DISSECTION OF THE PRECISE MECHANISMS OF RNA POLYMERASE II PAUSING AND ELONGATION USING NASCENT TRANSCRIPT ANALYSIS

A Dissertation
Presented to the Faculty of the Graduate School
of Cornell University
In Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

by
Hojoong Kwak
August 2013

DISSECTION OF THE PRECISE MECHANISMS OF RNA POLYMERASE II PAUSING AND ELONGATION USING NASCENT TRANSCRIPT ANALYSIS

Hojoong Kwak, M. D., Ph. D.
Cornell University 2013

Limiting RNA polymerase II (Pol II) at various stages of the transcription cycle is critical for gene regulation, which often occurs during the elongation stage at promoter proximal pause sites and in gene bodies. To determine the distribution of Pol II along genes, I used nascent transcript analysis as a general method. First, I identified the precise positions of Pol II pausing near promoters using a genome-wide nuclear run-on, called Precision Run-On sequencing (PRO-seq) in *Drosophila* embryonic cells. Using this, I revealed how the position of pausing is associated with initiation and promoter DNA elements. To further dissect the precise dynamics of paused Pol II, I probed the stability of paused Pol II and its termination by analyzing steady-state turn-over of the nascent transcript associated with *Drosophila Hsp70* promoter. This shows that paused Pol II on *Hsp70* is stable for around 5 min and can either terminate or elongate into the gene body, which is consistent with optical measurements of paused Pol II. I also examined how Pol II elongates during the time course of rapid and robust inhibition of pause escape in mouse embryonic stem cells. The analysis of the elongation rates in nearly 1,000 genes showed tight interplay between promoter proximal pausing, early elongation rates, and co-transcriptional splicing at the beginning of the genes. Finally, I demonstrate that the nascent transcriptome analysis methods can be directly extended into mammalian tissues, and show possibility of linking the study of the fundamental mechanism of Pol II into biomedical applications.

BIOGRAPHICAL SKETCH

Hojoong Kwak was born in a middle class family in Seoul, Korea. He was a curious young boy who liked to play sports outside. As he grew up, he showed interest in science and mathematics, and went to the Seoul Science High School, where he met more of his kind who shared their visions as future scientists. He was particularly interested in physics, and even won a gold medal in International Physics Olympiad for the first time from Korea, but still too young to commit himself as a scientist. Instead he went to the medical school.

During the medical school, he started to realize that he still wanted to do something more than a clinical practice. After graduating from the medical school, he spent a year in a hospital for intern training, during the time when he met Yoona Lee, a beautiful wise young woman who did not discourage him from pursuing science. He then went to do a master's degree in the biochemistry department of the medical school in Seoul National University, finish the military service at Korea Food and Drug Administration, and applied for a graduate school in the U. S. and was finally admitted to Cornell University.

At Cornell, he met Dr. John Lis, a passionate scientist who is a pioneer of the transcription field. Hojoong still remembers the first day when he met John enthusiastically explaining the newest GRO-seq data one by one. He eventually joined the Lis Lab, and started his early scientific career learning the cutting edge technologies, and developing solid background in molecular biology. He believes his graduate study at the Lis lab has been very successful, and looks forward to his next steps.

ACKNOWLEDGMENTS

First of all, I would like to thank John, my advisor for his mentorship during my graduate study. He is a great teacher, and a role model as a scientist to me. He is also a wonderful person, always enthusiastic and optimistic about everything, caring for everyone. Having him as my teacher was the luckiest thing and really changed my life. I also thank my committee members, Dr. Roberts and Dr. Pleiss, for their advices and supports.

Members of the Lis lab have been the greatest supporters throughout my stay in Cornell in every aspect. I thank Martin and Iris for sharing their data and allowing me to take a role in their fantastic projects. Nick was always there when someone needed his help. Most of all, I thank Leighton Core, who first mentored me during the rotation and guided me all throughout my graduate study.

Finally, I thank my family, my parents and especially my wife Yoona, and our princess Enhu Claire for their patience, trust, and everlasting love.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1[a]

INTRODUCTION : CONTROL OF TRANSCRIPTIONAL ELONGATION

Summary

Elongation is becoming increasingly recognized as a critically controlled step in transcriptional regulation. While traditional genetic and biochemical studies have identified major players of transcriptional elongation, our understanding of the importance and roles of these factors is evolving rapidly through the recent advances in genome-wide and single-molecule technologies. Here we focus on how elongation can modulate the transcriptional outcome through the rate-liming step of RNA polymerase II pausing near promoters, and how the participating factors were identified. Among the factors we describe are NELF and DSIF, the pausing factors, and P-TEFb, the key player in pause release. We also describe non-exclusive models for how pausing is achieved by making use of high resolution genome-wide mapping of paused Pol II relative to promoter elements and the first nucleosome. We also discuss Pol II elongation through the bodies of genes and the roles of FACT and Spt6, the factors that allow Pol II to move through nucleosomes.

Introduction

Advances in genome-wide, high-throughput technology have taken our analysis of transcriptional elongation into a new era. The distribution of RNA polymerases and their regulatory or auxiliary factors are being examined on all genes under steady-state cellular conditions or in response to regulatory signals and systematic perturbations. Computational analyses of these massive data sets along

[a] This chapter has been submitted to and is in press at *Annual Review of Genetics*, Volume 47 © Annual Reviews, http://www.annualreviews.org.

with more traditional genetic and biochemical analyses are transforming our understanding of transcription and its control. With all these advances, transcriptional elongation is becoming more widely recognized as a significant step in gene regulation. In this review, we describe and connect the ideas from traditional studies with results from newer methodologies, and we provide high-resolution models of how transcription is controlled at elongation.

We begin by illustrating one classic example of how traditional genetics have contributed to the study of transcriptional elongation. In early 80's, Winston and colleagues selected genes that are critical for the transcription of a retrotransposon element called *Ty* in yeast[1,2]. These Suppressor of *Ty* (*Spt*) genes turned out to be essential elements in Pol II transcription in general. In particular, a group of *Spt* genes characterized by genetic complementation patterns later turned out as genuine elongation factors acting directly on elongating Pol II using biochemical analyses[2]. Through this convergence between genetics and biochemistry, some of the key players of elongation have been identified, and are currently being extensively studied using the newest methods such as genome-wide or single-molecule technologies.

To further discuss the control mechanisms of transcriptional elongation in depth, we divided the stages of elongation into two parts, 'early elongation' and 'productive elongation'. After initiation, Pol II starts the iterative incorporation of nucleotides at its 3' end to extend the nascent RNA. This elongation by Pol II is not uniform throughout the gene. During the early elongation stage, Pol II first transcribes the initial 20~60 nucleotides (nt) of RNA and pauses[3]. Progressing beyond this point is rate-limiting for approximately half of all active Drosophila and mammalian genes, and the escape beyond the pause into productive elongation is often regulated.

Once Pol II makes the transition from the paused state into productive elongation, it then progresses through the body of the gene. In the gene body, Pol II still faces other barriers and requires additional elongation factors to overcome them[4]. Finally, Pol II transcribes through the 3' end of the gene, where nascent RNA cleavage and poly-adenylation defines the end of the mRNA transcript. The termination of a transcribing Pol II takes place on average 8 kb downstream of the 3' end of the gene in human cells[5]. This 3' end processing of the nascent RNA and Pol II termination can add extra dimensions to the co-transcriptional and post-transcriptional[6].

In this review, we will focus on the distinct stages of early elongation and the productive elongation and discuss the factors and their control mechanisms (Table 1.1). Other recent reviews deal with complementary aspects of transcription elongation control including the functions of promoter-proximal pausing, conflicts between transcription and DNA replication, and co-transcriptional RNA processing.

Identification of the Regulatory Step During Early Elongation

Early elongation can be defined as the transition of Pol II between promoter escape and fully productive elongation, and is often accompanied by an intermediate step of promoter proximal pausing[4]. This transition of initiated Pol II to paused occurs as nascent RNA with a length from ~10 nucleotides (nt) is extended to 20-60 nt, as measured most precisely in *Drosophila*[3,7,8], and the bulk of this paused Pol II in mammalian genes occurs in a similar position[9]. When the RNA chain is less than 10 nt long, Pol II is still considered to be within the initiation stage; it is associated with TFIIB and can terminate prematurely (abortive initiation), which may provide checkpoints for promoter control[10]. Once the RNA grows longer than 12 nt and TFIIB is displaced, the Pol II elongation complex is stably engaged[11] and is now in the early elongation stage. Early elongation is not a simple smooth transition of Pol II but involves critical regulatory steps as observed in a plethora of evidence from the earliest to most recent studies.

Table 1.1. A growing list of the factors participating in transcriptional elongation

Factor		Function	Related factors and notes
GAGA-factor		Generate nucleosome free region[12] and promoter structure for pausing[13]	NURF[14]
GTFs	TFIID	Generate promoter structure for pausing[13]	
	TFIIF	Increase elongation rate[15,16]	Near promoters[17]
	TFIIS	Rescue backtracked Pol II[15,18]	Pol III[19]
Pausing factors	NELF	Stabilize Pol II pausing[20,21]	
	DSIF	Stabilize Pol II pausing[22]/ facilitate elongation[23,24]	
P-TEFb		Phosphorylates NELF[25] , DSIF[26], Pol II CTD[27] for pause release	
Processivity factors	Elongin	Increase elongation rate[28]	
	ELL	Increase elongation rate[29]	AFF4[30]
	SEC	Contains P-TEFb and ELL[30]	Mediator[31], PAF[32]
Activator	c-Myc	Can directly recruit P-TEFb[33]	
	NF-κB	Can directly recruit P-TEFb[34]	
Coactivator	Brd4	Recruits P-TEFb[35,36]	
	Mediator	Recruits P-TEFb via SEC[31]	
Capping machinery	CE	Facilitates P-TEFb recruitment[37], counters NELF/DSIF[38]	
	RNMT [a]	Methylates RNA 5’ end to complete capping	Myc[39]
Premature termination factors	Dcp2	De-caps nascent RNA for Xrn2 digestion[40]	Dcp1a / Edc3 [40]
	Microprocessor	Cleaves hairpin structure for Xrn2 digestion[41]	Tat, Senx
	Xrn2	‘Torpedo’s Pol II with RNA 5’-3’ exonucleation[40,41]	
	TTF2	Release Pol II from DNA[40,41]	
Gdown1		Anti-termination and stabilize paused Pol II[42-44]	TFIIF[44], Mediator[42]
Histone chaperone	FACT	H2A/H2B eviction and chaperon[45,46]	Tracks with Pol II[47]
	Nap1 [a]	H2A/H2B chaperon[48]	RSC[49], CHD[50]
	Spt6	H3/H4 chaperon[51]	Tracks with Pol II[47]
	Asf1 [a]	H3/H4 chaperon[52]	H3K56ac[53]
Chromatin remodeler	RSC [a]	SWI/SNF remodeling in gene body[54]	H3K14ac[55]
	Chd1 [a]	Maintain gene body nucleosome organization[56]	FACT, DSIF[50]
	NURF	ISWI remodeling at promoter[14]	GAGA factor[12]
PARP		Transcription independent nucleosome loss[57]	Tip60[58]
PAF complex		Loading dock for elongation factors[59]	SEC[32], FACT[60]
Histone tail modifiers	MOF	H4K16ac, recruit Brd4[61]	H3S10ph, 14-3-3[61]
	Tip60	H2AK5ac, activate PARP[58]	
	Elongator [a]	Acetylates H3, facilitate nucleosomal elongation[62]	Also in cytoplasm[62]
	Rpd3C (Eaf3) [a]	Deacetylates and inhibit spurious initiation in gene body[63]	H3K36me3[63]
	Set1 [a]	H3K4me3[64,65]	MLL/ COMPASS
	Set2 [a]	H3K36me3[66], regulate acetylation-deacetylation cycle[63]	Rpd3C[63]
	PIM1	H3S10ph, recruit 14-3-3 and MOF[67]	
	RNF20/40	H2BK123ub1, facilitate nucleosomal DNA unwrapping[60]	UbcH6, PAF

[a] Not covered in the main text

Evidence suggesting a mode of transcription regulation in eukaryotes occurs early after initiation

Peaks of paused Pol II on proximal promoters. Studies of transcription regulation have for decades focused primarily on how regulatory signals and key transcription factors act at the level of the recruitment of Pol II to promoters or the initiation step of transcription[68]. This model gained strong support for all genes that were closely examined in *S. cerevisiae*. Activation of transcription generally produced a recruitment of general transcription factors (GTFs) and Pol II to promoters, and the recruited Pol II then began transcribing and produced full gene-length transcripts and a relatively even distribution of Pol II across each gene[69].

One set of observations that seemed at least partially at odds with all transcription being controlled at initiation was that genes in chicken[70], Drosophila heat shock genes[3,71] and human c-Myc and c-Fos[72-74], appeared to have much higher levels (peaks) of Pol II near their promoters - at levels that dwarf that on gene bodies. Moreover, this promoter-proximal Pol II was transcriptionally engaged, that is, it had already initiated transcription and had progressed beyond the point of known abortive initiation. These Pol II molecules were also shown to associate with melted DNA found in elongation complexes[75] and could elongate efficiently in nuclear run-on reactions if inhibitory factors were removed by high salt or sarkosyl treatments (both of which block new initiation)[76]. These results indicated that a ‘paused’ Pol II elongation complex can accumulate to high levels at least on some genes without efficiently entering into productive elongation.

Whereas promoter-proximal pausing of Pol II was first characterized in a handful of genes, it turns out to be a general mechanism for other genes across species. Pol II ChIP experiments using genome-wide methods[77-79] and GRO-seq[5,9,80] have shown that Pol II pausing is widespread near promoters. Furthermore, similar to *Hsp70*, other genes in *Drosophila* are sensitive to nuclear run-on with sarkosyl treatment, indicating that the promoter proximal Pol II in other genes are also physically tethered at the paused state by inhibitory factors[81].

Genes that have paused Pol II are also highly regulated and the escape of Pol II from the pause into productive elongation can be activated in the cases of heat shock genes by stress and in c-Myc by serum stimulation[4]. Thus, regulation of these genes was not solely at Pol II recruitment or initiation, but rather, a main component appeared to be at the escape of the paused Pol II into productive elongation.

In vitro transcription systems needed pausing factors and a kinase to recapitulate full regulation seen in metazoans in vivo. The early in vitro transcription systems derived from human whole nuclear extracts were shown to initiate properly and elongate on model DNA templates. All that seemed to be required to produce transcripts was the recruitment of GTFs and Pol II to promoters. Therefore, reconstituted systems composed of GTFs and Pol II were used, and many regulatory transcription factors were shown to facilitate the production of transcripts indicating a mode of regulation at the level of recruitment of GTFs and Pol II[82].

Interestingly, these early reconstituted transcription systems were insensitive to an inhibitor of transcription in mammalian cells, DRB[22]. DRB is a kinase inhibitor, and was postulated to act at the elongation stage to reduce the production of full length transcripts[83]. The search for factors that make a transcription system responsive to DRB inhibition led to the discovery and characterization of protein complexes that act to stabilize paused Pol II and a kinase that overcame this rate limiting pausing by phosphorylating the pausing complexes[20,22,84]

Wada et al. purified a complex from HeLa nuclei that conferred DRB sensitivity to the human in vitro transcription system[22]. This complex, DSIF, consists of two subunits, and remarkably, they were the homologs of yeast *Spt4* and *Spt5* genes identified from the early genetic studies by Winston and colleagues[22].

Studies by Price and colleagues helped clarify the mechanism by which DSIF inhibits elongation. Results from *Drosophila* nuclear extract suggested that Pol II is tightly associated with negative transcription elongation factor (N-TEF), and a positive transcription elongation factor (P-TEF) relieves the effect of N- TEF[85]. A series of biochemical complementation analyses showed that DRB inhibits P-TEF, and that N-TEF is epistatic to P-TEF inhibition[84,85]. Therefore, it was concluded that DSIF plays the role of N-TEF. However, in the presence of recombinant DSIF alone, transcription was not fully sensitive to DRB inhibition. Handa and colleagues purified an additional N-TEF protein factor that conferred complete DRB sensitivity and was named NELF[20]. NELF is a multiprotein complex composed of four subunits - A, B, C or D, and E[21] - and has been shown to physically interact with DSIF and Pol II[20].

In summary, the discovery of Pol II pausing near promoters and identifying several pausing factors that are essential for full regulation of transcription has revealed another critical layer of transcription control that occurs during early elongation.

Sequence of events leading to promoter proximal pausing

Follow-up on these pioneering studies revealed many factors and promoter features that are part of early elongation control. Additionally, the mechanisms of early elongation were found to be connected both temporally and spatially to that of initiation and the promoter structure.

The promoter is a dynamic structure that can direct transcription initiation and pausing. In *Drosophila*, promoter DNA sequences by themselves often have a tendency to position a nucleosome at the TSS[86], and these promoter regions are typically packaged by nucleosomes under repressed conditions[87] (Figure 1.1A). On the other hand, many yeast or human promoters have DNA features such as the A-T richness or CpG islands that can intrinsically favor an accessible chromatin structure[88]. In either case, for a gene to be poised for activation, the accessibility of the promoter to transcription factors is critical[87]. Transcription activators or other DNA binding proteins such as GAGA factor in *Drosophila* [12] can recruit chromatin remodelers to produce an open promoter structure[87] (Figure 1.1B). Importantly, prolonged transcription factor binding (35) and Pol II pausing itself[86] contribute to the maintenance of this open promoter structure .

General transcription factors (GTFs) are assembled on core promoters[87], and direct the recruitment and transcriptional initiation of Pol II (Figure 1.1C). The TFIID complex binds provides a strong foundation for binding to the promoter region. The interaction between the TATA box and TATA box binding protein (TBP), a major subunit of TFIID, as well as interactions of TFIID-associated proteins with elements downstream of the TSS contribute to TFIID binding in many promoters. TFIIA and TFIIB also associate with TBP to facilitate TFIID binding and Pol II recruitment[89]. Pol II in complex with TFIIF binds to the promoter followed by TFIIE and TFIIH. TFIIH contains the XPB DNA helicase subunit that melts the DNA and generates the open Pol II complex to expose the template strand of DNA and allow initiation of transcription[90].

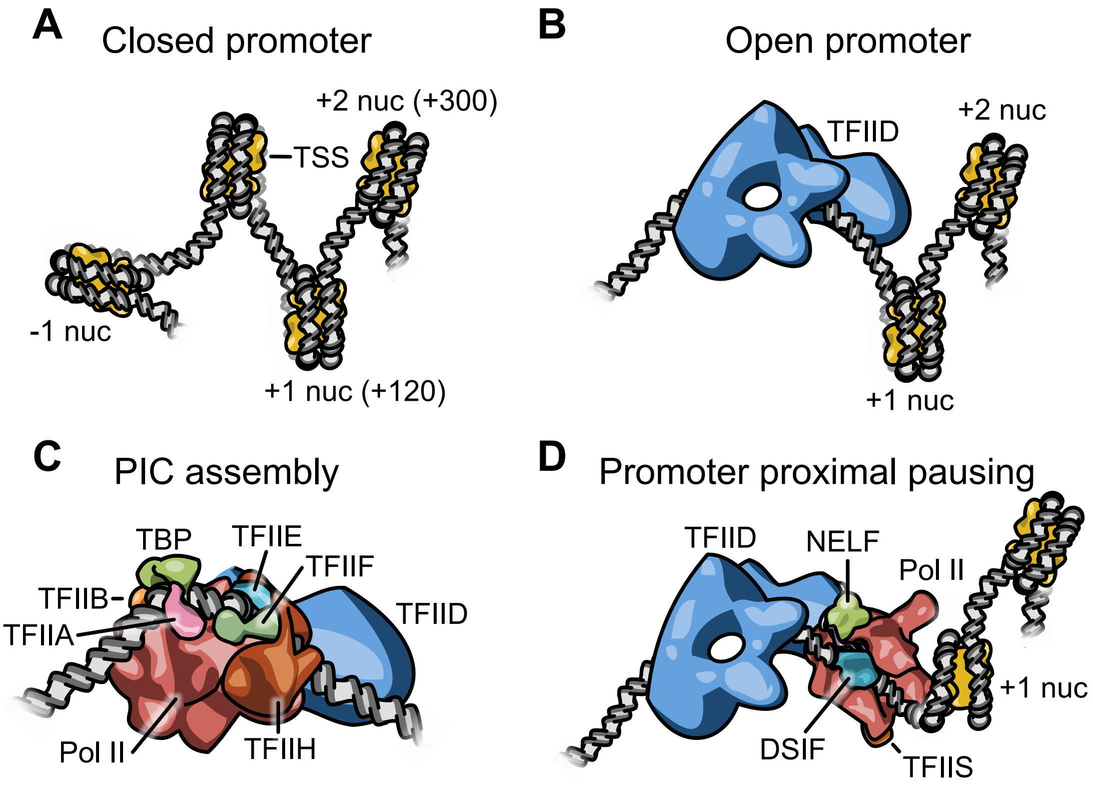


Figure 1.1. Structure of the promoter and Pol II before and during early elongation (**A**) The promoter is occupied by a nucleosome in a closed configuration. This conformation represents a *Drosophila* promoter unbound by Pol II before priming[86]. The nucleosome (nuc) positions are the distance between the dyad axes and the TSS in base-pairs (bp), based on the average micrococcal nuclease sequencing profile[86]. (**B**) The promoter is in an open configuration, unbound by nucleosome and occupied by general transcription factors[87]. (**C**) The pre-initiation complex (PIC) is assembled (zoomed in relative to other panels). Part of the TFIID structure is removed to visualize the assembly of other general transcription factors. TBP is a subunit of the TFIID structure but is not removed to illustrate its binding to the promoter DNA and Pol II. (**D**) Pol II is paused between the promoter and the first nucleosome (+1 nuc). The pausing position is at +40 from TSS, which is typical in *Drosophila* promoters. Pol II bound NELF, DSIF, and TFIIS are also shown. TFIID may or may not be present at the promoter depending on its residency in a re-initiation scaffold[91]. The illustrated molecules are based on crystal[89,92-94] and cryo-electron microscope[95,96] structures except for NELF (structure not available), and are scaled proportional to their actual dimensions. The length of each DNA turn is about 10 bp or 3.4 nm.

Another important function of TFIIH is to phosphorylate the C-terminal domain (CTD) of the Pol II Rpb1 subunit[90,97]. Pol II CTD is composed of a species-specific number of 7 amino acid repeats, Tyr-Ser2-Pro-Thr-Ser5-Pro-Ser7, ranging from 26 in yeast up to over 50 in mammals. A cyclin-dependent kinase Cdk7 subunit of TFIIH phosphorylates the fifth residue (Ser5)[27]. Ser2 and Ser7 can also be selectively phosphorylated later on, which allows the interaction of accessary elongation factors with Pol II.

After the very early elongation of nascent RNA (up to ~10 nt), Pol II escapes this initiation stage. Some GTFs are dissociated, but part of the initiation complex such as TFIID, TFIIA, and TFIIH may remain associated with the promoter providing a scaffold for re-initiation[91]. Moreover, TFIIF can remain associated with the Pol II as an elongation factor[15,16]. During the course of early elongation, Pol II is prone to arrests in a backtracked position, where TFIIS binds to Pol II facilitating the cleavage of protruded nascent RNA and rescue from the arrested state[15,94].

In higher eukaryotes, after transcribing the initial 20-60nt of RNA, Pol II becomes paused between the promoter and the first (+1) nucleosome (Figure 1.1D). Pausing factors NELF and DSIF can bind to paused Pol II and stabilize it in vivo[98]. However, the mechanistic details for these complexes are still being determined. A possible mechanism by which these pausing factors mediate their function could be related to their RNA binding property. NELF contains the subunit NELF-E which contains an RNA recognition motif[21,25], and DSIF has been shown to bind shorter nascent RNA as it emerges from the elongating Pol II complex[99]. Consistent with this, the crystal structure of DSIF bound to archaeal RNA polymerase (RNAP) shows that DSIF binds near the active center cleft and RNA exit channel[93] where it may interact with the transcription bubble and modulate the processivity of the RNA polymerase.

Mechanisms of Promoter Proximal Pausing

To better understand proximal pausing and its role as a rate-limiting step of transcriptional regulation, we need to review in detail the mechanics of early Pol II elongation. In addition to what is known about the factors involved, we need to consider how DNA sequence elements and nucleosomes influence Pol II pausing and its escape. In this section, we discuss the mechanisms of how elongating Pol II can become paused near promoters and what the fate of the paused Pol II is. Also, we describe genomic approaches to identify the properties of promoter that determine the extent and the pattern of promoter proximal pausing.

Molecular models of promoter proximal pausing

After initiation, the early elongating Pol II meets its first roadblock and pauses. The identity of this first roadblock is very important, because it provides the strongest rate-limiting step in the course of Pol II elongation along the gene. Various in vitro studies have identified many different components that comprise this stop sign. The mechanisms for establishing the pause in vivo can be summarized into three models that are not mutually exclusive (Figure 1.2A). The actual promoter environment may be a blend of these individual models, and different promoters may use different blends.

The first can be described as the kinetic model, where the balance between early elongation and the recruitment of pausing factors determines Pol II pausing. Intrinsic features of Pol II such as the early elongation rate, and DNA and RNA sequence specific features such as the DNA-RNA hybrid energy will affect the

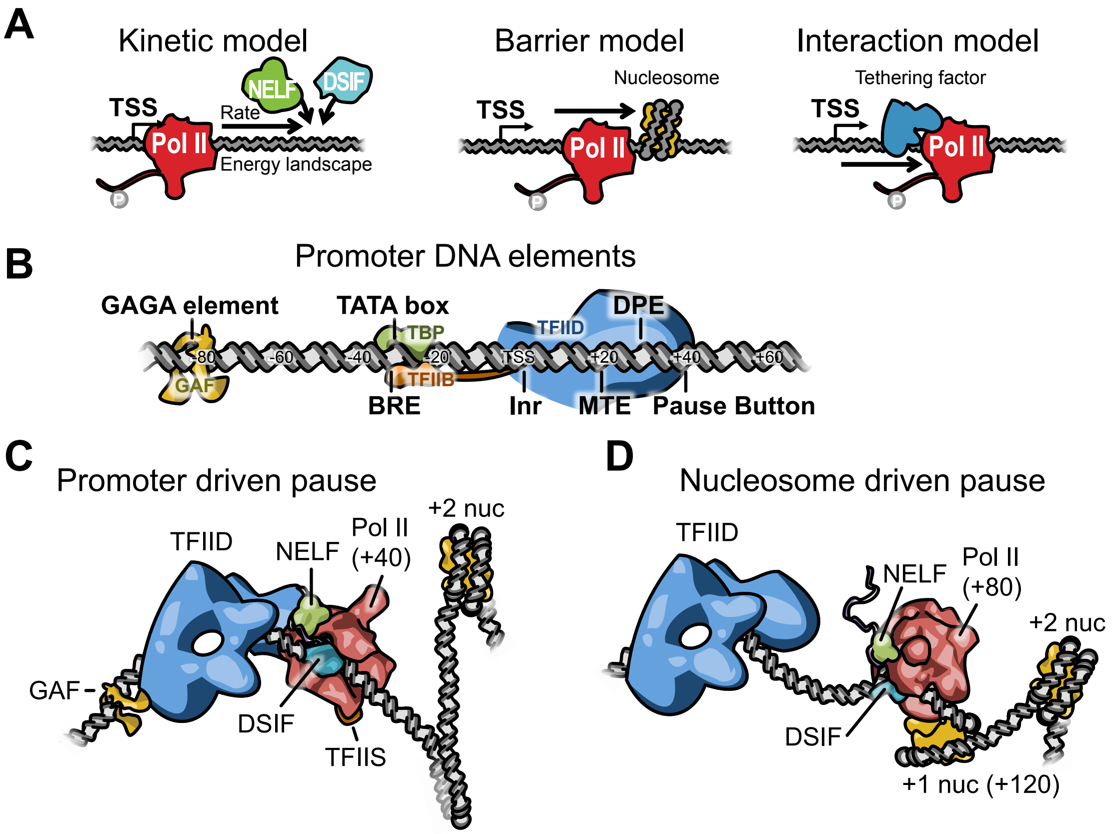


Figure 1.2. Models of promoter proximal Pol II pausing (**A**) The driving force for establishing the position and the extent of Pol II pausing near promoter is summarized in 3 models. (*left*) In the Kinetic model, the recruitment of the pausing factors competes with the elongation of Pol II. (*middle*) In the Barrier model, physical barriers such as nucleosomes impede with elongating Pol II. (*right*) In the Interaction model, a tethering factor that binds to DNA and Pol II at the same time drives Pol II pausing. (**B**) Promoter DNA elements in *Drosophila* are shown[100,101]. Protein factors such as GAGA-factor(GAF), TBP, and TFIIB can bind to their DNA elements. TFIID complex can make specific contacts with the DNA elements at multiple positions that are relatively downstream. (**C**) In promoters with stronger DNA elements, Pol II is more proximal to the TSS within the contact range of the promoter complex[8], and the interaction between the promoter and Pol II can drive the pause. Pausing is stronger in these promoters and the first nucleosome may be absent[86]. (**D**) In promoters with weaker DNA elements, Pol II is closer to the dyad axis of the first nucleosome[8], and is more compatible with pausing driven by nucleosome barriers. The interaction with the promoter complex and the downstream DNA is weaker and may generate less resistance to the Pol II. The illustrations in (**C**) and (**D**) are based on actual structures [92,95,96,102]

kinetics and the energy landscape of elongating Pol II complexes[103]. In regions with a higher energy state, Pol II may spontaneously backtrack, and need to be rescued from this backtrack by TFIIS, which is associated with paused Pol II[104]. These kinetic processes along with rate of pausing factor binding and the rate and energetics of Pol II will determine the level and the position of promoter proximal pausing[7].

Second is the barrier model where a physical barrier on the chromatin blocks Pol II elongation. Nucleosomes have long been known to provide barriers to Pol II elongation in vitro[105]. In *Drosophila*, it was shown that the first nucleosomes make contacts with paused Pol II and can be cross-linked together in vivo[106]. However, in some highly paused genes, the first nucleosome barriers may not be present, for example, the nucleosome free region around the *Drosophila Hsp70* promoter extends to +250 bp from TSS[86,87].

Third is the interaction model where sequence specific binding factors interact with elongating Pol II and determine pausing. This has been an underappreciated model in higher eukaryotes, but for bacteria, it successfully explains the sequence specific pausing by DNA or RNA binding factors that interact with RNAP. For instance, RNA secondary structure at the leader pause site of the *E. coli his* gene was shown to interact with Pol II[107], and at other genes, the initiation factor σ_{70} induces pausing by interacting with the downstream DNA elements during its retention within Pol II after the initiation event[108] . In higher eukaryotes, there is evidence that pausing may be determined by RNA binding of NELF or DSIF[99,109]. Additionally, a testable hypothesis is that like σ_{70}, the components of the initiation complex may remain bound to the promoter DNA and tether Pol II to induce pausing. Therefore, the general features of *E. coli* pausing may provide a framework of a molecular model for the promoter proximal pausing of Pol II.

High-resolution mapping allows integration of pausing models

Advances in the genome-wide approaches have made it possible to define the status of Pol II at the promoter for thousands of genes. In mammalian cells, GRO-seq identified about 40-46% of active genes as paused[5,9]. Likewise in a *Drosophila* embryonic cell line, 70%-89% of active promoters contained paused Pol II[81]. However, although a large fraction of the active genes are paused, the degree of pausing and distribution of pause sites within a promoter varies between genes depending on the features of the promoter[8], and here we consider two types of *cis*-acting elements: promoter DNA elements and the nucleosome.

The promoter region around a TSS contains several conserved sequence motifs recognized by the general transcription factors (GTFs) and Pol II (Figure 1.2B). In Drosophila, these include the TATA box, initiator (Inr), TFIIB recognition element (BRE), downstream promoter element (DPE)[100], and motif-ten element (MTE)[101] (Figure 1.3B). The extent to which promoter DNA elements contribute to GTF binding does not simply rely on true or false type of matching to the consensus motif but on the strength and the quality of the elements[110]. The search for the DNA motifs in paused genes in *Drosophila* embryo have shown the enrichment of some of the core promoter elements such as Inr and DPE, and more upstream GAGA element[111,112]. These elements were also found to be required for pausing in *Hsp70* in assays of transgenic mutant promoters[13]. A novel GC rich element bearing sequence similarity to DPE was also identified as enriched on paused genes and called the 'Pause Button' [112]. In addition, the strength of these and other core promoter DNA elements at the consensus position strongly correlated with the extent of pausing[8]. Therefore, the strength of a promoter complex as a whole may be a determinant of pausing as well as initiation[86].

The precise position of pausing also relates to the strength of the promoter DNA elements. In *Drosophila*, promoters with stronger DNA elements tend to have more proximal pausing (+40), while promoters with weaker elements have more distal dispersed pausing (up to +80)[8]. Since the average position of the center of the first nucleosome is around +120 from TSS[86], the distal pause position is more consistent with the nucleosome barrier mechanism[8,105,113].

These findings lead us to propose an amalgam of molecular models of pausing for different promoters that involves distinct inputs and mechanisms for a proximal, promoter driven pausing (Figure 1.2C) and for a distal, nucleosome driven pausing (Figure 1.2D). For the 'kinetic' model, the energy landscape and rates are defined not only by the DNA-RNA sequences but also by features of the 'interaction' and the 'barrier' models. Simply put, the 'kinetic' model is an integral part of the 'interaction' and 'barrier' models. For a promoter with stronger DNA elements, the 'interaction' model dominates the energy landscape and the output is a more proximal, promoter driven pausing (Figure 1.2C). On the other hand, for a promoter with weaker DNA elements, the nucleosome barrier defines the energy landscape more dominantly and the result is a more distal, nucleosome driven pausing (Figure 1.2D). The interaction and barrier models can even co-exist within a single promoter, and often one may mask the other depending on the balance between the two principal components.

Fate of paused Pol II

Promoter proximal paused Pol II can resume the NTP catalysis and enter productive elongation. It has been shown that the vast majority of paused genes have detectable levels of elongating Pol II in the gene bodies[9,81]. Measuring the ratio between paused and elongating Pol II will make it possible to estimate how frequently Pol II escapes into productive elongation and assess the activity of the promoter.

However, confounding this calculation is the loss of paused Pol II though termination at or near the pause site.

Premature termination from the pause or arrest has been extensively documented in the bacterial Rho dependent termination mechanism, but this mechanism usually comes into play downstream of any promoter-associated pause. In eukaryotic cells, stalled Pol II at DNA damage sites can be remove by the ubiquitination mediated degradation pathway during transcription coupled nucleotide excision repair[114,115]. This is not necessarily dependent on the DNA damage itself but is a response to prolonged Pol II stalling[116].

In mammals, mRNA de-capping and 5’→3’ exonuclease Xrn2 mediated termination, similar to the torpedo-like mechanism at the 3’ end of the gene, can take place in promoter proximal region. RNAi depletion of de-capping proteins or the termination factors result in the redistribution of Pol II from the proximal region into the gene bodies[40]. Although it is technically challenging to assess the balance between termination and pause escape, it will be critical to quantify the extent of each in order to understand their roles in the regulatory mechanism.

Escape Into Productive Elongation

Productive elongation is defined here as the efficient progression of Pol II through the gene body, and the ‘escape into productive elongation’ refers to the transition of the paused Pol II into this productive mode. During Pol II’s residence in the promoter proximal region, two important modifications to the major components of transcription elongation complex take place. One is the 5’ capping of nascent RNA, and the other is the phosphorylation of the paused Pol II complex. The latter is mediated by protein kinase complex P-TEFb that also phosphorylates the pausing factors DSIF and NELF. At this stage, Pol II and its associated nascent RNA are modified so as to allow recruitment of additional elongation and RNA processing factors that help polymerase overcome further barriers in the gene body and couple transcription and RNA processing[4].

5’ capping as a checkpoint for productive elongation

Transcription is coordinated with other essential events that lead to proper processing of RNA[6]. This requires the loading Pol II with an entourage of proteins that facilitate the RNA processing and help Pol II navigate properly through chromatin. RNA capping has been regarded as a mechanism dependent on early elongation, but accumulating data suggests that it may also have an active role as a control point.

Capping the 5’ end of the nascent RNA requires three enzymatic activities: digestion of the 5’ triphosphate of RNA, addition of a guanine base in 5’ to 5’ linkage, and methylation of the added guanine base[6]. In yeast, these activities depend on three different genes, whereas in metazoans, a single Capping Enzyme (CE) performs the first two activities. CE can be recruited to Pol II when the nascent RNA is about 30 nt long through the interaction with phosphorylated Ser5 and DSIF[38]. Recruitment of CE can relieve the action of NELF and provide a platform for P-TEFb loading[37,38].

Examination of the capped state of Pol II-associated RNAs at high resolution in the pause region of heat shock genes under non-heat shock conditions showed that capping occurs progressively as Pol II moves through the pause region – the most proximal paused RNAs are largely uncapped and the more distal (beyond +30) are completely capped[3]. Thus, capping could be required for escape into productive elongation, but given the low rate of escape in non-HS conditions, it appears that capping may not be sufficient on its own to drive this escape.

Major role of P-TEFb in pause escape

Identification of P-TEFb. As described previously, biochemical analysis of the elongation activity in *Drosophila* Kc nuclear extract showed that the majority of Pol II complexes were only able to produce short, paused transcripts. However, a fraction of the Pol II complexes were able to produce full-length transcripts from the template in a DRB sensitive manner[85]. This fraction was postulated to contain positive transcription elongation factor (P-TEF), and later a homogenous fraction was purified that contained P-TEFb[84].

P-TEFb is a heterodimer complex of Cyclin T (CycT) and Cyclin-dependent kinase 9 (Cdk9) subunits[117]. Its major role is to phosphorylate Ser2 of Pol II CTD[27], and the pausing factors DSIF[24,26] and NELF[25]. However, there is partial redundancy and non-specificity in Pol II CTD phosphorylation by P-TEFb. P-TEFb can also phosphorylate Ser5 in vitro[27], and Cdk12[118] or Brd4[119] can also perform Ser2 phosphorylation. Nevertheless, inhibition of P-TEFb kinase activity with another potent inhibitor flavopiridol resulted in a 5' to 3' clearing of Pol II within the gene body in *Drosophila Hsp70*[120] and globally in mouse embryonic stem cells[121]. Upon this inhibition, Pol II molecules that are already elongating at the time of inhibition continue elongation through the gene body and still maintain a high level of Ser2 phosphorylation level near the 3' end of *Hsp70*[120]. Therefore, P-TEFb plays a role in pause escape by phosphorylating the pausing factors and releasing NELF, but it is less clear whether Ser2 phosphorylation itself is required for pause escape.

The recruitment of active P-TEFb is a critical step in controlling gene expression, and is regulated at multiple stages. In human cells, the free form of P-TEFb is sequestered by 7SK snRNP in an inactive state[122,123] (Figure 1.3A). Upon activation, P-TEFb is released from the sequestration and recruited to the promoter in a variety of ways (Figure 1.3B). After completing its function, P-TEFb can be sequestered again by 7SK snRNP, but in some highly active genes, it remains bound with the elongating Pol II in Super Elongation Complexes (SECs)[124] (Figure 1.3C).

