TATA complexes and increases their kinetic stability. J Mol Biol *372*, 619-632.

Hnisz, D., Shrinivas, K., Young, R.A., Chakraborty, A.K., and Sharp, P.A. (2017). A Phase Separation Model for Transcriptional Control. Cell *169*, 13-23.

Hobson, D.J., Wei, W., Steinmetz, L.M., and Svejstrup, J.Q. (2012). RNA polymerase II collision interrupts convergent transcription. Mol Cell *48*, 365-374.

Hoffmann, N.A., Jakobi, A.J., Moreno-Morcillo, M., Glatt, S., Kosinski, J., Hagen, W.J., Sachse, C., and Müller, C.W. (2015). Molecular structures of unbound and transcribing RNA polymerase III. Nature *528*, 231-236.

Holstege, F.C., Fiedler, U., and Timmers, H.T. (1997). Three transitions in the RNA polymerase II transcription complex during initiation. Embo j *16*, 7468-7480.

Holstege, F.C., Tantin, D., Carey, M., van der Vliet, P.C., and Timmers, H.T. (1995). The requirement for the basal transcription factor IIE is determined by the helical stability of promoter DNA. EMBO J *14*, 810-819.

Holstege, F.C., van der Vliet, P.C., and Timmers, H.T. (1996). Opening of an RNA polymerase II promoter occurs in two distinct steps and requires the basal transcription factors IIE and IIH. Embo j *15*, 1666-1677.

Hsin, J.P., and Manley, J.L. (2012). The RNA polymerase II CTD coordinates transcription and RNA processing. Genes Dev *26*, 2119-2137.

Imbalzano, A.N., Zaret, K.S., and Kingston, R.E. (1994). Transcription factor (TF) IIB and TFIIA can independently increase the affinity of the TATA-binding protein for DNA. J Biol Chem *269*, 8280-8286.

Izban, M.G., and Luse, D.S. (1992). The RNA polymerase II ternary complex cleaves the nascent transcript in a 3'----5' direction in the presence of elongation factor SII. Genes Dev *6*, 1342-1356.

Jacob, F., and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol *3*, 318-356.

Jeronimo, C., Langelier, M.F., Bataille, A.R., Pascal, J.M., Pugh, B.F., and Robert, F. (2016). Tail and Kinase Modules Differently Regulate Core Mediator Recruitment and Function In Vivo. Mol Cell *64*, 455-466.

Jeronimo, C., and Robert, F. (2014). Kin28 regulates the transient association of Mediator with core promoters. Nat Struct Mol Biol *21*, 449-455.

Jiang, C., and Pugh, B.F. (2009). Nucleosome positioning and gene regulation: advances through genomics. Nat Rev Genet *10*, 161-172.

Kall, L., Canterbury, J.D., Weston, J., Noble, W.S., and MacCoss, M.J. (2007). Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nat Methods *4*, 923-925.

Kapanidis, A.N., Margeat, E., Ho, S.O., Kortkhonjia, E., Weiss, S., and Ebright, R.H. (2006). Initial transcription by RNA polymerase proceeds through a DNA-scrunching mechanism. Science *314*, 1144-1147.

Kays, A.R., and Schepartz, A. (2000). Virtually unidirectional binding of TBP to the AdMLP TATA box within the quaternary complex with TFIIA and TFIIB. Chem Biol *7*, 601-610.

Kedinger, C., Gniazdowski, M., Mandel, J.L., Jr., Gissinger, F., and Chambon, P. (1970). Alpha-amanitin: a specific inhibitor of one of two DNA-pendent RNA polymerase activities from calf thymus. Biochem Biophys Res Commun *38*, 165-171.

Keene, R.G., and Luse, D.S. (1999). Initially transcribed sequences strongly affect the extent of abortive initiation by RNA polymerase II. J Biol Chem *274*, 11526-11534.

Kelleher, R.J., 3rd, Flanagan, P.M., and Kornberg, R.D. (1990). A novel mediator between activator proteins and the RNA polymerase II transcription apparatus. Cell *61*, 1209-1215.

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

Khoshouei, M., Radjainia, M., Baumeister, W., and Danev, R. (2017). Cryo-EM structure of haemoglobin at 3.2 Å determined with the Volta phase plate. Nat Commun *8*, 16099.

Kim, J.L., Nikolov, D.B., and Burley, S.K. (1993a). Co-crystal structure of TBP recognizing the minor groove of a TATA element. Nature *365*, 520-527.

Kim, T.K., Lagrange, T., Wang, Y.H., Griffith, J.D., Reinberg, D., and Ebright, R.H. (1997). Trajectory of DNA in the RNA polymerase II transcription preinitiation complex. Proc Natl Acad Sci U S A *94*, 12268-12273.

Kim, Y., Geiger, J.H., Hahn, S., and Sigler, P.B. (1993b). Crystal structure of a yeast TBP/TATA-box complex. Nature *365*, 512-520.

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.

Koleske, A.J., and Young, R.A. (1994). An RNA polymerase II holoenzyme responsive to activators. Nature *368*, 466-469.

Komissarova, N., Kireeva, M.L., Becker, J., Sidorenkov, I., and Kashlev, M. (2003). Engineering of elongation complexes of bacterial and yeast RNA polymerases. Methods Enzymol *371*, 233-251.

Kornberg, R.D. (2007). The molecular basis of eukaryotic transcription. Proc Natl Acad Sci U S A *104*, 12955-12961.

Kostrewa, D., Zeller, M.E., Armache, K.J., Seizl, M., Leike, K., Thomm, M., and Cramer, P. (2009). RNA polymerase II-TFIIB structure and mechanism of transcription initiation. Nature *462*, 323-330.

Krishnamurthy, S., He, X., Reyes-Reyes, M., Moore, C., and Hampsey, M. (2004). Ssu72 Is an RNA polymerase II CTD phosphatase. Mol Cell *14*, 387-394.

Kuehner, J.N., and Brow, D.A. (2006). Quantitative analysis of in vivo initiator selection by yeast RNA polymerase II supports a scanning model. J Biol Chem *281*, 14119-14128.

Kugel, J.F., and Goodrich, J.A. (2002). Translocation after synthesis of a four-nucleotide RNA commits RNA polymerase II to promoter escape. Mol Cell Biol *22*, 762-773.

Lidschreiber, M., Leike, K., and Cramer, P. (2013). Cap completion and C-terminal repeat domain kinase recruitment underlie the initiation-elongation transition of RNA polymerase II. Mol Cell Biol *33*, 3805-3816.

Liebschner, D., Afonine, P.V., Baker, M.L., Bunkoczi, G., Chen, V.B., Croll, T.I., Hintze, B., Hung, L.W., Jain, S., McCoy, A.J.*, et al.* (2019). Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. Acta Crystallogr D Struct Biol *75*, 861-877.

Lin, Y.C., and Gralla, J.D. (2005). Stimulation of the XPB ATP-dependent helicase by the beta subunit of TFIIE. Nucleic Acids Res *33*, 3072-3081.

Liu, X., Bushnell, D.A., Wang, D., Calero, G., and Kornberg, R.D. (2010). Structure of an RNA polymerase II-TFIIB complex and the transcription initiation mechanism. Science *327*, 206-209.

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

Lu, X., Zhu, X., Li, Y., Liu, M., Yu, B., Wang, Y., Rao, M., Yang, H., Zhou, K., Wang, Y.*, et al.* (2016). Multiple P-TEFbs cooperatively regulate the release of promoter-proximally paused RNA polymerase II. Nucleic Acids Res *44*, 6853-6867.

Luse, D.S. (2013). Promoter clearance by RNA polymerase II. Biochim Biophys Acta *1829*, 63-68.

Luse, D.S. (2019). Insight into promoter clearance by RNA polymerase II. Proc Natl Acad Sci U S A *116*, 22426-22428.

Ly, E., Powell, A.E., Goodrich, J.A., and Kugel, J.F. (2020). Release of Human TFIIB from Actively Transcribing Complexes Is Triggered upon Synthesis of 7- and 9-nt RNAs. J Mol Biol *432*, 4049-4060.

Mandal, S.S., Chu, C., Wada, T., Handa, H., Shatkin, A.J., and Reinberg, D. (2004). Functional interactions of RNA-capping enzyme with factors that positively and negatively regulate promoter escape by RNA polymerase II. Proc Natl Acad Sci U S A *101*, 7572-7577.