Recruitment of P-TEFb by activator, Brd4, and Mediator. There are several ways for P-TEFb to be recruited to paused Pol II. The first is directly by activators. It has been shown that c-Myc can physically interact with the CycT subunit of P-TEFb during transcriptional activation[33,125]. Treating mouse ES cells with the chemical inhibitor of c-Myc/Max complex resulted in a decrease of pause release preferentially in known c-Myc targets[121]. Similarly, the RelA subunit of NF-κB also binds to CycT and recruits P-TEFb to TNF-α target genes[34]. Finally, the HIV Tat transactivator can bind CycT and directly recruit P-TEFb to the 5' LTR region through Tat binding to TAR hairpin RNA, adjacent to where Pol II is paused[117].

A second way that P-TEFb is recruited is through the co-activator Brd4[35,36]. Brd4 is a bromodomain protein that can recognize acetylated histone tails, and can also bind to the CycT subunit of P-TEFb once it is released from 7SK snRNP[35]. Recruitment of Brd4 may also require a 'histone crosstalk'[61] triggered by PIM1

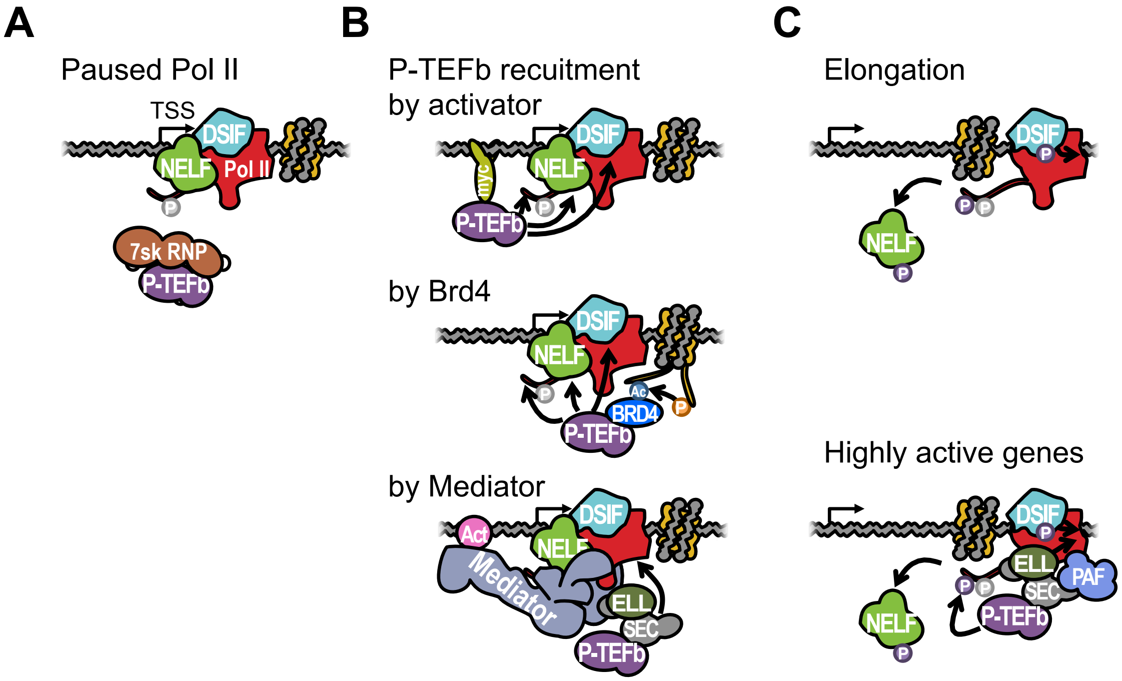


Figure 1.3. Productive transcription elongation complex (**A**) Paused Pol II is bound by NELF and DSIF. The CTD is phosphorylated at Ser5. P-TEFb is held inactive by 7SK RNP. (**B**) P-TEFb is activated and recruited to the paused Pol II by various mechanisms. The first is directly by activator that binds to the DNA. Recruited P-TEFb can phosphorylate NELF, DSIF, and Ser2 of Pol II CTD. The second is through Brd4 that binds to acetylated histone tails. In human cells, Brd4 can bind to H4K16 acetylation which is dependent on H3S10 phosphorylation through the 'histone crosstalk'[61]. The third is indirectly through the Mediator complex, which links the activator (Act) and Pol II. A Mediator subunit MED26 can recruit SEC, which also contains P-TEFb[31]. (**C**) Pol II escapes pausing. Phosphorylated NELF is dissociated from Pol II, and DSIF turns into a positive elongation factor after being phosphorylated by P-TEFb. Alternatively, P-TEFb can remain bound to Pol II by SEC that also interacts with PAF in genes with highly active elongation[124]. P-TEFb can continuously phosphorylate Ser2, and a SEC component ELL can facilitate elongation. PAF can also recruit additional elongation factors. RNA is not shown for the clarity of viewing the complexes.

kinase to phosphorylate H3Ser10 (H3S10ph)[67]. 14-3-3 adaptor protein recognizes H3S10ph and recruits histone acetyl-transferase complex MOF to acetylate H4Lys16 which serves as a binding platform for Brd4[61]. Brd4 was also identified as a therapeutic target of acute myeloid leukemia that can control the expression of Myc at the elongation level[126]. Thus Brd4-mediated recruitment of P-TEFb may be a key mechanism of gene regulation.

P-TEFb can also be recruited through the Mediator complex. Human Mediator subunit MED26 can serve as a binding platform for SEC, which contains P-TEFb. During initiation and pausing, MED26 is associated with TFIID, but MED26 can be switched to recruit SEC instead[31]. Recruitment of P-TEFb through Mediator is believed to be more sensitive in a subpopulation of Pol II that contains a tightly associated factor Gdown1[42]. Gdown1 can compete with TFIIF and therefore might reduce the efficiency of initiation[44]. In contrast, after Pol II initiates, Gdown1 stabilizes Pol II in a paused configuration[43]. In addition to P-TEFb recruitment, an in vitro study also showed that Mediator helps Pol II overcome the +1 nucleosome barrier on a mononucleosomal template[127], further strengthening its positive function during the early elongation steps of the transcription cycle.

Factors required for more efficient elongation

There is a set of factors associated with elongating Pol II that is critical for its efficient transcription. Historically, some of these factors were discovered from a class of frequent genetic mutations involving mixed lineage leukemia (MLL) gene in a type of hematopoietic malignancy[128]. More surprisingly, all of these factors turned out to be associated with each other to form a single complex, both functionally and physically [124].

The first factor discovered among these is Elongin. Elongin was purified from a fractionation of rat liver nuclear extract that enhanced the in vitro transcription rate of Pol II from the adenovirus 2 major late promoter (AdML) template, composed of three subunits A, B, and C. Elongin's function in transcription elongation was further confirmed when an Elongin related gene in humans, eleven-nineteen lysine-rich leukemia 1 (ELL1), was found to be a frequent fusion partner MLL in translocation t(11;19)(q23;p13.1)[29]. Functional characterization showed that the Elongin and ELL family of proteins increase the net catalytic rate of Pol II, presumably by reducing the transient pauses in gene bodies[124].

ELL can interact with the elongating Pol II as a part of a multi-protein complex named super elongation complex (SEC). SEC consists of AFF1 or AFF4 (AF4/FMR2 family member 1 and 4), ENL (eleven-nineteen leukemia) or AF9 (ALL fused gene from chromosome 9), ELL1 or ELL2, and active P-TEFb[30,124]. The possible combinations of the factors generate diverse subtypes SECs that can activate distinct sets of genes under various physiological contexts[124]. In addition, AFF1, AF9, and ENL also are frequent translocation partners of MLL[128]. MLL can bind to Menin and LEDGF (lens epithelial-derived growth factor), and these two factors are responsible for the aberrant tethering of MLL-fused SEC to promoters of Hox genes, resulting in misregulation at the elongation stage and disease progression into leukemia[124].

Elongation Through the Nucleosome and Gene Body

Even after pause escape, Pol II still has to break further roadblocks in the gene body, most of which are nucleosomes (Figure 1.4). Genes can use general mechanisms to remove or weaken the nucleosome barriers, such as nucleosome remodeling, exchange of histone variants, and histone tail modifications. Elongating Pol II can make use of an entourage of its associated chaperones to overcome the nucleosome barriers, which can be detected by assaying pausing of Pol II or the loss of the evicted barriers.

Overcoming the nucleosome barrier

FACT. Nucleosomes provide a barrier to elongation. In vitro assembled nucleosomal templates strongly inhibit the generation of full-length transcripts by Pol II[45,105]. However, adding HeLa nuclear extract allowed transcription through these chromatinized templates, and FACT was identified from the biochemical fraction that allowed Pol II to overcome the nucleosome barriers[45]. FACT has 2 subunits, Spt16 and SSRP1[129]. It functions to disassemble an H2A-H2B dimer from nucleosomes and Pol II can transcribe through the remaining histone hexamer without being displaced[46]. Afterwards, the disassembled H2A-H2B dimer can be re-deposited by FACT, leaving the reassembled histone octamer in a more dynamic state[130]. FACT's in vivo association with chromatin was also confirmed on *Drosophila Hsp70*, and it showed localization and recruitment kinetics that are consistent with elongating Pol II[47].

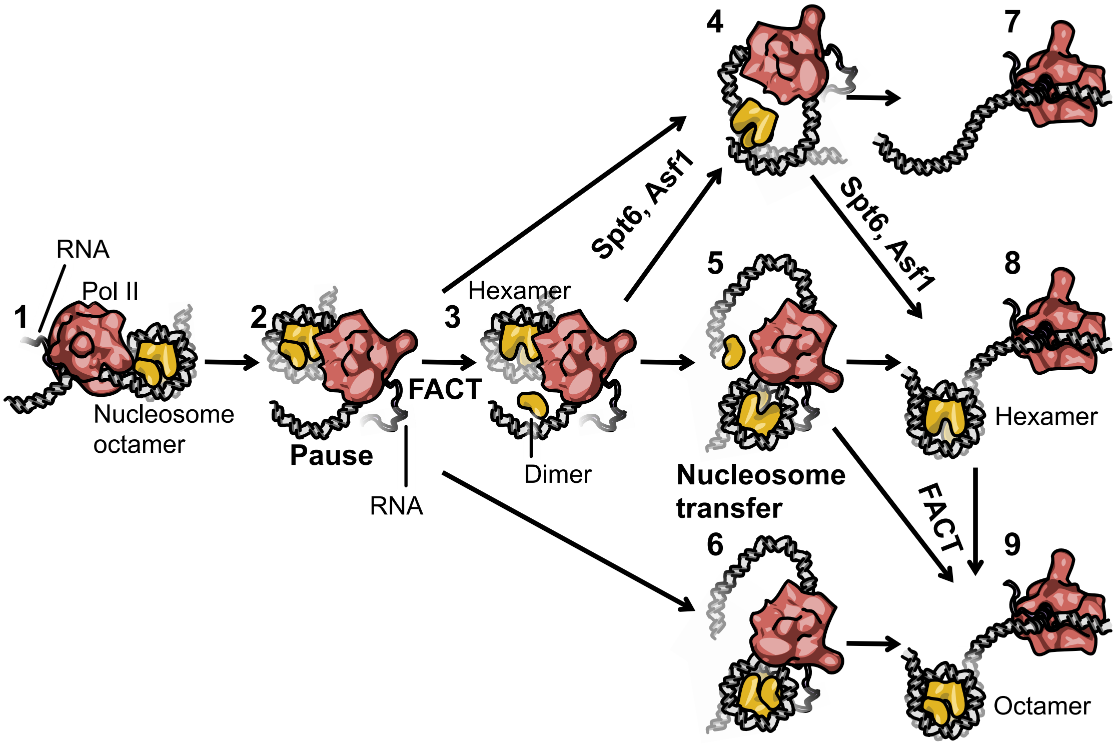


Figure 1.4. Pol II transcribing through gene body nucleosomes**.** Pol II may use multiple mechanisms to get through a nucleosome, and not all the steps are used. **Step 1**, Pol II approaches and makes contact with a nucleosome. **Step 2**, the outer wrap of nucleosomal DNA can be easily unwrapped, and Pol II moves into the nucleosome near the dyad axis. Pol II active site is at around -40 from the dyad axis[8,105,113]. The nucleosome binding is strong at this point and Pol II often pauses transiently. **Step 3**, H2A/H2B dimer is dissociated from the DNA and the nucleosome is now a hexamer. A dissociated dimer can still remain through its association with FACT and be re-deposited later[46]. **Step 4**, H3/H4 core nucleosomal particle is evicted from DNA. H3 can remain associated and be re-deposited by Spt6 or Asf[151,52]. **Step 5**, nucleosome hexamer transfers upstream of Pol II while Pol II transcribes into downstream region. A looping intermediate may form during the transfer. **Step 6**, nucleosome octamer transfers upstream of Pol II, which can be facilitated by histone chaperones. **Step 7**, Pol II evicts the nucleosome by transcribing through it. **Step 8**, Pol II transcribes through the nucleosome leaving a hexamer. **Step 9**, Pol II transcribes through the nucleosome leaving an octamer.

Spt6. Spt6 was initially identified from the *Suppressor of Ty* genetic screen in yeast[1,2]. Spt6 can physically interact with histones H3 and H4 directly, and has histone chaperone activity[51]. Spt6 travels with Pol II, as it shows an in vivo association with chromatin on *Drosophila Hsp70* that is consistent with elongating Pol II[47]. Its function is proposed to displace the nucleosome in front of Pol II and reassemble it in the back, allowing Pol II to transcribe through[51]. However, Spt6 can also increase the elongation rate of Pol II on naked DNA in vitro[131] and independent of the nucleosomes in vivo[57,132]. Depletion of Spt6 by RNAi decreases the elongation rate on *Drosophila Hsp70* under heat-shock conditions[132], when the nucleosomes on *Hsp70* are already lost[57,58]. Therefore, Spt6 may both increase the intrinsic elongation rate of Pol II and remove nucleosome barriers at the same time.

PAF complex. PAF complex is another Pol II associated complex that has a role in elongation, and was first identified from the immuno-precipitating yeast whole cell extract using an antibody to the Pol II CTD[59]. Like Spt6, PAF also is associated with elongating Pol II in vivo[47], suggesting that it travels with Pol II. PAF complex by itself is not known to contain any enzymatic activity, but serves as a platform for recruiting elongation factor complexes such as SEC[1,2,32], FACT[60], and histone modifying enzymes to elongating Pol II[133]. In particular, PAF complex was required for H2BK123 monoubiquitination by RNF20/40 (Rad6) and UbcH6 *in vitro*, and this H2B modification shows a cooperative effect with FACT to allow Pol II to transcribe through nucleosomes[60].

PARP. All the factors above in this section relieve the nucleosome barriers in transcription-dependent manner, associated with elongating Pol II. However, recent studies suggested that PARP (poly(ADP-ribose) polymerase) can achieve this in a transcription-independent way[57,58]. PARP is required for chromatin loosening at transcriptionally active puffs in *Drosophila* polytene chromosomes[134] and is bound to the promoters of active genes in a pattern that is reciprocal to histone H1[135]. On *Drosophila Hsp70*, heat shock factor (HSF) response recruits Tip60 histone acetyl-transferase complex to acetylate H2AK5 and subsequently activate PARP. This results in a creation of poly(ADP-ribose) chains and rapid nucleosome loss that precedes the elongation of Pol II through the gene body[58].

Single molecule studies provide insights to mechanism

Using biochemical experiments, it was shown that SP6 RNAP can transcribe into partially unwrapped DNA that rewraps on the same nucleosome as soon as the polymerase has transcribed through, and the nucleosome 'steps around' (~ 50 bp upstream) the polymerase[136]. However, Pol II is about 1.7 times larger than SP6 RNAP in dimension, and in other in vitro studies, most Pol II molecules were unable to overcome the nucleosome barrier by itself[136]. The precise mechanism could be assessed by examining the Pol II molecules individually to characterize their elongation movements and by defining the conditions under which Pol II can transcribe through the nucleosomes.

Using optical tweezers, it was shown that Pol II could transcribe through the nucleosome under higher ionic strength conditions. The kinetics of Pol II movements were consistent with the model that Pol II transcribes into locally unwrapped DNA[137], similar to SP6 RNAP. Atomic force microscopy made it possible to identify the intermediate structure with locally unwrapped DNA containing Pol II looping back to the same nucleosome. Also, some nucleosomal particles have lost H2A/H2B dimers to become hexamers, and even the whole octamers were lost when the Pol II elongation rates were faster under higher NTP concentrations[138].

When the histones were modified, the kinetics of Pol II through the nucleosomes change. The position of Pol II pausing relative to the nucleosome could be at the entry, central, and exit regions, but reached the maxima at the central region (-35 to +5bp from dyad)[15,137,139]. This is consistent with in vitro[105] and in vivo[8,113] positions of Pol II relative to gene body nucleosomes (not the first nucleosome) where the pause is on average ~ 30 bp inside the upstream margin and ~ -40 bp relative to the dyad of the nucleosomes. When the histone tails are acetylated or absent, the pause density at the entry region is decreased, indicating that the unacetylated tails normally interact with the outer turn of the DNA surrounding the nucleosome. When the core histone H3 and H4 are mutated, the pause density in the central region is decreased indicating that the relaxed interaction near the dyad allows Pol II to transcribe through better[139].

One limitation of these methods is that the experiments were carried out under ionic strengths that are higher than the physiological conditions[137,139]. However, by adding reconstituted elongation factors to the system and reproducing in vivo conditions, these sets of single molecule methods will be powerful tools to define precisely how Pol II overcomes the nucleosome barrier and how elongation factors cooperate with this process.

Elongation rates

The elongation rate of Pol II is an important property of the transcriptional machinery that can direct the level, timing, and the processing of nascent RNA into mature mRNA. It has two components: the pause-free velocity and the frequency/duration of the pauses. The catalytic rate of Pol II sets up the pause-free velocity. It depends on the translocation of the catalytic active site, and this rate is known to be governed by the ‘trigger loop’ structure of Pol II around the active site[140].

Therefore, elongation factors that can bind to this region of Pol II may affect the trigger loop structure and alter the translocation rate. The net elongation rate of Pol II would then be a composite of this pause-free velocity and the frequency/duration of pausing or arrests due to underlying sequences and nucleosomes.

The in vitro assays that were used to identify factors related to elongation were basically net elongation rate assays over time courses to determine how long it takes to transcribe the full-length templates. These rates do not represent the true elongation rates in vivo, since rate can be affected by limiting NTP concentrations and the presence of additional elongation factors. To measure the elongation rate in vivo, the traditional approach was to use a rapid induction or shut-down of a gene and measure the time for the "first wave" or the "last wave" of Pol II to reach a certain position on the gene[141]. A collection of these rates from yeast to *Drosophila* and mammals range anywhere between 1-5 kilobases per minute (kb min^{-1})[141]. However these measurements are limited in numbers and types of genes examined, and a systematic comparative analysis between the genes was difficult.

Another approach to measuring the elongation rate is by fluorescence recovery after photobleaching (FRAP). Upon photobleaching fluorescently tagged Pol II, the time for full recovery is the time it takes for the newly initiated polymerase to finish transcribing the whole gene. In *Drosophila*, salivary gland polytene chromosomes were used to identify the endogenous *Hsp70* locus under heat shock condition for the FRAP[142]. Similarly, nascent RNA synthesis rate can be measured by FRAP using fluorescently labeled RNA binding proteins. In diploid cells, transgenic genes were used that had structured RNA regions to recruit fluorescently tagged RNA binding proteins[143,144].

Spt6 was the first among known elongation factors for which the role in elongation rate was robustly confirmed in vivo. Using ChIP, it was shown that depletion of Spt6 in *Drosophila* embryonic cells delayed the traversal of the first wave of heat shock response in *Hsp70* gene[132]. In addition, FRAP in *Drosophila* polytene chromosomes showed that the elongation rates on induced *Hsp70* in Spt6 knock-down animals is more than two fold slower than the wild type[132,142].

More recently, GRO-seq is being used as a tool to measure elongation rates at a large number of genes simultaneously[145]. This was made possible because the resolution and the sensitivity of GRO-seq allow the detection elongating Pol II in gene bodies. By stimulating human cells with estrogen or TNFα and measuring the time-course of Pol II induction wave in longer genes, the elongation rates of over 160 responsive genes could be determined. Elongation rates differed between cell types and genes, ranging from 0.4-3.6 kb min^{-1} in estrogen responsive genes[145]. Also genes with higher levels of Pol II in the gene body had faster elongation rates consistent with the in vitro model where cooperative Pol II interactions help rescue arrested Pol II[145-147].

Although the use of stimulus responsive genes yielded elongation rates in one to two hundred genes, using a more general inhibitor of elongation may greatly expand the number. For example, a DRB washout strategy has been used to induce an artificial first wave of Pol II and pre-mRNA level at various positions were detected to measure the elongation rate in several long human genes[148]. Applying genome-wide technology to this type of inhibitor approach will be highly informative in determining the rates and the role of elongation factors in vivo[141].

Conclusion

Traditional genetic and biochemical studies have revealed that elongation is as critical as initiation in transcriptional regulation of some genes. Recent genome-wide analyses of the distribution of transcription complexes reveal that on many genes in metazoans significant regulated barriers to elongation exist, most prominently near promoters. This promoter-proximal pausing and its regulated escape to productive elongation are being better understood with the advent of high resolution methods for precisely mapping Pol II relative to promoter elements and nucleosome barriers. The pausing intensity and position depend both on core promoter complex interactions with Pol II and the first nucleosome barrier, both appear to contribute to differing extents on different promoters. New single molecule technologies are also providing unprecedented views of individual RNA polymerases during elongation and are providing further insights to the mechanics of elongation and the transit of Pol II through nucleosome barriers during elongation. Further studies in the field will be directed towards testing how mechanistically the promoter-proximal pausing is regulated by transcription factors and the how factors influence the efficiency and rate of elongation in cells.

REFERENCES

1. Winston, F., Chaleff, D. T., Valent, B. & Fink, G. R. Mutations affecting Ty-mediated expression of the HIS4 gene of Saccharomyces cerevisiae. *Genetics* **107,** 179–197 (1984).

2. Winston, F. 47 Analysis of SPT Genes: A Genetic Approach toward Analysis of TFIID, Histones, and Other Transcription Factors of Yeast. *Cold Spring Harbor Monograph Archive* **22B,** 1271–1293 (1992).

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

4. Saunders, A., Core, L. J. & Lis, J. T. Breaking barriers to transcription elongation. *Nat. Rev. Mol. Cell Biol.* **7,** 557–567 (2006).

5. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. *Science* **322,** 1845–1848 (2008).

6. Perales, R. & Bentley, D. 'Cotranscriptionality': the transcription elongation complex as a nexus for nuclear transactions. *Mol Cell* **36,** 178–191 (2009).

7. Nechaev, S. *et al.* Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila. *Science* **327,** 335–338 (2010).

8. Kwak, H., Fuda, N. J., Core, L. J. & Lis, J. T. Precise Maps of RNA Polymerase Reveal How Promoters Direct Initiation and Pausing. *Science* **339,** 950–953 (2013).

9. Min, I. M. *et al.* Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells. *Genes Dev* **25,** 742–754 (2011).

10. Liu, X., Bushnell, D. A., Silva, D.-A., Huang, X. & Kornberg, R. D. Initiation complex structure and promoter proofreading. *Science* **333,** 633–637 (2011).

11. Sainsbury, S., Niesser, J. & Cramer, P. Structure and function of the initially transcribing RNA polymerase II–TFIIB complex. *Nature* **493,** 437–440 (2012).

12. Tsukiyama, T. T., Becker, P. B. P. & Wu, C. C. ATP-dependent nucleosome disruption at a heat-shock promoter mediated by binding of GAGA transcription factor. *Nature* **367,** 525–532 (1994).

13. Shopland, L. S. L., Hirayoshi, K. K., Fernandes, M. M. & Lis, J. T. J. HSF access to heat shock elements in vivo depends critically on promoter architecture defined by GAGA factor, TFIID, and RNA polymerase II binding sites. *Genes Dev* **9,** 2756–2769 (1995).

14. Tsukiyama, T. & Wu, C. Purification and properties of an ATP-dependent nucleosome remodeling factor. *Cell* **83,** 1011–1020 (1995).

15. Bengal, E., Flores, O., Krauskopf, A., Reinberg, D. & Aloni, Y. Role of the mammalian transcription factors IIF, IIS, and IIX during elongation by RNA polymerase II. *Mol Cell Biol* **11,** 1195–1206 (1991).

16. Tan, S., Conaway, R. C. & Conaway, J. W. Dissection of transcription factor TFIIF functional domains required for initiation and elongation. *Proc Natl Acad Sci U S A* **92,** 6042–6046 (1995).

17. Pokholok, D. K., Hannett, N. M. & Young, R. A. Exchange of RNA polymerase II initiation and elongation factors during gene expression in vivo. *Mol Cell* **9,** 799–809 (2002).

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

19. Ghavi-Helm, Y. *et al.* Genome-wide location analysis reveals a role of TFIIS in RNA polymerase III transcription. *Genes Dev* **22,** 1934–1947 (2008).

20. Yamaguchi, Y. Y. *et al.* NELF, a Multisubunit Complex Containing RD, Cooperates with DSIF to Repress RNA Polymerase II Elongation. *Cell* **97,** 11–11 (1999).

21. Narita, T. *et al.* Human transcription elongation factor NELF: identification of novel subunits and reconstitution of the functionally active complex. *Mol Cell Biol* **23,** 1863–1873 (2003).

22. Wada, T. T. *et al.* DSIF, a novel transcription elongation factor that regulates RNA polymerase II processivity, is composed of human Spt4 and Spt5 homologs. *Genes Dev* **12,** 343–356 (1998).

23. Hartzog, G. A., Wada, T., Handa, H. & Winston, F. Evidence that Spt4, Spt5, and Spt6 control transcription elongation by RNA polymerase II in Saccharomyces cerevisiae. *Genes Dev* **12,** 357–369 (1998).

24. Yamada, T. *et al.* P-TEFb-mediated phosphorylation of hSpt5 C-terminal repeats is critical for processive transcription elongation. *Mol Cell* **21,** 227–237 (2006).

25. Fujinaga, K. *et al.* Dynamics of human immunodeficiency virus transcription: P-TEFb phosphorylates RD and dissociates negative effectors from the transactivation response element. *Mol Cell Biol* **24,** 787–795 (2004).

26. Kim, J. B. & Sharp, P. A. Positive transcription elongation factor B phosphorylates hSPT5 and RNA polymerase II carboxyl-terminal domain independently of cyclin-dependent kinase-activating kinase. *J Biol Chem* **276,** 12317–12323 (2001).

27. Ramanathan, Y. *et al.* Three RNA polymerase II carboxyl-terminal domain kinases display distinct substrate preferences. *J Biol Chem* **276,** 10913–10920 (2001).

28. Aso, T., Lane, W., Conaway, J. & Conaway, R. Elongin (SIII): a multisubunit regulator of elongation by RNA polymerase II. *Science* **269,** 1439–1443 (1995).

29. Shilatifard, A., Lane, W. S., Jackson, K. W., Conaway, R. C. & Conaway, J. W. An RNA Polymerase II Elongation Factor Encoded by the Human ELL Gene. *Science* **271,** 1873–1876 (1996).

30. Lin, C. *et al.* AFF4, a component of the ELL/P-TEFb elongation complex and a shared subunit of MLL chimeras, can link transcription elongation to leukemia. *Mol Cell* **37,** 429–437 (2010).

31. Takahashi, H. *et al.* Human mediator subunit MED26 functions as a docking site for transcription elongation factors. *Cell* **146,** 92–104 (2011).

32. He, N. *et al.* Human Polymerase-Associated Factor complex (PAFc) connects the Super Elongation Complex (SEC) to RNA polymerase II on chromatin. *Proc Natl Acad Sci U S A* **108,** E636–45 (2011).

33. Eberhardy, S. R. & Farnham, P. J. Myc recruits P-TEFb to mediate the final step in the transcriptional activation of the cad promoter. *J Biol Chem* **277,** 40156–40162 (2002).

34. Barboric, M., Nissen, R. M., Kanazawa, S., Jabrane-Ferrat, N. & Peterlin, B. M. NF-kappaB binds P-TEFb to stimulate transcriptional elongation by RNA polymerase II. *Mol Cell* **8,** 327–337 (2001).

35. Jang, M. K. *et al.* The bromodomain protein Brd4 is a positive regulatory component of P-TEFb and stimulates RNA polymerase II-dependent transcription. *Mol Cell* **19,** 523–534 (2005).

36. Yang, Z. *et al.* Recruitment of P-TEFb for stimulation of transcriptional elongation by the bromodomain protein Brd4. *Mol Cell* **19,** 535–545 (2005).

37. Lenasi, T., Peterlin, B. M. & Barboric, M. Cap-binding protein complex links pre-mRNA capping to transcription elongation and alternative splicing through positive transcription elongation factor b (P-TEFb). *J Biol Chem* **286,** 22758–22768 (2011).

38. Mandal, S. S. S. *et al.* 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 (2004).

39. Cowling, V. H. & Cole, M. D. The Myc transactivation domain promotes global phosphorylation of the RNA polymerase II carboxy-terminal domain independently of direct DNA binding. *Mol Cell Biol* **27,** 2059–2073 (2007).

40. Brannan, K. *et al.* mRNA decapping factors and the exonuclease Xrn2 function in widespread premature termination of RNA polymerase II transcription. *Mol Cell* **46,** 311–324 (2012).

41. Wagschal, A. *et al.* Microprocessor, Setx, Xrn2, and Rrp6 co-operate to induce premature termination of transcription by RNAPII. *Cell* **150,** 1147–1157 (2012).

42. Hu, X. *et al.* A Mediator-responsive form of metazoan RNA polymerase II. *Proc Natl Acad Sci U S A* **103,** 9506–9511 (2006).

43. Cheng, B. *et al.* Functional association of Gdown1 with RNA polymerase II poised on human genes. *Mol Cell* **45,** 38–50 (2012).

44. Jishage, M. *et al.* Transcriptional Regulation by Pol II(G) Involving Mediator and Competitive Interactions of Gdown1 and TFIIF with Pol II. *Mol Cell* **45,** 51–63 (2012).

45. Orphanides, G., LeRoy, G., Chang, C. H., Luse, D. S. & Reinberg, D. FACT, a factor that facilitates transcript elongation through nucleosomes. *Cell* **92,** 105–116 (1998).

46. Belotserkovskaya, R. *et al.* FACT facilitates transcription-dependent nucleosome alteration. *Science* **301,** 1090–1093 (2003).

47. Saunders, A. *et al.* Tracking FACT and the RNA polymerase II elongation complex through chromatin in vivo. *Science* **301,** 1094–1096 (2003).

48. Ishimi, Y. & Kikuchi, A. Identification and molecular cloning of yeast homolog of nucleosome assembly protein I which facilitates nucleosome assembly in vitro. *J Biol Chem* **266,** 7025–7029 (1991).

49. Lorch, Y., Maier-Davis, B. & Kornberg, R. D. Chromatin remodeling by nucleosome disassembly in vitro. *Proc Natl Acad Sci U S A* **103,** 3090–3093 (2006).

50. Walfridsson, J., Khorosjutina, O., Matikainen, P., Gustafsson, C. M. & Ekwall, K. A genome-wide role for CHD remodelling factors and Nap1 in nucleosome disassembly. *EMBO J.* **26,** 2868–2879 (2007).

51. Bortvin, A. & Winston, F. Evidence That Spt6p Controls Chromatin Structure by a Direct Interaction with Histones. *Science* **272,** 1473–1476 (1996).

52. Schwabish, M. A. & Struhl, K. Asf1 mediates histone eviction and deposition during elongation by RNA polymerase II. *Mol Cell* **22,** 415–422 (2006).

53. Rufiange, A., Jacques, P.-E., Bhat, W., Robert, F. & Nourani, A. Genome-wide replication-independent histone H3 exchange occurs predominantly at promoters and implicates H3 K56 acetylation and Asf1. *Mol Cell* **27,** 393–405 (2007).

54. Carey, M., Li, B. & Workman, J. L. RSC exploits histone acetylation to abrogate the nucleosomal block to RNA polymerase II elongation. *Mol Cell* **24,** 481–487 (2006).

55. Kasten, M. *et al.* Tandem bromodomains in the chromatin remodeler RSC recognize acetylated histone H3 Lys14. *EMBO J.* **23,** 1348–1359 (2004).

56. Gkikopoulos, T. *et al.* A role for Snf2-related nucleosome-spacing enzymes in genome-wide nucleosome organization. *Science* **333,** 1758–1760 (2011).

57. Petesch, S. J. & Lis, J. T. Rapid, transcription-independent loss of nucleosomes over a large chromatin domain at Hsp70 loci. *Cell* **134,** 74–84 (2008).

58. Petesch, S. J. & Lis, J. T. Activator-induced spread of poly(ADP-ribose) polymerase promotes nucleosome loss at Hsp70. *Mol Cell* **45,** 64–74 (2012).

59. Shi, X. *et al.* Paf1p, an RNA polymerase II-associated factor in Saccharomyces cerevisiae, may have both positive and negative roles in transcription. *Mol Cell Biol* **16,** 669–676 (1996).

60. Pavri, R. *et al.* Histone H2B monoubiquitination functions cooperatively with FACT to regulate elongation by RNA polymerase II. *Cell* **125,** 703–717 (2006).

61. Zippo, A. *et al.* Histone crosstalk between H3S10ph and H4K16ac generates a histone code that mediates transcription elongation. *Cell* **138,** 1122–1136 (2009).

62. Kim, J.-H., Lane, W. S. & Reinberg, D. Human Elongator facilitates RNA polymerase II transcription through chromatin. *Proc Natl Acad Sci U S A* **99,** 1241–1246 (2002).

63. Carrozza, M. J. *et al.* Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription. *Cell* **123,** 581–592 (2005).

64. Ng, H. H., Robert, F., Young, R. A. & Struhl, K. 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 (2003).

65. Kim, T. & Buratowski, S. Dimethylation of H3K4 by Set1 Recruits the Set3 Histone Deacetylase Complex to 5′ Transcribed Regions. *Cell* **137,** 259–272 (2009).

66. Krogan, N. J. *et al.* Methylation of histone H3 by Set2 in Saccharomyces cerevisiae is linked to transcriptional elongation by RNA polymerase II. *Mol Cell Biol* **23,** 4207–4218 (2003).

67. Zippo, A., De Robertis, A., Serafini, R. & Oliviero, S. PIM1-dependent phosphorylation of histone H3 at serine 10 is required for MYC-dependent transcriptional activation and oncogenic transformation. *Nat. Cell Biol.* **9,** 932–944 (2007).

68. Ptashne, M. & Gann, A. Transcriptional activation by recruitment. *Nature* **386,** 569–577 (1997).

69. Stargell, L. A. & Struhl, K. Mechanisms of transcriptional activation in vivo: two steps forward. *Trends Genet.* **12,** 311–315 (1996).

70. Gariglio, P., Bellard, M. & Chambon, P. Clustering of RNA polymerase B molecules in the 5' moiety of the adult beta-globin gene of hen erythrocytes. *Nucleic Acids Res.* **9,** 2589–2598 (1981).

71. Gilmour, D. S. D. & Lis, J. T. J. RNA polymerase II interacts with the promoter region of the noninduced hsp70 gene in Drosophila melanogaster cells. *Mol Cell Biol* **6,** 3984–3989 (1986).

72. Strobl, L. J. & Eick, D. Hold back of RNA polymerase II at the transcription start site mediates down-regulation of c-myc in vivo. *EMBO J.* **11,** 3307–3314 (1992).

73. Krumm, A., Meulia, T., Brunvand, M. & Groudine, M. The block to transcriptional elongation within the human c-myc gene is determined in the promoter-proximal region. *Genes Dev* **6,** 2201–2213 (1992).

74. Plet, A., Eick, D. & Blanchard, J. M. Elongation and premature termination of transcripts initiated from c-fos and c-myc promoters show dissimilar patterns. *Oncogene* **10,** 319–328 (1995).

75. Giardina, C., Pérez-Riba, M. & Lis, J. T. Promoter melting and TFIID complexes on Drosophila genes in vivo. *Genes Dev* **6,** 2190–2200 (1992).

76. Rougvie, A. E. A. & Lis, J. T. J. The RNA polymerase II molecule at the 5' end of the uninduced hsp70 gene of D. melanogaster is transcriptionally engaged. *Cell* **54,** 795–804 (1988).

77. Muse, G. W. *et al.* RNA polymerase is poised for activation across the genome. *Nat Genet* **39,** 1507–1511 (2007).

78. Zeitlinger, J. *et al.* RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo. *Nat Genet* **39,** 1512–1516 (2007).

79. Guenther, M. G., Levine, S. S., Boyer, L. A., Jaenisch, R. & Young, R. A. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell* **130,** 77–88 (2007).

80. Larschan, E. *et al.* X chromosome dosage compensation via enhanced transcriptional elongation in Drosophila. *Nature* **471,** 115–118 (2011).

81. Core, L. J. *et al.* Defining the status of RNA polymerase at promoters. *Cell Rep* **2,** 1025–1035 (2012).

82. Roeder, R. G. The role of general initiation factors in transcription by RNA polymerase II. *Trends Biochem. Sci.* **21,** 327–335 (1996).

83. Chodosh, L. A., Fire, A., Samuels, M. & Sharp, P. A. 5,6-Dichloro-1-beta-D-ribofuranosylbenzimidazole inhibits transcription elongation by RNA polymerase II in vitro. *J Biol Chem* **264,** 2250–2257 (1989).

84. Marshall, N. F. & Price, D. H. Purification of P-TEFb, a Transcription Factor Required for the Transition into Productive Elongation. *J Biol Chem* **270,** 12335–12338 (1995).

85. Marshall, N. F. & Price, D. H. Control of formation of two distinct classes of RNA polymerase II elongation complexes. *Mol Cell Biol* **12,** 2078–2090 (1992).

86. Gilchrist, D. A. D. *et al.* Pausing of RNA Polymerase II Disrupts DNA-Specified Nucleosome Organization to Enable Precise Gene Regulation. *Cell* **143,** 12–12 (2010).

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

88. Guertin, M. J. & Lis, J. T. Mechanisms by which transcription factors gain access to target sequence elements in chromatin. *Curr. Opin. Genet. Dev.* (2012). doi:10.1016/j.gde.2012.11.008

89. Carey, M. PICking apart Pol II initiation. *Nat Struct Mol Biol* **19,** 737–738 (2012).

90. Tirode, F., Busso, D., Coin, F. & Egly, J.-M. Reconstitution of the Transcription Factor TFIIH. *Mol Cell* **3,** 87–95 (1999).

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

92. Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* **389,** 251–260 (1997).

93. Martinez-Rucobo, F. W., Sainsbury, S., Cheung, A. C. & Cramer, P. Architecture of the RNA polymerase–Spt4/5 complex and basis of universal transcription processivity. *EMBO J.* **30,** 1302–1310 (2011).

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

95. Grob, P. *et al.* Cryo-electron microscopy studies of human TFIID: conformational breathing in the integration of gene regulatory cues. *Structure* **14,** 511–520 (2006).

96. Kostek, S. A. *et al.* Molecular architecture and conformational flexibility of human RNA polymerase II. *Structure* **14,** 1691–1700 (2006).

97. Spangler, L., Wang, X., Conaway, J. W., Conaway, R. C. & Dvir, A. 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 (2001).

98. Lee, C. *et al.* NELF and GAGA factor are linked to promoter-proximal pausing at many genes in Drosophila. *Mol Cell Biol* **28,** 3290–3300 (2008).