Mao, X., Schwer, B., and Shuman, S. (1995). Yeast mRNA cap methyltransferase is a 50-kilodalton protein encoded by an essential gene. Mol Cell Biol *15*, 4167-4174.

Marshall, N.F., and Price, D.H. (1995). Purification of P-TEFb, a transcription factor required for the transition into productive elongation. J Biol Chem *270*, 12335-12338.

Martinez-Rucobo, F.W., Kohler, R., van de Waterbeemd, M., Heck, A.J., Hemann, M., Herzog, F., Stark, H., and Cramer, P. (2015). Molecular Basis of Transcription-Coupled Pre-mRNA Capping. Mol Cell *58*, 1079-1089.

Mavrich, T.N., Jiang, C., Ioshikhes, I.P., Li, X., Venters, B.J., Zanton, S.J., Tomsho, L.P., Qi, J., Glaser, R.L., Schuster, S.C.*, et al.* (2008). Nucleosome organization in the Drosophila genome. Nature *453*, 358-362.

McCracken, S., Fong, N., Rosonina, E., Yankulov, K., Brothers, G., Siderovski, D., Hessel, A., Foster, S., Shuman, S., and Bentley, D.L. (1997). 5'-Capping enzymes are targeted to pre-mRNA by binding to the phosphorylated carboxy-terminal domain of RNA polymerase II. Genes Dev *11*, 3306-3318.

Meyer, K.D., Lin, S.C., Bernecky, C., Gao, Y., and Taatjes, D.J. (2010). p53 activates transcription by directing structural shifts in Mediator. Nat Struct Mol Biol *17*, 753-760.

Mindell, J.A., and Grigorieff, N. (2003). Accurate determination of local defocus and specimen tilt in electron microscopy. J Struct Biol *142*, 334-347.

Missra, A., and Gilmour, D.S. (2010). Interactions between DSIF (DRB sensitivity inducing factor), NELF (negative elongation factor), and the Drosophila RNA polymerase II transcription elongation complex. Proc Natl Acad Sci U S A *107*, 11301-11306.

Mitra, A.K. (2019). Visualization of biological macromolecules at near-atomic resolution: cryo-electron microscopy comes of age. Acta Crystallogr F Struct Biol Commun *75*, 3-11.

Mosley, A.L., Pattenden, S.G., Carey, M., Venkatesh, S., Gilmore, J.M., Florens, L., Workman, J.L., and Washburn, M.P. (2009). Rtr1 is a CTD phosphatase that regulates RNA polymerase II during the transition from serine 5 to serine 2 phosphorylation. Mol Cell *34*, 168-178.

Murakami, K., Calero, G., Brown, C.R., Liu, X., Davis, R.E., Boeger, H., and Kornberg, R.D. (2013a). Formation and fate of a complete 31-protein RNA polymerase II transcription preinitiation complex. J Biol Chem *288*, 6325-6332.

Murakami, K., Calero, G., Brown, C.R., Liu, X., Davis, R.E., Boeger, H., and Kornberg, R.D. (2013b). Formation and fate of a complete 31-protein RNA polymerase II transcription preinitiation complex. J Biol Chem *288*, 6325-6332.

Murakami, K., Elmlund, H., Kalisman, N., Bushnell, D.A., Adams, C.M., Azubel, M., Elmlund, D., Levi-Kalisman, Y., Liu, X., Gibbons, B.J.*, et al.* (2013c). Architecture of an RNA polymerase II transcription pre-initiation complex. Science *342*, 1238724.

Murakami, K., Mattei, P.J., Davis, R.E., Jin, H., Kaplan, C.D., and Kornberg, R.D. (2015a). Uncoupling Promoter Opening from Start-Site Scanning. Mol Cell *59*, 133-138.

Murakami, K., Tsai, K.L., Kalisman, N., Bushnell, D.A., Asturias, F.J., and Kornberg, R.D. (2015b). Structure of an RNA polymerase II preinitiation complex. Proc Natl Acad Sci U S A *112*, 13543-13548.

Murata, K., and Wolf, M. (2018). Cryo-electron microscopy for structural analysis of dynamic biological macromolecules. Biochim Biophys Acta Gen Subj *1862*, 324-334.

Muse, G.W., Gilchrist, D.A., Nechaev, S., Shah, R., Parker, J.S., Grissom, S.F., Zeitlinger, J., and Adelman, K. (2007). RNA polymerase is poised for activation across the genome. Nat Genet *39*, 1507-1511.

Nagai, S., Davis, R.E., Mattei, P.J., Eagen, K.P., and Kornberg, R.D. (2017). Chromatin potentiates transcription. Proc Natl Acad Sci U S A *114*, 1536-1541.

Nakane, T., Kimanius, D., Lindahl, E., and Scheres, S.H. (2018). Characterisation of molecular motions in cryo-EM single-particle data by multi-body refinement in RELION. Elife *7*.

Nechaev, S., Fargo, D.C., dos Santos, G., Liu, L., Gao, Y., and Adelman, K. (2010). Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila. Science *327*, 335-338.

Neyer, S., Kunz, M., Geiss, C., Hantsche, M., Hodirnau, V.V., Seybert, A., Engel, C., Scheffer, M.P., Cramer, P., and Frangakis, A.S. (2016). Structure of RNA polymerase I transcribing ribosomal DNA genes. Nature *540*, 607-610.

Ng, H.H., Robert, F., Young, R.A., and Struhl, K. (2003). Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity. Mol Cell *11*, 709-719.

Nguyen, V.Q., Ranjan, A., Liu, S., Tang, X., Ling, Y.H., Wisniewski, J., Mizuguchi, G., Li, K.Y., Jou, V., Zheng, Q.*, et al.* (2020). Spatio-Temporal Coordination of Transcription Preinitiation Complex Assembly in Live Cells. bioRxiv, 2020.2012.2030.424853.

Nilson, K.A., Guo, J., Turek, M.E., Brogie, J.E., Delaney, E., Luse, D.S., and Price, D.H. (2015). THZ1 Reveals Roles for Cdk7 in Co-transcriptional Capping and Pausing. Mol Cell *59*, 576-587.

Noe Gonzalez, M., Sato, S., Tomomori-Sato, C., Conaway, J.W., and Conaway, R.C. (2018). CTD-dependent and -independent mechanisms govern co-transcriptional capping of Pol II transcripts. Nat Commun *9*, 3392.

Ohkuma, Y., and Roeder, R.G. (1994). Regulation of TFIIH ATPase and kinase activities by TFIIE during active initiation complex formation. Nature *368*, 160-163.

Osman, S., Mohammad, E., Lidschreiber, M., Stuetzer, A., Bazsó, F.L., Maier, K.C., Urlaub, H., and Cramer, P. (2021). The Cdk8 kinase module regulates interaction of the Mediator complex with RNA polymerase II. J Biol Chem *296*, 100734.

Pal, M., McKean, D., and Luse, D.S. (2001). Promoter clearance by RNA polymerase II is an extended, multistep process strongly affected by sequence. Mol Cell Biol *21*, 5815-5825.

Pal, M., Ponticelli, A.S., and Luse, D.S. (2005). The role of the transcription bubble and TFIIB in promoter clearance by RNA polymerase II. Mol Cell *19*, 101-110.

Parvin, J.D., and Sharp, P.A. (1993). DNA topology and a minimal set of basal factors for transcription by RNA polymerase II. Cell *73*, 533-540.

Patel, A.B., Louder, R.K., Greber, B.J., Grünberg, S., Luo, J., Fang, J., Liu, Y., Ranish, J., Hahn, S., and Nogales, E. (2018). Structure of human TFIID and mechanism of TBP loading onto promoter DNA. Science *362*.

Petrenko, N., Jin, Y., Wong, K.H., and Struhl, K. (2016). Mediator Undergoes a Compositional Change during Transcriptional Activation. Mol Cell *64*, 443-454.

Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., and Ferrin, T.E. (2004). UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem *25*, 1605-1612.

Pettersen, E.F., Goddard, T.D., Huang, C.C., Meng, E.C., Couch, G.S., Croll, T.I., Morris, J.H., and Ferrin, T.E. (2021). UCSF ChimeraX: Structure visualization for researchers, educators, and developers. Protein Sci *30*, 70-82.

Plaschka, C., Hantsche, M., Dienemann, C., Burzinski, C., Plitzko, J., and Cramer, P. (2016). Transcription initiation complex structures elucidate DNA opening. Nature *533*, 353-358.