99. Missra, A. A. & Gilmour, D. S. D. 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 (2010).

100. Butler, J. E. F. J. & Kadonaga, J. T. J. The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes Dev* **16,** 2583–2592 (2002).

101. Ohler, U., Liao, G.-C., Niemann, H. & Rubin, G. M. Computational analysis of core promoters in the Drosophila genome. *Genome Biol.* **3,** RESEARCH0087 (2002).

102. Omichinski, J. G., Pedone, P. V., Felsenfeld, G., Gronenborn, A. M. & Clore, G. M. The solution structure of a specific GAGA factor-DNA complex reveals a modular binding mode. *Nat. Struct. Biol.* **4,** 122–132 (1997).

103. Bai, L., Shundrovsky, A. & Wang, M. D. Sequence-dependent kinetic model for transcription elongation by RNA polymerase. *J. Mol. Biol.* **344,** 335–349 (2004).

104. Adelman, K. *et al.* Efficient release from promoter-proximal stall sites requires transcript cleavage factor TFIIS. *Mol Cell* **17,** 103–112 (2005).

105. Izban, M. G. & Luse, D. S. Transcription on nucleosomal templates by RNA polymerase II in vitro: inhibition of elongation with enhancement of sequence-specific pausing. *Genes Dev* **5,** 683–696 (1991).

106. Mavrich, T. N. T. *et al.* Nucleosome organization in the Drosophila genome. *Nature* **453,** 358–362 (2008).

107. Wang, D., Severinov, K. & Landick, R. Preferential interaction of the his pause RNA hairpin with RNA polymerase beta subunit residues 904-950 correlates with strong transcriptional pausing. *Proc Natl Acad Sci U S A* **94,** 8433–8438 (1997).

108. Ring, B. Z., Yarnell, W. S. & Roberts, J. W. Function of E. coli RNA polymerase sigma factor sigma 70 in promoter-proximal pausing. *Cell* **86,** 485–493 (1996).

109. Yamaguchi, Y., Inukai, N., Narita, T., Wada, T. & Handa, H. Evidence that negative elongation factor represses transcription elongation through binding to a DRB sensitivity-inducing factor/RNA polymerase II complex and RNA. *Mol Cell Biol* **22,** 2918–2927 (2002).

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

111. Wang, Y. V., Tang, H. & Gilmour, D. S. Identification in vivo of different rate-limiting steps associated with transcriptional activators in the presence and absence of a GAGA element. *Mol Cell Biol* **25,** 3543–3552 (2005).

112. Hendrix, D. A., Hong, J.-W., Zeitlinger, J., Rokhsar, D. S. & Levine, M. S. Promoter elements associated with RNA Pol II stalling in the Drosophila embryo. *Proc Natl Acad Sci U S A* **105,** 7762–7767 (2008).

113. Churchman, L. S. & Weissman, J. S. Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature* **469,** 368–373 (2011).

114. Bregman, D. B. *et al.* UV-induced ubiquitination of RNA polymerase II: a novel modification deficient in Cockayne syndrome cells. *Proc Natl Acad Sci U S A* **93,** 11586–11590 (1996).

115. Ratner, J. N. Ultraviolet Radiation-induced Ubiquitination and Proteasomal Degradation of the Large Subunit of RNA Polymerase II. *J Biol Chem* **273,** 5184–5189 (1998).

116. Somesh, B. P. *et al.* Multiple mechanisms confining RNA polymerase II ubiquitylation to polymerases undergoing transcriptional arrest. *Cell* **121,** 913–923 (2005).

117. Price, D. H. P-TEFb, a cyclin-dependent kinase controlling elongation by RNA polymerase II. *Mol Cell Biol* **20,** 2629–2634 (2000).

118. Bartkowiak, B. *et al.* CDK12 is a transcription elongation-associated CTD kinase, the metazoan ortholog of yeast Ctk1. *Genes Dev* **24,** 2303–2316 (2010).

119. Devaiah, B. N. *et al.* BRD4 is an atypical kinase that phosphorylates serine2 of the RNA polymerase II carboxy-terminal domain. *Proc Natl Acad Sci U S A* **109,** 6927–6932 (2012).

120. Ni, Z. *et al.* P-TEFb is critical for the maturation of RNA polymerase II into productive elongation in vivo. *Mol Cell Biol* **28,** 1161–1170 (2008).

121. Rahl, P. B. *et al.* c-Myc regulates transcriptional pause release. *Cell* **141,** 432–445 (2010).

122. Nguyen, V. T., Kiss, T., Michels, A. A. & Bensaude, O. 7SK small nuclear RNA binds to and inhibits the activity of CDK9/cyclin T complexes. *Nature* **414,** 322–325 (2001).

123. Yang, Z., Zhu, Q., Luo, K. & Zhou, Q. The 7SK small nuclear RNA inhibits the CDK9/cyclin T1 kinase to control transcription. *Nature* **414,** 317–322 (2001).

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

125. Kanazawa, S., Soucek, L., Evan, G., Okamoto, T. & Peterlin, B. M. c-Myc recruits P-TEFb for transcription, cellular proliferation and apoptosis. *Oncogene* **22,** 5707–5711 (2003).

126. Zuber, J. *et al.* RNAi screen identifies Brd4 as a therapeutic target in acute myeloid leukaemia. *Nature* **478,** 524–528 (2011).

127. Nock, A., Ascano, J. M., Barrero, M. J. & Malik, S. Mediator-regulated transcription through the +1 nucleosome. *Mol Cell* **48,** 837–848 (2012).

128. Rowley, J. D. The Critical Role of Chromosome Translocations in Human Leukemias. *Annu. Rev. Genet.* **32,** 495–519 (1998).

129. Orphanides, G., Wu, W. H., Lane, W. S., Hampsey, M. & Reinberg, D. The chromatin-specific transcription elongation factor FACT comprises human SPT16 and SSRP1 proteins. *Nature* **400,** 284–288 (1999).

130. Xin, H. *et al.* yFACT induces global accessibility of nucleosomal DNA without H2A-H2B displacement. *Mol Cell* **35,** 365–376 (2009).

131. Endoh, M. *et al.* Human Spt6 stimulates transcription elongation by RNA polymerase II in vitro. *Mol Cell Biol* **24,** 3324–3336 (2004).

132. Ardehali, M. B. *et al.* Spt6 enhances the elongation rate of RNA polymerase II in vivo. *EMBO J.* **28,** 1067–1077 (2009).

133. Sims, R. J., Belotserkovskaya, R. & Reinberg, D. Elongation by RNA polymerase II: the short and long of it. *Genes Dev* **18,** 2437–2468 (2004).

134. Tulin, A. & Spradling, A. Chromatin loosening by poly(ADP)-ribose polymerase (PARP) at Drosophila puff loci. *Science* **299,** 560–562 (2003).

135. Krishnakumar, R. *et al.* Reciprocal binding of PARP-1 and histone H1 at promoters specifies transcriptional outcomes. *Science* **319,** 819–821 (2008).

136. Studitsky, V. M., Clark, D. J. & Felsenfeld, G. A histone octamer can step around a transcribing polymerase without leaving the template. *Cell* **76,** 371–382 (1994).

137. Hodges, C., Bintu, L., Lubkowska, L., Kashlev, M. & Bustamante, C. Nucleosomal fluctuations govern the transcription dynamics of RNA polymerase II. *Science* **325,** 626–628 (2009).

138. Bintu, L. *et al.* The elongation rate of RNA polymerase determines the fate of transcribed nucleosomes. *Nat Struct Mol Biol* **18,** 1394–1399 (2011).

139. Bintu, L. *et al.* Nucleosomal elements that control the topography of the barrier to transcription. *Cell* **151,** 738–749 (2012).

140. Larson, M. H. *et al.* Trigger loop dynamics mediate the balance between the transcriptional fidelity and speed of RNA polymerase II. *Proc Natl Acad Sci U S A* **109,** 6555–6560 (2012).

141. Ardehali, M. B. & Lis, J. T. Tracking rates of transcription and splicing in vivo. *Nat Struct Mol Biol* **16,** 1123–1124 (2009).

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

143. Brody, Y. *et al.* The in vivo kinetics of RNA polymerase II elongation during co-transcriptional splicing. *PLoS Biol.* **9,** e1000573 (2011).

144. Yunger, S., Rosenfeld, L., Garini, Y. & Shav-Tal, Y. Single-allele analysis of transcription kinetics in living mammalian cells. *Nat. Methods* **7,** 631–633 (2010).

145. Danko, C. G. *et al. Signaling Pathways Differentially Affect RNA Polymerase II Initiation, Pausing, and Elongation Rate in Cells*. (Mol Cell, 2013).

146. Epshtein, V. & Nudler, E. Cooperation between RNA polymerase molecules in transcription elongation. *Science* **300,** 801–805 (2003).

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

148. Singh, J. & Padgett, R. A. Rates of in situ transcription and splicing in large human genes. *Nat Struct Mol Biol* **16,** 1128–1133 (2009).

CHAPTER 2[b]

PRECISE MAPS OF RNA POLYMERASE REAVEAL HOW PROMOTERS DIRECT INITIATION AND PAUSING

Summary

Transcription regulation occurs frequently through promoter-associated pausing of RNA polymerase II (Pol II). We developed a Precision nuclear Run-On and sequencing assay (PRO-seq) to map the genome-wide distribution of transcriptionally-engaged Pol II at base-pair resolution. Pol II accumulates immediately downstream of promoters, at intron-exon junctions that are efficiently used for splicing, and over 3' poly-adenylation sites. Focused analyses of promoters reveal that pausing is not fixed relative to initiation sites nor is it specified directly by the position of a particular core promoter element or the first nucleosome. Core promoter elements function beyond initiation, and when optimally positioned they act collectively to dictate the position and strength of pausing. We test this 'Complex Interaction' model with insertional mutagenesis of the *Drosophila Hsp70* core promoter.

Introduction

Tracking the accumulation of Pol II along genes reveals potential points of regulation[1]. For example, a rate limiting step in early elongation, known as promoter-proximal pausing, has revealed a major regulatory block in the transition to productive elongation in Drosophila and mammals[2-8]. Also, less extensive, but significant accumulation of Pol II over the 3' cleavage/polyA region of genes is proposed to facilitate 3' processing and transcription termination[9,10]. Finally, the interplay of

[b] From Kwak *et al*., Precise Maps of RNA Polymerase Reveals How Promoters Direct Initiation and Pausing, *Science* **339**, 950-953 (2013). Reprinted with permission from AAAS.

transcription rate and splicing efficiency[11] might be reflected in the selective accumulation of Pol II at splice junctions.

Promoter-associated Pol II pausing is a culmination of intrinsic interactions between Pol II and the underlying DNA, as well as extrinsic stabilization by protein complexes[12]. Protein factors such as Negative Elongation Factor (NELF) and DRB Sensitivity Inducing Factor (DSIF)[3,13], DNA elements[14,15], DNA sequence composition[16], nascent RNA processing[16], and nucleosomes[17] can influence pausing. Understanding how these elements and factors function mechanistically requires a high resolution view of their spatial relationship. Current tools for precise tracking of the location and status of Pol II *in vivo* have distinct limitations[18]. ChIP-based methods that collect Pol II or associated RNAs do not distinguish paused Pol II from other Pol II-RNA complexes[16,18,19]. The genome-wide nuclear run-on approach (GRO-seq method)[6-8] circumvents these issues by enriching nascent transcripts only associated with actively engaged polymerase with high sensitivity, but it has a resolution of only 30-50 bases[16].

Development of PRO-seq and Validation of the Pausing Sites

We developed a genome-wide, nuclear run-on assay called PRO-seq that has the sensitivity of GRO-seq, but maps Pol II with base-pair resolution. PRO-seq uses biotin-labeled ribonucleotide triphosphate analogs (biotin-NTP) for nuclear run-on reactions, allowing the efficient affinity purification of nascent RNAs for high throughput sequencing from their 3' ends (Figure. 2.1A). Supplying only one of the four biotin-A/C/G/UTP restricts Pol II to incorporate a single or at most a few identical bases, resulting in sequence reads that have the same 3' end base within each library (Table 2.1). Moreover, the incorporation of the first biotin-base inhibits further transcript elongation, ensuring base-pair resolution (data not shown).

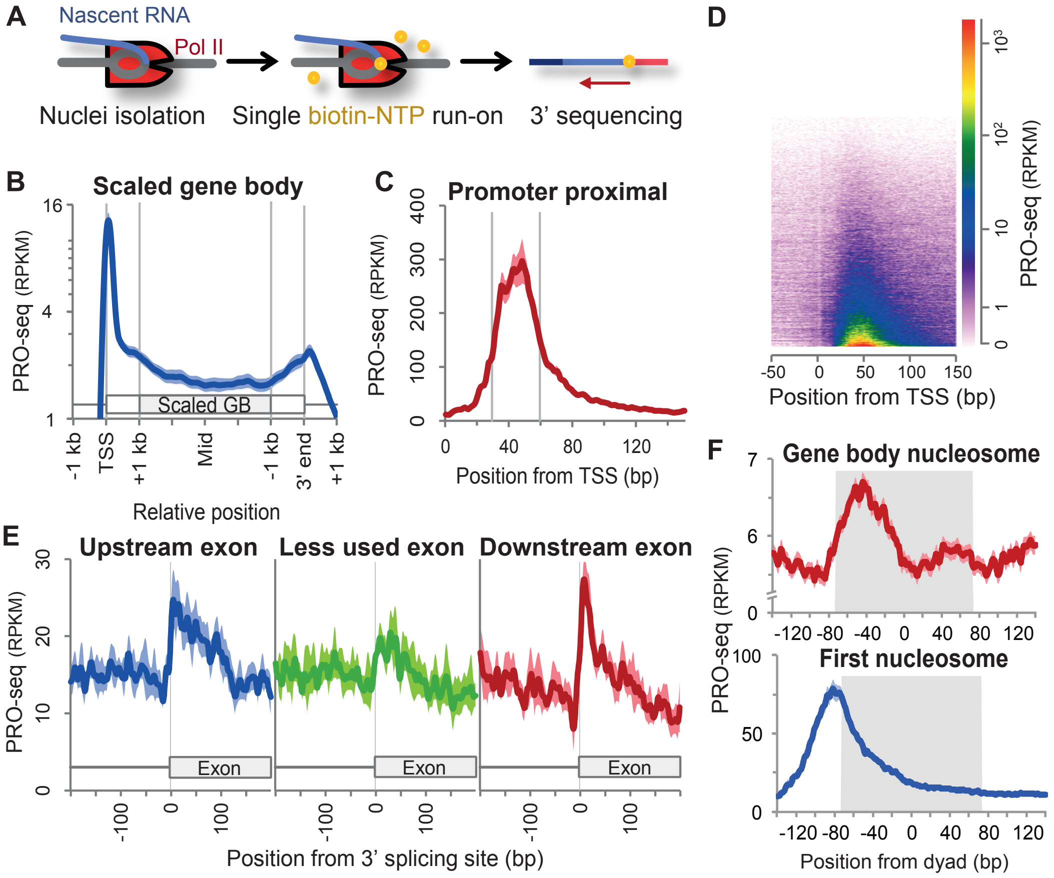


Figure 2.1. Accumulation of Pol II at promoters, 3' ends, 3' splice sites, and nucleosomes. (**A**) Schematic of PRO-seq. (**B**) Average PRO-seq profile of non-overlapping genes (n=6,309) for the sense strand. Gene body regions (+1 kb from the 5' end to -1 kb from the 3' end) are scaled to 4 kb. Read counts are adjusted to per kilobase per million mapped (RPKM). Shaded margins surrounding the average plot show standard errors of the means. (**C**) High resolution PRO-seq profile from the TSS to +150 bp (n=16,746). (**D**) Heatmap visualization of PRO-seq profile of the annotated genes. Genes are arranged by their increasing PRO-seq density. (**E**) Average PRO-seq profile at 3' splicing sites of less used and their flanking exons (n=242 each). Less used exons have RNA-seq densities less than 5% of their flanking exons. (**F**) Average PRO-seq profile relative to the dyad centers of gene body and first nucleosomes. Region occupied by nucleosome is in grey shades.

Table 2.1. 3’ end uniformity in PRO-seq sequence reads. For each of the 4 libraries, we counted the number of sequences reads ending in each of the 4 bases. Then we calculated the percentage of the reads ending in the same base as the nuclear run-on substrate. High substrate base percentage indicates that the majority of 3’ ends of the sequences were at the active sited of engaged polymerases.

Libraries	3’ end base				Substrate base percentage
	A	C	G	T	
PRO-seq$_{ATP}$	11,198,117	270,067	831,090	314,045	88.8%
PRO-seq$_{CTP}$	765,809	11,760,234	402,994	423,382	88.1%
PRO-seq$_{GTP}$	1,447,593	439,723	9,900,845	275,139	82.1%
PRO-seq$_{UTP}$	1,965,563	1,903,897	465,259	8,915,790	67.3%

The average profile of PRO-seq density (Figure 2.1B) reveals pausing of Pol II immediately downstream of the transcription start site (TSS) (Figure 2.1C, Figure 2.1D), and accumulation of Pol II at 3’ cleavage/PolyA sites, consistent with previous studies[20,21]. Interestingly, Pol II also accumulates near 3’ splicing sites at spliced exons, but less often at skipped exons (Figure 2.1E), suggesting that splicing decisions are connected to differential rates of Pol II elongation through splice junctions[11]. Although we have insufficient sequencing coverage to quantify Pol II accumulation at particular 3’ splice sites, our composite analyses support a functional coupling between elongation and splicing.

The highest density of PRO-seq reads map within +30 to +60 from the TSS (Figure 2.1C, Figure 2.1D), providing a higher resolution view of paused Pol II mapped by GRO-seq[21](Figure 2.2D). Moreover, the pattern of pausing by PRO-seq is consistent with the positions and levels of short nuclear-capped RNAs (scRNAs, Figure 2.2A)[16]. Additionally, we demonstrate that PRO-seq maps correspond precisely to positions of engaged Pol II observed in intact cells seen by previous permanganate footprints of transcription bubbles (Figure 2.3).

Nucleosomes are known to act as barriers to Pol II[12]. In the bodies of genes (nucleosomes at position +2 and greater), the average PRO-seq density shows a relative increase at ~ -40 from the previously mapped[22] nucleosome centers (Figure 2.1F). This is consistent with measurements of strong DNA-nucleosome interactions[23] and measured impediments to Pol II transcription through nucleosomes measured *in vitro* and in yeast[19,24]. However, the PRO-seq density relative to the first (+1) nucleosome is different (Figure 2.1F), with the average PRO-seq density at a maximum ~ -80 from the nucleosome centers. Thus, the bulk of promoter-proximal pausing is inconsistent with a standard nucleosome barrier model, and is more consistent with tethering of polymerases near the promoter[21].

Pausing Sites Are Not Fixed, but Linked to Initiation Sites

Whereas the average promoter-associated pause location is approximately at +40 from the TSS, pausing is far from uniform. Some genes have more proximal and focused pausing, while others have distal and dispersed pausing (Figure 2.4A). We systematically assessed genome-wide pausing positions relative to the TSS and their dispersion to identify two characteristic groups of promoters: focused-proximal (*Prox*), and dispersed-distal (*Dist*) promoters (Figure 2.4B).

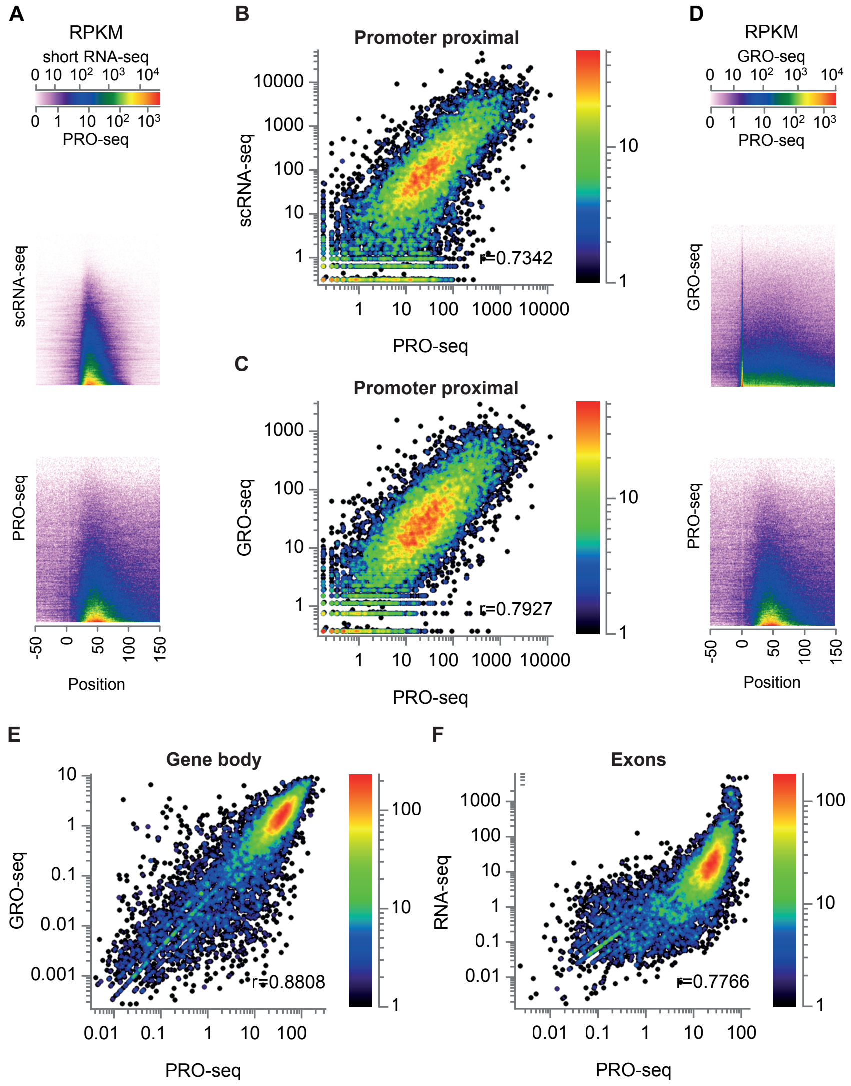
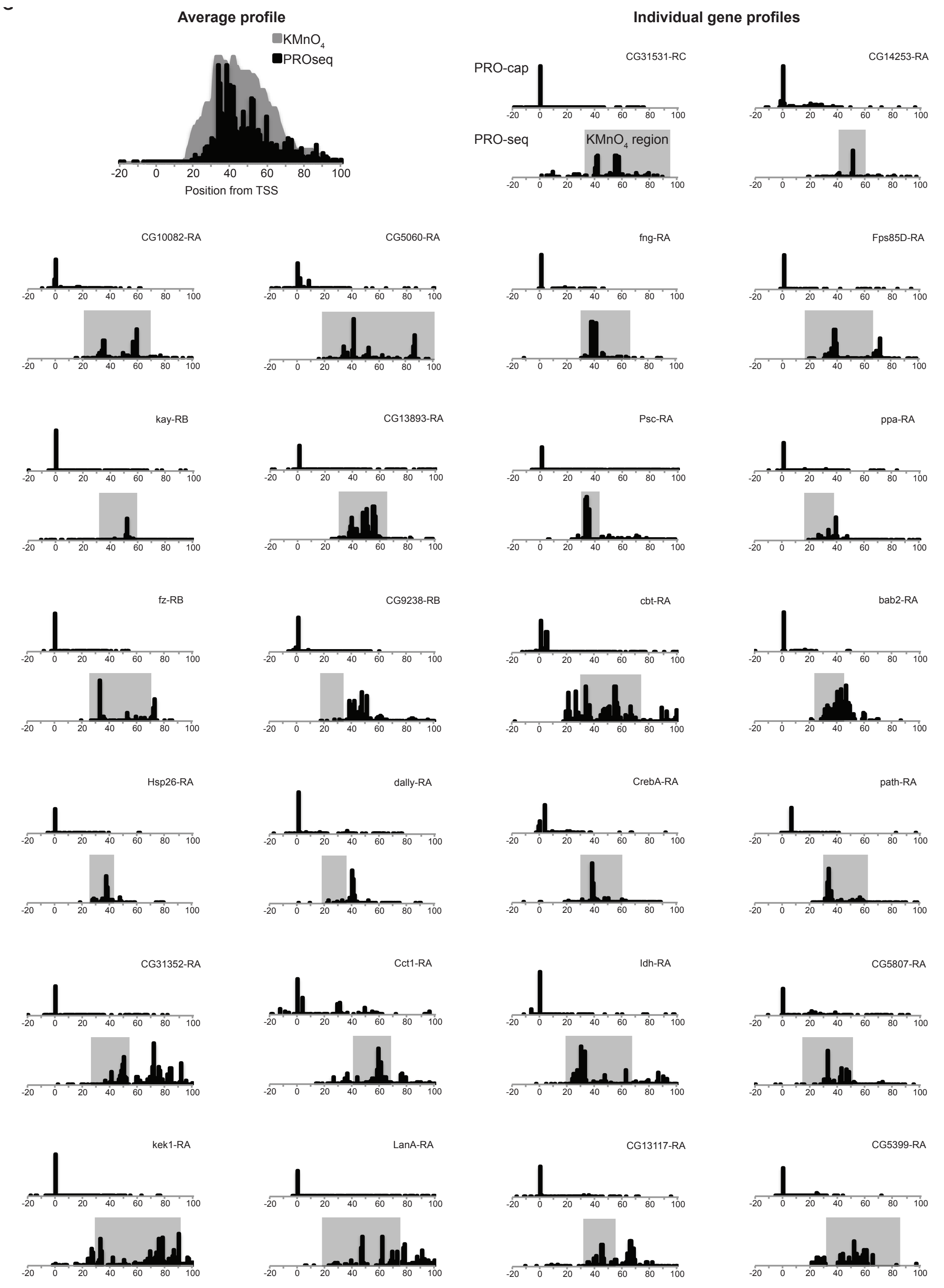


Figure 2.2. PRO-seq compared with other genome-wide transcription assays. (**A**) High resolution heatmap of short capped nuclear RNA-seq (scRNA-seq) and PRO-seq profiles near the TSS. Each profile is sorted by the sum of the reads in the region for each gene. (**B,C**) Scatterplots of promoter-proximal densities between PRO-seq and scRNA-seq (**B**) or GRO-seq (**C**). Promoter-proximal densities are the read densities from -50 to + 150 around TSS, RPKM normalized. (**D**) High resolution heatmap of GRO-seq and PRO-seq profiles near TSS. (**E**) PRO-seq and GRO-seq density correlations in gene body regions. (**F**) PRO-seq and poly-A RNA-seq (modENCODE_3138; GSM461182) densities in exons. The upward concave curvature of the distribution suggests the possibility of elongation rate, nascent RNA processing efficiency, or mRNA stability being positively correlated with density of transcribing polymerases.



Permanganate footprints were obtained from Lee *et. al.* at http://mcb.asm.org/content/suppl/2008/04/18/28.10.3290.DC1/S1_ChIP_KMnO4_59genes_111607.zip

Figure 2.3. Comparison of the PRO-seq , PRO-cap, and permanganate footprint average profiles in selected genes and their individual profiles.

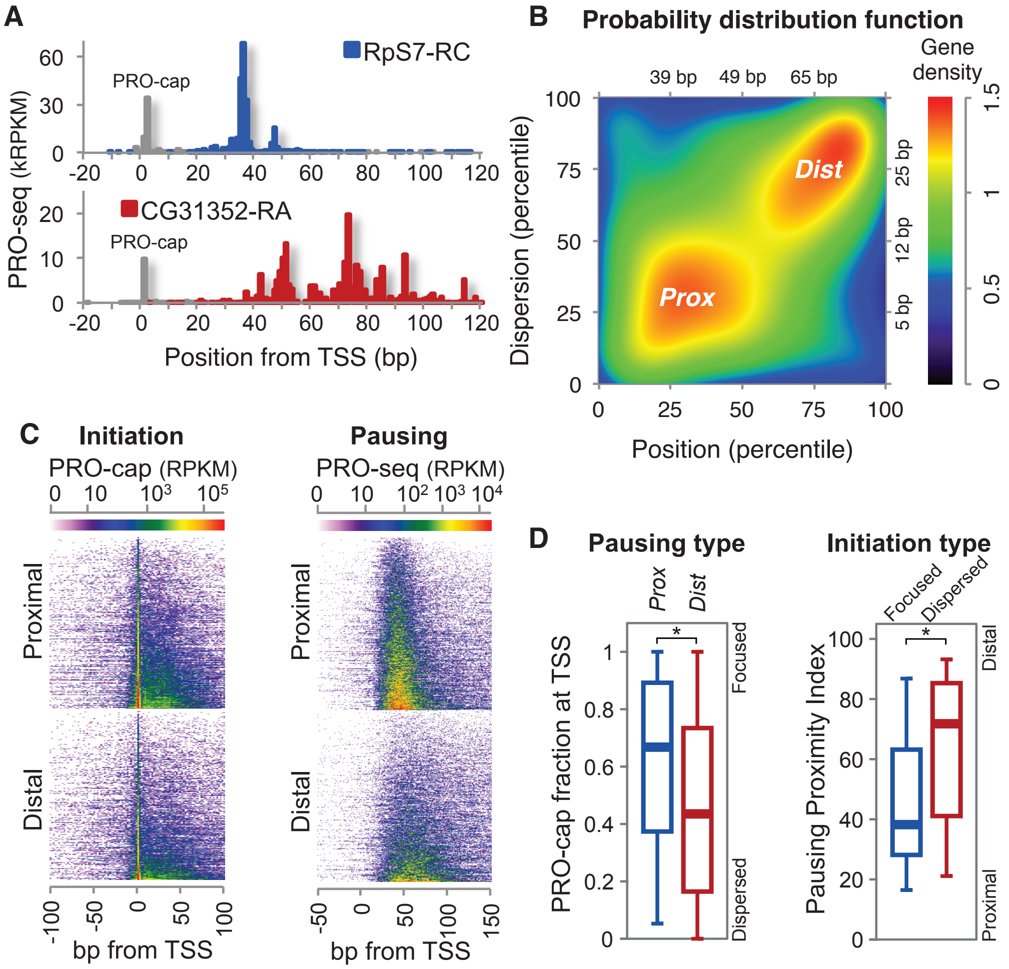


Figure 2.4. Variations of the pause sites and TSSs. (**A**) Examples of highly paused genes with different pausing patterns. Initiation sites from PRO-cap mapping are shown in grey. (**B**) Distribution of paused genes (n=3,225) by the pausing position and dispersion percentiles. Focused-proximal (*Prox,* n=848) and dispersed-distal (*Dist,* n=846) groups are indicated and axis units in base-pairs are also shown. (**C**) Heat map of initiation (PRO-cap) and pausing (PRO-seq) for *Prox* and *Dist* genes. (**D**) Association between initiation and pausing patterns. 'TSS focusing' in *Prox vs Dist* genes (left) and 'Pausing proximity' in focused *vs* dispersed initiation genes (right). The fraction of PRO-cap reads at (±1 bp) the TSS over the sum of reads around (±50 bp) the TSS represents the 'TSS focusing'[22]. Focused and dispersed initiation genes are the quartiles of the paused genes with the highest and the lowest TSS focusing respectively. Pausing Proximity Index is defined by the average of pausing position and dispersion percentiles. Boxes represent 25th, 50th(median), and 75th percentiles; whiskers are 5th and 95th percentiles. Asterisks (*) indicate p<0.001 by Kolmogorov-Smirnov(KS) test.

The *Prox* and *Dist* pausing patterns could arise from a fixed length of elongation from initiation sites that have the same dispersion, or from variable lengths of elongation from more focused initiation sites. Distinguishing between these possibilities requires precise mapping of the initiation sites using the same pool of Pol II-engaged nascent RNAs. Therefore, we modified the PRO-seq method to detect initiation sites (PRO-cap) and compared the degree of variation in the initiation and the pause sites. We observed that both *Prox* and *Dist* genes have relatively focused initiation in general (Figure 2.4A, Figure 2.4D), and that pausing is overall more dispersed than initiation. Nonetheless, the degree of the focused initiation, the fraction of initiation arising exactly at a single TSS, is higher for *Prox* genes, and genes with more focused initiation also have more proximal pausing (Figure 2.4D). These findings indicate that although pausing is not fixed to initiation, the mechanisms that produce focused initiation affect the resulting pattern of pausing.

In an effort to otherwise explain the differential patterns of pausing, we first compared the nucleosome occupancy around *Prox* and *Dist* promoters. *Prox* promoters have less nucleosome occupancy than *Dist* promoters (Figure 2.5A, Figure 2.5B), and some Pol II at *Dist* promoters appear to have more intimate contact with the first nucleosome (Figure 2.5C). These results (and Figure 2.1F) support a nucleosome independent mechanism of pausing for *Prox* promoters, whereas a subset of *Dist* promoters could have a component of pausing that is established by direct nucleosome barriers. Because nucleosome position and occupancy does not explain the bulk of Pol II pausing, we investigated the underlying DNA elements around promoters.

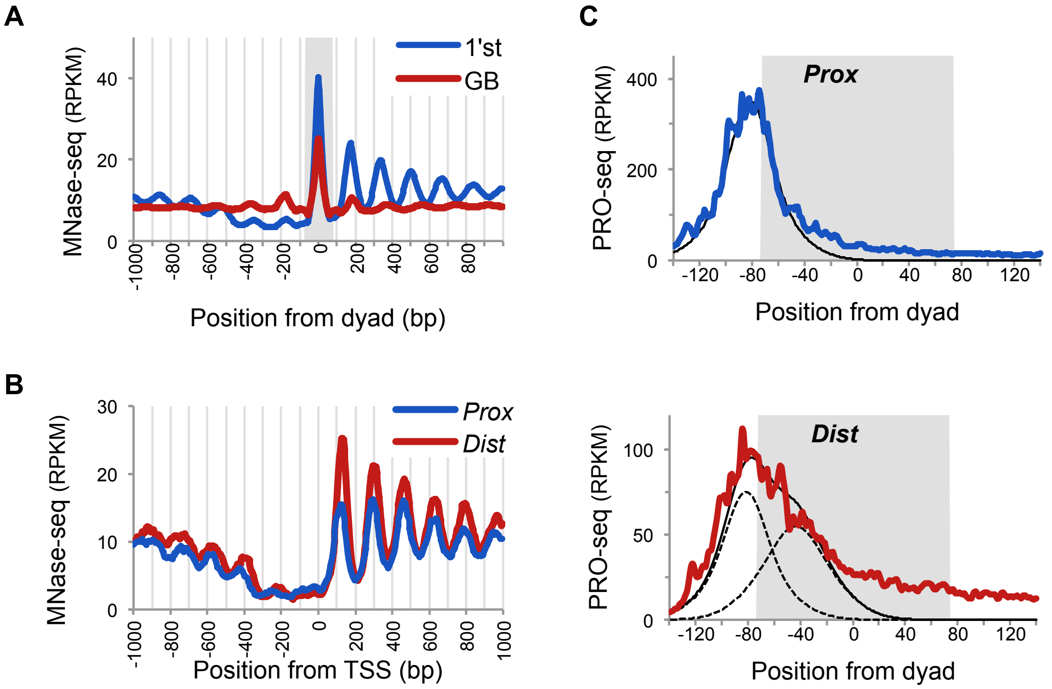


Figure 2.5. PRO-seq relative to nucleosomes. (**A**) Micrococcal nuclease (MNase)-seq profile relative to nucleosome centers in gene body and the first nucleosome. Nucleosome centers were defined from a local Gaussian fit of the MNase-seq data[22] within 175 bp windows. Nucleosome center positions that were more than 2 kb downstream of the TSS were considered gene body, and those that are between 0 to +200 bp from the TSS were considered the first nucleosomes. Regions of the nucleosome occupancy are shaded in grey. (**B**) MNase-seq profile relative to TSS in *Prox* and *Dist* genes. (**C**) Average PRO-seq profile relative to the first nucleosome center in *Prox* and *Dist* genes. Fit curves with broken strokes for the *Dist* genes show possible combination of the *Prox* fit and the gene body nucleosome fit (Figure 2.1F) that are centered at -83 and -44 from nucleosome centers, respectively.

Promoter DNA Elements Shape the Pausing Pattern

Critical DNA sequence elements within the core promoter direct the position, direction and efficiency of transcription initiation[25]. These include the TATA box, Initiator (Inr), Motif Ten Element (MTE), the Downstream Promoter Element (DPE)[25] and a recently discovered element implicated in pausing, the Pause Button[15] (Figure 2.6A). Core promoter elements are more enriched on *Prox* than on *Dist* promoters (Figure 2.7A, Figure 2.6B-D). Additionally, when we searched within the extended promoter regions of *Prox* and *Dist* groups for the presence of 232 additional functional DNA elements[26] (Figure 2.6E), only the GAGA element, residing ~80 bp upstream of the TSS[15,27] shows strong associations with *Prox* genes (Figure 2.7B), as does the level of GAGA-factor binding. Thus, core promoter elements and GAGA-factor appear to play a significant role in the mechanism of pausing.

Promoter DNA elements could be causing the variable pausing positions through direct tethering of elongating Pol II to the elements. Alternatively a 'Complex Interaction' model - where pausing is mediated through protein complexes that function best when cognate elements are located at specific positions in the core promoter. Thus, if we examine the association of the positions of the DNA elements and the pausing sites in this model, we expect a 'V'-shaped plot of association rather than a simple linear correlation: displacement of the element from the optimal position will weaken the interactions within the core complex, resulting in downstream scattering and reduced level of pausing (Figure 2.7C). To test this, we examined genes where a particular promoter DNA element occurs only once, and divided genes into three subsets: the optimal consensus position, upstream and downstream. Genes with the DNA elements nearest to the consensus positions have more proximal pausing.