Plaschka, C., Larivière, L., Wenzeck, L., Seizl, M., Hemann, M., Tegunov, D., Petrotchenko, E.V., Borchers, C.H., Baumeister, W., Herzog, F.*, et al.* (2015). Architecture of the RNA polymerase II-Mediator core initiation complex. Nature *518*, 376-380.

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.

Qiu, C., Jin, H., Vvedenskaya, I., Llenas, J.A., Zhao, T., Malik, I., Visbisky, A.M., Schwartz, S.L., Cui, P., Čabart, P.*, et al.* (2020). Universal promoter scanning by Pol II during transcription initiation in Saccharomyces cerevisiae. Genome Biol *21*, 132.

Ranish, J.A., Hahn, S., Lu, Y., Yi, E.C., Li, X.J., Eng, J., and Aebersold, R. (2004). Identification of TFB5, a new component of general transcription and DNA repair factor IIH. Nat Genet *36*, 707-713.

Rasmussen, E.B., and Lis, J.T. (1993). In vivo transcriptional pausing and cap formation on three Drosophila heat shock genes. Proc Natl Acad Sci U S A *90*, 7923-7927.

Rawat, P., Boehning, M., Hummel, B., Aprile-Garcia, F., Pandit, A.S., Eisenhardt, N., Khavaran, A., Niskanen, E., Vos, S.M., Palvimo, J.J.*, et al.* (2021). Stress-induced nuclear condensation of NELF drives transcriptional downregulation. Mol Cell *81*, 1013-1026.e1011.

Rengachari, S., Schilbach, S., Aibara, S., Dienemann, C., and Cramer, P. (2021). Structure of human Mediator-RNA polymerase II pre-initiation complex. Nature.

Rhee, H.S., and Pugh, B.F. (2012). Genome-wide structure and organization of eukaryotic pre-initiation complexes. Nature *483*, 295-301.

Rinner, O., Seebacher, J., Walzthoeni, T., Mueller, L.N., Beck, M., Schmidt, A., Mueller, M., and Aebersold, R. (2008). Identification of cross-linked peptides from large sequence databases. Nat Methods *5*, 315-318.

Robinson, P.J., Bushnell, D.A., Trnka, M.J., Burlingame, A.L., and Kornberg, R.D. (2012). Structure of the mediator head module bound to the carboxy-terminal domain of RNA polymerase II. Proc Natl Acad Sci U S A *109*, 17931-17935.

Robinson, P.J., Trnka, M.J., Bushnell, D.A., Davis, R.E., Mattei, P.J., Burlingame, A.L., and Kornberg, R.D. (2016a). Structure of a Complete Mediator-RNA Polymerase II Pre-Initiation Complex. Cell *166*, 1411-1422.e1416.

Robinson, P.J., Trnka, M.J., Bushnell, D.A., Davis, R.E., Mattei, P.J., Burlingame, A.L., and Kornberg, R.D. (2016b). Structure of a Complete Mediator-RNA Polymerase II Pre-Initiation Complex. Cell *166*, 1411-1422 e1416.

Rodriguez, C.R., Cho, E.J., Keogh, M.C., Moore, C.L., Greenleaf, A.L., and Buratowski, S. (2000). Kin28, the TFIIH-associated carboxy-terminal domain kinase, facilitates the recruitment of mRNA processing machinery to RNA polymerase II. Mol Cell Biol *20*, 104-112.

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

Roeder, R.G., and Rutter, W.J. (1970). Specific nucleolar and nucleoplasmic RNA polymerases. Proc Natl Acad Sci U S A *65*, 675-682.

Rosen, G.A., Baek, I., Friedman, L.J., Joo, Y.J., Buratowski, S., and Gelles, J. (2020). Dynamics of RNA polymerase II and elongation factor Spt4/5 recruitment during activator-dependent transcription. Proc Natl Acad Sci U S A *117*, 32348-32357.

Rossi, M.J., Kuntala, P.K., Lai, W.K.M., Yamada, N., Badjatia, N., Mittal, C., Kuzu, G., Bocklund, K., Farrell, N.P., Blanda, T.R.*, et al.* (2021). A high-resolution protein architecture of the budding yeast genome. Nature *592*, 309-314.