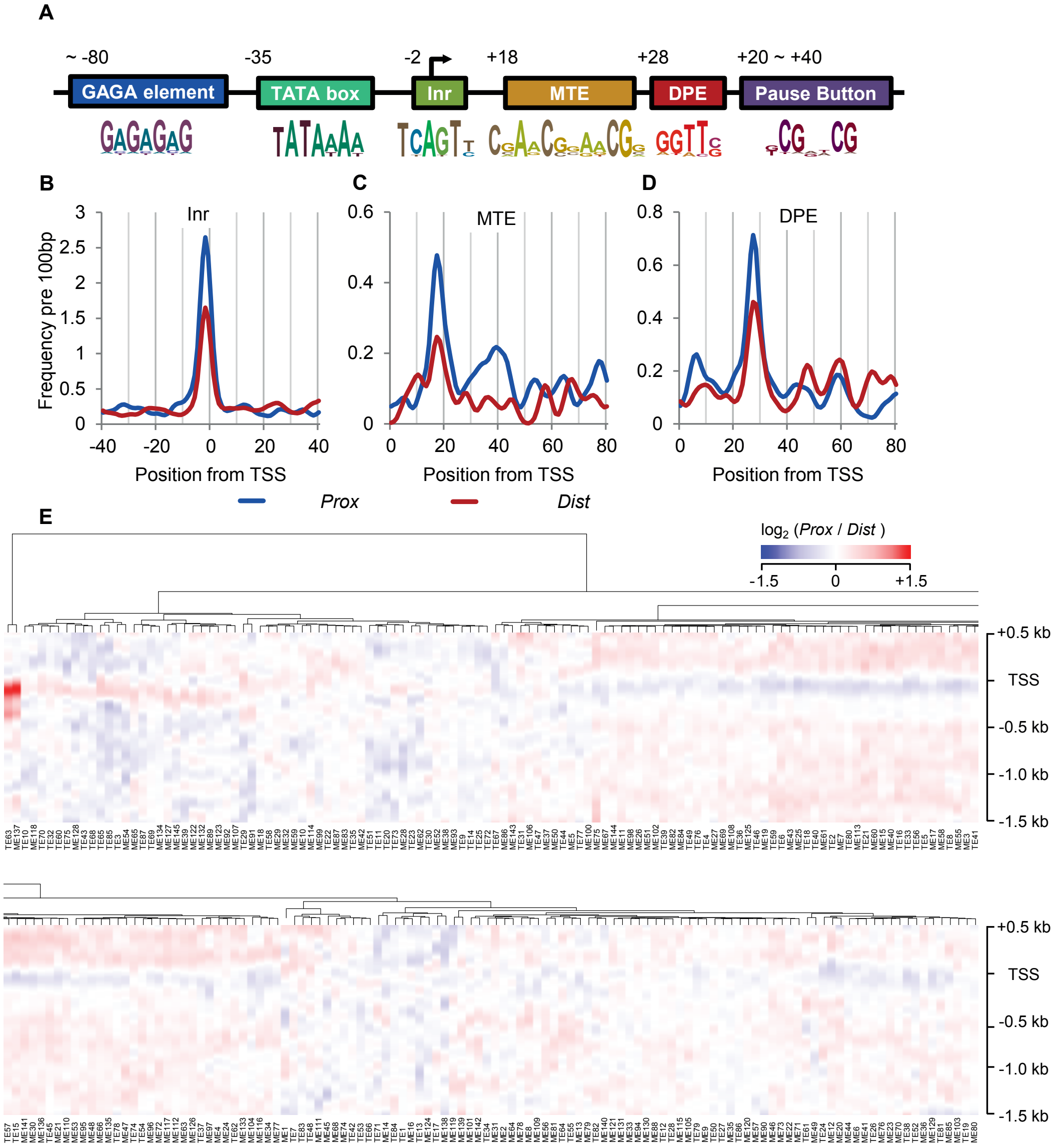


Figure 2.6. DNA elements in different pausing subsets
(**A**) Structure of promoter DNA elements. GAGA element, TATA box, Initiator element (Inr), Motif Ten Element (MTE), Downstream Promoter Element (DPE), and Pause Button (PB)1 are shown at the positions of their peak occurrences. Sequence logo representation of each element is also shown. (**B**, **C**, **D**) Frequency of the core promoter elements in gene subsets of Prox or Dist pausing. Inr, MTE and DPE are shown respectively. (**E**) Heatmap of the difference in the occurrence of sequence elements between Prox and Dist groups. 232 regulatory DNA elements in two sub-panels from Stark et al.1 were examined. The elements are ordered according to a hierarchical clustering. TE63 and ME137, both of which match the consensus for GAGA element, are the only elements showing noticeable enrichments.

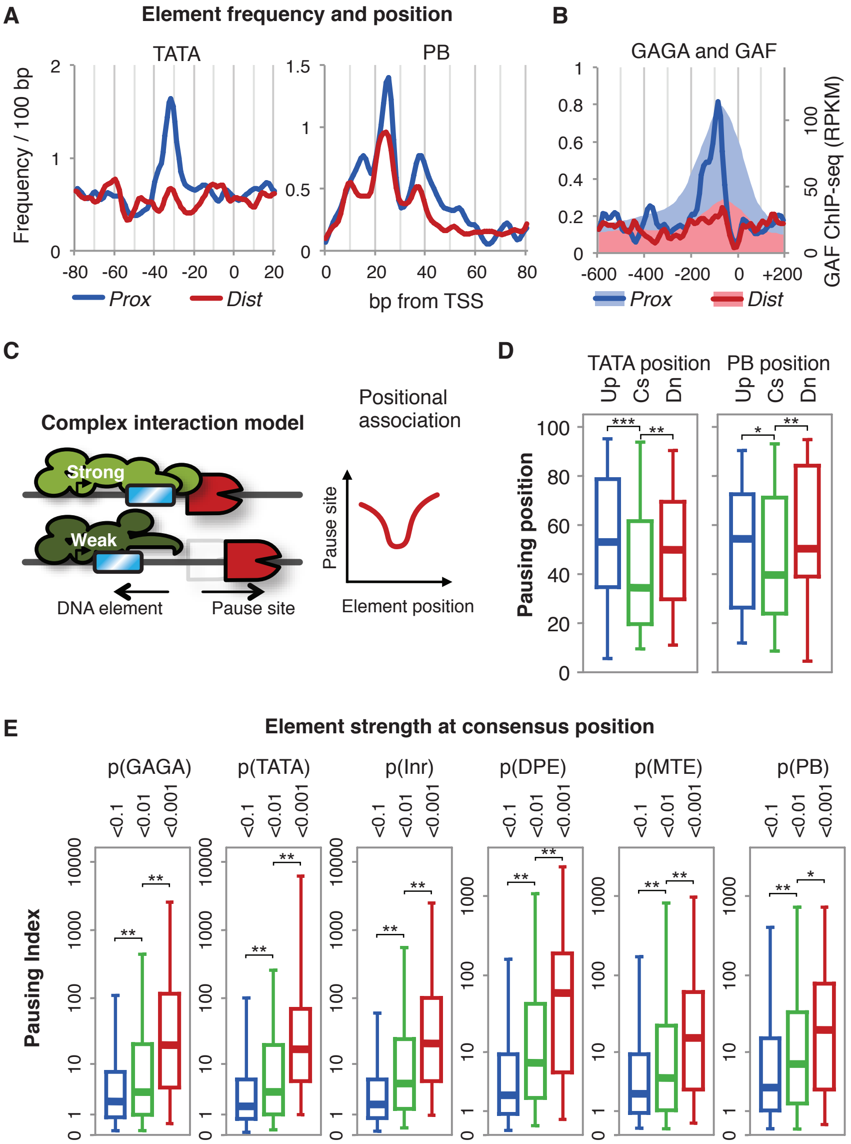


Figure 2.7. Relationship between promoter DNA elements and Pol II pausing. (**A**) Frequency of TATA box and Pause Button (PB) in *Prox* and *Dist* subsets. The average frequency per gene is shown. (**B**) Frequency of GAGA element (lines) and GAF binding (shades) in *Prox* and *Dist* subsets. (**C**) 'Complex interaction model' between DNA elements and paused Pol II. The DNA elements (blue) are at their consensus (Strong) or slightly upstream (Weak) positions and the expected changes of the pausing positions are plotted. (**D**) Pattern of the positional association between DNA elements and pausing positions. Pausing position percentiles are shown for gene subsets by the element positions (Cs: optimal consensus position, Up: upstream, Dn: downstream). Asterisks indicate p-values for the KS test(***: p<0.002, **: p<0.06, *: p<0.14). (**E**) Association of promoter DNA element strength at consensus positions with Pausing Index*(6)*. Active genes (n=5,471) are divided into three subsets according to the distance-weighted p-values of the DNA elements to the consensus positions. Asterisks indicate p-values for the KS test

Genes with TATA at around -30 have more proximal pausing than the genes with TATA at positions of -40 or more, showing a 'V'-shaped association (Figure 2.7D). This 'V' pattern was observed in both the upstream elements TATA and Inr (Figure 2.8A), and the downstream elements PB (Figure 2.7D) and MTE (Figure 2.8B). Also, pausing tends to be stronger in genes with the elements at the optimal positions (Figure 2.8D). Furthermore, the extent of pausing shows strong dependency on the match of the DNA elements to their consensus sequence and consensus positions (data not shown). Together, these association patterns between core promoter elements and pausing support the 'Complex Interaction' model, and explain the strong and focused pausing on *Prox* promoters.

Testing the Positional Association of the DNA Elements in Hsp70

The 'Complex Interaction' model depends on both the presence and the correct positioning of core promoter elements. We disrupted the positional relationship of core elements in the well-studied *Drosophila* gene *Hsp70*[1]. Transgenic fly lines were generated that carry mutant *Hsp70* promoters with spacers inserted at the +15 position between the upstream and downstream promoter elements (Figure 2.9A), and analyzed by PRO-seq. The initiation sites remain constant in these mutant promoters, indicated by the 5' ends of the PRO-seq reads (Figure 2.9B). The transgenic *Hsp70* without spacers inserted shows a strong pause peak mainly at +31 (Figure 2.9C). When 5 bp are inserted, the pause peak is shifted 5-7 bp downstream from the original site. The additional bases transcribed before pausing again demonstrates that the position of pausing is not predetermined by elongation distance. When 10 bp are inserted, pausing sites become scattered between +20 to +60 (Figure 2.9D) and have fewer reads (Figure 2.9C).

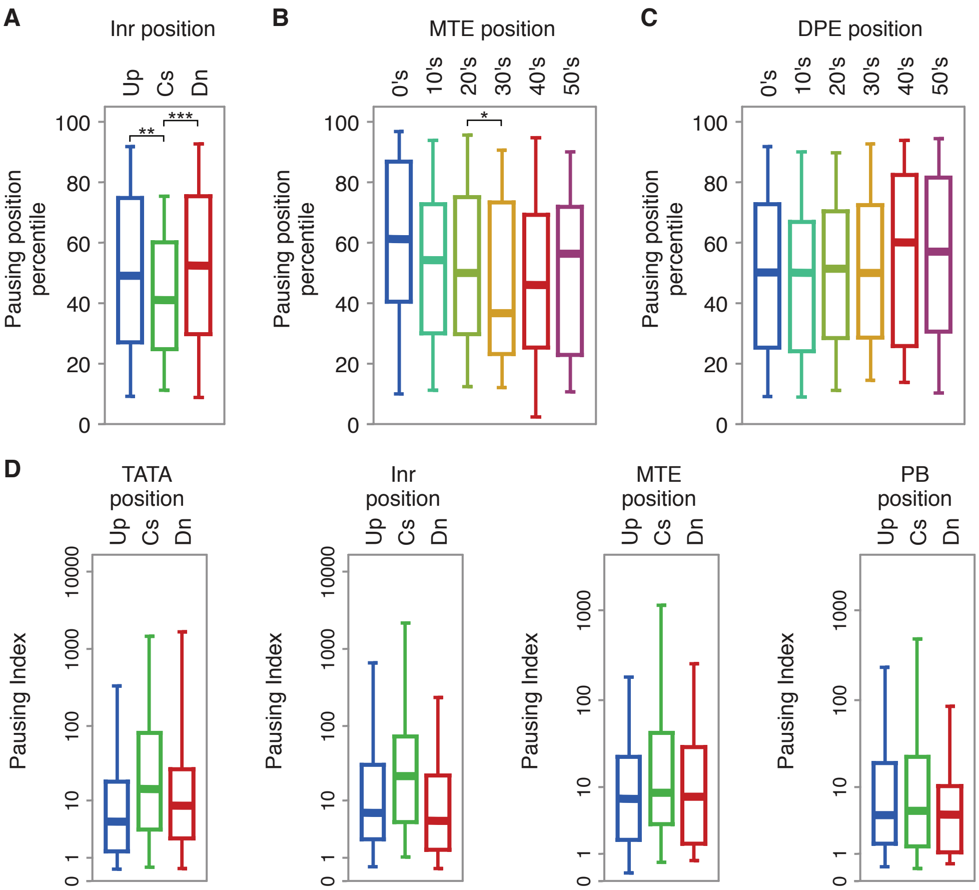


Figure 2.8. Association between the strength or the position of promoter DNA elements and Pol II pausing or initiation (A) Pattern of positional association between Inr and pausing. Inr positions are -24 to -4 from TSS for upstream genes (Up, n=111), -4 to +1 from TSS for consensus positioned genes (Cs, n=131), +1 to +21 for downstream genes (Dn, n=110). (B) Pattern of positional association between MTE and pausing. Genes were divided based on the position of MTE in bins of 10 bp from TSS to +60 (n=56, 90, 70, 64, 52, 51). (C) Pattern of positional association between DPE and pausing. Genes were divided based on the position of DPE in bins of 10 bp from TSS to +60 (n=148, 102, 142, 105, 101, 97). Asterisks indicate p-values for the Kolmogorov-smirnov (KS) test (*** - p<0.0022, ** - p<0.021, * - p<0.066). (D) Pausing indices in genes with DNA elements at Up, Cs, and Dn positions. TATA and PB gene subsets are the same sets used in Figure. 2.8E; Inr subsets are the same sets used above; for MTE subsets, subsets containing MTE at 20's, 30's, and 40's were used for Up, Cs, and Dn subsets respectively.

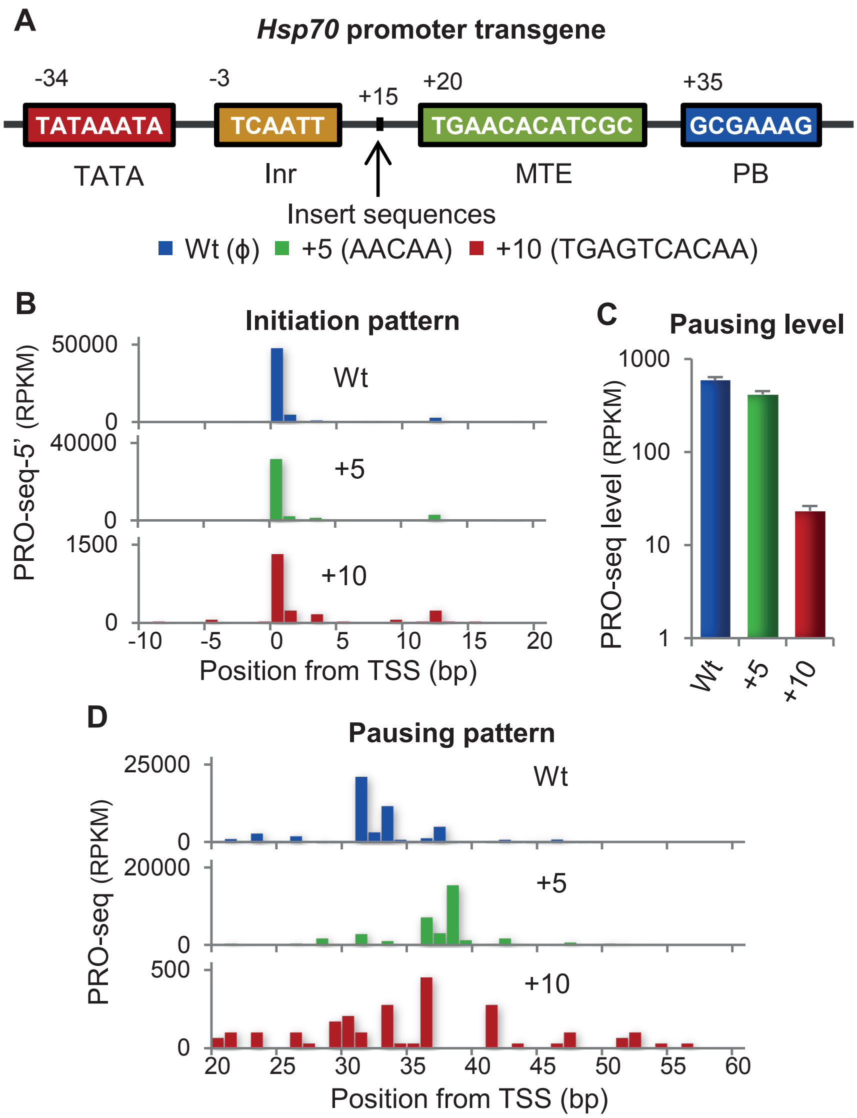


Figure 2.9. Disruption in the position of downstream DNA elements in *Hsp70* promoter. (**A**) Structure of the *Drosophila Hsp70* promoter. Mutant promoters have different length inserts at +15. (**B**) Initiation patterns at transgenic *Hsp70* promoters. 5' end counts of PRO-seq reads in each transgenic adult fly line are shown to compare the position and pattern of initiation. (**C**) Level of pausing in transgenic *Hsp70* promoters. The sum of the read counts within the pausing region is normalized to the total mapped reads. (**D**) Positions of pausing in transgenic *Hsp70* promoters.

Collectively, these results support the core interaction model and suggest that the interaction complex can accommodate a small change (5 bp) in the positional context of the DNA sequences, but a larger change (10 bp) results in reduced and dispersed pausing.

Conclusion

The advances in resolution provided by PRO-seq enabled the precise and genome-wide assessment of the relationship between promoter-proximal pausing and the core promoter structure. For the strong and tightly clustered pausing of the *Prox* genes, we provide support for a 'Complex Interaction' model involving the promoter initiation complex which can extend up to 30 bp from the TSS[28], physically contacting and tethering the pausing complexes. This may share a kinship with bacterial initiation factor σ that is retained within the early elongation complex and interacts with promoter proximal DNA during transcription pausing in *E. coli*[29]. Interestingly the *Prox* genes are expressed on average at a lower level but show a broader range of expression, and the *Dist* genes are enriched in constitutively active genes (Table 2.2). These results suggest that the mechanistic distinctions have regulatory consequences. A well-structured core promoter may strongly recruit Pol II; however, it can also effectively retain Pol II in a paused configuration close to the TSS, until activation signals allow its escape into productive elongation.

Table 2.2. Enrichment of *Dist* pausing group among constitutive genes. A list of constitutive genes from a developmental transcriptome study[30] was examined for the number of genes in each *Prox* and *Dist* pausing groups. As defined in the original study, genes with short poly(A)+ RNA-seq levels greater than 1.0 FPKM in all 30 developmental conditions were considered constitutively active[30]. In addition, genes that have significant PRO-seq reads from upstream genes were removed and only the 'upstream clear' genes were selected as the 'active' genes. Expected counts of the constitutive genes are derived from the proportions of each group in all genes. Asterisk (*) indicate p-value < 0.0043 by χ^2 test.

	All genes	Constitutive genes (expected count)	
Active	5,471	3,557	
Paused	3,225	2,022 (2,097)	
Prox	848	492 (532)	] *
Dist	846	583 (530)	

Materials and Methods

Cell culture and nuclei isolation

Drosophila S2 cells were maintained in Shields and Sang M3 insect medium supplemented with 10% Fetal Bovine Serum, Bacto-Peptone and Yeast Extract at 30°C. At 16~20 passages, nuclei were isolated as described previously with modifications[6,20,21]. All temperatures were at 4°C or ice cold unless otherwise specified. Briefly, cells were washed in PBS and resuspended in Buffer S (10 mM Tris-HCl pH 7.5, 10% glycerol, 3 mM $CaCl_2$, 2 mM $MgCl_2$, 0.5 mM DTT, protease inhibitors cocktail (Roche), 4 u/ml RNase inhibitor (SUPERaseIN, Ambion) at the cell density of 2×10^7 cells/ml. After 5 min of incubation, 9× volume of Buffer L was added and immediately homogenized using a tight fitting pestle until over 90% of the nuclei were released. Nuclei were fractionated by centrifugation at 1000 g for 4 min and recovered from the pellet fraction. Recovered nuclei were washed twice in Buffer

L and once in Buffer D (50 mM Tris-HCl pH 8.0, 25% glycerol, 5 mM MgAcetate$_2$, 0.1 mM EDTA, 5 mM DTT). Washed nuclei were finally resuspended in Buffer D at a density of 2×10^7 nuclei/100 µl) and immediately frozen in liquid nitrogen. Nuclei were stored in -80˚C until usage.

PRO-seq and PRO-cap library preparations

Four parallel run-on reactions of PRO-seq$_{ATP}$, PRO-seq$_{CTP}$, PRO-seq$_{GTP}$ and PRO-seq$_{UTP}$ were carried out as follows. 2×10^7 nuclei were added to the same volume of 2× Nuclear Run-On (NRO) reaction mixture (10 mM Tris-HCl pH 8.0, 300 mM KCl, 1% Sarkosyl, 5 mM MgCl$_2$, 1 mM DTT, 500 µM biotin-11-A/C/G/UTP (Perkin-Elmer), 0.8 u/µl RNase inhibitor) and incubated for 3 min at 30˚C. Alternatively, 375 µM of each of all 4 biotin-11-NTPs were supplemented in the reaction for an abbreviated protocol (PRO-seq$_{4NTP}$) or PRO-cap. Nascent RNA was extracted using Trizol and precipitated in 75% ethanol. Extracted nascent RNA was fragmented by base hydrolysis in 0.2 N NaOH on ice for 10~12 min, and neutralized by adding 1× volume of 1 M Tris-HCl pH 6.8. For PRO-cap, the fragmentation step was omitted. Excessive salt and residual NTPs were removed by using P-30 column (Bio-rad). Fragmented nascent RNA was bound to 30 µl of Streptavidin M-280 magnetic beads (Invitrogen) following the manufacturer's instructions. The beads were washed once in high salt (2 M NaCl, 50 mM Tris-HCl pH 7.4, 0.5% Triton X-100), once in medium salt (300 mM NaCl, 10 mM Tris-HCl pH 7.4, 0.1% Triton X-100), and once in low salt (5 mM Tris-HCl pH 7.4, 0.1% Triton X-100). Bound RNA was extracted from the bead using Trizol (Invitrogen) in two consecutive extractions, and the RNA fractions were pooled, followed by ethanol precipitation.

For the first ligation reaction, fragmented nascent RNA was redissolved in H$_2$O and incubated with 10 pmol of reverse 3' RNA adaptor (5'p-rGrArUrCrGrUrCrG rGrArCrUrGrUrArGrArArCrUrCrUrGrArArC-/3'InvdT/) and T4 RNA ligase I (NEB) under manufacturer's condition for 6 hr at 20˚C. For PRO-cap, the standard 3' RNA adaptor (Illumina) was used. Ligated RNA was enriched with biotin-labeled products by another round of Streptavidin bead binding and extraction. To repair 5' ends, the RNA products were treated with Tobacco Acid Pyrophosphatase (TAP, Epicentre) and Polynucleotide Kinase (PNK, NEB). Each reaction was followed by an ethanol precipitation step. For PRO-cap, PNK treatment step was omitted, and Antarctic phosphatase (AP, NEB) was used to treat the RNA preparation prior to TAP treatment to enrich for 5' capped RNA. Since these procedures repair 5' ends after the 3' ligation, self-circularized products were not expected during the first ligation step. 5' repaired RNA was ligated to reverse 5' RNA adaptor (5'-rCrUrGrArArCrArArGrC rArGrArArGrArCrGrGrCrArUrArCrGrA-3' or 5'-rCrCrUrUrGrGrCrArCrCrCrGrArG rArArUrUrCrCrA-3' for using TruSeq barcodes (Illumina)). Standard 5' RNA adaptors were used for PRO-cap. Ligated RNA products were further enriched for biotin-labels by the third round of streptavidin bead binding and extraction. Adaptor ligated nascent RNA was reverse transcribed using 25 pmol RT primer (5'-AATGAT ACGGCGACCACCGACAGGTTCAGAGTTCTACAGTCCGA-3'(GX2 primer, Illumina) or 5'- AATGATACGGCGACCACCGAGATCTACACGTTCAGAGTTCT ACAGTCCGA-3' for TRU-seq barcodes (RP1 primer, Illumina). Standard Illumina RT primers were used for PRO-cap.

A portion of the RT product was removed and used for trial amplifications to determine the optimal number of PCR cycles. For the final amplification, 12.5 pmol of GX1 primer (Illumina) or RPI-index primers (for TRU-seq barcodes, Illumina) was added to the RT product with Phusion polymerase (NEB) under standard PCR condition. Excess RT primer served as one primer of the pair used for the PCR. The product was amplified 12~18 cycles and PAGE purified before being analyzed by Illumina's GenomeAnalyzer 2 or HiSeq 2000 machines.

Processing raw sequence data for polymerase active site mapping

Raw sequences were preprocessed using FASTX-Toolkit. Adaptor sequences were removed from the raw sequences using 'fastx_clipper', and the first 26 bases were trimmed with 'fastx_trimmer'. Sequence reads shorter than 16 bases were removed. The first bases, which were the reverse complements of 3' end bases, were counted for each library to verify that the 3' ends represent the polymerase active sites (Table 2.1). Indeed, the majority of the sequences had the same 3' end base as the biotin-NTP that was added in the run-on reaction, indicating that the identified sequences define the 3' ends that are exactly at or near the Pol II active sites.

Reverse complements of the sequence reads, which were the sense sequences of nascent RNA, were generated using 'fastx_reverse_complement'. Each of the 4 biotin-NTP libraries was aligned to the *Drosophila melanogaster* (Dm3) reference genome using Bowtie[31] allowing 2 mismatches and excluding any non-uniquely aligned reads. The histograms of the 3' end positions in base pair resolution were generated in the 'bedgraph' format.

For the normalization of the 4 biotin-NTP libraries to generate a composite profile, we first assumed that the probability of finding polymerase on difference bases in the bodies of the genes (GB) were uniform. Under this assumption, a normalization factor should be multiplied to a library such that the sum of the normalized reads mapped to the gene body divided by the corresponding base counts in the gene body regions becomes uniform throughout different base libraries. The normalization factor for each library is calculated as follows,

$$Normalization\ factor_{base} = \sum_{i=A,C,G,U} \frac{Reads\ mapped\ to\ GB_i}{Base\ counts\ in\ GB_i} \Big/ 4\frac{Reads\ mapped\ to\ GB_{base}}{Base\ counts\ in\ GB_{base}}$$

, where *GB* is the set of all gene body positions (table S4).

Using the normalization factors, composite PRO-seq histogram (in bedgraph format) was generated.

$$PROseq(pos) = \sum_{base=A,C,G,U} Normalization\ factor_{base} \cdot PROseq_{base}(pos)$$

This composite PRO-seq dataset was used for the downstream analysis unless specified otherwise.

Analysis of pausing level and gene activity

For the analysis of pausing level and gene activity, we first generated a list of genes for which the PRO-seq densities could be measured without having interference from other genes. From the scRNA-seq based re-annotated gene list[16], we defined promoter upstream, promoter downstream, and 5' genic regions as -300 to -100 bp, +300 to +500 bp, and 0 to +500 bp from TSS respectively. For each region of the individual genes, we calculated 'active site coverage', which is the fraction of positions covered by 3' end of PRO-seq reads within each region. Because of the normalization, some positions have read counts less than 1 and we considered these positions partially covered. Active site coverage can be formulated as follows.

$$Active\ site\ coverage(region) = \sum_{pos \in region} \max\left(PROseq(pos), 1\right) \Big/ length\ of\ the\ region$$

We called genes 'upstream clear' if promoter upstream region had the active site coverage of less than 0.01, or less than one fourth of the downstream region active site coverage (n=11,584). This was intended to filter out genes that have polymerase transcribing through from the upstream genes that can interfere with the downstream levels. Among the ‘upstream clear’ genes, we called genes ‘active’ if the active site coverage in 5’ genic regions was greater than 0.01 (n=5,471).

To calculate the pausing level, PRO-seq read counts per million normalized mapped reads (RPM) from -50 to +150 relative to TSS were obtained and the sum of the read count was divided by the length of the region (0.2 kb) to generate RPM per kb, or RPKM which equivalent to the commonly used definition of RPKM in RNA-seq. For the consistency of the unit usage, we also used RPKM to describe PRO-seq profiles for individual genes along the positions on the genome, regarding that a RPM read count on a single base position can be considered as a RPKM density for a 0.001 kb region.

For the gene body activity, we used active site coverage instead of read counts to minimize the effects of unexpected spikes or unannotated transcription initiation within the gene body region. This modified PRO-seq density was calculated by multiplying a conversion factor to the active site coverage from +300 from the TSS to the 3’ end of the gene and RPKM normalized. For genes that contain another annotated TSS within the gene body, we truncated their gene body region to -300 bp from any downstream annotated gene starts. The conversion factor is given below.

$$Conversion\, factor = \frac{\sum_{pos \in GB} PROseq(pos)}{Active\, site\, coverage(GB) \cdot length\, of\, GB}$$

All the densities were adjusted by the mappability of 26 bp sequence uniquely to the genome.

Data visualization using scatterplots, average profiles and heatmaps

The scatterplots were generated using in-house scripts. Briefly, on a 1,000×1,000 pixel bitmap, each data-point was represented as filler circles with 11 pixel diameter on log axes. For each pixel, data-point counts were stackable. After plotting all the data-points, the counts in each pixel was converted to a color code, and the image was anti-aliased. The scale-bars for the color code were shown together with the colored scatterplot images. Pearson coefficients were presented on the scatterplots

Average profiles relative to position lists were generated using modified bootstrap methods and permutation tests. Briefly, position lists of N genes were randomly partitioned into [N/100] subsets (integer part of N/100) each containing ~100 members. The average profile of each subset was calculated removing 2 greatest and least outliers per relative positions. The average and the standard error of the subset profiles were calculated and usually plotted together respectively as a line and margins surrounding the line in shades. Gaussian smoothing was applied to the profiles if necessary using the formula below,

$$smoothed\,profile(pos) = \sum_{i=-3b/2}^{3b/2} \phi(2i/b)\cdot raw\,profile(pos+i) \Bigg/ \sum_{i=-3b/2}^{3b/2} \phi(2i/b)$$

where $\phi(x)=exp(-x^2/2)$ is the Gaussian density function, and b is the bandwidth of smoothing which is twice the standard deviation of regular Gaussian distribution. A data value is smoothed over 3 bandwidths around the data point. For most profiles, smoothing bandwidth of 2 bp was used unless specified otherwise.

Scaled ‘metagene’ profile was generated as described previously[20], with modification to the scaled region of the gene body being TSS +1 kb to 3’ end −1 kb of the gene. A smoothing bandwidth of 200 bp was used.

Heatmaps were generated using in-house scripts. Briefly, a data matrix of PRO-seq read counts, with genes on the rows and relative position to each TSS on the columns, was scaled to a 200×1,000 matrix with an algorithm that uses incremental accumulators for each pixel. The data values were converted to color codes and the image was anti-aliased. Typically, this generated moderately averaged profiles for gene lists containing more than 10,000 genes, but represented individual genes relatively well for comparing gene subsets containing up to 1,000 genes.

Pause peak identification and paused gene clustering by pausing pattern

With the 'active' genes (n=5,471) listed above, we defined PRO-seq peaks using a clustering algorithm. Briefly, for each gene, we scaled the number of reads to 1,000 pseudo-reads maintaining their relative positions in the promoter proximal region (-50 to +150 from TSS), and applied a k-means clustering algorithm by their positions to identify the peaks. The number of clusters (k), *i.e.* the number of peaks, was determined by taking the minimum k for which the variance to the cluster centroid was less than 5 (bp×bp). The k value was modified within ±1 range to have the local maximum of the average silhouette. For each peak, the total read count of the actual reads was calculated, and major peaks greater than one fourth of the maximum peak of the region were selected. Each peak is assigned with two parameters, average position and total read count. We repeated the same peak calling algorithm in promoter downstream regions (+300 to +500), and called a gene 'paused' if the total read count of the maximum peak at promoter proximal region is greater than 4 times the read count of the maximum peak at the promoter downstream region (n=3,225). These cut-offs are chosen for the purpose of relative comparison between groups, but they do not necessary define pausing *per se*[6]. For each paused gene, we calculated the median position of the peaks and the average dispersion of the peaks weighted by the read counts. We calculated the percent rank of the median position and the average dispersion within the paused genes subset and defined them as ‘position percentile’ and ‘dispersion percentile’ of the peaks respectively for each gene.

For the 2D heatmap representation of the pausing pattern on the position-dispersion axes, we used an in-house script (Fig. 2.4B). Briefly, each gene was added as a 2D Gaussian peak,

$$\phi(x,y)=\frac{1}{2\pi(b/2)^2}e^{-(x^2+y^2)/2(b/2)^2}$$

where *x* and *y* are the relative position to the position vs dispersion percentile coordinate of the gene on a 2D space [0,1]×[0,1], and *b* is the bandwidth of the peak (20%). The overall density on the 2D space was normalized by dividing by the total gene number. Therefore, the integral over an area reflects the probability of finding a gene in the corresponding position-dispersion range, and the integral over the whole 2D space, which is the probability of finding a gene over the whole region, equals 1. To further identify the two apparent clusters of the genes- ‘Clustered proximal’ (*Prox*) on the lower left quadrant and ‘Dispersed distal’ (*Dist*) on the upper right quadrant- in Fig. 2B, we employed an Expectation-Maximization algorithm. This was done using the ‘Mclust’ package in R software. Briefly, we performed Mclust on the position-dispersion dataset with the prior specification of 2 clusters with ellipsoidal model (variable volume, shape, and orientation of covariance matrix: ‘VVV’ model), and initialization of a Poisson noise model (p=0.25). This was done iteratively and a representative set was chosen. Two clusters were generated allowing outliers, and we determined the cut-offs for their z-scores (*Prox*: 0.15, *Dist*: 0.08) to have maximum number of non-overlapping elements and least difference in cluster sizes. Genes were assigned to *Prox* or *Dist* clusters if their z-scores were smaller than the cut-offs (n=848, 846 respectively). The average profiles show similar pausing distributions as anticipated by individual cases. Since the majority of the genes lie on the diagonal of the position-dispersion space, we defined the Pausing Proximity Index (PPI) as the average of the pausing position percentile and the pausing dispersion percentile.

Analysis of the initiation from PRO-cap

PRO-cap results were processed in the same way as PRO-seq, except that the 5' ends of non-reverse complemented sequence reads were used and the promoter proximal window was set to be -100 to +100 from the TSS. Average relative profile was generated by first dividing the read counts by total number of reads in the promoter proximal window for each gene, and then calculating the average plots afterwards. This allowed us to examine the average pattern of initiation at the TSS without having the pattern be overly affected by genes with the highest read counts.

To compare the dispersion pattern of initiation and pausing, we used the identical script to identify the initiation peaks. For a direct comparison, the dispersions were shown in the actual number of base pairs instead of the percentiles in boxplots.

To assess the focusing of initiation, we defined the 'read fraction at TSS' parameter (frTSS) for each gene as described previously with modifications[22].

$$frTSS = \frac{\sum_{pos \in TSS \pm 1bp} PROcap(pos)}{\sum_{pos \in TSS \pm 25bp} PROcap(pos)}$$

Scoring the positions and the strengths of the DNA elements

Promoter DNA elements were identified from promoter proximal regions using existing position weight matrices (PWM) or consensus sequences by a fast permutated string-match algorithm. First we extracted a sequence substring from the promoter proximal sequence of each gene on every position, and calculated the PWM score. The score was compared to a cumulative distribution function (CDF) of the scores of 100,000 permutated sequences that were randomly generated using the same background letter frequencies. From the CDF, p-values were obtained for every position on the promoter proximal region of a gene. When the PWM was not available, we built a PWM from the log-likelihood of the consensus match at matched letters and 0 at non-matched letters.

We identified the positions and the strengths of DNA elements by generating a log likelihood array (-log p-value) for each gene at every position (Figure 2.6B-D). The array was smoothed with a modified Gaussian smoothing,

$$Smoothed \log L(pos) = \max\left(\phi(2i/b)\cdot \log L(pos+i); -\tfrac{3}{2}b < i < \tfrac{3}{2}b\right)$$

where *Smoothed* log$L$(pos) is the smoothed profile for relative position *pos*, log$L$(pos) is the log likelihood profile at position *pos*, *b* is the bandwidth of the Gaussian curve for smoothing, and $\phi(x)=exp(-x^2/2)$ is the Gaussian density function. This type of smoothing reports the strongest element nearby that is modified by a distance factor dependent on the bandwidth, and reflects the probability of finding a factor binding at the position within the bandwidth. A bandwidth of 10 bp was used for core promoter elements and 40 bp for GAGA elements.

Generation of the fly lines with sequence modified Hsp70 promoter transgenes (This part was done by Dr. Nick Fuda)

Adult flies carrying sequence modified *Hsp70* promoter transgenes (Figure 2-9A) were made as described below. First, the *Hsp70* gene was amplified and cloned into pBluescriptII KS+. The 5 bp insert and 10 bp insert transgene were created using site-directed mutagenesis with the following primer sets; +5 bp forward: 5’-CGACGG AGAGTCAATTCAATTCAAACAAAACAAGCAAAGTGAACACATCG C-3’; +5 bp reverse: 5’-GCGATGTGTTCACTTTGCTTGTTTTGTTTGAATTGAATTGACT CTCCGTCG-3’; +10 bp forward: 5’-CGACGGAGAGTCAATTCAATTCAAACAA

TGAGTCACAAGCAAAGTGAACACATCGC-3'; +10 bp reverse: 5'-GCGATGTG TTCACTTTGCTTGTGACTCATTGTTTGAATTGAATTGACTCTCCGTCG-3'.

The gene was cut out of the wild type (*Hsp70wt*), 5 bp insert (*Hsp70+5*), and 10 bp insert (*Hsp70+10*) plasmids with XbaI to yield a fragment from -245 to +1863 (relative to the TSS), and cloned into a modified pCasper4 containing the *attB* site[32]. The ΦC31-mediated transformation was performed by Best Gene Inc. to insert each transgene into 22A3 (*PBac{yellow[+]-attP-3B}VK00037*). The lines were balanced and crossed to *Hsp70* null flies[33] (Bloomington 8841: *w1118*; *Df(3R)Hsp70A*, *Df(3R)Hsp70B*) to create homogenous stocks.

PRO-seq and the analysis of transgenic Hsp70 promoter fly lines

The nuclear isolation from adult flies was adapted as described previously[27]. One gram of flies were homogenized in 15 ml cold Buffer A (10 mM Tris-HCl pH 8.0, 30 0mM sucrose, 3 mM $CaCl_2$, 2 mM $MgOAc_2$, 0.1% TritonX-100, 0.5 mM DTT) for 1 minute using the Omni-mixer, the homogenate was filtered through 100 um nylon mesh into a 40 ml Dounce homogenizer. After 40 strokes in the homogenizer, the homogenate was filtered through 35 um nylon mesh and mixed with an equal volume of Buffer B (10 mM Tris-HCl pH 8.0, 2 M sucrose, 5 mM $MgOAc_2$, 0.5 mM DTT). The homogenate was then layered over 10 ml Buffer B in 35 ml Ultracentrifuge tube, and centrifuged through the Buffer B cushion at 12 krpm for 25 minutes at 4˚C in a SW28 swinging bucket rotor. The supernatants were removed, and the nuclei were resuspended to the density of 1×10^8 nuclei/ml in 1 ml buffer C (50 mM Tris-HCl pH 8.0, 25% glycerol, 5 mM $MgOAc_2$, 0.1 mM EDTA, 5 mM DTT).

REFERENCES

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

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

3. Lee, C. *et al.* NELF and GAGA factor are linked to promoter-proximal pausing at many genes in Drosophila. *Mol Cell Biol* **28,** 3290–3300 (2008).

4. Muse, G. W. *et al.* RNA polymerase is poised for activation across the genome. *Nat Genet* **39,** 1507–1511 (2007).

5. Zeitlinger, J. *et al.* RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo. *Nat Genet* **39,** 1512–1516 (2007).

6. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. *Science* **322,** 1845–1848 (2008).

7. Hah, N. N. *et al.* A Rapid, Extensive, and Transient Transcriptional Response to Estrogen Signaling in Breast Cancer Cells. *Cell* **145,** 13–13 (2011).

8. Min, I. M. *et al.* Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells. *Genes Dev* **25,** 742–754 (2011).

9. Core, L. J. & Lis, J. T. Transcription regulation through promoter-proximal pausing of RNA polymerase II. *Science* **319,** 1791–1792 (2008).

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

11. Kornblihtt, A. R. A. Chromatin, transcript elongation and alternative splicing. *Nat Struct Mol Biol* **13,** 5–7 (2006).