Sabari, B.R., Dall'Agnese, A., Boija, A., Klein, I.A., Coffey, E.L., Shrinivas, K., Abraham, B.J., Hannett, N.M., Zamudio, A.V., Manteiga, J.C.*, et al.* (2018). Coactivator condensation at super-enhancers links phase separation and gene control. Science *361*.

Saeki, H., and Svejstrup, J.Q. (2009). Stability, flexibility, and dynamic interactions of colliding RNA polymerase II elongation complexes. Mol Cell *35*, 191-205.

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

Scheres, S.H. (2012). RELION: implementation of a Bayesian approach to cryo-EM structure determination. J Struct Biol *180*, 519-530.

Schilbach, S., Aibara, S., Dienemann, C., Grabbe, F., and Cramer, P. (2021). Structure of RNA polymerase II pre-initiation complex at 2.9 Å defines initial DNA opening. Cell.

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.

Schroeder, S.C., Schwer, B., Shuman, S., and Bentley, D. (2000). Dynamic association of capping enzymes with transcribing RNA polymerase II. Genes Dev *14*, 2435-2440.

Schroeder, S.C., Zorio, D.A., Schwer, B., Shuman, S., and Bentley, D. (2004). A function of yeast mRNA cap methyltransferase, Abd1, in transcription by RNA polymerase II. Mol Cell *13*, 377-387.

Shaevitz, J.W., Abbondanzieri, E.A., Landick, R., and Block, S.M. (2003). Backtracking by single RNA polymerase molecules observed at near-base-pair resolution. Nature *426*, 684-687.

Shen, P.S. (2018). The 2017 Nobel Prize in Chemistry: cryo-EM comes of age. Anal Bioanal Chem *410*, 2053-2057.

Sheridan, R.M., Fong, N., D'Alessandro, A., and Bentley, D.L. (2019). Widespread Backtracking by RNA Pol II Is a Major Effector of Gene Activation, 5' Pause Release, Termination, and Transcription Elongation Rate. Mol Cell *73*, 107-118.e104.

Sigurdsson, S., Dirac-Svejstrup, A.B., and Svejstrup, J.Q. (2010). Evidence that transcript cleavage is essential for RNA polymerase II transcription and cell viability. Mol Cell *38*, 202-210.

Smale, S.T., and Kadonaga, J.T. (2003). The RNA polymerase II core promoter. Annu Rev Biochem *72*, 449-479.

Søgaard, T.M., and Svejstrup, J.Q. (2007). Hyperphosphorylation of the C-terminal repeat domain of RNA polymerase II facilitates dissociation of its complex with mediator. J Biol Chem *282*, 14113-14120.

Soutourina, J. (2018). Transcription regulation by the Mediator complex. Nat Rev Mol Cell Biol *19*, 262-274.

Spangler, L., Wang, X., Conaway, J.W., Conaway, R.C., and Dvir, A. (2001). TFIIH action in transcription initiation and promoter escape requires distinct regions of downstream promoter DNA. Proc Natl Acad Sci U S A *98*, 5544-5549.

Staby, L., O'Shea, C., Willemoës, M., Theisen, F., Kragelund, B.B., and Skriver, K. (2017). Eukaryotic transcription factors: paradigms of protein intrinsic disorder. Biochem J *474*, 2509-2532.

Stevens, J.L., Cantin, G.T., Wang, G., Shevchenko, A., Shevchenko, A., and Berk, A.J. (2002). Transcription control by E1A and MAP kinase pathway via Sur2 mediator subunit. Science *296*, 755-758.

Suh, M.H., Meyer, P.A., Gu, M., Ye, P., Zhang, M., Kaplan, C.D., Lima, C.D., and Fu, J. (2010). A dual interface determines the recognition of RNA polymerase II by RNA capping enzyme. J Biol Chem *285*, 34027-34038.

Szentirmay, M.N., and Sawadogo, M. (1994). Sarkosyl block of transcription reinitiation by RNA polymerase II as visualized by the colliding polymerases reinitiation assay. Nucleic Acids Res *22*, 5341-5346.

Takahashi, H., Parmely, T.J., Sato, S., Tomomori-Sato, C., Banks, C.A., Kong, S.E., Szutorisz, H., Swanson, S.K., Martin-Brown, S., Washburn, M.P.*, et al.* (2011). Human mediator subunit MED26 functions as a docking site for transcription elongation factors. Cell *146*, 92-104.

Takase, Y., Takagi, T., Komarnitsky, P.B., and Buratowski, S. (2000). The essential interaction between yeast mRNA capping enzyme subunits is not required for triphosphatase function in vivo. Molecular and cellular biology *20*, 9307-9316.

Tan, S., Hunziker, Y., Sargent, D.F., and Richmond, T.J. (1996). Crystal structure of a yeast TFIIA/TBP/DNA complex. Nature *381*, 127-151.

Tatomer, D.C., Elrod, N.D., Liang, D., Xiao, M.S., Jiang, J.Z., Jonathan, M., Huang, K.L., Wagner, E.J., Cherry, S., and Wilusz, J.E. (2019). The Integrator complex cleaves nascent mRNAs to attenuate transcription. Genes Dev *33*, 1525-1538.

Thompson, C.M., Koleske, A.J., Chao, D.M., and Young, R.A. (1993). A multisubunit complex associated with the RNA polymerase II CTD and TATA-binding protein in yeast. Cell *73*, 1361-1375.

Tome, J.M., Tippens, N.D., and Lis, J.T. (2018). Single-molecule nascent RNA sequencing identifies regulatory domain architecture at promoters and enhancers. Nat Genet *50*, 1533-1541.

Tomko, E.J., Fishburn, J., Hahn, S., and Galburt, E.A. (2017). TFIIH generates a six-base-pair open complex during RNAP II transcription initiation and start-site scanning. Nat Struct Mol Biol *24*, 1139-1145.

Tran, K., and Gralla, J.D. (2008). Control of the timing of promoter escape and RNA catalysis by the transcription factor IIb fingertip. J Biol Chem *283*, 15665-15671.

Tsai, K.L., Sato, S., Tomomori-Sato, C., Conaway, R.C., Conaway, J.W., and Asturias, F.J. (2013). A conserved Mediator-CDK8 kinase module association regulates Mediator-RNA polymerase II interaction. Nat Struct Mol Biol *20*, 611-619.

Tuttle, L.M., Pacheco, D., Warfield, L., Luo, J., Ranish, J., Hahn, S., and Klevit, R.E. (2018). Gcn4-Mediator Specificity Is Mediated by a Large and Dynamic Fuzzy Protein-Protein Complex. Cell Rep *22*, 3251-3264.

Ujvari, A., Pal, M., and Luse, D.S. (2002). RNA polymerase II transcription complexes may become arrested if the nascent RNA is shortened to less than 50 nucleotides. J Biol Chem *277*, 32527-32537.

van Eeuwen, T., Li, T., Kim, H.J., Gorbea Colón, J.J., Parker, M.I., Dunbrack, R.L., Garcia, B.A., Tsai, K.L., and Murakami, K. (2021a). Structure of TFIIK for phosphorylation of CTD of RNA polymerase II. Sci Adv *7*.

van Eeuwen, T., Shim, Y., Kim, H.J., Zhao, T., Basu, S., Garcia, B.A., Kaplan, C., Min, J.-H., and Murakami, K. (2021b). Cryo-EM structure of TFIIH/Rad4-Rad23-Rad33 in damaged DNA opening in Nucleotide Excision Repair. Nature Communications *in press*.

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.

Verger, A., Monté, D., and Villeret, V. (2019). Twenty years of Mediator complex structural studies. Biochem Soc Trans *47*, 399-410.

Vervoort, S.J., Welsh, S.A., Devlin, J.R., Barbieri, E., Knight, D.A., Offley, S., Bjelosevic, S., Costacurta, M., Todorovski, I., Kearney, C.J.*, et al.* (2021). The PP2A-Integrator-CDK9 axis fine-tunes transcription and can be targeted therapeutically in cancer. Cell *184*, 3143-3162.e3132.

Viswanath, S., Chemmama, I.E., Cimermancic, P., and Sali, A. (2017). Assessing Exhaustiveness of Stochastic Sampling for Integrative Modeling of Macromolecular Structures. Biophys J *113*, 2344-2353.

Vo Ngoc, L., Wang, Y.L., Kassavetis, G.A., and Kadonaga, J.T. (2017). The punctilious RNA polymerase II core promoter. Genes Dev *31*, 1289-1301.