12. Saunders, A., Core, L. J. & Lis, J. T. Breaking barriers to transcription elongation. *Nat. Rev. Mol. Cell Biol.* **7,** 557–567 (2006).

13. Wu, C.-H. *et al.* NELF and DSIF cause promoter proximal pausing on the hsp70 promoter in Drosophila. *Genes Dev* **17,** 1402–1414 (2003).

14. Wu, C. H. *et al.* Analysis of core promoter sequences located downstream from the TATA element in the hsp70 promoter from Drosophila melanogaster. *Mol Cell Biol* **21,** 1593–1602 (2001).

15. Hendrix, D. A., Hong, J.-W., Zeitlinger, J., Rokhsar, D. S. & Levine, M. S. Promoter elements associated with RNA Pol II stalling in the Drosophila embryo. *Proc Natl Acad Sci U S A* **105,** 7762–7767 (2008).

16. Nechaev, S. *et al.* Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila. *Science* **327,** 335–338 (2010).

17. Mavrich, T. N. T. *et al.* Nucleosome organization in the Drosophila genome. *Nature* **453,** 358–362 (2008).

18. Nechaev, S. & Adelman, K. Pol II waiting in the starting gates: Regulating the transition from transcription initiation into productive elongation. *Biochim. Biophys. Acta* **1809,** 34–45 (2011).

19. Churchman, L. S. & Weissman, J. S. Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature* **469,** 368–373 (2011).

20. Larschan, E. *et al.* X chromosome dosage compensation via enhanced transcriptional elongation in Drosophila. *Nature* **471,** 115–118 (2011).

21. Core, L. J. *et al.* Defining the status of RNA polymerase at promoters. *Cell Rep* **2,** 1025–1035 (2012).

22. Gilchrist, D. A. D. *et al.* Pausing of RNA Polymerase II Disrupts DNA-Specified Nucleosome Organization to Enable Precise Gene Regulation. *Cell* **143,** 12–12 (2010).

23. Hall, M. A. *et al.* High-resolution dynamic mapping of histone-DNA interactions in a nucleosome. *Nat Struct Mol Biol* **16,** 124–129 (2009).

24. Bondarenko, V. A. *et al.* Nucleosomes can form a polar barrier to transcript elongation by RNA polymerase II. *Mol Cell* **24,** 469–479 (2006).

25. Juven-Gershon, T., Hsu, J.-Y., Theisen, J. W. & Kadonaga, J. T. The RNA polymerase II core promoter - the gateway to transcription. *Current Opinion in Cell Biology* **20,** 253–259 (2008).

26. Stark, A. *et al.* Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. *Nature* **450,** 219–232 (2007).

27. Lee, H., Kraus, K. W., Wolfner, M. F. & Lis, J. T. DNA sequence requirements for generating paused polymerase at the start of hsp70. *Genes Dev* **6,** 284–295 (1992).

28. Emanuel, P. A. & Gilmour, D. S. Transcription factor TFIID recognizes DNA sequences downstream of the TATA element in the Hsp70 heat shock gene. *Proc Natl Acad Sci U S A* **90,** 8449–8453 (1993).

29. Ring, B. Z., Yarnell, W. S. & Roberts, J. W. Function of E. coli RNA polymerase sigma factor sigma 70 in promoter-proximal pausing. *Cell* **86,** 485–493 (1996).

30. Graveley, B. R. *et al.* The developmental transcriptome of Drosophila melanogaster. *Nature* **471,** 473–479 (2011).

31. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).

32. Maheshwari, S. & Barbash, D. A. Cis-by-Trans regulatory divergence causes the asymmetric lethal effects of an ancestral hybrid incompatibility gene. *PLoS Genet.* **8,** e1002597 (2012).

33. Gong, W. J. & Golic, K. G. Genomic deletions of the Drosophila melanogaster Hsp70 genes. *Genetics* **168,** 1467–1476 (2004).

CHAPTER 3[c]

RNA POLYMERASE II PAUSING AND TERMINATION KINETICS AT THE DROSOPHILA *HSP70* PROMOTER IN VIVO

Summary

Many metazoan promoters have a high occupancy of transcriptionally engaged RNA polymerase II (Pol II) relative to their gene bodies[1-5]. Although it is widely accepted that Pol II is paused near promoters, this accumulation could also be a consequence of polymerases that are rapidly initiating and then terminating prematurely similar to some prokaryotic and viral promoters[6,7]. Notably, in addition to pausing factors, termination factors are also enriched at metazoan promoters, and the depletion of each can alter the distribution of Pol II[8]. These findings raise the possibility that polymerases might undergo cycles of initiation and non-productive elongation. However, current Pol II occupancy measurements are insufficient to define the fate of promoter proximal Pol II *in vivo*. Here we show the kinetic status of the promoter proximal Pol II through a combination of complementing optical and biochemical strategies in *Drosophila*. By tracking the fate of photo-activatable-GFP tagged Pol II at the *Hsp70* locus on polytene chromosomes, we found that Pol II is stably paused and has a half-life of 5 minutes on the uninduced gene. Biochemical analysis of short nascent RNA from *Hsp70* showed that this half-life is composed of Pol II that has the both probabilities of escaping into productive elongation and non-productively terminating. We further confirm this termination rate using optical

[c] Works in this chapter was done in collaboration with Dr. Martin S. Buckley. Extents of contributions are noted in the figure legends.

measurements after blocking the productive elongation with a chemical inhibitor, Flavopiridol[9]. These findings show that the competing fates of paused Pol II to termination and to escape into productive elongation can modulate transcription levels.

Optical Measurement of Pol II Stability in Hsp70

The *Drosophila Hsp70* heat shock gene possesses a promoter-proximal Pol II that has been extensively characterized[10]. Notably, many of the features of the paused Pol II at uninduced *Hsp70* are similar to the large number (70% of active Drosophila genes) containing paused Pol II[11]. In addition, many of the proteins identified to be involved in *Hsp70* gene regulation have corresponding activities at other genes in various organisms[10]. These findings indicate that the mechanisms governing *Hsp70* gene regulation are general. Therefore, in order to gain insights into the status of the promoter-proximal paused Pol II, we focused our efforts on measuring the stability of paused Pol II at *Drosophila Hsp70* (Figure 3.1).

To address this, we first made use of an optical approach to measure the stability of photoactivatable GFP (paGFP)-tagged Pol II at the *Hsp70* locus on *Drosophila* salivary gland polytene chromosomes (Figure 3.2A). These interphase-like giant chromosomes have been used as a platform for high-resolution imaging[12]. Dynamics of Pol II and other transcription factors can be measured by optical pulse-chase experiments such as fluorescence recovery after photobleaching (FRAP) at induced *Hsp70* gene loci (87A and 87C) in living cells[13-15]. These FRAP studies rely on the fact that the activated 87A/C heat shock (HS) loci can be unambiguously localized using a distinctive puff doublet pattern formed by fluorescently labeled Pol II subunits such as eGFP-Rpb3[13-15]. However, under basal conditions, identifying the *Hsp70* gene loci is technically challenging, since paused eGFP-Rpb3 signal from

Hsp70 is too weak to reliably identify from other bands and perform photobleaching analysis.

To circumvent this problem, we engineered a transgenic fly line that contains the full length *Hsp70* gene marked with 256 repeats of Lac operator sites (LacO), that when bound by the mCherry tagged-Lac repressor (mCherry-LacI), allows us to easily identify the uninduced *Hsp70* loci in salivary gland nuclei (Figure 3.2A).

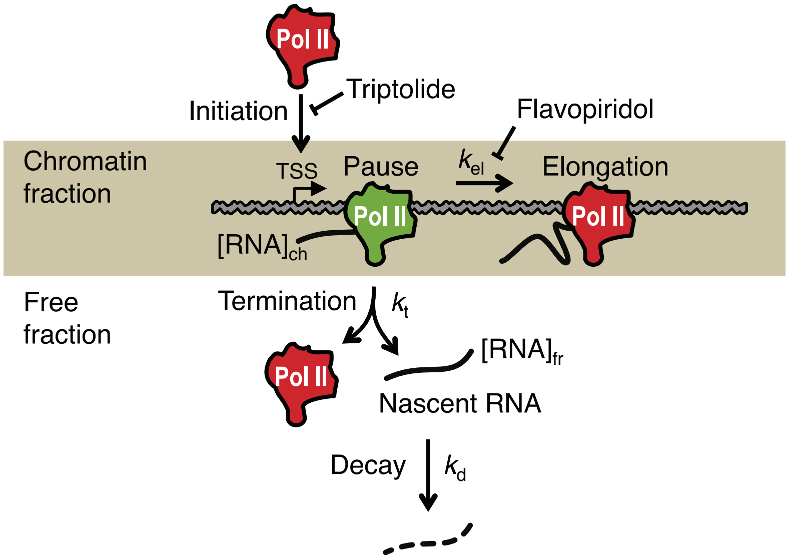


Figure 3.1. Schematics of Pol II kinetics at the pause site. Pol II near TSS can be described by a mixture of multiple first order kinetics. The half-life of paused Pol II can be directly measured by the decay of photoactivated Pol II (green) on the polytene chromosome. This paused Pol II can elongate into the gene body with a rate constant of k_{el} or terminate with a rate constant of k_t. The short nascent RNA associated with paused Pol II or terminated Pol II can represent the amount of Pol II molecules themselves. Under the assumption that the short nascent RNA at the free fraction is at steady-state, the production rate of free nascent RNA, $k_t[RNA]_{ch}$ should be equal to the decay rate of the free nascent RNA, $k_d[RNA]_{fr}$. k_d can be directly calculated by inhibiting the initiation with Triptolide, and measuring the time-course of the nascent RNA decrease. From these values, k_t could be determined as $k_d([RNA]_{fr} / [RNA]_{ch})$.

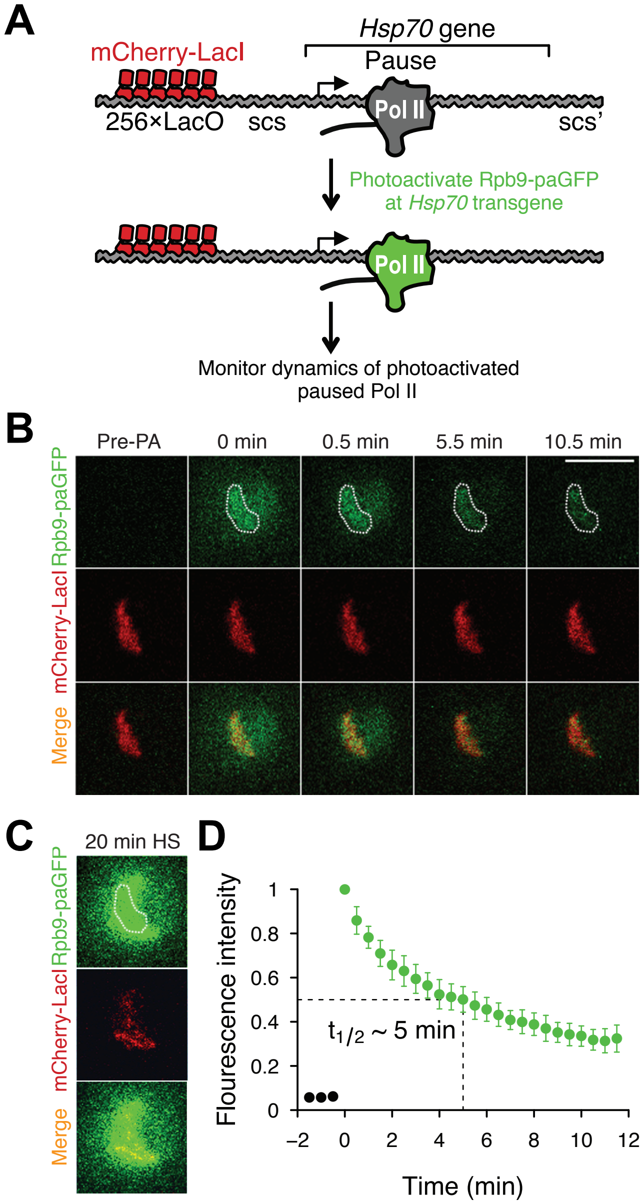


Figure 3.2. Imaging the stability of paused Pol II at uninduced *Hsp70* transgene in living cells**.** (**A**) Schematic diagram of live-cell imaging experimental set-up. The transgenic construct contains 256 LacO repeats followed by *Hsp70* sequence which is flanked by the insulator elements scs and scs'. (**B**) During NHS, images of *Hsp70* transgene marked with mCherry-LacI before and after photoactivation of Rpb9-paGFP and the time course that follows. The strongest paGFP-Pol II signal is outlined in white. (**C**) During HS, images of mCherry-LacI, Rpb9-paGFP, and merge at *Hsp70* transgene. (**D**) Normalized fluorescence intensities of Rpb9-paGFP flourescence decay after photoactivation (FDAP within outlined region in b) under NHS conditions; n=9. Error bars indicate SD. Scale bar, 10 μm. Data for this figure was generated by Dr. Martin S. Buckley.

HS induction of the transgene leads to a response similar to the native HS loci (decondensation and additional recruitment of Pol II)[15], indicating that it is functional. Using this system, we designed an optical pulse-chase experiment by photoactivating Pol II subunits instead of photobleaching[16], at the marked *Hsp70* locus, using a transgenic fly that expresses a paGFP tagged subunit of Pol II (Rpb9-paGFP) in the salivary gland nuclei. Rpb9-paGFP distributes on chromosomes bands in a pattern identical to the fluorescently tagged Rpb3 subunit (RFP-Rpb3), under basal conditions and HS conditions. Moreover, like Rpb3, Rpb9-paGFP reliably locates to the mCherry-LacI tagged *Hsp70* transgene before and after HS.

Rpb9-paGFP is fluorescently inert at the *Hsp70* transgene before photoactivation (Figure 3.2B). In order to directly assay the dynamics of paused Pol II at the *Hsp70* loci, laser scanning confocal microscopy (LSCM) was used to specifically photoactivate Rpb9-paGFP (~14 fold increase in fluorescence) at the mCherry-LacI marked *Hsp70* transgene under non-heat shock (NHS) conditions (Fig. 1a). We then obtained a time series of the fluorescence decay of Rpb9-paGFP at the *Hsp70* transgene (Figure 3.2B). To validate that the locus examined was the Hsp70 transgene and not some other gene near the single mCherry-LacI band in nuclei, we determined that Rpb9-paGFP signal is specific is heat shock inducible (Figure 3.2C). We also demonstrated that the decay of signal is due to loss of the photoactivated Rpb9-paGFP from the locus and not photobleaching by showing that photobleaching is minimal during the time course of imaging. The resulting decay plot corresponds approximately to a first order kinetics with the half-life of around 5 min (Figure 3.2D). This finding indicates that Pol II is relatively stable at this locus. Because the main form of Pol II at the uniduced *Hsp70* locus is paused Pol II, this suggests that this Pol II is stably associated with a turnover half-life of 5 min.

Biochemical Measurement of Pol II Stability by Nascent RNA Fractionation

Although paused Pol II at *Hsp70* appears to be relatively stable, it does not have an infinite lifetime. We reasoned that the clearance of the paused Pol II would be the sum of escape into elongation and premature termination. To measure their contributions to the stability of paused Pol II, we used an independent biochemical and kinetic approach in *Drosophila* S2 cells (Figure 3.3A, Figure 3.1). First, we evaluated the escape into elongation by calculating the kinetic constant of the polymerase escaping into the gene body. The rate of escaping Pol II can be derived from the gene body Pol II density and the speed of Pol II elongation. By dividing this rate by the total amount of paused Pol II, the kinetic constant of escape (fraction of paused Pol II that escapes into elongation per minute) can be calculated (Figure 3.4).

For this measurement, we used pre-existing nuclear run-on sequencing (GRO-seq or PRO-seq) datasets in S2 cells (Figure 3.3B), and the elongation speed of 1.5 kb/min from previous FRAP study in *Hsp70*[13] as independent measurements of the Pol II elongation rate on Drosophila genes. The run-on sequencing results show that at *Hsp70*, about 86% of engaged Pol II are restricted to promoter proximal region on average while 13% are distributed in the 2.4 kb gene body region (Figure 3.3C). From these numbers, we calculated that about 9.4% of paused Pol II escapes into elongation every minute, or in other words, each Pol II has an escape half-life of 7.3 min. This measurement agrees well with the half-life estimate derived from the density of transcribing RNA polymerase at Hsp70 loci in electron micrographs of Miller spreads (data not published). Interestingly, this half-life estimate shows minor discordance from the optical measurement (5 min), indicating that Pol II escaping to elongation composes a significant fraction, but probably not all of the loss of paused Pol II seen optically.

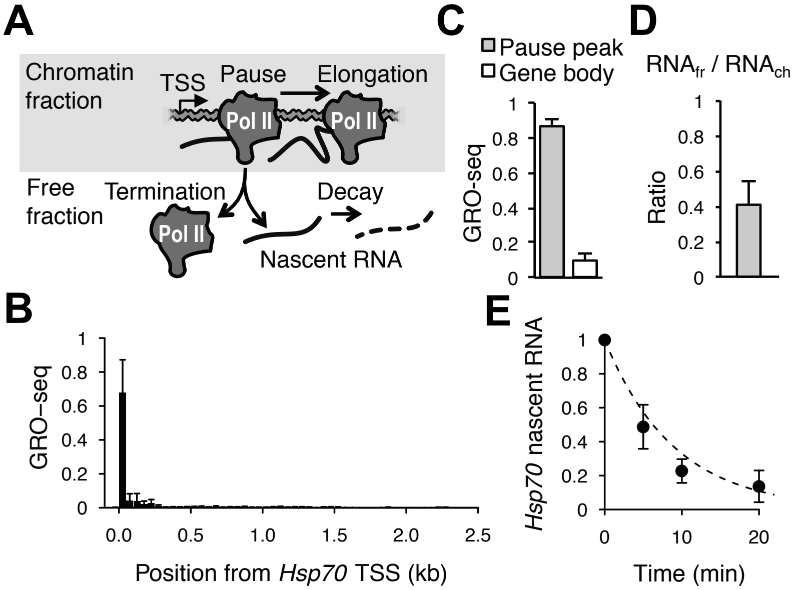


Figure 3.3. Biochemical analysis of steady state paused Pol II kinetics. (**A**) Schematic diagram of paused Pol II and associated short nascent RNA kinetics on chromatin and in free fractions. (**B**) Average GRO-seq profile on *Hsp70* gene. Vertical axis is the fraction of GRO-seq read counts in 50 bp bins relative to the total read counts across *Hsp70* gene (n=14 datasets). Error bars indicate SD. (**C**) Fraction of GRO-seq reads mapped to pause peak (-50 to +250 from TSS) and gene body (+300 to 3' end). (**D**) Ratio between free and chromatin bound fractions of short nascent RNA associated with *Hsp70* TSS (n=9). (**E**) Time-course of free *Hsp70* nascent RNA after Triptolide addition (10 μM). Each time-point is normalized to pre-treatment level. Error bars indicate SEM unless noted otherwise.

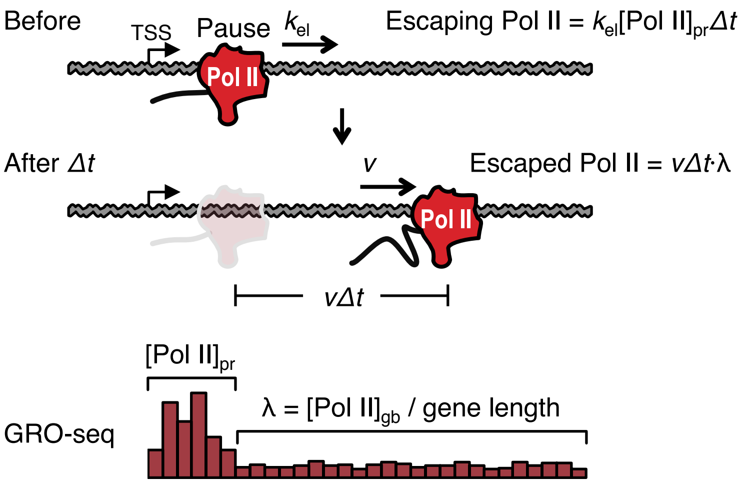

Before
TSS
Pause
k_{el}
Escaping Pol II = k_{el}[Pol II]$_{pr}$$\Delta t$
Pol II
After Δt
v
Escaped Pol II = $v\Delta t{\cdot}\lambda$
Pol II
$v\Delta t$
[Pol II]$_{pr}$
GRO-seq
λ = [Pol II]$_{gb}$ / gene length

Figure 3.4. Estimation of the rate constant of elongation (k_{el}) from pausing. By the definition of rate constant, the amount of Pol II escaping during a short time period Δt is k_{el}[Pol II]$_{pr}$$\Delta t$, where [Pol II]$_{pr}$ is the level of Pol II at the promoter. During Δt, the first escaped Pol II travels a distance of $v\Delta t$, where v is the elongation speed of Pol II. Pol II that escaped during Δt will be distributed within this range, and the amount of escaped Pol II will be $v\Delta t{\cdot}\lambda$, where is λ is the Pol II density in the gene body. Pol II level at the promoter and gene body is obtained from GRO-seq datasets on *Hsp70* gene. From these values, k_{el} could be determined as $v\lambda$ / [Pol II]$_{pr}$.

Next we evaluated the rate of premature termination. Terminated Pol II is no longer associated with chromatin and will be freely released, along with the short nascent RNA derived from the same Pol II molecule (Figure 3.3A, Figure 3.1). The ratio between terminated and paused RNAs will then reflect the rate of termination (see methods). We measured the amount of short nascent RNA arising from *Hsp70* TSS in both chromatin-associated (paused) and free (terminated) fractions. The free fraction was on average 41% of the chromatin bound fraction (Figure 3.3D), suggesting that a significant amount of Pol II may be terminating from the *Hsp70* pause site.

To estimate the kinetic constant of termination, we analyzed the flux of short nascent RNA in and out of the free fraction (Figure 3.3A, Figure 3.1). At steady state, since free nascent RNA is generated from terminating Pol II, the rate of free nascent RNA production should be equal to the rate of the free RNA decay. In this case, the kinetic constant of termination equals the constant for RNA decay multiplied by the ratio between free and chromatin bound short nascent RNA levels (Figure 3.3A, Figure 3.1). To estimate the decay constant, we blocked the RNA production at the initiation step by using Triptolide, a potent chemical inhibitor of TFIIH helicase[17], and measured the time-course of free nascent RNA decay. The time course showed a decay half-life around 6 min (Figure 3.3E), and the calculation of the termination kinetic constant implies that 4.6% of paused Pol II undergoes termination per every minute. After combining these estimates of termination (4.6%) and elongation (9.4%), about 14% of paused Pol II is cleared from the pause site per minute, which corresponds to the half-life of 5.0 min. It is reassuring that the two independent methods - biochemical steady-state kinetics and the optical pulse-chase measurements - are in such close agreement (Table 3.1).

Table 3.1. Rate constants for the kinetics of promoter proximal Pol II

Method	Rate	Description	Value (min^{-1})	Half life (min)
Optical puls e-chase	k_{ctrl}	Overall stability	0.13	5.4
	k_{FP}	Elongation blocked by FP	0.072	9.6
Biochemical steady-stat e	k_{el}	Elongation by GRO-seq	0.094	7.3
	k_t	Termination in steady-state	0.046	15
	k_{cl}	Overall stability (= $k_{el} + k_t$)	0.14	5.0

Both Pause Escape and Termination Can Be Optically Measured Separately.

The Pol II CTD-kinase P-TEFb is required for the escape of paused Pol II into productive elongation at most genes including *Drosophila Hsp70*[3,9,18]. Therefore, to further confirm the early termination rate of paused Pol II *in vivo*, salivary glands were treated with the P-TEFb kinase inhibitor Flavopiridol to block transcription elongation (Figure 3.6A). The termination rate was then optically measured by assaying the fluorescence decay after photoactivation of Rpb9-paGFP at the uninduced Hsp70 transgene (Figure 3.6B). The decay curves show the stabilization of Pol II (Figure 3.6C) and near doubling of the half-life with Flavopiridol treatment (Figure 3.6D, Table 3.1), indicating that the elongation and termination can also be distinctly measured using optical methods *in vivo*. Together, the optical and biochemical analyses provide a strong indication that paused Pol II is relatively stable and undergoes both productive elongation and termination *in vivo*.

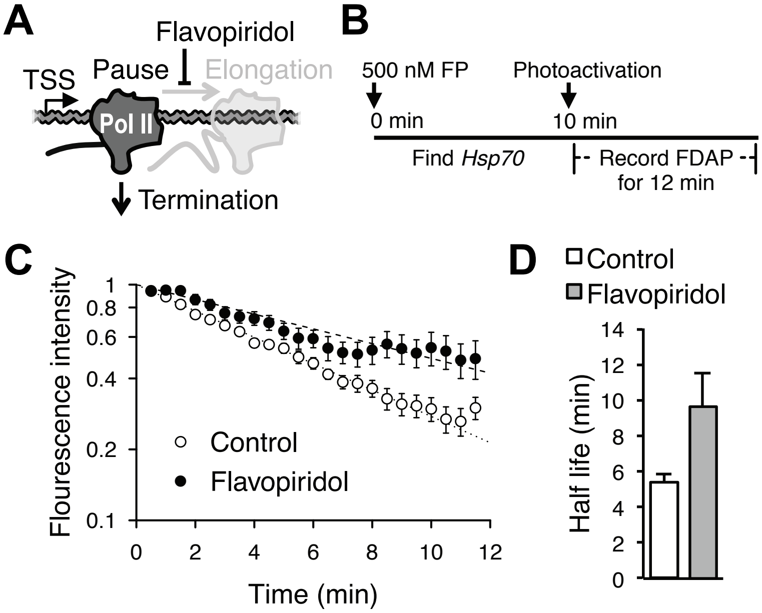


Figure 3.6. Optical measurement of the termination rate of paused Pol II at *Hsp70* transgene in living cells. (**A**) Schematic diagram outlining overall logic of using Flavopiridol to measure paused Pol II termination kinetics. (**B**) Illustration of the experimental scheme. (**C**) Semi-log plot of normalized fluorescence intensities of paGFP-Rpb9 FDAP under NHS condition in the presence (filled circles) and the absence (empty circles) of Flavopiridol treatment; n=7 and n=9, respectively. Data corrected for background Pol II signal, see methods. Error bars indicate SD. (**D**) Half-lives of paused Pol II decay with Flavopiridol treatment and control using the data in panel c. Error bars indicate SEM.

Conclusion

The prevailing model for regulation of paused Pol II proposes that recruitment of elongation factors such as P-TEFb facilitate the escape of Pol II into productive elongation[19]. This mode of regulation implies that paused Pol II stably associates with promoters until factors are recruited to trigger its release. Here we provide evidence that the canonical paused Pol II at the *Drosophila Hsp70* gene is dynamic and undergoes both termination and elongation. This raises the intriguing possibility that premature termination plays a role in metazoan gene regulation. Importantly, these findings provide a new avenue for future studies to assess the termination status of paused Pol II genome-wide and importantly, its potential contribution to gene regulation.

Materials and Methods

Generations of the transgenic fly lines

Fly lines are generated and maintained by Dr. Martin S Buckley in the Lis lab. *Hsp70 transgenic fly line:* The *Hsp70* gene region (-252 to +2439) was PCR amplified from the 56H8 plasmid[20], and cloned into pBSIIKS vector (Agilent Technologies) creating pBSIIKS-*Hsp70*. The SCS/SCS' boundary elements were amplified from w^{1118} flies and sequentially cloned into pBSIIKS-*Hsp70* to generate pBSIIKS-SCS-*Hsp70*-SCS'. LacO repeats (256x) were digested from the vector pps8.32 (ref) and cloned into the pSTBlue-1 (Millipore) using XhoI/SalI sites creating pSTBlue-1-LacO. The SCS-*Hsp70*-SCS' fragment was digested from PBSIIKS-SCS-*Hsp70*-SCS' and cloned into pSTBlue-1-LacO using ApaI/SpeI sites to create pSTBlue-1-LacO-SCS-*Hsp70*-SCS'. Internal restrictions sites for SphI/NotI were introduced to pCasperII(attb) vector by inserting annealed oligos between

BamHI/BglII sites, creating pCasperII(InsertF/InsertR). The LacO-SCS-*Hsp70*-SCS' fragment was digested with pSTBlue-1LacOSCS*Hsp70*SCS' and cloned into pCasper II(InsertF/InsertR) using SphI/NotI sites creating pCasperII(InsertF/InsertR)-LacO-SCS-*Hsp70*-SCS'. The attb fragment was digested from the vector pCasper II(attb) and cloned into the pCasperII(InsertF/InsertR)-LacO-SCS-*Hsp70*-SCS' using NotI sites to generate pCasperII(attb)-LacO-SCS-*Hsp70*-SCS'. This final plasmid was transformed into the *Drosophila* genome using the PHiC31 platform line attp16 (Genetic Services Inc.).

Rpb9-paGFP transgenic fly line: The *paGFP* gene was PCR amplified from the *mPAGFP*-pRK5 plasmid[16] and cloned into a Gateway (Invitrogen) based UAST P-element insertion vector, pTWG, creating pTW-paG. The *Rpb9* cDNA was amplified from *Drosophila* cDNA (Open Biosystems) and the Rpb9 coding region was cloned into the pDONR221 vector using the Gateway system generating pDONOR221-*Rpb9*. The clone was transferred to pTW-paG generating the construct pT-*Rpb9*-paG. The *Rpb9-paGFP* fusion construct was introduced into the *Drosophila* germ line by P-element mediated transformation (Best Gene Inc.).

mCherry-LacI transgenic fly line: The minimal *Sgs3* promoter (drives gene expression in salivary gland) was PCR amplified from w^{1118} flies and cloned into pCaSpeR-4 vector creating pCaSpeR-*Sgs3*. *mCherry* was PCR amplified from the pRSET-B-*mCherry* (Dr. Roger Tsien) and cloned into pAFS144 vector (contains *LacI* ORF) creating pAFS-*mCherry-LacI*. The *mCherry-LacI* fragment was digested from the pAFS-*mCherry-LacI* and cloned into pCaSpeR-*Sgs3* using XhoI/XbaI sites creating pCaSpeR-*Sgs3*-*mCherry-LacI*. The *mCherry-LacI* fusion construct was introduced into the *Drosophila* germ line by P-element mediated transformation (Best Gene Inc.).

Fluorescence Decay After Photoactivation (FDAP) of polytene nuclei (This experiment was carried out by Dr. Martin Buckley)

Dr. Martin S. Buckley generated the primary imaging data. Briefly, intact *Drosophila* salivary glands were dissected from third instar larvae and transferred to a MaTek glass bottom dish (P35G-1.0-14-C) containing Grace's media (Sigma-Aldrich) and a glass coverslip was placed on the sample to reduce movement of the glands. For drug experiments, glands were transferred to 500 nM Flavopiridol (Sigma) diluted in media. Laser scanning confocal microscopy of salivary glands was carried out on a Zeiss 710 microscope using a Zeiss 63x C-Apochromat objective (numerical aperture 1.2, water immersion). The RFP-LacI tagged Hsp70 transgene was identified using a 561 nm laser. Samples were photoactivated using a circular region of interest limited (ROI) to the dimensions of the RFP-LacI spot using a 405 nm. The fluorescent of both the RFP-LacI and paGFP-Rpb9 was imaged using a 488 nm every 30 sec for 12 min. To confirm that the *Hsp70* gene was targeted, an objective pre-heated to 37°C (Bioptechs) was used to heat-shock samples for 20 min. The images were analyzed with ImageJ.

Correction of background genomic Pol II signal.

The FDAP measurement within a region of interest (ROI) can have non-*Hsp70* Pol II background signal depending on the resolution and the orientation of the polytene chromosome. To correct for this background, we used a mixed linear decay model for Pol II signal from adjacent genes. For each gene within 300 kb from the *Hsp70* transgene, Pol II level, gene length, and the distance from *Hsp70* transgene were obtained using GRO-seq datasets. Then assuming that Pol II signals clear from the gene bodies with the elongation rate of 1.5kb/min, we derived a linear decay function for each gene. We introduced a resolution parameter, and took the distance-dependent Gaussian-weighted (exp(-distance^2/2/resolution^2)) sum of individual decay functions as the background signal decay function. We fitted each FDAP time-course curve to a two-component model composed of mixed linear decay with a resolution parameter, and an exponential decay with a half-life parameter. The best fit resolution and half-life parameter combination was found using a grid search algorithm (Figure 3.7).

Biochemical analysis of steady state kinetics.

The rate constant of elongating Pol II from pausing (k_{el}) was derived from GRO-seq and PRO-seq data in *Drosophila* S2 cells at the *Hsp70* gene:

$$k_{\mathrm{el}} = v\lambda / [\text{Pol II}]_{\mathrm{pr}}$$

where v = Pol II elongation speed (kb/min), λ = gene body Pol II density (reads/kb), $[\text{Pol II}]_{\mathrm{pr}}$ = level of promoter proximal Pol II (reads) (Supplementary Fig. 9). The kinetic constant of Pol II termination (k_{t}) was determined from nascent RNA fractionation in S2 cells:

$$k_{\mathrm{t}} = k_{\mathrm{d}} ([\text{RNA}]_{\mathrm{fr}} / [\text{RNA}]_{\mathrm{ch}})$$

where $[\text{RNA}]_{\mathrm{ch}}$ = nascent RNA in chromatin fraction, $[\text{RNA}]_{\mathrm{fr}}$ = nascent RNA in free fraction, k_{d} = free nascent RNA decay constant (min^{-1}) (Figure 3.1).

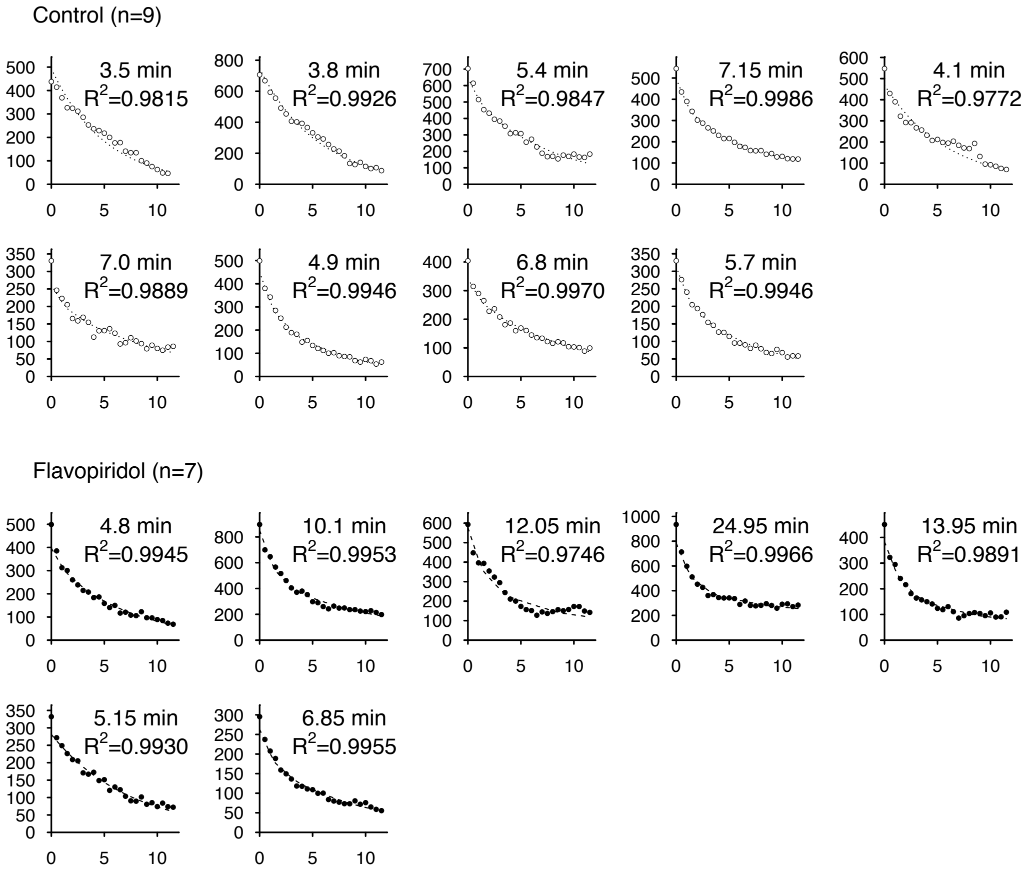


Figure 3.7. Half life estimation after background correction of FDAP. FDAP measurements and fits for control and flavopiridol treatment conditions are shown. The grid search algorithm was used to find the resolution and half-life pair applying the least square method. The search space was [5, 100]×[0.05, 50] and the grid size was 0.05×5 (min×kb). First, the parameter pairs were limited to those that have the sum of squares for error (SSE) within the +5% margin of the least SSE. Then, the one with least resolution parameter was selected. Finally, the background component of the mixed linear decay model with the resolution parameter was subtracted from the FDAP measurement, and was fit to an exponential decay model. The highest and the lowest values from each condition were removed for the half-life analysis.

Nascent RNA fractionation.

Nascent RNA fractionations were performed as described by Wuarin and Schibler[21], with minor modifications. Briefly, 0.5~1×10^7 S2 cell was lysed directly in the 1 M urea lysis buffer and fractionated by ultracentrifugation for 20 min at 45,000 g. RNA from supernatant fraction was phenol-chloroform extracted. Chromatin pellet was resuspended in Trizol (Ambion) and disrupted by short bursts of sonication, followed by RNA extraction. A short *in vitro*-transcribed spike-in RNA from an *Arabidopsis* gene (*RCP1*) sequence (5×10^{-2} fmol) was added before ethanol precipitation as a normalization control. Each RNA fraction was hybridized to a mixture of biotin labeled DNA probes complementary to *Hsp70* and *RCP1* sequences[22], and specific RNA was enriched. RNA was sequentially treated with 3' RNA adaptor ligation (T4 RNA ligase I; NEB), 5' phosphate dependent exonuclease (Terminator; Epicentre), Tobacco acid pyrophosphatase (TAP; Epicentre), 5' RNA adaptor ligation, and reverse transcription (Superscript; Invitrogen) following the manufacturer's instructions. *Hsp70* and *RCP1* products were quantified by qPCR using primers that span across the insert and adaptor junctions. *Hsp70* products of different lengths (25 ~ 40 nt) were normalized to *RCP1* level and the normalized values from different lengths were averaged. RNA adaptor and DNA primer sequences are listed in Table 3.2.

Table 3.2. DNA and RNA sequences for nascent RNA fractionation. (**A**) RNA adaptor sequences. (**B**) Organization of *Hsp70* qPCR primers. (**C**) *Hsp70* qPCR primers. (**D**) *RCP1* control sequences. /idT/ : inverted dT. /bio/ : 3'-biotin.

A

5'-RNA adaptor	5′-GUUCAGAGUUCUACAGUCCGACGAUC-3′
3'-RNA adaptor	5′-pUCGUAUGCCGUCUUCUGCUUGU/idT/-3′

B

Hsp70 insert (25nt)	5′-AUUCUAUUCAAACAAGCAAAGUGAA-3′
Adaptor ligated	5′-GUUCAGAGUUCUACAGUCCGACGAUC AUUCUAUUCAAACAAGCAAAGUGAA UCGUAUGCCGUCUUCUGCUUGU-3′
Hsp70-F1	5′-TCCGACGATC ATTCTATTCAAA-3′
Hsp70-R25	3′-CGTTTCACTT AGCATACGGC-5'

C

Hsp70-R25	5′-CGGCATACGATTCACTTTGC-3′
Hsp70-R26	5′-GGCATACGAGTTCACTTTGC-3′
Hsp70-R27	5′-GGCATACGATGTTCACTTTGC-3′
Hsp70-R28	5′-CGGCATACGAGTGTTCACTTT-3′
Hsp70-R29	5′-CGGCATACGATGTGTTCACT-3′
Hsp70-R30	5′-CGGCATACGAATGTGTTCA-3′
Hsp70-R31	5′-CGGCATACGAGATGTGTTCA-3′
Hsp70-R32	5′-GGCATACGACGATGTGTTCA-3′
Hsp70-R33	5′-GCATACGAGCGATGTGTTCA-3′
Hsp70-R34	5′-GCATACGAAGCGATGTGTTC-3′
Hsp70-R35	5′-GGCATACGATAGCGATGTGTT-3′
Hsp70-R36	5′-GGCATACGATTAGCGATGTG-3′
Hsp70-R37	5′-GGCATACGACTTAGCGATGTG-3′
Hsp70-R38	5′-GGCATACGAGCTTAGCGATGTG-3′
Hsp70-R39	5′-ACCGGCATACGACGCTTAG-3′
Hsp70-R40	5′-ACCGGCATACGATCGCTTAG-3′

D

T7template-*RCP1*-F	5′-TAATACGACTCACTATAGGGAGATGGTGGACTCTCCGTTCTTC-3′
T7template-*RCP1*-R	5′-CGGATGGATCCACTGATTTGAGGAAGAACGGAGA-3′
RCP1 RNA (48nt)	5′-GGGAGA UGGUGGACUCUCCGUUCUUCCUCAAAUCAGUGGAUCCAUCCG-3′
RCP1-probe	5′-TGATTTGAGGAAGAACGGAGAGTCCACCACCC/biotin-3′
RCP1-F1	5′-TGGTGGACTCTCCGTTCTTC-3′
RCP1-F2	5′-ATCGGGAGATGGTGGACTCT-3′
RCP1-F3	5′-GACGATCGGGAGATGGTG-3′
RCP1-R1	5′-GGCATACGACGGATGGAT-3′
RCP1-R2	5′-CGGATGGATCCACTGATTTG-3′

REFERENCE

1. Muse, G. W. *et al.* RNA polymerase is poised for activation across the genome. *Nat Genet* **39,** 1507–1511 (2007).

2. Nechaev, S. *et al.* Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila. *Science* **327,** 335–338 (2010).

3. Rahl, P. B. *et al.* c-Myc regulates transcriptional pause release. *Cell* **141,** 432–445 (2010).

4. Zeitlinger, J. *et al.* RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo. *Nat Genet* **39,** 1512–1516 (2007).

5. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. *Science* **322,** 1845–1848 (2008).

6. Gollnick, P. & Babitzke, P. Transcription attenuation. *Biochim. Biophys. Acta* **1577,** 240–250 (2002).

7. Kao, S. Y., Calman, A. F., Luciw, P. A. & Peterlin, B. M. Anti-termination of transcription within the long terminal repeat of HIV-1 by tat gene product. *Nature* **330,** 489–493 (1987).

8. Brannan, K. *et al.* mRNA decapping factors and the exonuclease Xrn2 function in widespread premature termination of RNA polymerase II transcription. *Mol Cell* **46,** 311–324 (2012).

9. Chao, S. H. & Price, D. H. Flavopiridol inactivates P-TEFb and blocks most RNA polymerase II transcription in vivo. *J Biol Chem* **276,** 31793–31799 (2001).

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

11. Core, L. J. *et al.* Defining the status of RNA polymerase at promoters. *Cell Rep* **2,** 1025–1035 (2012).

12. Lis, J. T. Imaging Drosophila gene activation and polymerase pausing in vivo. *Nature* **450,** 198–202 (2007).

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

14. Yao, J., Munson, K. M., Webb, W. W. & Lis, J. T. Dynamics of heat shock factor association with native gene loci in living cells. *Nature* **442,** 1050–1053 (2006).

15. Zobeck, K. L., Buckley, M. S., Zipfel, W. R. & Lis, J. T. Recruitment timing and dynamics of transcription factors at the Hsp70 loci in living cells. *Mol Cell* **40,** 965–975 (2010).

16. Patterson, G. H. & Lippincott-Schwartz, J. A photoactivatable GFP for selective photolabeling of proteins and cells. *Science* **297,** 1873–1877 (2002).

17. Titov, D. V. *et al.* XPB, a subunit of TFIIH, is a target of the natural product triptolide. *Nat. Chem. Biol.* **7,** 182–188 (2011).

18. Lis, J. T., Mason, P., Peng, J., Price, D. H. & Werner, J. P-TEFb kinase recruitment and function at heat shock loci. *Genes Dev* **14,** 792–803 (2000).

19. Zhou, Q., Li, T. & Price, D. H. RNA Polymerase II Elongation Control. *Annu. Rev. Biochem.* **81,** 119–143 (2012).

20. Török, I. & Karch, F. Nucleotide sequences of heat shock activated genes in Drosophila melanogaster. I. Sequences in the regions of the 5‘ and 3’ ends of the hsp 70 gene in the hybrid plasmid 56H8. *Nucleic Acids Res.* **8,** 3105–3123 (1980).

21. Wuarin, J. & Schibler, U. Physical isolation of nascent RNA chains transcribed by RNA polymerase II: evidence for cotranscriptional splicing. *Mol Cell Biol* **14,** 7219–7225 (1994).

22. Rasmussen, E. B. & Lis, J. T. Short transcripts of the ternary complex provide insight into RNA polymerase II elongational pausing. *J. Mol. Biol.* **252,** 522–535 (1995).

CHAPTER 4[d]

GENOME-WIDE TRANSCRIPTION DYNAMICS REVEAL THE INTERPLAY OF PAUSING AND ELONGATION RATES ON PRE-MRNA SPLICING

Summary

The transcription cycle of RNA polymerase II (Pol II) consists of many potentially rate-limiting steps that can be regulated. One major step in regulation is the escape of promoter-proximal paused Pol II into productive elongation driven by Positive Transcription Elongation Factor (P-TEFb) kinase. To ascertain which genes require P-TEFb kinase activity and when during the transcription cycle it is required, we inhibited P-TEFb with flavopiridol (FP) and examined the effects on transcribing and paused Pol II distributions using the highly sensitive GRO-seq assay. Transcription of nearly all genes is dramatically inhibited but the peak of paused Pol II often persists or is enhanced, showing that escape from pause is a ubiquitous and crucial step in the transcription cycle. Furthermore, we tracked the progression of elongating Pol II following inhibition of pause escape by FP and derived elongation rates on a genome-wide scale. Systematic analysis of nearly 1,000 genes shows that Pol II rates vary between and within genes. Notably, Pol II accelerates as it transcribes through genes and slows when splicing occurs. Furthermore, splicing is less efficient in 5' regions of genes, but its fidelity is enhanced by promoter-proximal pausing, revealing the interplay between Pol II elongation and splicing.

[d] Works in this chapter was done in collaboration with Dr. Iris H. Jonkers. Extents of contributions are noted in the figure legends.

Introduction

Many steps throughout the transcription cycle of RNA polymerase II (Pol II) can be regulated, and modulation of each step has the potential to alter the timing, total output, and the isoforms of mRNA production. First is initiation, which is mostly dependent on the transcription factor recruitment, pre-initiation complex assembly, and formation of the transcription bubble[1]. While initiation is very important, it is by no means the only step at which transcription can be regulated. After initiation, Pol II can be paused by Negative Elongation Factor (NELF), DRB-sensitivity Inducing Factor (DSIF) and components of the core promoter[2,3]. The escape of paused Pol II into productive elongation can be rate-limiting, and is dependent on the positive elongation factor P-TEFb, which consists of protein kinases Cdk9 and CyclinT1[4-6]. P-TEFb is recruited directly or indirectly to the paused Pol II complex by transcription activators, where it phosphorylates the CTD, as well as DSIF and NELF to transform DSIF into a positive elongation factor and evict NELF[7]. P-TEFb appears to be both necessary and sufficient for paused Pol II escape into productive elongation.

The importance of transcription regulation at pausing has become increasingly evident over the last few years. Genome-wide Run-on sequencing (GRO-seq)[8,9], sequencing of short capped nascent RNAs[10], and Pol II Chromatin Immunoprecipitation (ChIP) experiments[11,12] all have shown in organisms ranging from *Drosophila* to humans that a large fraction of actively transcribed genes have accumulations of paused Pol II within the promoter proximal region, indicating a rate-limiting step during early elongation that for many genes appears P-TEFb dependent[13]. Also, expressed genes that do not have a peak of paused Pol II in one cell type, may acquire one in another[9], indicating that many genes have the potential of becoming regulated at the step of escape into productive elongation. However, it is unclear if all genes undergo this P-TEFb kinase dependent step.

Several functions of Pol II pausing have been proposed and have variable levels of support in the literature[2]. One unproven, but attractive hypothesis is that pausing might function as a checkpoint to ensure that the elongation complex is fully assembled and modified for efficient elongation and co-transcriptional RNA processings, i.e. capping and splicing. Capping occurs when the nascent RNA protrudes from the paused Pol II located 20-30 nucleotides from the TSS[14,15], and the cap can promote co-transcriptional splicing by recruitment of splicing factors via Cap Binding Complex CBC[16-18]. Also, increased phosphorylation of the CTD and the association of additional RNA processing components coincides with increased Pol II escape to productive elongation[19]. These studies indicate that RNA capping and the maturation of Pol II may form a link between pausing and co-transcriptional mRNA processing.

In addition to the maturation of Pol II during pausing, the rate of Pol II elongation has also been proposed to influence co-transcriptional processes such as transcription splicing[20-22], 3' end processing[23], and termination[24]. Transcription with a slow mutant of Pol II promotes alternative splicing in human and yeast cells[21,22]. Slowing of Pol II around exons is thought to facilitate assembly of the spliceosome and recognition of the splice site, and thereby, splicing of the associated intron[25,26].

Many studies have measured elongation rates at individual genes in various organisms, producing a range from 1 to 4 kb/min[27]. Recently, elongation rates for over one hundred genes were measured simultaneously by following the induction wave of Pol II after Estrogen or TNF-α treatment[28]. Interestingly, this study showed a broad range of elongation rates within single cell types and different rates of sets of genes responding to different activators.

Here, we use FP, a specific inhibitor of the P-TEFb kinase[29], and the highly sensitive GRO-seq assay[8] to demonstrate that productive transcription of virtually all

Pol II genes in mouse embryonic stem cells (mESCs) requires P-TEFb kinase activity. The added sensitivity and transcription orientation information provided by GRO-seq confirms and extends earlier results using a ChIP-seq assay of Pol II[13]. Thus, P-TEFb is generally required for promoter-proximal Pol II to escape into a productive elongation, even on genes where the escape of paused Pol II is apparently not a rate-limiting step. Furthermore, this striking inhibition of Pol II caused a 'wave' of elongating Pol II that, when assayed as a function of time following FP addition, allowed measurement of elongation rates of nearly a thousand genes simultaneously, and at various regions within genes. These systematic analyses show that Pol II elongation rates increase within the gene body, and decrease during splicing events. Finally, we show that proximal pausing correlates with efficient splicing at the 5' end of genes, supporting a model where pausing has a role in allowing the establishment of a functionally mature elongation complex that can couple transcription and RNA processing.

Generation of GRO-seq Libraries Treated with FP and Trp

P-TEFb kinase is recruited to promoters and phosphorylates the paused Pol II•NELF•DISF complex, thereby allowing the paused Pol II to escape into productive elongation. To identify all genes regulated by P-TEFb, we inhibited P-TEFb kinase activity with the drug FP, and compared this to inhibition of pre-initiation complex formation by Triptolide (Trp), a drug that interferes with the XPB helicase function of TFIIH[30,31]. We isolated replicates of nuclei of untreated mESCs and cells treated for 2, 5, 12.5, 25 and 50 min. with 300nM FP, as well as nuclei treated for 12.5, 25, and 50 min. with 500nM Trp (Figure 4.1A). To minimize off target effects, the concentrations of drugs are at the lower spectrum of concentrations used in previous experiments[29-31].

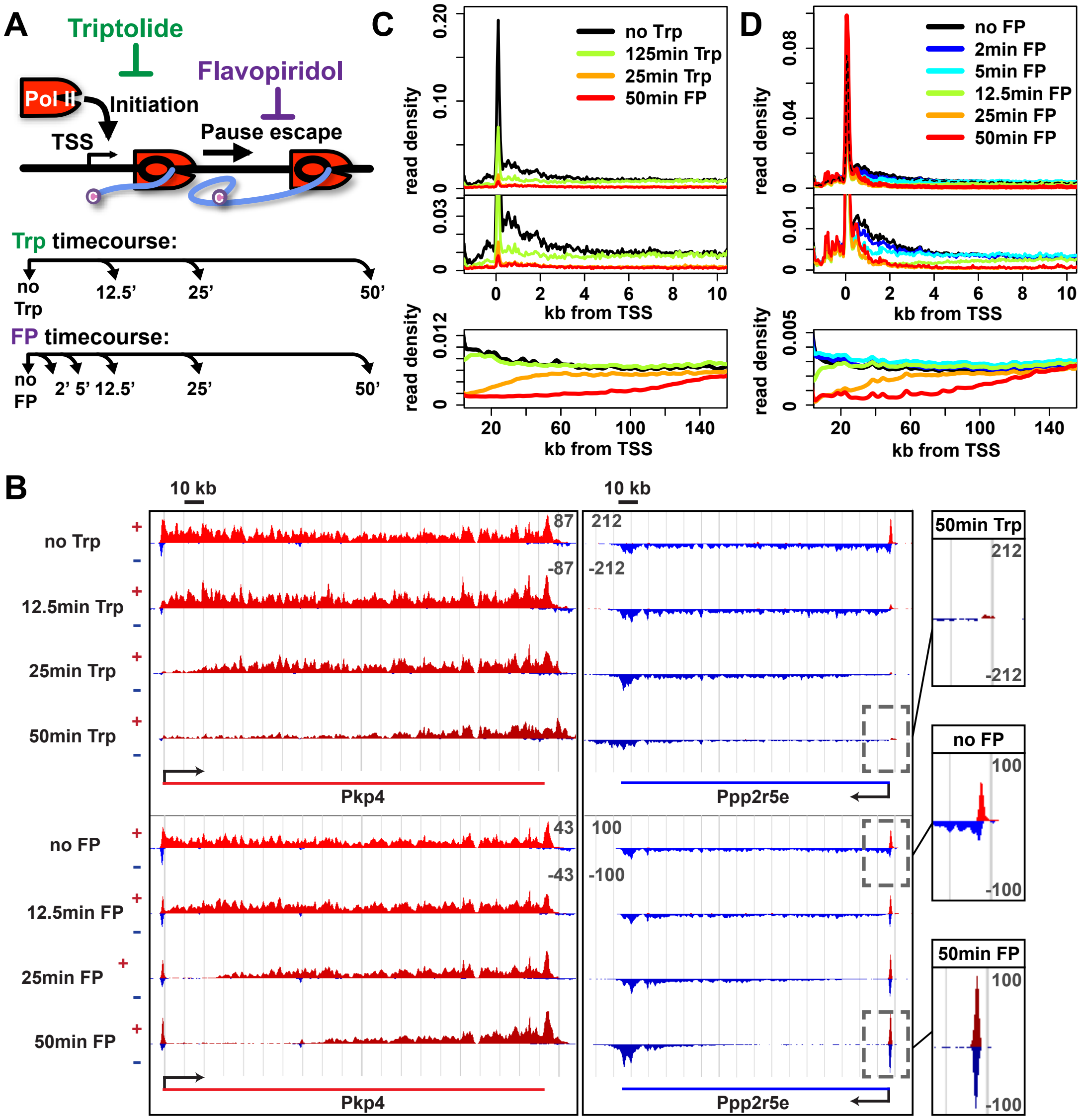


Figure 4.1. (**A**) Experimental set-up, 300 nM flavopiridol (FP) and 500 nM of triptolide (Trp) were used to block pause escape or transcription initiation in mES cells. Nuclei were isolated at timepoints after treatment as specifiied, control nuclei were untreated for the FP samples, and treated for 50 min with DMSO for the Trp samples. (**B**) Screenshot of genes Pkp4 and Ppp2r5e with or without Trp or FP treatment for 12.5, 25 or 50 min, with sense reads in red and antisense reads in blue. Inset numbers are the number of normalized reads per kb per unit area for each gene and treatment timecourse. (**C**) Composite profile of GRO-seq read density of all genes >12.5 kb (top panel) or >150 kb (bottom panel) after treatment with Trp for various durations of time. The middle panel is a zoom-in of the top panel of the region downstream of the TSS. (**D**) As C, but after treatment with FP for various periods of time. The experimental design, GRO-seq experiments were performed by Dr. Iris Jonkers.

Sequencing results	Total reads	% trimmed	% ribosomal	aligned	% aligned	Spike-In	% Spike-In
Control #1	39054383	23.3%	29.5%	5016163	12.8%	219906	0.6%
12.5min Trp #1	35515502	24.6%	24.3%	5152395	14.5%	338370	1.0%
25min Trp #1	34107998	51.2%	13.9%	613603	1.8%	73777	0.2%
50min Trp #1	37503054	22.3%	40.0%	1824854	4.9%	253943	0.7%
Control #2	74380410	20.3%	31.7%	13074860	17.6%	393978	0.5%
12.5min Trp #2	45737802	20.4%	25.1%	9147442	20.0%	358560	0.8%
25min Trp #2	76937683	24.0%	33.2%	8798933	11.4%	805210	1.0%
50min Trp #2	40672137	21.0%	39.7%	2639097	6.5%	304272	0.7%

Sequencing results	Total reads	% trimmed	% ribosomal	aligned	% aligned	Spike-In	% Spike-In
untreated #1	104889824	16.2%	34.8%	25539711	24.3%	3046032	2.9%
2min FP #1	92103568	16.1%	34.7%	22843152	24.8%	2627811	2.9%
5min FP #1	82094962	16.7%	30.0%	24548283	29.9%	2454930	3.0%
12.5min FP #1	84600449	15.3%	34.2%	21607659	25.5%	2980280	3.5%
25min FP #1	94235237	16.7%	41.1%	17356268	18.4%	3055098	3.2%
50min FP #1	101614542	14.0%	49.1%	15403548	15.2%	2721538	2.7%
untreated #2	63997320	15.4%	33.5%	17869848	27.9%	1295280	2.0%
2min FP #2	41265136	12.4%	35.5%	12447634	30.2%	901113	2.2%
5min FP #2	50758538	14.9%	36.9%	12912212	25.4%	1286001	2.5%
12.5min FP #2	54675137	10.4%	35.4%	16926587	31.0%	1873970	3.4%
25min FP #2	74393287	9.6%	48.6%	15992442	21.5%	2544132	3.4%
50min FP #2	83154700	9.2%	55.2%	14019123	16.9%	2124287	2.6%

Pearson correlation	promoter	gene body
Control replicates	1.000	0.994
12.5min Trp replicates	0.999	0.972
25min Trp replicates	-	-
50min Trp replicates	1.000	0.999

Spearman correlation	promoter	gene body
Control replicates	0.958	0.986
12.5min Trp replicates	0.937	0.988
25min Trp replicates	-	-
50min Trp replicates	0.772	0.964

Pearson correlation	promoter	gene body
untreated replicates	0.999	0.992
2min FP replicates	0.998	0.997
5min FP replicates	0.993	0.944
12.5min FP replicates	0.999	0.993
25min FP replicates	0.999	0.998
50min FP replicates	1.000	0.999

Spearman correlation	promoter	gene body
untreated replicates	0.976	0.993
2min FP replicates	0.975	0.991
5min FP replicates	0.973	0.990
12.5min FP replicates	0.977	0.990
25min FP replicates	0.976	0.986
50min FP replicates	0.971	0.980

Table 4.1. Sequencing statistics report. The GRO-seq library and sequencing data was generated by Dr. Iris Jonkers.

Nuclear run-on and library preparation was performed, and replicates were sequenced (Table 4.1). Biological replicates correlated extremely well and were combined for all further analysis. Because inhibition of escape from elongation and initiation were anticipated to have large genome-wide effects on Pol II transcription, we chose to normalize treated and control libraries using spike-in controls of in-vitro transcribed *Arabidopsis thaliana* RNA added during the run-on.

Two long genes, Pkp4 and Ppp2r5e (Figure 4.1B), illustrate the effects of inhibition of P-TEFb kinase or PIC formation on the kinetics of Pol II distribution. Clearly these genes show a distinct time dependent wave of Pol II leaving the body of the gene after either FP or Trp treatment, while the effects at the promoter were opposite for the two drugs. The Pol II pause peaks increased after FP treatment, but disappeared after Trp treatment (Figure 4.1B, zoom box). These are as expected, because inhibition of P-TEFb is known to inhibit escape from pause, but not Pol II initiation nor productively elongating Pol II[6,29]. In contrast, Trp prevents melting of DNA at the promoter, thereby preventing transcription initiation, but not the downstream steps.

To assess the global effects on total Pol II and phosphorylation levels of the CTD in the cell, we fractionated the insoluble chromatin from the soluble fractions of untreated mESCs and cells treated with 300 nM FP, 500 nM Trp, or Trp-vehicle DMSO for 50 min, and performed western blots with antibodies against N-terminus of Rpb1, and Ser5 or Ser2 phosphorylated CTD. Overall, chromatin bound Pol II is generally reduced after treatment with FP or Trp. However, phosphorylation of the CTD was reduced only after FP treatment, but not Trp, indicating that FP and Trp exert the intended effects.

P-TEFb Dependent Escape from the Pause Occurs at All Active TSSs

The entry of Pol II into productive transcription for genes shown in Figure 4.1B clearly are dependent on P-TEFb kinase activity. However, to assess if P-TEFb is utilized generally in Pol II escape from the promoter proximal region, we examined composite GRO-seq profiles of selected genes for all experimental conditions (Figure 4.1B and C) or heat maps of gene-specific change in Pol II density around the TSS after 50 min treatment of FP or Trp (Figure 4.2).

After inhibiting the initiation step with Trp, levels of promoter proximal Pol II and Pol II downstream of the TSS dropped within 12.5 min, and depletion progressed in time for the majority of genes (Figure 4.1C). After 50 min of Trp treatment, Pol II is depleted to near background density at the promoter proximal region and up to 120kb downstream, showing that Trp effectively blocks Pol II transcription initiation but allows elongation. Treatment with FP does not reduce promoter proximal Pol II, but like Trp, it causes a rapid decrease of Pol II density immediately downstream of the TSS, and Pol II density was reduced farther into the gene body with increasing time (Figure 4.1D). Elongating Pol II forms an 'inhibition wave' that is very similar after Trp or FP treatment (Figure 4.1C and D, lower panels). The reduced Pol II density region expands in a time dependent fashion, while the region where the inhibition wave has not yet passed remains at equal Pol II density levels throughout the time course, indicating that Pol II in the gene body at the start of treatment is not affected. These results indicate that Trp blocks transcription initiation and causes the time dependent clearing of Pol II from both the promoter and gene body, while FP's prominent effect is to block escape from the pause, causing a time dependent gene body clearance but enhanced pausing of Pol II at the promoter.

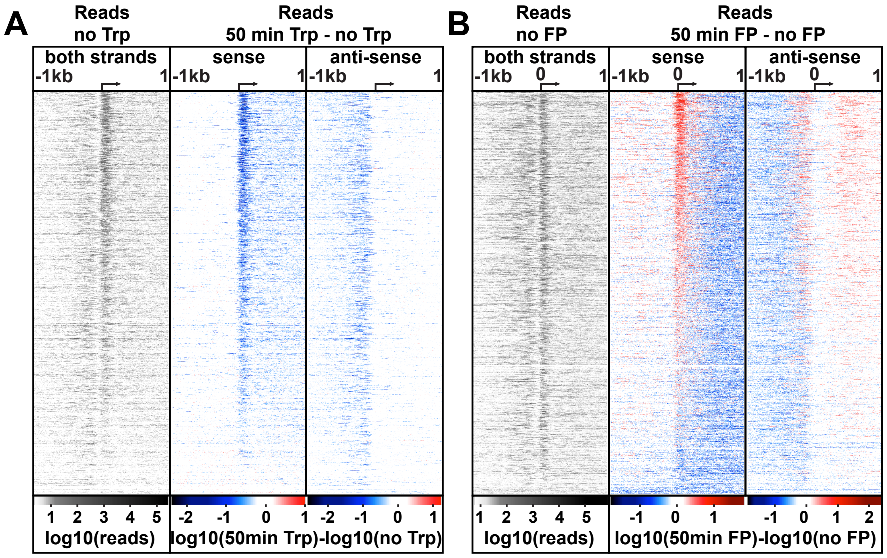


Figure 4.2. (**A**) Left panel shows a density plot of the log10 of reads of the no Trp dataset, in 10 bp windows +/-1 kb around TSS of 6380 active genes. Genes are ordered by the maximum decrease after Trp treatment at the promoter proximal region. The two right panels show the log10 difference in 10 bp windows after treatment with Trp for 50 min on the sense strand (middle) and antisense strand (right panel) with decrease in blue and increase in red. Genes are ordered as in the left panel. The density scales and color code for each panel are depicted the bottom. (**B**) same as in A, log10 reads in 10 bp windows around the TSS of the no FP control dataset (left panel), and the log10 difference in reads after 50 min FP treatment on the sense (middle panel) and antisense strand (right panel). Genes are ordered by maximum increase after FP treatment at the promoter proximal region. This Figure was generated by Dr. Iris Jonkers.

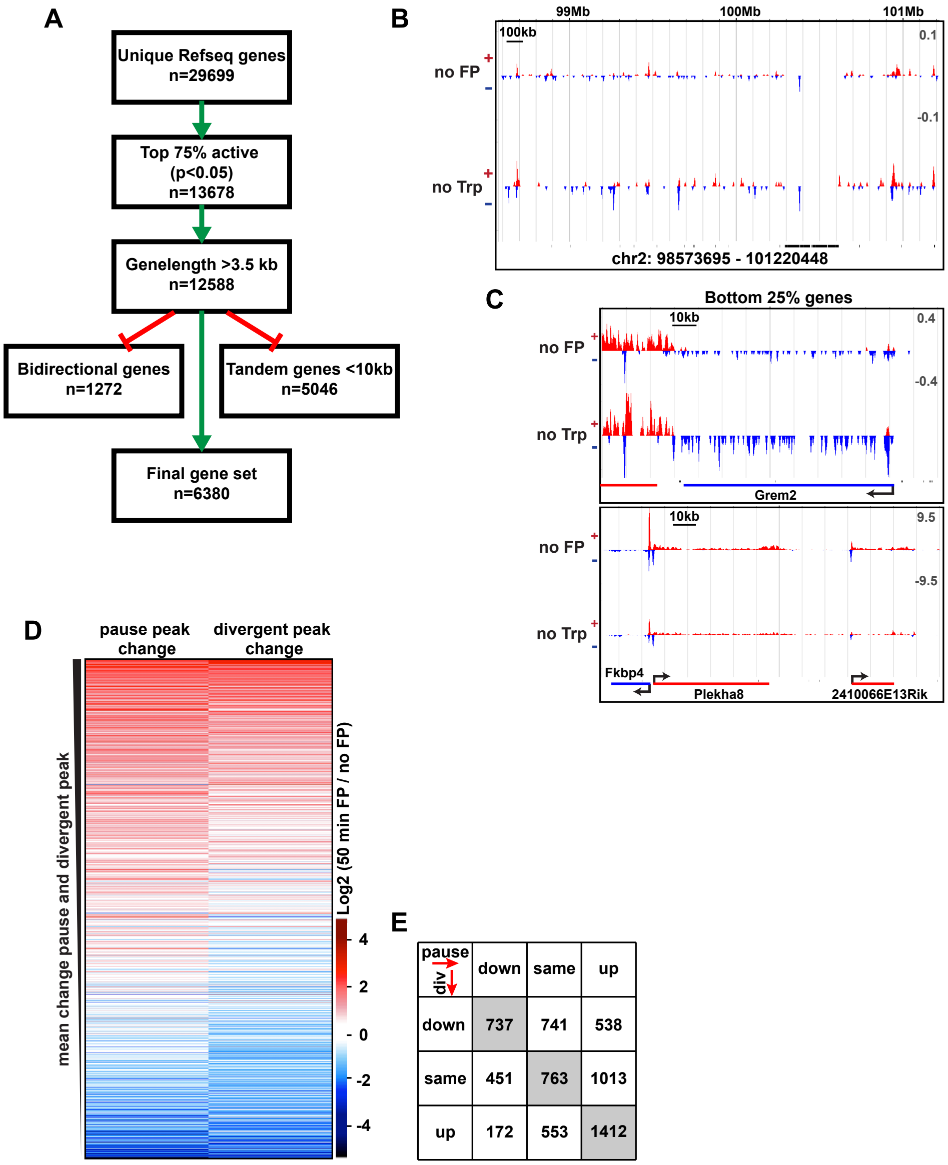


Figure 4.3. (**A**) Scheme of the selection of genes used for analysis of FP or Trp effect. (**B**) Example of a ~200 Mb genepoor region used to establish background transcription levels. (**C**) Screenshots of two example genes that are defined as active but are excluded from general FP and Trp analysis as they are within the bottom 25% of expressed genes. (**D**) Log2 of change within the promoter proximal region on the sense strand (pause peak) or antisense strand (divergent peak) after 50 min FP treatment. Color scale is on the right and genes are ordered by the maximum to minimum of average change within the pause and divergent peak of the genes. (**E**) Quantification of the number of genes that have a pause (top) or divergent (left) peak that has a 1.5 fold decrease (down), 1.5 fold increase (up) or no change larger than 1.5 fold (same) of normalized read density after 50 min of FP treatment.

To facilitate the quantitative analysis of transcription, we selected the top 75% most actively transcribed genes that are long enough (over 3.5kb, to obtain a good measure of gene body transcription), and well resolved (>10kb from another tandemly oriented gene and not a member of a bidirectional pair of genes)(Figure 4.3A; n=6,380). The heat maps in Figure 4.2A display the Trp-induced changes on individual members of this gene set. The leftmost heat map shows the Pol II density in the absence of drug, while the center and right maps show a drug induced change in density on the sense and anti-sense strands after 50 min. Trp causes a reduction in the promoter and downstream regions of both the annotated gene (sense strand), and the upstream divergent region (antisense strand). This upstream divergent transcription is a well-documented feature of mammalian promoters[8,32,33] (Figure 4.2A). These results indicate that Trp blocks the entry of Pol II on all promoters in both directions.

The results of FP inhibition on the promoter regions are in sharp contrast. Figure 2B shows that while the Pol II density in the gene body is decreased after FP treatment, the Pol II at the promoter proximal region and divergently transcribed region increase at the majority of genes (Figure 4.2B and Table 4.2). This is consistent with FP blocking the entry of Pol II into productive elongation. Notably, a smaller fraction (~20%) of genes displayed an unexpected decrease in paused and divergent Pol II peaks (Table 4.2, and Figure 4.2B), indicating that FP can also affect initiation on a smaller subset of genes. Generally, the paused and divergent peaks at a single gene change in concert (both up or both down) after FP treatment (Figure 4.3D and 4.3E), indicating that levels of paused and divergent Pol II at promoters are regulated similarly.

Significantly changed genes (p < 0.05)	12.5min Trp	25min Trp	50min Trp	2min FP	5min FP	12.5min FP	25min FP	50min FP
gene body decrease	4013 (62.9%)	5849 (91.7%)	6101 (95.6%)	4065 (63.7%)	5425 (85.0%)	6047 (94.8%)	6118 (95.9%)	6071 (95.2%)
gene body increase	178 (2.8%)	28 (0.4%)	7 (0.1%)	270 (4.2%)	25 (0.4%)	44 (0.7%)	40 (0.6%)	85 (1.3%)
pause peak decrease	4684 (73.4%)	5500 (86.2%)	5689 (89.2%)	1948 (30.5%)	1726 (27.1%)	1377 (21.6%)	1353 (21.2%)	1161 (18.2%)
pause peak increase	23 (0.4%)	8 (0.1%)	5 (0.1%)	499 (7.8%)	373 (5.8%)	1787 (28.0%)	3058 (47.9%)	3601 (56.4%)
divergent peak decrease	3733 (58.5%)	4735 (74.2%)	4914 (77.0%)	1373 (21.5%)	1512 (23.7%)	1358 (21.3%)	1848 (29.0%)	1629 (25.5%)
divergent peak increase	13 (0.2%)	9 (0.1%)	9 (0.1%)	524 (8.2%)	289 (4.5%)	1234 (19.3%)	2042 (32.3%)	2670 (41.8%)
pausing index decrease (no p-value)	5162 (80.9%)	4896 (76.7%)	4982 (78.1%)	2059 (32.3%)	1174 (18.4%)	431 (6.8%)	359 (5.6%)	320 (5.0%)
pausing index increase (no p-value)	1218 (19.1%)	1484 (23.3%)	1398 (21.9%)	4321 (67.7%)	5206 (81.6%)	5949 (93.2%)	6021 (94.4%)	6060 (95.0%)

Table 4.2. Number of genes with significant changes in GRO-seq profiles

Quantification of the decrease in read density in the gene body region from 1 to 3.5 kb shows that 96% of genes are significantly decreased after Trp treatment, while 95% of genes are decreased after FP treatment (Table 4.2). This demonstrates that inhibition of P-TEFb causes a block in the transcription cycle that is as universal as inhibition of initiation.

Measurement of the Elongation Rates and the Acceleration of Elongating Pol II in the Gene Body

Upon the addition of FP to block Pol II entry into productive elongation, Pol II that are already transcribing generate a clearly distinguishable wave of elongation. We could track the rate of this wave's progress at more than 1000 genes that are sufficiently long and actively transcribed (Figure 4.1D, 4.4A). Determining the distance traveled by the wave between two time points provides a measure of elongation rates. To do this in an unbiased and systematic way, we adapted a Hidden Markov Model (HMM) to detect the transition of the wave, i.e., the point in the gene where Pol II density changes (Figure 3B and S3B). Similarly, we tracked the wave in the time-course following Trp addition to ESCs as an independent strategy for blocking Pol II entry into elongation.

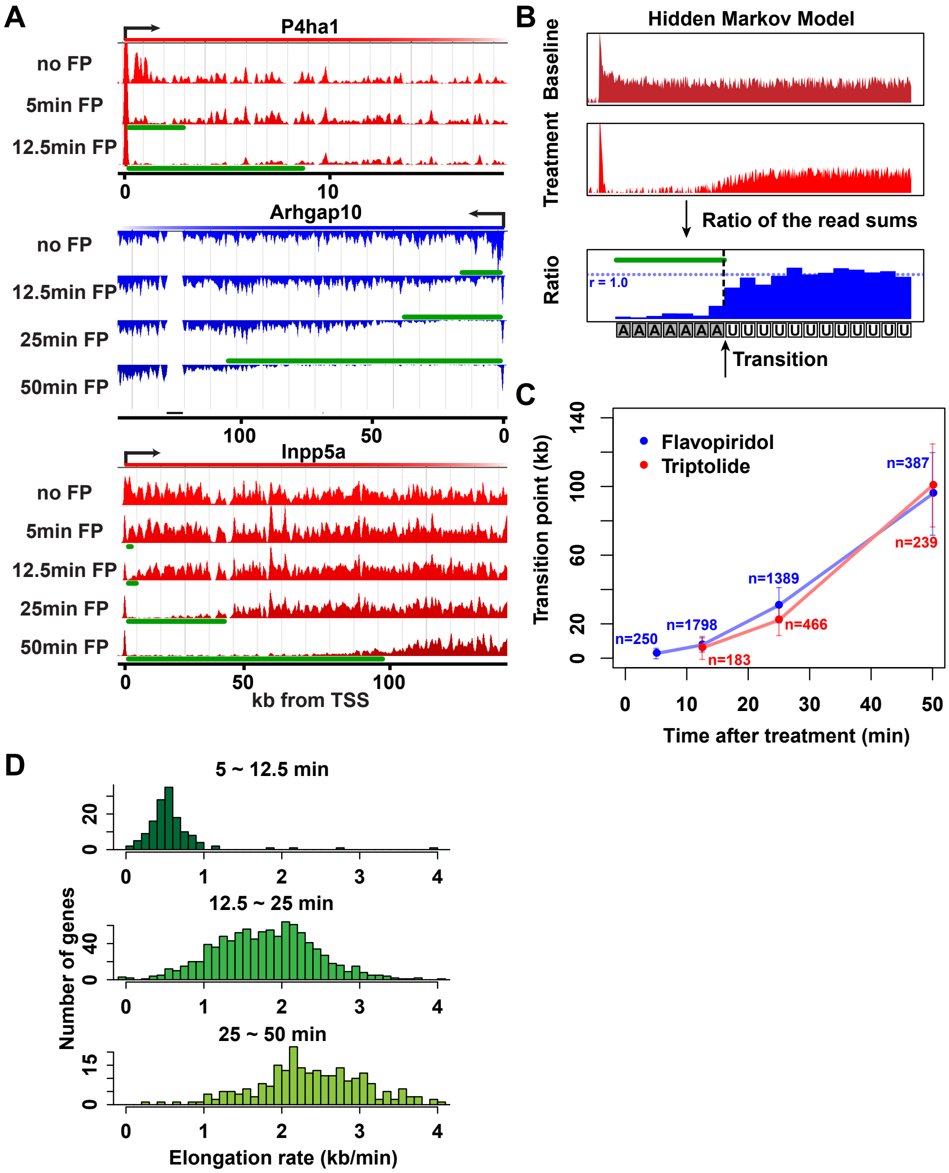


Figure 4.4. (**A**) Three representative genes used for measurement of transition points with the HMM. In green, the affected region after FP treatment as established by the HMM, the transition point being the endpoint of this region. (**B**) General scheme of the Hidden Markov Model: The ratio of the read sums of the normalized control and treated data in windows are taken, where a ratio near 1 is unaffected. The HMM runs from 5' to 3' through the gene and defines the most probabilistic point of transition from Affected (grey bins) to Unaffected (white bins). (**C**) Average of HMM derived transition points of the FP timecourse (blue) or Trp timecourse (red) plotted against the time of drug treatment. Error bars are standard deviation from the mean. (**D**) Elongation rates derived from the FP timecourse. Elongation rate is the ratio of the distance traveled in the time spanning 5 to 12.5 min (top; n=141), 12.5 to 25 min (middle; n=937) and 25 to 50 min (bottom; n=245).