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., Linden, A., Urlaub, H., and Cramer, P. (2020). Structure of complete Pol II-DSIF-PAF-SPT6 transcription complex reveals RTF1 allosteric activation. Nat Struct Mol Biol *27*, 668-677.

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

Wade, J.T., and Struhl, K. (2008). The transition from transcriptional initiation to elongation. Curr Opin Genet Dev *18*, 130-136.

Wagner, T., Merino, F., Stabrin, M., Moriya, T., Antoni, C., Apelbaum, A., Hagel, P., Sitsel, O., Raisch, T., Prumbaum, D.*, et al.* (2019). SPHIRE-crYOLO is a fast and accurate fully automated particle picker for cryo-EM. Commun Biol *2*, 218.

Wang, D., Bushnell, D.A., Huang, X., Westover, K.D., Levitt, M., and Kornberg, R.D. (2009). Structural basis of transcription: backtracked RNA polymerase II at 3.4 angstrom resolution. Science *324*, 1203-1206.

Warfield, L., Ramachandran, S., Baptista, T., Devys, D., Tora, L., and Hahn, S. (2017). Transcription of Nearly All Yeast RNA Polymerase II-Transcribed Genes Is Dependent on Transcription Factor TFIID. Mol Cell *68*, 118-129.e115.

Watanabe, T., Hayashi, K., Tanaka, A., Furumoto, T., Hanaoka, F., and Ohkuma, Y. (2003). The carboxy terminus of the small subunit of TFIIE regulates the transition from transcription initiation to elongation by RNA polymerase II. Mol Cell Biol *23*, 2914-2926.

Webb, B., and Sali, A. (2016). Comparative Protein Structure Modeling Using MODELLER. Curr Protoc Bioinformatics *54*, 5 6 1-5 6 37.

Weiner, A., Hughes, A., Yassour, M., Rando, O.J., and Friedman, N. (2010). High-resolution nucleosome mapping reveals transcription-dependent promoter packaging. Genome Res *20*, 90-100.

Westover, K.D., Bushnell, D.A., and Kornberg, R.D. (2004). Structural basis of transcription: nucleotide selection by rotation in the RNA polymerase II active center. Cell *119*, 481-489.

Wong, K.H., Jin, Y., and Struhl, K. (2014). TFIIH phosphorylation of the Pol II CTD stimulates mediator dissociation from the preinitiation complex and promoter escape. Mol Cell *54*, 601-612.

Yang, C., Bolotin, E., Jiang, T., Sladek, F.M., and Martinez, E. (2007). Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. Gene *389*, 52-65.

Yudkovsky, N., Ranish, J.A., and Hahn, S. (2000). A transcription reinitiation intermediate that is stabilized by activator. Nature *408*, 225-229.

Yue, Z., Maldonado, E., Pillutla, R., Cho, H., Reinberg, D., and Shatkin, A.J. (1997). Mammalian capping enzyme complements mutant Saccharomyces cerevisiae lacking mRNA guanylyltransferase and selectively binds the elongating form of RNA polymerase II. Proc Natl Acad Sci U S A *94*, 12898-12903.

Zeitlinger, J., Stark, A., Kellis, M., Hong, J.W., Nechaev, S., Adelman, K., Levine, M., and Young, R.A. (2007). RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo. Nat Genet *39*, 1512-1516.

Zhang, H., Chen, D.H., Mattoo, R.U.H., Bushnell, D.A., Wang, Y., Yuan, C., Wang, L., Wang, C., Davis, R.E., Nie, Y.*, et al.* (2021). Mediator structure and conformation change. Mol Cell *81*, 1781-1788.e1784.

Zhao, H., Young, N., Kalchschmidt, J., Lieberman, J., El Khattabi, L., Casellas, R., and Asturias, F.J. (2021a). Structure of mammalian Mediator complex reveals Tail module architecture and interaction with a conserved core. Nat Commun *12*, 1355.

Zhao, T., Vvedenskaya, I.O., Lai, W.K.M., Basu, S., Pugh, B.F., Nickels, B.E., and Kaplan, C.D. (2021b). Ssl2/TFIIH function in Transcription Start Site Scanning by RNA Polymerase II in <em>Saccharomyces cerevisiae</em>. bioRxiv, 2021.2005.2005.442816.

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.