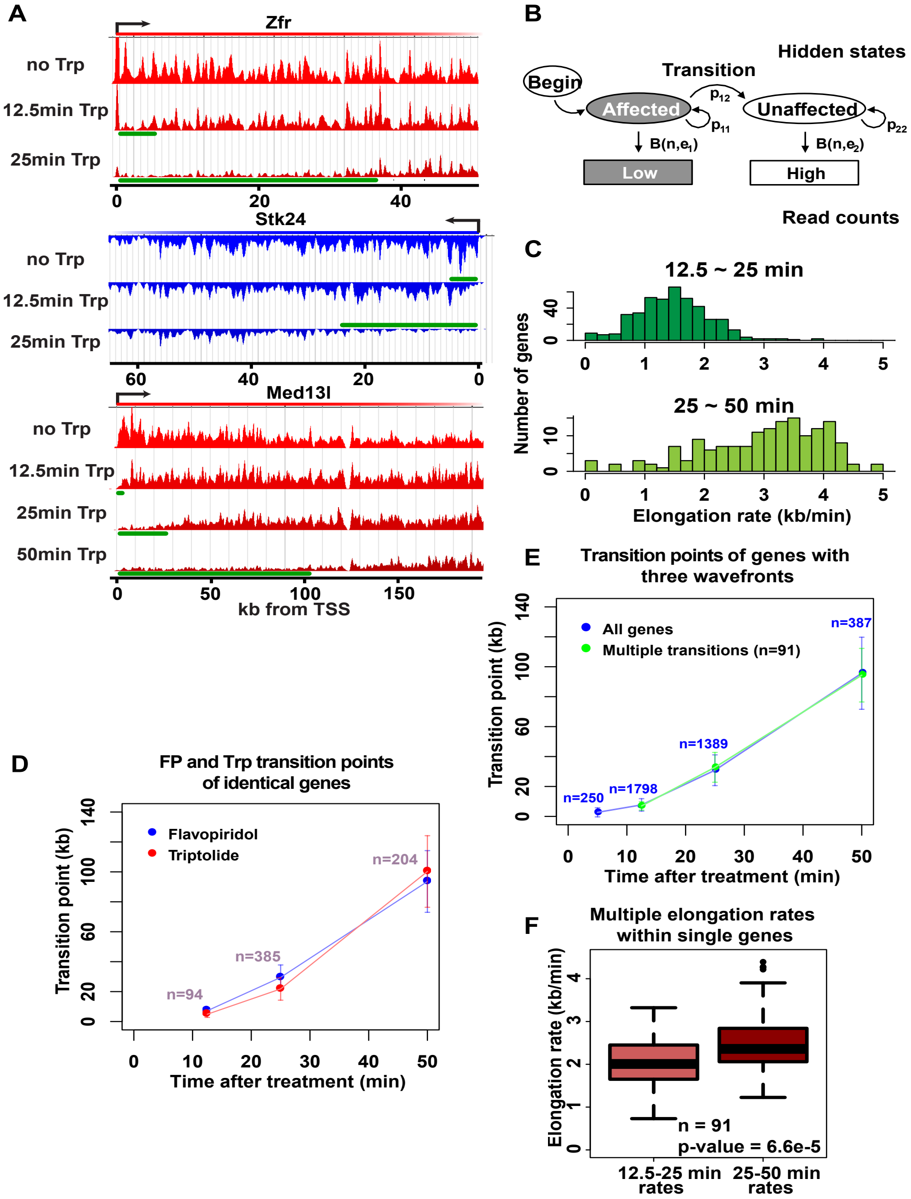


Figure 4.5. (**A**) Three representative genes used for measurement of transition points with the HMM. In green, the affected region after Trp treatment as established by the HMM, the transition point being the endpoint of this region. (**B**) Scheme of the Hidden Markov Model (HMM). (**C**) Elongation rates derived from the Trp timecourse. (**D**) Average of HMM derived transition points of the FP timecourse (blue) or Trp timecourse (red). (**E**) Average of HMM derived transition points of all the transition points in the FP timecourse (blue) or of genes (n=91) that had three consecutive transition points (green) plotted gainst the time of drug treatment. (**F**) Boxplot of the 12.5-25 min and 25-50 min elongation rates of genes that had both elongation rates (n = 91; same genes as in E).

Although the inhibition wave using Trp was not as robust (Figure 4.5A), and the HMM produced transition points for fewer genes than with the FP time-course (Figure 4.4C), the resulting clearing wave of Pol II had kinetics indistinguishable from the FP treatment (Figure 4.5C and D).

On average, the wave of Pol II travels about 100 kb during the 50 minutes of FP treatment. This corresponds to an elongation rate of about 2 kb/min, which is within the range of Pol II rates from previous studies[27,28]. The average Pol II inhibition wave travels equally far after Trp or FP treatment, indicating that inhibition of P-TEFb does not have a significant effect on elongation rates downstream of the promoter proximal region (Figure 4.4C). Also, the average transition points of genes with distinctive inhibition waves at three different time-points behaved the same as those of the genes that the HMM derived only one or two time-point data (Figure 4.5E), indicating that average transition points were not biased by how the groups of the genes were selected.

Interestingly, we found that the Pol II wave did not propagate linearly, but in an accelerating manner, both in FP and Trp time-courses (Figure 4.4D). Subsequently, elongation rates derived from different regions within the same gene increased further downstream (Figure 4.5F; p-value = 6.6×10^{-5}), showing that the increase in elongation rates happens within most genes and is not a consequence of elongation rate variation between genes. The mean elongation rates increase from 0.5 kb/min to 1.8 kb/min and then 2.4 kb/min for the intervals 5-12.5 min, 12.5-25 min, and 25-50 min respectively.

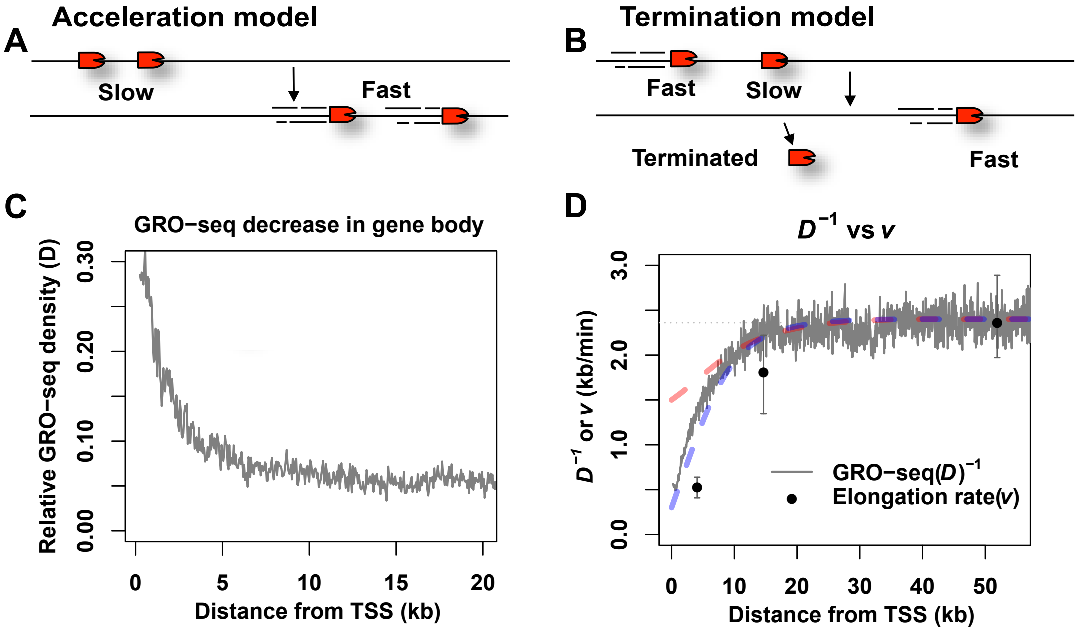


Figure 4.6. The apparent increase in elongation rate can be described by two models. (**A**) Acceleration model; Pol II becomes more transcriptionally competent and accelerates (**B**) Termination model; Population of Pol II consists of a relatively slow, poorly active fraction and a relatively fast, highly active fraction, and termination of the slow fraction leads to an apparent increase of elongation rate. (**C**) Composite profile of the no FP GRO-seq of genes >60 kb from the TSS until 20 kb from the TSS. Note that density remains constant after ~15 kb. (**D**) Inverse plot of the density composite profile in C (D^{-1} plot), scaled by the density of the constant downstream region. The red line indicates the projected average elongation rates considering parameters in line with a termination model; the blue line depicts the projected average elongation rates when the acceleration model is more likely. In black, the actual mean elongation rates at the average distance traveled are depicted, with error bars indicating the standard deviation.

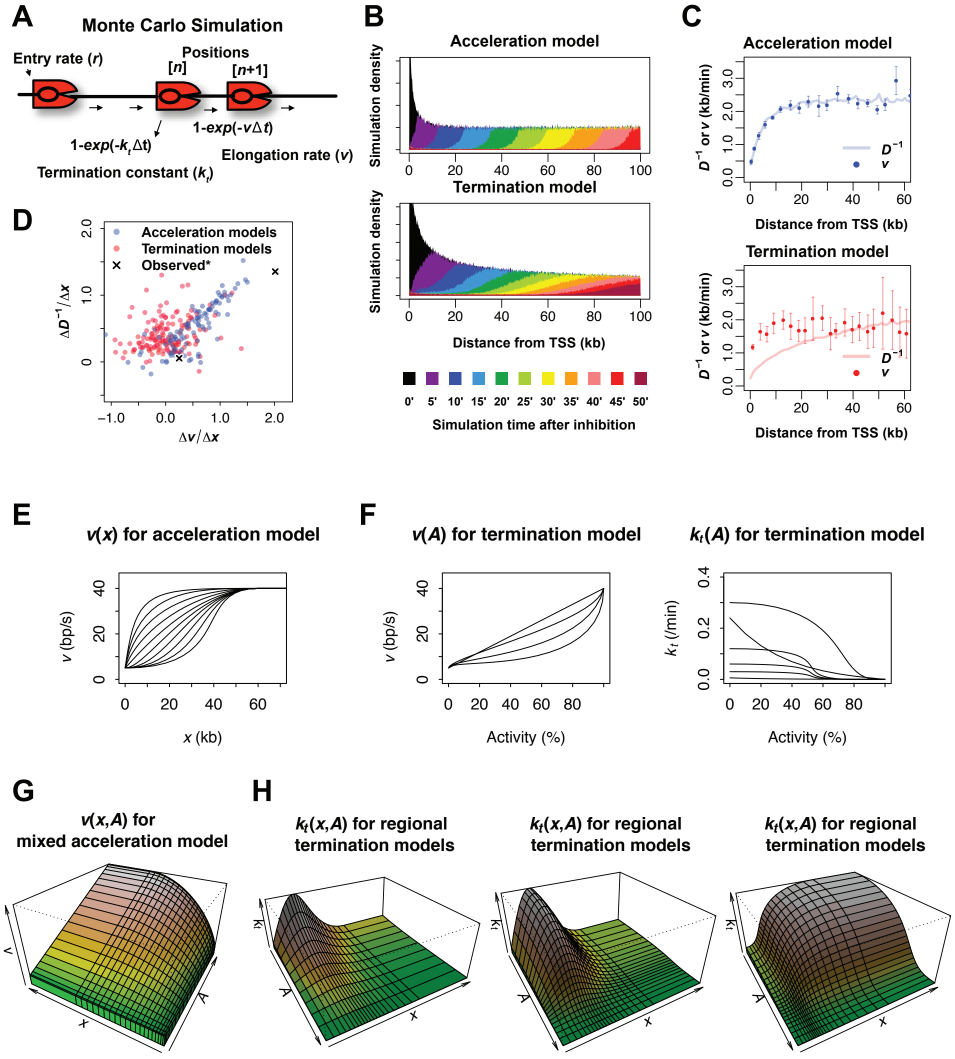


Figure 4.7. Modeling of the acceleration and the termination hypotheses. (**A**) Scheme of the elongation simulation. (**B**) Simulation time-course for an acceleration model and a termination model. (**C**) D^{-1} plot for the two models in panel B. (**D**) Scatterplot of the slope of v and the slope of D^{-1} for multiple acceleration models and termination models with varying parameters. The observed datapoints fit to the distribution of the acceleration models. (**E**) Elongation rate parameters for the acceleration models. (**F**) Elongation and termination rate parameters for the termination models. (**G**) Elongation rate parameter as a two dimensional function of position and activity in a mixed acceleration model. The scale of the v axis is 0-40 bp/s. (**H**) Termination rate parameters as functions of the position and activity in regional termination models. The scale of the k_t axes are 0-0.1s^{-1}, 0-0.125s^{-1}, 0-0.125s^{-1} for each subpanels. The scale of the x axis is 0-100kb, and A axis is 0-100% for all 2D subpanels.

Modeling Acceleration of Pol II in the Gene Body

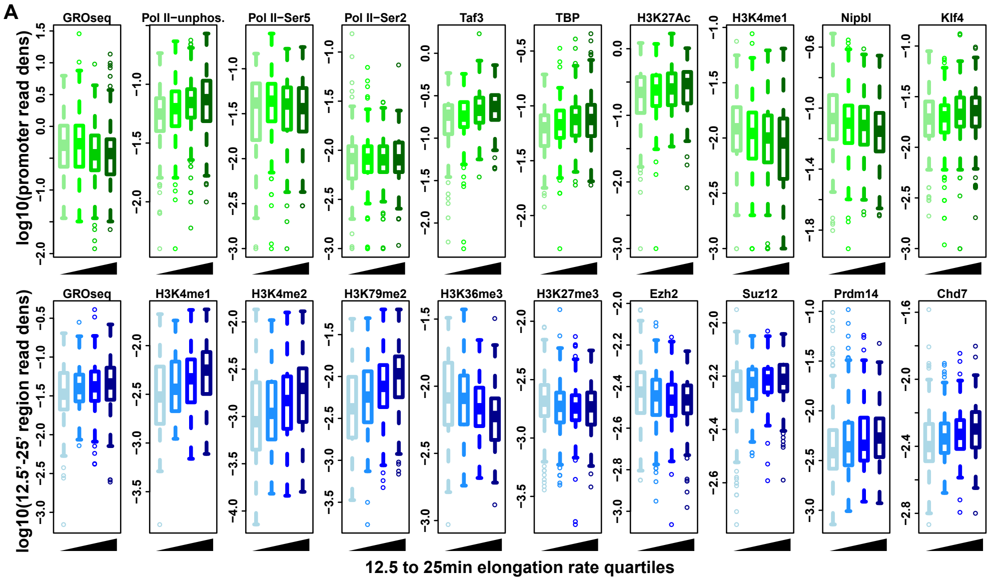
The apparent acceleration of Pol II in the gene body can be interpreted in two ways. The total population of elongating Pol II transcribes slowly at the 5’ end of the gene, but increases in speed along the gene body, ramping up to ~2.4 kb/min on average after traveling approximately 30 kb downstream of the TSS (Figure 4.6A). Alternatively, elongating Pol II could consist of different populations, fast, transcribing at maximum speed throughout the entire gene body, and slow. It has been recently shown that termination of Pol II in the gene body can have regulatory roles[34], and if the slow population terminates prematurely leaving the fast Pol IIs, then the overall apparent elongation rate at the 5’ end of the gene would be slower than at the 3’ end (Figure 4.6B). To distinguish between these two models, we compared the measured elongation rates vs. inferred elongation rates from the steady-state GRO-seq density within the gene. This comparison allows one to address whether premature termination of Pol II plays a significant role in the apparent elongation rates.

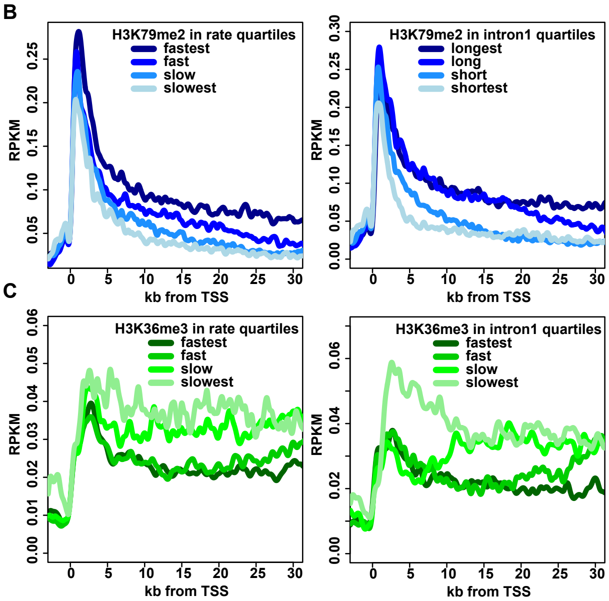
In the first model, because faster Pol II molecules will appear less dense on the GRO-seq profile, the elongation rate (v) will be proportional to the inverse of the GRO-seq density (D). Since D decreases to a constant level upon increasing distance from the TSS (x) (Figure 4.6C), the inverse of D (D^{-1}) can be adjusted by a scaling factor to match the increasing trend of v (Figure 4.6D; blue dashed line). On the other hand, if termination takes place, D will decrease more than what is explained by the acceleration alone, and the slope of D^{-1} curve will not match but will be steeper than the increase of v (Figure 4.6D; red dashed line). In other words, if premature termination plays a major role at the 5’ end of genes, than the increase in elongation rates will be less steep than the increase of D^{-1} in the same region. While measured average elongation rates at 5-12.5 min and 12.5-25 min fall slightly below this curve, the increasing trend is consistent with D^{-1} curve pattern (Figure 4.6D). This indicates that on average, the progressive increase in the elongation rate can largely explain the gradual decrease of the GRO-seq density and that termination of slowly transcribing Pol II populations plays a minor role in increasing the apparent elongation rate.

The acceleration model is further supported by a thorough examination of parameter space using a Monte Carlo simulation describing the dynamics of Pol II movement through the gene body (Figure 4.7A). We simulated the time course of the inhibition wave of Pol II and used the same HMM to define the transition points of inhibition waves and the elongation rates under the two model scenarios (Figure 4.7B). The relationship between D^{-1} and v shows a clear difference between the two models (Figure 4.7C and 4.7D). After scaling the simulated D^{-1} to converge to the simulated v, v falls right at the D^{-1} curve in the acceleration model, while v falls above the D^{-1} curve in the termination model. When we plotted the slope of v vs. the slope of D^{-1} in various simulated acceleration and termination models along with actual observations, the observation fits better with the distribution of the acceleration models than the termination models (Figure 4.7D).

Correlations between elongation rates and transcription factors, histone marks and modifiers

There is a wide variation in elongation rate of Pol II between genes (Figure 4.4D), and we set out to find factors that contribute to this variation by looking at correlations between elongation rates and existing ChIP-seq data of transcription factors, chromatin modifiers, and histone modifications (see for full list of factors Figure 4.8D). We used 12.5-25 min elongation rates (n=937), and divided the genes into quartiles.





D

ChIP-seq datasets	Reference
Pol II-Ser2, Pol II-Ser5, Pol II-60' DMSO, Pol II-60' FP, NelfA, Ctr9, Spt5	Rahl et al., 2010
Oct4, Sox2, Nanog, Tcf3, Suz12, H3K4me3, H3K36me3, H3K79me2	Marson et al., 2008
H3K4me1, H3K27Ac	Creyghton, 2010
c-Myc, n-Myc, Oct4, Klf4, Nanog, Sox2, Esrrb, p300, CTCF, Smad1, Tcfcp2l1, E2f1, Zfx	Chen et al., 2008
Cdk9, Ell2, Aff4	Lin et al., 2011
Ell	Smith et al., 2011
Ezh2, Suz12, Ring1B	Ku et al., 2008
Pol II, H4K4me3, H3K27me3, H3K36me3, H3K9me3, H4K20me3	Mikkelsen et al., 2007
H3K4me1, H3K4me2	Meissner et al., 2008
Smc1, Smc3, Med12, Med1, Nipbl	Kagey et al., 2010
TBP, Taf1, Taf3, Pol II-unphosphorylated CTD	Liu et al., 2011
Chd7	Schnetz et al., 2010
SetDB1, H3K9me3	Bilodeau et al., 2009
Dpy-30	Jiang et al., 2011
Atrx	Law et al., 2010
YY1	Mendenhall et al., 2010
Prdm14	Ma et al., 2010

Figure 4.8. (**A**) Boxplots of the log10 of read density of ChIP-seq datasets in promoter region (-750 to 250 bp from the TSS) in green, or in the 12.5-25 min transition region in blue in elongation rate quartiles from low rates in light shade to fast rates in dark shades. Only the most striking correlations of ChIP-seq data and 12.5-25 min elongation rates are shown. (**B**) Composite profile of H3K79me2 read density from -2 to 30 kb from the TSS in quartiles of 12.5-25 min elongation rates (left), or intron1 size quartiles (right). (**C**) As in C, but with H3K36me3 read density. (**D**) Table of ChIP-seq datasets and the associated references used to correlate with elongation rates. This Figure was generated by Dr. Iris Jonkers

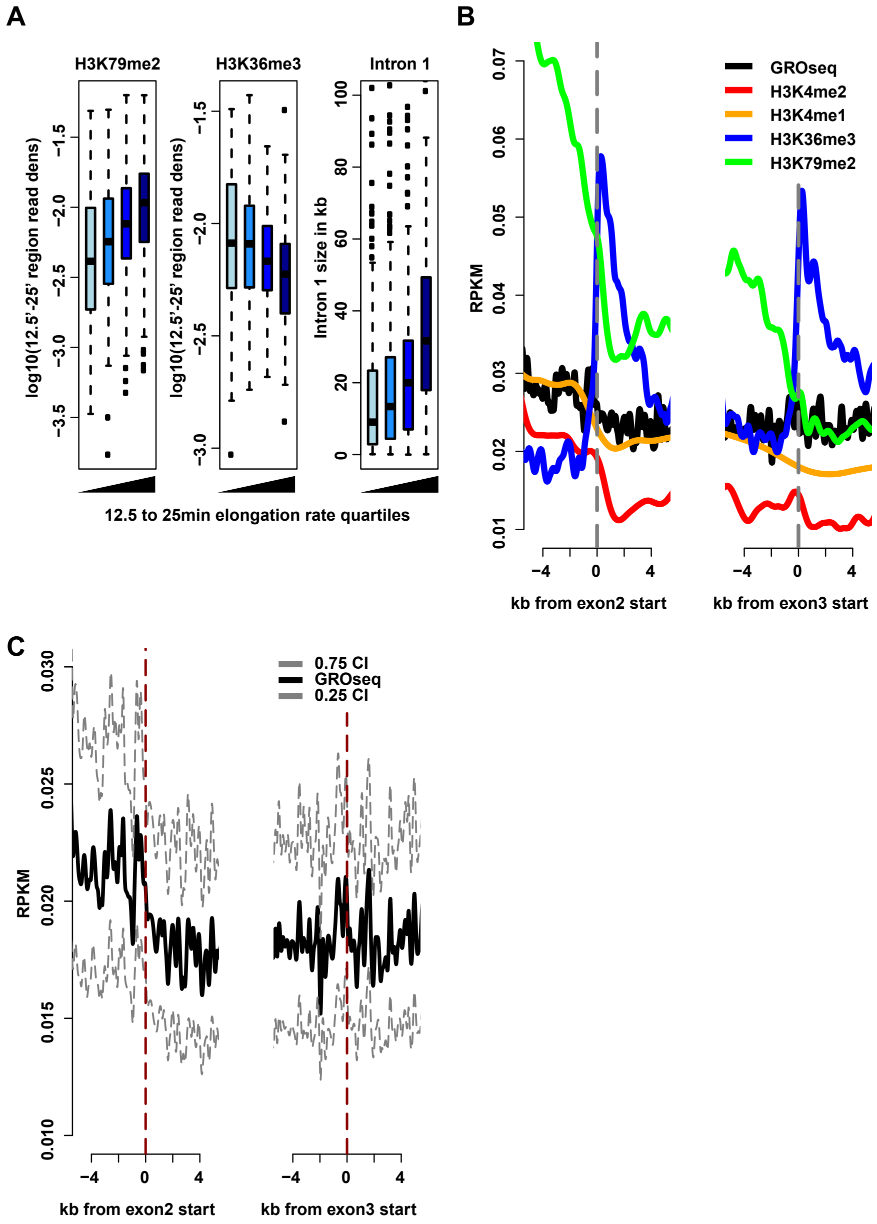


Figure 4.9. (**A**) Boxplots of H3K79me2 (left) and H3K36me3 (middle) log10 read density within the 12.5-25 min transit (early) region of genes corresponding to quartiles of the 12.5-25 min elongation rates. The right boxplot shows the size of the first intron in kb of genes within quartiles of the 12.5-25 min elongation rates. Elongation rate quartiles increase in speed from left to right. (**B**) Histone mark and no FP GRO-seq read density composite profiles of genes with 12.5-25 min elongation rates and an intron1 and 2 of >5 kb (n=400). Profiles are centered on intron1-exon2 or intron2-exon3 junctions. (**C**) The median of the subsampled no FP GRO-seq composite profile centered on intron1-exon2 and intron2-exon3 junctions, including 0.75 and 0.25 confidence intervals after subsampling. This Figure was made by Dr. Iris Jonkers

Then we looked for enrichment of factors within the early region (the region from the 12.5 to the 25 min transition points) of the individual genes, or in the promoter region (-750 to 250 bp around the TSS) and plotted the factor density between the quartiles (Figure 4.9A, 4.8A and data not shown). Although correlations between elongation rate and chromatin marks or transcription factors were modest in general (Spearman's rho < 0.3), several factors showed intriguing correlations.

Specific histone methyl marks showed a reasonably strong positive (H3K4me1, H3K4me2, and H3K79me2) or negative (H3K36me3) correlation with elongation rates (Figure 4.9A and 4.8A). Although H3K36me3 has been shown to be enriched in the gene body of actively transcribed genes, the enrichment is not uniform; H3K36me3 increases at the intron1-exon2 junction, while H3K79me2 decreases at the junction[35]. Because the early regions often contain intron1, we addressed whether elongation rate correlations with these histone marks were a consequence of intron1-exon2 position. Again, we divided the genes into quartiles based on the intron 1 sizes as well as the elongation rates. The composite profiles of the histone modifications between the quartiles reveales that the H3K79me2 (Figure 4.8B, right), H3K36me3 (Figure 4.8C, right) and H3K4me1 and 2 (data not shown) distributions depend strongly on the size of intron1. Indeed, elongation rates and intron1 size correlate better than any of the histone modifications (Spearman's rho = 0.31, Figure 4.9A).

Interestingly, we also observed that GRO-seq levels showed a similar drop at the intron1-exon2 junction (p-value = 0.008) that was not present at the intron2-exon3 junction (p-value = 0.58)(Figure 4.9C). GRO-seq density within a gene is a measure of elongation rate, as slow moving Pol II will increase in density, while fast moving Pol II decrease. Therefore, we interpret the drop in GRO-seq density as an increase in elongation rate at the intron1-exon2 junction. The apparent speed-up could be a consequence of the abrupt changes in chromatin structure at the intron-exon junction, or a conformational change of the transcription complex as a consequence of coupled cotranscriptional-splicing event at the first intron-exon.

Elongation Rates and Splicing

Elongation rates have long been implicated in regulation of splicing[21,22,26]. However, most evidence supporting the hypothesis that Pol II slows down around exons has been measured indirectly by looking at accumulation of Pol II around individual exons with ChIP[36,37]. Our observation that 12.5-25 min elongation rates tend to be higher in genes with large first introns could point to a direct correlation between splicing and transcription speed, because genes with large first introns will have a lower exon density and fewer splicing events within the early region. To examine in more detail whether exons have a direct effect on elongation rates, we plotted elongation rates against exon density of the early region. We saw a striking negative correlation (Figure 4.10A), and calculated that adding an exon to the early region resulted in a reduction of elongation rate of 60 bp/min (roughly 3% delay of 2kb/min) or a 20-25 sec delay (3% delay of 12.5-25min interval; p-value = 3×10^{-14}).

To further establish that elongation rates are directly related to splicing, and not to the presence of exons alone, we used RNA-seq data[38] to map exons, and looked for alternatively spliced exons within the early regions. Exons were defined as 'skipped' when read density in the exon was significantly (p-value <0.01) 3 fold lower than the average exon density of the gene, or 'over-included' when exon density was significantly 3 fold higher than the average. When looking at the splice sites of exons in active genes larger than 3.5 kb, we saw that over-included exons had a more conserved 3' polypyrimidine tract, which interacts with splice factor U2AF and enhances splice site recognition[39].

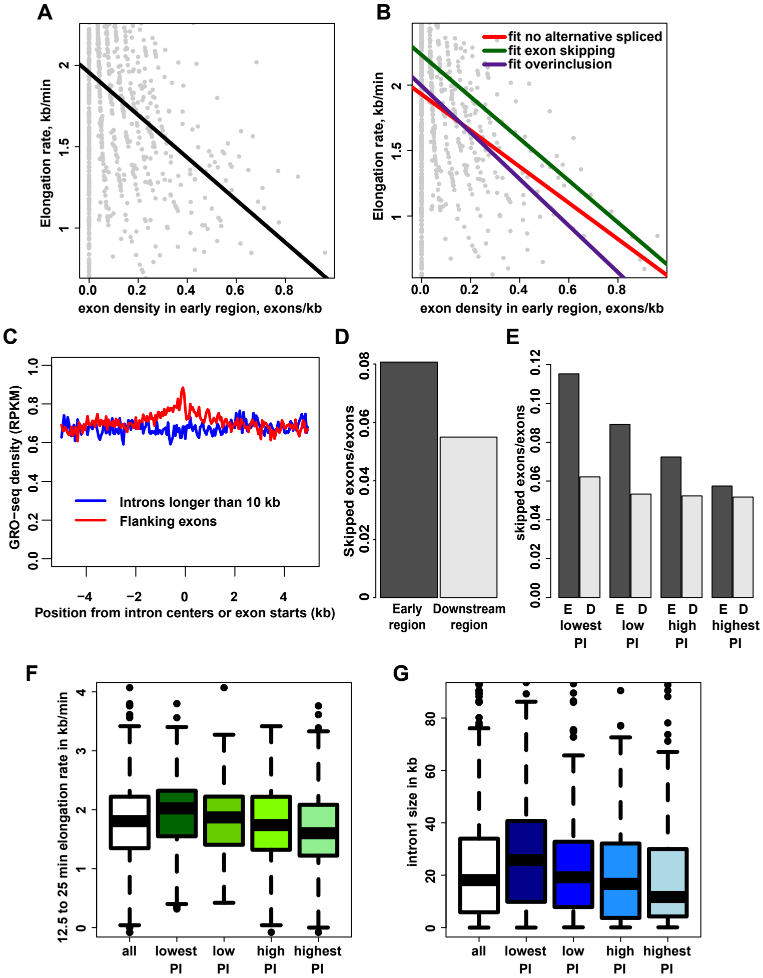


Figure 4.10. (**A**) 12.5-25 min elongation rates (n=937) are plotted as a function of exon density within the gene-specific 12.5-25 min transit region (grey points). Linear regression of the points is shown in black. (**B**) Same points are plotted as in A in grey, but linear regression is done over elongation rates taken over the transit region in which no exons were alternatively spliced (red; n= 759), exons were skipped (green; n=131), or exons were over-included (purple; n=50). (**C**) Composite profile of no FP GRO-seq centered on introns of > 20 kb (blue) and their surrounding exon start sites (red). (**D**) Skip ratio (skipped exon density /exon density) of all 12.5-25 min elongation rate genes (n=937) in the 12.5-25 min transit region (Early region; dark grey), or downstream of the 25 min transition point until 1 kb upstream of the stop (Downstream region; light grey). (**E**) Same as in D, but skip ratios are taken from 12.5-25 min rate genes in quartiles of pausing index, with Early (E) and Downstream (D) region skip ratios of the least paused genes left, and 5' and 3' region skip ratios of the most paused genes right. (**F**) Box plots of 12.5-25 min elongation rates of genes in pausing index quartiles (shades of green, with dark green being the least paused quartile) or of all 12.5-25 min rate genes (white; n=937). (**G**) Box plots of intron1 size in kb of genes in pausing quartiles (shades of blue, with dark blue being the least paused quartile) or of all 12.5-25 min rate genes (white; n=937).

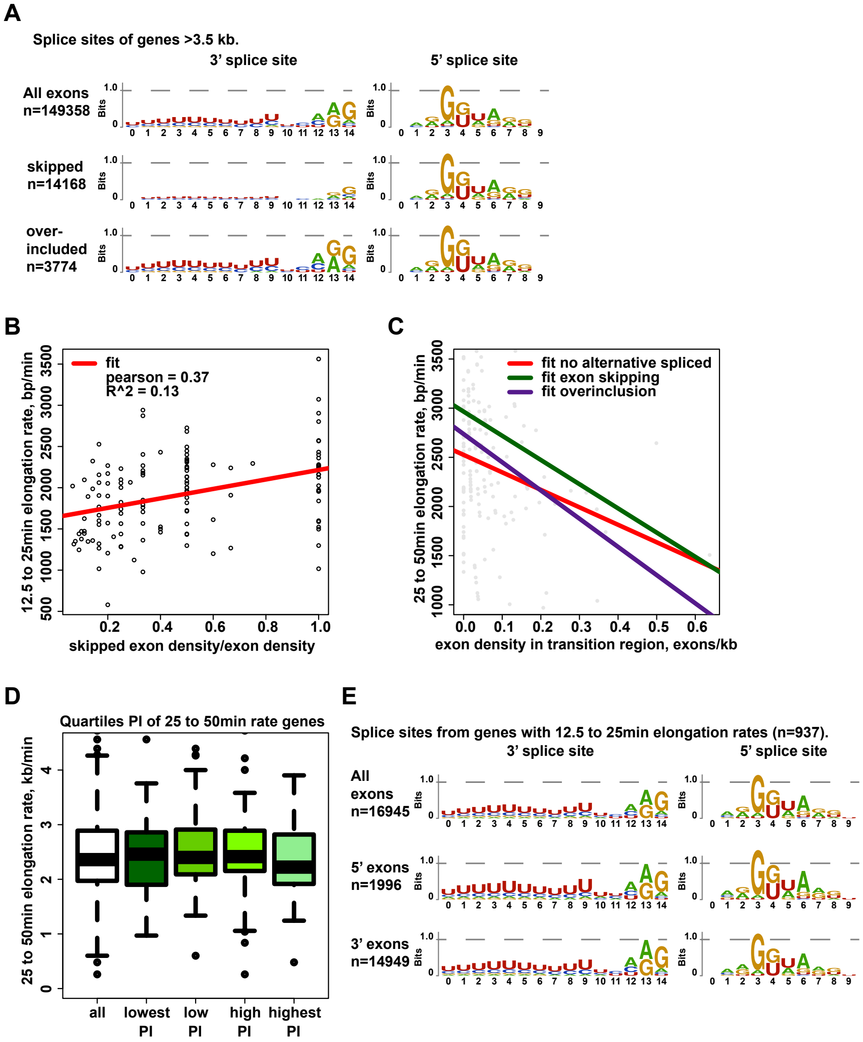


Figure 4.11. (**A**) Splice sites logo's of exons of top 75% expressed genes longer than 3.5 kb separated by all exons (top), skipped exons (middle), or over-included exons (bottom). (**B**) As Figure 6B, but with 25-50 min elongation rates and accompanying exon density in the region from the 25 to 50 min transition points. (**C**) Skip ratio in the early region of genes with a skip ratio > 0 plotted against 12.5-25 min elongation rate. Linear fit of the data points is plotted in red. (**D**) As Figure 6B, but with boxplots of 25-50 min elongation rates in pausing index quartiles of genes (n=245). (**E**) Splice site logo's of exons of 12.5-25 min elongation rate genes separated by all exons (top), exons located upstream of the 25 min transition point (middle, 5' exons), or downstream of the 25 min transition point (bottom, 3' exons).

On the other hand, skipped exons had a less conserved polypyrimidine tract (Figure 4.11A). Thus, splicing efficiency correlates generally with the quality of the splice sites in mESCs.

If splicing has a direct negative effect on elongation rates, than intervals with one or more skipped exons should have faster elongation rates, while over-included exons should have slower elongation rates. Indeed, effects of exon skipping and exon over-inclusion significantly contributed to elongation rates (p-value < 0.05; Figure 4.10B). Furthermore, elongation rates positively correlated with the extent of exon skipping within the early region as measured by the skip ratio (skipped exon density / exon density)(Figure 4.11B, $R^2 = 0.13$). These correlations could be found clearly within regions transited in the 12.5-25 min interval, but also in the 25-50 min interval, although results were less significant as a consequence of fewer measurements (Figure4.11C). Overall, these results show definitively that elongation rates directly correlate with splicing.

Reduced elongation rates as a result of splicing should also be reflected on the GRO-seq profile as an increase of read density at the exons. To measure the GRO-seq density of the introns that is not affected by adjacent exons, we selected the midpoints of the introns >10 kb. We compared these intronic GRO-seq density profiles to GRO-seq densities at the exons that are adjacent to these introns. As expected, the GRO-seq densities are increased around the exons compared to the introns (Figure 4.10C). Not only did we observe a local increase right at the exon-intron junction, but also a regional increase (~10%) in the ± 2kb region surrounding the exons. Together the increased Pol II density corresponds to about 12 seconds of delay (10% of the 2 min interval for Pol II to pass the 4 kb region) per exon, which corroborates reasonably well with the observed reduction of elongation rates as a result of transcription through an exon.

Pausing Increases Splicing Efficiency

Because slow elongation rates are associated with less exon skipping (Figure 1.10B and 4.11B and C), we hypothesized that fewer exons would be skipped at the 5' regions of genes, as elongation rates are significantly slower in this region (Figure 4.4D, 4.5E and F). Surprisingly, significantly more exons were skipped in the early region than further downstream in the gene (p-value = 3.4×10^{-5}; Figure 4.10D), pointing to less accurate co-transcriptional splicing close to the TSS.

To address this apparent paradox, we searched for factors near the TSS that correlate with improved splicing accuracy. We hypothesized that pausing could have an effect, as capping is thought to be facilitated by pausing[14,15,40]. The 5'-methyl-guanidine cap interacts with CBC, which has been hypothesized to improve splicing by recruiting splice factors[16-18]. To see if pausing has an effect on splicing fidelity, we looked at skip ratios in pausing index quartiles of genes with 12.5-25 min elongation rates (Figure 4.10E). Interestingly, the skip ratio in early regions of poorly paused genes is significantly higher compared to those of the highly paused genes (p-value = 0.003), and to those in the downstream regions (p-value = 5×10^{-4}). Nevertheless, skipping in downstream regions in genes of all pausing index quartiles is constant (Figure 4.10E). Thus, pausing correlates with fidelity of splicing of the first few exons. If splicing is more efficient in highly paused genes, elongation rates should be slower due to the reduction of Pol II speed at spliced exon regions. Indeed, the 12.5-25 min elongation rates in the highly paused gene quartile are clearly slower than in the poorly paused genes (Figure 4.10F; p-value = 1×10^{-7}). Furthermore, this effect of pausing on the elongation rate is reduced in downstream regions of genes (Figure 4.11D), where skip ratios are lower and relatively constant between pausing quartiles (Figure 4.10D and E).

While pausing seems to facilitate splicing in 5' regions of genes, another way to reduce splicing mistakes in this region could be by increasing splice site fidelity, or by moving exons out of the region close to the TSS. Indeed, 3' splice sites are more conserved in the upstream region of genes compared to downstream regions (Figure 4.11E). Furthermore, it has been shown that the first intron in mice is larger than average introns[41], although no satisfactory reason for this phenomenon has been proposed previously. We show that poorly paused genes have significantly larger first introns than highly paused genes (Figure 4.10G), consistent with the idea that selective pressure caused by erroneous splicing in poorly paused genes has lead to exons residing farther downstream from the TSS.

Discussion

P-TEFb kinase activity has been known to be critical for release of paused Pol II into productive elongation. By using GRO-seq to map the amount and orientation of transcribing polymerase at high sensitivity, we show that not just genes with pausing as a rate-limiting step are dependent on P-TEFb, but that all active Pol II undergoes an intermediate, P-TEFb dependent step from initiation to productive elongation. Furthermore, analysis of GRO-seq signals at times following P-TEFb inhibition allows highly sensitive measurement of elongation rates at ~1,000 genes, and within different regions of genes. Analysis of elongation rates shows that Pol II accelerates while transcribing in the gene body, and that elongation rates are inversely correlated to splicing. Furthermore, we report a link between promoter proximal pausing, elongation rates and splicing efficiency at the 5' region of genes.

Pol II undergoes a P-TEFb dependent pause escape step at all actively transcribed genes

Previous studies have shown that promoter proximal pausing of Pol II is a rate-limiting step at many genes in both Drosophila and mammals, and productive elongation depends on P-TEFb recruitment and its kinase activity[6,8,42]. Rahl et al., (2010) showed that about 75% of active genes show increased pausing after treatment with FP, and here we establish that at least 95% of actively transcribed genes undergo a P-TEFb regulated 'pause' step. In fact, inhibition of initiation by Trp has an effect equal to that of FP on gene bodies, showing that Pol II pausing and P-TEFb recruitment and kinase activity is an integral step of the transcription cycle of virtually all genes.

The widespread effect of FP on P-TEFb action is further exemplified by the dramatic loss of phosphorylation at the CTD after treatment with FP for 50 min. P-TEFb has generally been proposed to phosphorylate Ser2 of the CTD; however, we also observed an equal loss of Ser5 phosphorylation. One explanation to this could be the cross reactivity of the antibodies to Ser2 and Ser5, underrepresenting the complex and dynamic profile of the phosphorylated CTD[43,44]. Another possibility is that specificity of the CTD kinases may not be simple one-to-one projections. Recent evidence has indicated Brd4 as well as Cdk12 also phosphorylate Ser2 of the CTD and may have a redundant role together with P-TEFb[45,46]. Moreover, studies of P-TEFb kinase specificity that do not rely on antibodies point to Ser5 as its primary target[47,48]. Overall, the prevalent theory that P-TEFb phosphorylates only the Ser2 residue of the Pol II CTD may be an oversimplification.

Molecular basis of differences in elongation rates within genes

FP mediated inhibition of P-TEFb in a temporal fashion allowed us to measure elongation rates of Pol II. Because P-TEFb inhibition affects all active genes, we could determine over 1,200 individual elongation rates of Pol II in many long genes, often even at multiple regions within the same gene. Danko et al. (2013) has used GRO-seq following a rapid induction of transcription to examine the rate of movement of the Pol II wave front. Our elongation rates measured here after inhibition of P-TEFb led to slightly lower elongation rates in comparison and may point to a model in which lagging Pol II pushes leading Pol II thereby increasing overall elongation rates[49]. Indeed, elongation rates in highly expressed genes are faster compared to rates in poorly expressed genes in both studies (Figure 4.8B)[28].

Strikingly, both Danko et al. (2013) and we demonstrate that elongation rates of Pol II are apparently slower upstream compared to downstream in a gene. Furthermore, composite profiles of GRO-seq reads revealed that Pol II density is higher near the 5' end of the gene, indicating more sluggish transcription within this region. A simple acceleration model of Pol II elongation rate could explain this, which was supported by our simulation models. It could do so by accumulating positive elongation factors, by post-transcriptional modifications that facilitate elongation, or by encountering a more preferable chromatin state as it transcribes through the gene. Indeed, an apparent increase Pol II elongation rate can be seen at the intron1-exon2 junction, where dramatic changes in chromatin landscape take place, as well as the first co-transcriptional splicing event (Figure 4.9B and C). This suggests maturation of Pol II occurs, which starts at the promoter proximal pause step and increases as it transcribes through the gene and completes its first co-transcriptional splice event.

A recent study implies that Xrn2 mediated and decapping dependent termination takes place early on in the transcription of a gene[50]. Our data is not consistent with widespread Xrn2 mediated termination. Instead, we postulate that Xrn2 executes termination of a fraction of Pol II that fails to meet checkpoints at the pause site for escape into productive elongation or fails to efficiently splice the first intron.

Pausing, elongation rates and splicing

It is a longstanding hypothesis that Pol II slows down at exons to facilitate splicing[26]. The most direct evidence for this hypothesis comes from studies that use slow mutant forms of Pol II, in which skipped exons are more likely to become included[21,22]. Other studies that indicate elongation rates have an effect on alternative splicing use ChIP of Pol II at exons to draw conclusions[36,37]. However, a direct and quantitative correlation between splicing, exon skipping, and elongation rates at multiple genes has not been demonstrated. Here, we show directly that Pol II slows when co-transcriptional splicing occurs by showing a striking inverse correlation between our kinetically measured elongation rates over specific regions of genes and the density of spliced exons measured within these regions.

We also show that despite relatively slow elongation rates near the TSS, splicing is less efficient in this region compared to downstream regions, as indicated by increased exon skipping (Figure 4.10D). Interestingly, genes that have a high promoter proximal pause can splice exons more efficiently regardless of their position within the gene, while poorly paused genes display less efficient splicing at the 5’ end of the gene. This coincides with a slower elongation rates in highly paused genes, and higher elongation rates and lower exon density in 5’ regions of poorly paused genes. Promoter-proximal pausing of Pol II might serve as a time interval or checkpoint to ensure phosphorylation of the CTD, capping of nascent RNA and recruitment of splice factors[16-18]. Capping occurs during pausing[14,15], and the CTD is phosphorylated both before the pause and during escape from the pause, thereby creating a pause dependent platform for the recruitment of splice factors. Recently, the SR protein class of splice factors has been reported to interact with the inactive 7SK-RNA P-TEFb complex, and it was proposed that these splice factors have a role in recruiting P-TEFb to paused Pol II[51]. Naturally, the SR proteins may also assist early spliceosome formation at highly paused genes.

Together, we provide direct evidence that splicing and elongation rate are highly correlated. Also, we show that splicing is less efficient at the first few exons, especially in genes without a rate-limiting step at the promoter proximal region, providing a first link between pausing of Pol II and co-transcriptional splicing. By defining the elongation rates genome-wide, this study takes great strides to elucidate what processes can directly influence elongation rate. The collection of genome-wide data for various factors in mouse ES cell is growing. Analysis of our data in conjunction with these more comprehensive datasets in the future is warranted to identify the splicing and elongation factors and underlying mechanism that interplay with elongation rate of Pol II genome-wide.

Material and methods

Cell culture, nuclei isolation and GRO-seq library preparation (This part was carried out by Dr. Iris Jonkers)

Cell culturing of the V6.5 mES cell line was done as described previously, and drug treatment was performed on pre-plated mES cells to remove irradiated MEF-feeder cells, grown for one passage on 15 cm^2 plates up to ~70% confluence before isolation of nuclei. Drugs were added to the cells by replacing ES medium with pre-heated ES medium containing 300 nM FP, 500 nM Trp or 0.0125% DMSO as no Trp control. Timed incubation of drug treatment was done at 37°C, after which cells were put on ice and rinsed twice with ice-cold PBS, pH 7.4, followed by immediate nuclei isolation. Nuclei isolation, nuclear run-on and nascent RNA library preparation was performed as described previously[8]. In brief, after rinsing the 15 cm^2 plates with drug-treated cells, 15 ml cell lysis buffer (10 mM Tris-Cl, pH 7.5, 300 mM Sucrose, 3mM $CaCl_2$, 2 mM $MgAc_2$, 0.5% NP-40, 5 mM DTT, 1 mM PMSF, protease inhibitors) was added to the plates and the cells were scraped off, and spun down at 4°C and 185 xg for 5 min in a GS-6R Beckman swing-bucket centrifuge, after which supernatant was discarded. 5 ml fresh cell lysis buffer was added and cells were dounced 50 times in a 5 ml douncer on ice and spun down at 328 xg for 5 min, after which supernatant was discarded and nuclei were taken up in ~250 µl of glycerol storage buffer (50 mM Tris-Cl, pH 8.3, 40% glycerol, 0.1 mM EDTA, 5 mM $MgAc_2$, 5 mM DTT, 1 mM PMSF, protease inhibitors) and snap frozen.

For each nuclear run-on (NRO), 10^7 nuclei were used. An NRO master mix was added containing nucleotides ATP, GTP, CTP and nucleotide analogue bromo-UTP, as well as sarkosyl and 300 mM KCl to remove all impediments that pause Pol II and prevent NRO and to prevent re-initiation of Pol II. The NRO was performed at 37°C for 5 min, which results on ~100 nt addition to the nascent RNA. Immediately after the NRO, a population of in-vitro transcribed spike-in RNAs with and without Br-UTP was added to the nascent RNA as a way to assess experimental differences between each library, and to use as a normalization tool between timepoints within each timecourse. The nascent RNA was fragmented to ~150 nts and NRO BrU-RNA was isolated 3 consecutive times with BrdU-antibody beads (sc-32323, Santa Cruz), with enzymatic TAP and PNK treatments to remove the cap and 3'-phosphate and to add a 5'-phosphate, as well as Illumina adaptor ligations between the BrU-RNA isolation steps. The three consecutive isolation steps lead to an approximate 500.000x enrichment of BrU-RNA over background RNA. BrU-RNA was reverse transcribed, amplified, barcoded and sent for sequencing. Each dataset was done in replicate.

Sequence alignments and normalization (This part was done by Dr. Iris Jonkers)

All the GRO-seq libraries were sequencing in 50 nt runs on the Illumina HiSeq and split by barcode. Sequences that contained the Illumina adaptor were clipped with the cutadapt tool (https://code.google.com/p/cutadapt/) and remaining sequences were trimmed down to 32-mers, aligned to the ribosomal genome to remove rRNAs, and then aligned uniquely to the mm9 reference genome, with up to 2 mismatches with bowtie (http://bowtie-bio.sourceforge.net/index.shtml). Replicates were highly correlated and were pooled for further analysis, with exception of one of the 25 min Trp-treated #1 replicate, which suffered from extensive RNA degradation during the library preparation, and was subsequently discarded. Spike-in RNA controls were aligned similarly to a spike-in reference genome. The number of read counts of individual datasets differed significantly between timepoints and treatments, but instead of using total read counts to normalize between treatments within either the FP or Trp timecourse, we chose to normalize with the spike-in RNA read counts instead. This was done because we expect the overall Pol II transcriptional activity of cells to decrease significantly after treatment with either FP or Trp, resulting in an overrepresentation of Pol II activity in the 25 and 50 min timepoints if we would normalize to total reads. The general decrease in Pol II transcriptional activity is reflected in the increasing percentage of rRNA reads in each timecourse (Table 4.1). Sequencing and alignment information for each library is presented in Table 4.1.

Analysis of FP and Trp drug treatment effects in time on the genome-wide Pol II distribution (This part was done by Dr. Iris Jonkers)

Calculation of gene activity, pausing index, pause peak locations and enrichment, divergent peak location and enrichment, and gene body enrichment were done largely as in Core e tal., with small modifications. The mm9 RefSeq genelist was used as reference genelist for all analysis and contained 29,699 unique gene entries mappable to specific locations on chromosomes. To assess the general pattern of Pol II distribution after drug treatment we made composite profiles. Composite profile plots were made by taking all genes >12.5 kb or >150 kb and taking the average read density from the coding strand in 50 bp windows in the region from -5,000 to +12,500 bp, or +10 kb to +150 kb from the TSS, respectively. Profiles were plotted in R and smoothed with smooth.spline.

To establish which genes were active and enriched above background, we mapped reads in genepoor regions (Figure 4.3B) of the mouse genome from the no FP and no Trp control datasets, took the average read density of these two datasets and multiplied it 10x as a safe threshold for background read density (5×10^{-4} reads/bp). Unmappable regions of the genome were identified and excluded by aligning the genome to itself in 30-mers and sequence reads mapping to these regions were not used in analysis. Genes were called active when they were enriched within mappable regions above the 5×10^{-4} reads/bp background cut-off with a p-value < 0.05. To ascertain whether these thresholds were reasonable we looked at screen shots of some of the least active genes and saw that most poorly transcribed genes could clearly be distinguished from background (Figure 4.3C). However, some genes that were characterized as active were the result of artifacts such as spurious peaks within the coding region, or read-through transcription of upstream tandem genes. To ensure that we used a clean gene set for further analysis of changes at the pause and divergent peaks as well as changes within the gene body, we defined a genelist based on the following criteria (Figure 4.3A): The genelist for analysis of FP and Trp sensitive genes contains the top 75% of active genes larger than 3.5 kb present in both the FP as the Trp control dataset, from which genes that have a TSS within 1000 bp on the opposite strand of its own TSS (bidirectional genes) and genes that have a stop site upstream of its TSS within 10 kb (tandem genes) are excluded (Figure 4.3A). This stringently selected list contained 6,380 genes.

Pause peak locations were found by searching for maximum coding strand read density in 10 bp windows from -500 to +500 around the annotated TSS of the 6,380 selected genes. The promoter proximal pause site was defined as a 250 bp region centered on the maximum 10 bp window. The same was done for the divergent region, but the search area was enlarged to -1000 to +500 bp from the TSS and only non-coding sequence reads were taken into account. Gene body regions were defined as starting +1,000 bp downstream of the annotated TSS and ending at the annotated stop. The pausing index was calculated by taking the ratio of the read density in the pause region and the gene body region, and genes were called paused when read density in the pause region was significantly higher than in the gene body region (Fisher Exact, p-value < 0.01). Changes of the pause peaks and divergent peaks after treatment with either FP or Trp was established by counting the reads mapped within the promoter proximal pause and divergent regions at each timepoint, adding a pseudo-count, calculating the density in the mappable region at the promoter and taking the log2 of the ratio of the read density after treatment and the no FP or Trp control. The change was significant if the Fisher Exact p-value calculated over read counts within the pause or divergent regions before and after treatment was < 0.05.

During the relatively short treatments with FP and Trp, the effects of the drugs could be observed close to the TSS, but not further downstream in the body of the gene (Figure 4.1B, C and D). The unaffected downstream regions would thus camouflage the effect the drugs have on the gene body and therefore deflate our analysis. Therefore we decided to look at gene body effects of the drug treatments in the region from +1,000 to 3,500 bp of the TSS of the 6,380 selected genes only, to ensure the results were not affected by the promoter proximal pause region, while still looking at the region where drug-related effects could be expected from as early as 2 to 5 min onwards. Reads were counted within the 1-3.5 kb region for each dataset, a pseudo-count was added and read density and ratios within the mappable region were calculated as above. The change in gene body density was significant if the Fisher Exact p-value was < 0.05 for read counts within gene body regions between control and treated datasets.

Density plots in the -1 to +1 kb region around the TSS of the 6,380 selected genes were made by counting reads in 10 bp windows from -1,000 to +1,000 from the annotated TSS, adding a pseudo-count to each window and taking the log10 of the read count. This was done separately for the coding and non-coding strands for the no FP and no Trp controls as well as for the 50 min FP and Trp treated datasets. The overall Pol II distribution on both strands is the sum of the log10 read count of the coding and non-coding 10 bp windows. The increase or decrease of read counts within each 10 bp window was calculated by subtracting the no FP or Trp log10 read count from the 50 min FP or Trp read count. Genes in the density plots were ordered by the maximum decrease or increase after 50 min Trp or FP treatment in read density at the promoter proximal pause peak, respectively. The density in windows was plotting using heatmap.2 of the R packages gplots together with RColorBrewer.

Protein fractionation and western blot analysis (This part was done by Dr. Iris Jonkers)

Extractions of chromatin bound and unbound proteins were done after treatment of pre-plated mES cells for 50 min with 300 nM FP, 500 nM Trp or 0.0125% DMSO as no Trp control. Cells were rinsed twice in ice-cold PBS, pH 7.4, scraped off the plates and spun down for 5 min at 1,000x g. Cells were resuspended in nuclei lysis buffer (20mM Tris-Cl pH 7.5, 3mM EDTA, 10% glycerol, 150mM KAc, 1.5mM $MgCl_2$, 1mM DTT, 0.1% NP-40, and phosphatase and protease inhibitors), dounced 60 times on ice and centrifuged at 13,000 rpm for 5 min at 4°C. Supernatant was snap frozen as the unbound fraction, while the remaining pellet was resuspended in the same volume of nuclei lysis buffer as above and sonicated to break up the chromatin and help solubilize the pellet. The chromatin fraction was also snap frozen until further use. Western blot analysis was done in triplicate with antibodies against the Ser5 (3E8, Millipore) or Ser2 (3E10, Millipore) phosphorylated CTD, or N-terminal Pol II (N-20, Santa Cruz).

Estimation of the inhibition wave-front and elongation rates using Hidden Markov Model (HMM) (This part was done by Hojoong Kwak)

In summary, we selected genes that are long enough, start at the annotated TSS, and do not have interfering transcription in the middle of the gene bodies for the determination of elongation rates. Then we used an HMM to identify the position of the inhibition wave fronts at each time-point for individual genes. These positions of transition points are used to calculate the elongation rate between the time points. First, we selected genes that are longer than 30kb for 5min to 12.5 min, 60kb for 12.5 min to 25 min, and 150kb for 25 min to 50 min analysis. Then we removed the genes that do not have corresponding transcription units at the annotated TSS or contain intragenic transcription units defined by a genome-wide HMM (script named ‘tar2’, described below). Also, genes that have premature termination before the annotated 3’ ends and/or the 60kb/150kb mark are removed using a regional HMM (script named ‘tar7’, described below). After filtering, the number of selected genes are n=4,461 for genes longer than 30kb, n=2,769 for genes longer than 60kb, and n=571 for genes longer than 150kb.

We used a HMM algorithm (script named ‘hmm2’, described below) for each replicate of GRO-seq Flavopirodol (FP) and Triptolide (Trp) time courses and determined the transition points from the drug affected (inhibited) region to the drug unaffected (uninhibited) region of the gene body. We removed the genes that have the replicate transition points more than 20% different from the average of the two. For FP data after this filtering, we had 250 genes for 5min transition point and 1,798 genes for 12.5 min transition points out of 4,461 genes that are longer than 30 kb. Similarly, 1,389 genes remained from 2,769 genes longer than 60 kb and 387 genes out of 571 genes remained from genes longer than 150 kb. For Trp data, the HMM worked less efficiently, and 183, 466, and 239 genes passed this filtering from the >30 kb, >60kb and >150kb gene lists. The elongation rates were calculated using the differences in the transition points. For FP, n=141 for rates between 5 and 12.5 min, n=937 between 12.5 and 25 min, and n=245 between 25 and 50 min. For Trp, n=432 between 12.5 and 25 min, and n=157 between 25 and 50 min.

The followings are the descriptions of the HMM scripts.

1. hmm2

For the selected genes, GRO-seq read coverage from the TSS to 60kb or 150kb positions are binned for each time-point, and were divided by the untreated read counts at the same bins (Figure 3B). The bins sizes are 500bp, 1kb, 2kb, and 5 kb respectively for analyzing 5min, 12.5min, 25min and 50min time points. These sizes were selected so that each gene will have approximately 30-40 bins throughout the gene body regions, which was the optimal number for the HMM results. For each gene, the ratios in individual bins were internally normalized, by dividing by the average ratios of last 5 bins at the 3' ends. These bins are considered as the Markov process with the ratios as the observations (Figure 4.5B). For efficient HMM calculation, the normalized ratios were digitized, ranging from 0 to 2.0, with the step size of 0.05. Therefore, each binned position contains can have 20 observed states of the read ratio, and the probability of each state follows a binominal distribution of $B(n, e)$, where n=20 and e is the emission probability. We assumed 2 hidden states 'inhibition affected' and 'inhibition unaffected', with two emission probabilities e_1 and e_2. The transition probabilities between the 'affected' and 'unaffected' states are p_{11}, p_{12}, and p_{21} (Figure 4.5B), and are unidirectional. The Baum-Welch algorithm is used to estimate the transition and emission probabilities, by iterative calculations until the probabilities converge. If the iteration is over 200 cycles without convergence, we dropped the gene from further analysis. The transition point is calculated from these probabilities.

2. tar2

The untreated GRO-seq data is used similarly as hmm2 with some modifications. First, instead of using ratio relative to the reference dataset, we used binary observation values for each 200bp bin of the Markov process; 1 if there is a GRO-seq read within the bin and 0 if there are no reads. In this case, the binomial distribution becomes $B(2, e)$. Second, the full length of each chromosome was used rather than individual genes. Third, the Baum-Welch algorithm was allowed to run up to 1,000 iterations. Finally, a Viterbi path was calculated to define transcriptionally active regions.

Using this de-novo transcription unit (TU) calling, we selected TUs that have both sense and divergent pairs starting within 2 kb from each other. These paired TUs can indicate annotated TSSs, unannotated TSSs, or other regulatory transcription activities such as lncRNA or enhancer transcripts. We compared these TUs to the annotated long genes and removed the genes that did not have paired TUs near the annotated TSS (<2 kb). Also, we dropped genes that contain divergent TUs within the gene body, since these paired sense TUs arising at the middle of the gene bodies can indicate alternative start sites that may interfere with the transcription signal arising from the annotated TSS of the gene.

3. tar7

Using untreated GRO-seq data, we generated binary bins from TSSs defined by tar2 to 50 kb downstream of the annotated 3' end in each individual gene, and used these values to estimate the probability parameters of the HMM. This is similar to tar2 in that it uses binary observation values, but also similar to hmm2 in that it is done in individual genes. One difference from hmm2 is that it starts from the active region and detects the transition point into the inactive region or in other words the end of the TU, which is opposite of hmm2. Genes that have the transition point before the 60kb/150kb point are dropped, as this may interfere with detection of the inhibition wave by hmm2.

Kinetic Monte Carlo simulation of the acceleration and the termination models (This part was done by Hojoong Kwak)

The dynamics of elongating Pol II in gene body is simulated using a newly designed modeling program to describe Pol II transcription through a gene (Figure 4.7A). First, we modeled a Pol II transcription complex entering the gene body region with an entry rate (*r*) as a function of time. For the steady state assumption, *r* is a constant over time (t), while for the simulation of the inhibition wave, r is an exponential decay function of t.

Each Pol II molecule was generated with the rate r, and has a randomly assigned activity parameter (A), from 0 to 100 as a percentile. This activity parameter A is an intrinsic value of each Pol II molecule that determines the efficiencies of the elongation and the termination rates. The termination constant (k_t) and the elongation rate (v) are the functions dependent on the activity (A) as well as the position (x) within the gene body of the Pol II molecule. For instance, in a simple acceleration model where all Pol II molecules accelerate uniformly, $k_t = 0$ for all A and x, while v is an increasing function of x but a constant function for A. In a termination model, intrinsically active polymerase molecules elongate faster while less active ones elongate at a slower rate and terminate more frequently. In this case, k_t is a decreasing function of A, and v is an increasing function of A regardless of x.

For running the simulation, entry, termination, and progression events are assessed after each time increment of Δt. For each event, a pseudorandom number between 0 and 1 is generated and compared to the probability of initiation as described by $1\text{-}exp(-r\Delta t)$, termination as described by $1\text{-}exp(-k_t\Delta t)$, and processive elongation as described by $1\text{-}exp(-v\Delta t)$ respectively. If the number is less than the probability of any of the processes, the status of the polymerase changes accordingly. For the approximation of the progression event, the polymerase can move k bases following the Poisson distribution if the pseudorandom number is in the range $[F(k; \lambda), F(k+1; \lambda))$, where F is the cumulative Poisson distribution function and $\lambda = v\Delta t$. If there is a collision event between two polymerases, the leading polymerase terminates.

The distribution of simulated Pol II in N=1,000 DNA templates are equilibrated for 10,000s. The average Pol II distribution at this point is recorded as D. Upon the simulation of the decay of entry, average Pol II distribution is recorded every

100s over 100,000 bp region. The average distribution at each timepoint is analyzed using the HMM, and the transition points are estimated. From the timecourse of transition points, the apparent elongation rates (v_a) are calculated as a function of the position (x). The slope plot (Figure 4.7D) is generated by calculating $\Delta D^{-1}/\Delta x$ and $\Delta v_a/\Delta x$ using linear interpolation between x=5, 15, 25, 35, 45, 55 kb or t=5, 12.5, 25, and 50 min.

We tested the following parameter spaces for the simulation. Only two examples of the results from simple acceleration and simple terminations models are shown in Figure 4.7C. However, all the described models were used to generate the scatterplot in Figure 4.7D.

1. Single component models

1) Simple acceleration model

$v(x{=}0\text{kb}) = 5$ bp/s, $v(x{=}60\text{kb})=40$ bp/s ; Pol II starts at 5 bp/s (300 bp/min) and accelerates to reach 40 bp/s (2.4 kb/min) at the 60 kb position and downstream. The interpolating $v(x)$ values are generated using cubic Bézier functions (Figure 4.7E). The termination rate k_t is 0.

2) Simple termination model

$v(A{=}0) = 5$ bp/s, $v(A{=}100)=40$ bp/s; the interpolating $v(A)$ values are made with the cubic Bézier functions (Figure S4F). $k_t(A{=}0)=0.002 – 0.0005$ (/s), $k_t(A{=}100)=0$; the interpolating $k_t(A)$ values are made with cubic Bézier functions (Figure S4F). The $k_t(A{=}0)$ values are chosen so that the simulated D curve reflects the observed D curve (Figure 4.7B) which gradually decreases in regions x<20kb and becomes nearly a constant where x>20kb. The less active (and slower) population of Pol II is mostly terminated within the first 20 kb region with selected k_t parameters. Finally, combinations of v and k_t functions are used.

2. Complex models

1) Mixed acceleration model

Models with mixed acceleration and termination are also tested. $v(x,A)$ is generated as a combination of the $v(x)$ of the acceleration model and $v(A)$ of the termination model (Figure S4G). The k_t functions of the simple termination models are used.

2) Regional termination model

The termination rate (k_t) can also have an added dimension and can be treated as a function of both x and A (Figure 4.7H). The interpretation of this is that termination takes place at some preferred positions along the gene. We adjusted the shape of the function so that slower Pol II is relatively stable near the 5' side of the gene to make the apparent elongation rate slower at the beginning. The increased termination at more downstream position leaves only the faster polymerase populations and the apparent elongation rates are higher.

ChIP-seq correlation study (This part was done by Dr. Iris Jonkers)

ChIPseq datasets listed in Figure 4.8D were downloaded and aligned to the mm9 genome when necessary. Reads were extended by 200 bp, and read density was derived in the mappable regions of interest. We correlated read density of ChIPseq factors in the promoter or elongation rate transition region with elongation rates and made boxplots of the elongation rate quartiles. For factors that correlated with elongation rates, we made additional composite profiles from -2 to 30 kb from the TSS in 25 bp windows of these factors in the elongation rate genes, and split the composite profiles up in quartiles of elongation rates or intron1 size. To further elucidate the relationship of intron1 size, histone modifications of interest and GRO-seq Pol II distribution, we made composite profiles centered on exon2 or exon3 start sites of genes with 12.5-25 min elongation rates, and introns 1 and 2 larger than 5 kb. GRO-seq distribution at these exon-intron junctions was further validated by subsampling 1000 times over 10% of the genes at the time, and plotting the median of the subsampled composite profile, with confidence intervals at 0.25 and 0.75.

Splicing analysis (This part was done in collaboration with Hojoong Kwak and Dr. Iris Jonkers)

We aligned the RNA-seq dataset from Sigova et al.[52] to the mm9 genome and established the read density within all exons of genes > 3.5 kb and expressed within the top 75% of activity, as well as the average read density in all exons of each gene. We next established whether read density within the individual exons is significantly different from the average read density in the coding region, and the fold difference of read density in individual exons compared to the average read density. We called an exon skipped when its read density was 3 fold less that the average density, and was significantly different with a p-value < 0.01. We also significantly different exons (p-value < 0.01) with 3 fold increase in read density compared to the average and called these over-included exons, in other words, exons that are always present within the coding region and highly spliced. LOGO's of splice sites for each of these exon groups were derived from the frequency matrix of the base compositions as described previously[53].

Next, we specifically looked at genes with elongation rates associated with them, and calculated the exon density of all exons (exons/kb) within the transition regions of these genes. We plotted exon density against the elongation rates and did a linear regression between elongation rates and total exon density, or split up the linear regression fit analysis of elongation rates and exon density into elongation rate genes without alternative spliced exons, with skipped exons or with over-included exons within the transition region. We also calculated the skip ratio within the transition region (skipped exon density/exon density) and plotted genes with a skip ratio > 0 against elongation rates to see if increased skipping of exons correlates with higher elongation rates.

To investigate whether splicing is influenced by the position of exons in the gene, we looked at the skip ratio in the region from the TSS until the 25 min transition point of genes with 12.5-25 min elongation rates, and compared it to skip ratio in the same genes in the region from the 25 min transition point until the stop (excluding the last exon). We did this for all genes with 12.5-25 min elongation rates, or for quartiles of pausing index, to assess the role of pausing with respect to splicing. Both 12.5-25 and 25-50 min elongation rates and intron1 size were also assessed with respect to pausing index quartiles. LOGO's of splice sites for exons upstream of the 25 min transition point or downstream of 25 min transition point were derived as described above.

REFERENCES

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

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

3. Kwak, H., Fuda, N. J., Core, L. J. & Lis, J. T. Precise Maps of RNA Polymerase Reveal How Promoters Direct Initiation and Pausing. *Science* **339,** 950–953 (2013).

4. Lis, J. T., Mason, P., Peng, J., Price, D. H. & Werner, J. P-TEFb kinase recruitment and function at heat shock loci. *Genes Dev* **14,** 792–803 (2000).

5. Marshall, N. F., Peng, J., Xie, Z. & Price, D. H. Control of RNA polymerase II elongation potential by a novel carboxyl-terminal domain kinase. *J Biol Chem* **271,** 27176–27183 (1996).

6. Ni, Z. *et al.* P-TEFb is critical for the maturation of RNA polymerase II into productive elongation in vivo. *Mol Cell Biol* **28,** 1161–1170 (2008).

7. Peterlin, B. M. & Price, D. H. Controlling the elongation phase of transcription with P-TEFb. *Mol Cell* **23,** 297–305 (2006).

8. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. *Science* **322,** 1845–1848 (2008).

9. Min, I. M. *et al.* Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells. *Genes Dev* **25,** 742–754 (2011).

10. Nechaev, S. *et al.* Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila. *Science* **327,** 335–338 (2010).

11. Muse, G. W. *et al.* RNA polymerase is poised for activation across the genome. *Nat Genet* **39,** 1507–1511 (2007).

12. Zeitlinger, J. *et al.* RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo. *Nat Genet* **39,** 1512–1516 (2007).

13. Rahl, P. B. *et al.* c-Myc regulates transcriptional pause release. *Cell* **141,** 432–445 (2010).

14. Moteki, S. & Price, D. Functional coupling of capping and transcription of mRNA. *Mol Cell* **10,** 599–609 (2002).

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

16. Izaurralde, E. *et al.* A nuclear cap binding protein complex involved in pre-mRNA splicing. *Cell* **78,** 657–668 (1994).

17. Lenasi, T., Peterlin, B. M. & Barboric, M. Cap-binding protein complex links pre-mRNA capping to transcription elongation and alternative splicing through positive transcription elongation factor b (P-TEFb). *J Biol Chem* **286,** 22758–22768 (2011).

18. Lewis, J. D., Izaurralde, E., Jarmolowski, A., McGuigan, C. & Mattaj, I. W. A nuclear cap-binding complex facilitates association of U1 snRNP with the cap-proximal 5' splice site. *Genes Dev* **10,** 1683–1698 (1996).

19. Gu, B. B., Eick, D. D. & Bensaude, O. O. CTD serine-2 plays a critical role in splicing and termination factor recruitment to RNA polymerase II in vivo. **41,** 1591–1603 (2013).

20. Luco, R. F. & Misteli, T. More than a splicing code: integrating the role of RNA, chromatin and non-coding RNA in alternative splicing regulation. *Curr. Opin. Genet. Dev.* **21,** 366–372 (2011).

21. la Mata, de, M. M. *et al.* A Slow RNA Polymerase II Affects Alternative Splicing In Vivo. *Mol Cell* **12,** 8–8 (2003).

22. Howe, K. J., Kane, C. M. & Ares, M. Perturbation of transcription elongation influences the fidelity of internal exon inclusion in Saccharomyces cerevisiae. *RNA (New York, N.Y.)* **9,** 993–1006 (2003).

23. Nag, A., Narsinh, K. & Martinson, H. G. The poly(A)-dependent transcriptional pause is mediated by CPSF acting on the body of the polymerase. *Nature Structural &#38; Molecular Biology* **14,** 662–669 (2007).

24. Hazelbaker, D. Z., Marquardt, S., Wlotzka, W. & Buratowski, S. Kinetic Competition between RNA Polymerase II and Sen1-Dependent Transcription Termination. *Mol Cell* **49,** 55–66 (2013).

25. la Mata, de, M., Lafaille, C. & Kornblihtt, A. R. First come, first served revisited: factors affecting the same alternative splicing event have different effects on the relative rates of intron removal. *RNA (New York, N.Y.)* **16,** 904–912 (2010).

26. Luco, R. F., Allo, M., Schor, I. E., Kornblihtt, A. R. & Misteli, T. Epigenetics in alternative pre-mRNA splicing. *Cell* **144,** 16–26 (2011).

27. Ardehali, M. B. & Lis, J. T. Tracking rates of transcription and splicing in vivo. *Nat Struct Mol Biol* **16,** 1123–1124 (2009).

28. Danko, C. G. *et al.* Signaling Pathways Differentially Affect RNA Polymerase II Initiation, Pausing, and Elongation Rate in Cells. *Mol Cell* **50,** 212–222 (2013).

29. Chao, S. H. & Price, D. H. Flavopiridol inactivates P-TEFb and blocks most RNA polymerase II transcription in vivo. *J Biol Chem* **276,** 31793–31799 (2001).

30. Titov, D. V. *et al.* XPB, a subunit of TFIIH, is a target of the natural product triptolide. *Nat. Chem. Biol.* **7,** 182–188 (2011).

31. Wang, Y., Lu, J.-J., He, L. & Yu, Q. Triptolide (TPL) Inhibits Global Transcription by Inducing Proteasome-Dependent Degradation of RNA Polymerase II (Pol II). *PLoS ONE* **6,** e23993 (2011).

32. Flynn, R. A., Almada, A. E., Zamudio, J. R. & Sharp, P. A. Antisense RNA polymerase II divergent transcripts are P-TEFb dependent and substrates for the RNA exosome. *Proc. Natl. Acad. Sci. U.S.A.* **108,** 10460 (2011).

33. Seila, A. C., Core, L. J., Lis, J. T. & Sharp, P. A. Divergent transcription: a new feature of active promoters. *Cell cycle (Georgetown, Tex.)* **8,** 2557–2564 (2009).

34. Wagschal, A. *et al.* Microprocessor, Setx, Xrn2, and Rrp6 co-operate to induce premature termination of transcription by RNAPII. *Cell* **150,** 1147–1157 (2012).

35. Huff, J. T., Plocik, A. M., Guthrie, C. & Yamamoto, K. R. Reciprocal intronic and exonic histone modification regions in humans. *Nat Struct Mol Biol* **17,** 1495–1499 (2010).

36. Close, P. *et al.* DBIRD complex integrates alternative mRNA splicing with RNA polymerase II transcript elongation. *Nature* **484,** 386–389 (2012).

37. Montes, M. M. *et al.* TCERG1 regulates alternative splicing of the Bcl-x gene by modulating the rate of RNA polymerase II transcription. *Mol Cell Biol* **32,** 751–762 (2012).

38. Sigova, A. A. *et al.* Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells. *Proceedings of the National Academy of Sciences of the United States of America* **110,** 2876–2881 (2013).

39. Brow, D. A. A LLOSTERICC ASCADE OFS PLICEOSOMEA CTIVATION. *Annu. Rev. Genet.* **36,** 333–360 (2002).

40. Pei, Y. & Shuman, S. Interactions between fission yeast mRNA capping enzymes and elongation factor Spt5. *The Journal of biological chemistry* **277,** 19639–19648 (2002).

41. Bradnam, K. R. & Korf, I. Longer first introns are a general property of eukaryotic gene structure. *PLoS ONE* **3,** e3093 (2008).

42. Core, L. J. *et al.* Defining the status of RNA polymerase at promoters. *Cell Rep* **2,** 1025–1035 (2012).

43. Hintermair, C. *et al.* Threonine-4 of mammalian RNA polymerase II CTD is targeted by Polo-like kinase 3 and required for transcriptional elongation. *EMBO J.* **31,** 2784–2797 (2012).

44. Chapman, R. D. *et al.* Transcribing RNA polymerase II is phosphorylated at CTD residue serine-7. *Science* **318,** 1780–1782 (2007).

45. Devaiah, B. N. *et al.* BRD4 is an atypical kinase that phosphorylates serine2 of the RNA polymerase II carboxy-terminal domain. *Proc Natl Acad Sci U S A* **109,** 6927–6932 (2012).

46. Bartkowiak, B. *et al.* CDK12 is a transcription elongation-associated CTD kinase, the metazoan ortholog of yeast Ctk1. *Genes Dev* **24,** 2303–2316 (2010).

47. Czudnochowski, N., Bösken, C. A. & Geyer, M. Serine-7 but not serine-5 phosphorylation primes RNA polymerase II CTD for P-TEFb recognition. *Nature Communications* **3,** 842 (2012).

48. Ramanathan, Y. *et al.* Three RNA polymerase II carboxyl-terminal domain kinases display distinct substrate preferences. *J Biol Chem* **276,** 10913–10920 (2001).

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

50. Brannan, K. *et al.* mRNA decapping factors and the exonuclease Xrn2 function in widespread premature termination of RNA polymerase II transcription. *Mol Cell* **46,** 311–324 (2012).

51. Ji, X. *et al.* SR Proteins Collaborate with 7SK and Promoter-Associated Nascent RNA to Release Paused Polymerase. **153,** 855–868 (2013).

52. Sigova, A. A. *et al.* Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells. *Proc. Natl. Acad. Sci. U.S.A.* **110,** 2876–2881 (2013).

53. Schneider, T. D. & Stephens, R. M. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* **18,** 6097–6100 (1990).

CHAPTER 5

ANALYSIS OF THE ORGAN SPECIFIC NASCENT TRANSCRIPTOME FROM SOLID TISSUES

Summary

RNA polymerase (polymerase) mapping is an important strategy of measuring the activity of transcription at the level of RNA synthesis. The use of nuclear run-on is emerging as a prominent method to monitor genome wide transcription. One of the obstacles for the wide application of nuclear run-on is the requirement of efficient nuclei isolation procedure from cells or tissues. We solved this problem by performing run-on reactions directly on the suspension of pelleted chromatin without isolating the nuclei. Using this method, we first demonstrated the genome-wide nuclear run-on assays in tissues from mouse organs, and acquired high-resolution tissue specific transcriptional landscape of the nascent transcriptome that defines the tissue identity as well as novel potential regulatory elements.

Introduction

Transcription can be measured in several ways. The simplest is to analyze the steady state total RNA within the cell[1]. However this preferentially detects processed forms of RNA, and not all of the transcriptional activity results in stable mature RNA. In contrast, transcription can be measured at the level of RNA synthesis by mapping the amount and the distribution of active RNA polymerases on the genome[2]. This has unique advantages over other transcription assays in that it detects all forms of transcriptional activity including long non-coding RNA, small regulatory RNA, and enhancer transcripts. Therefore, the application of the polymerase mapping method can become a valuable tool in both research and diagnostic purposes.

Currently, Chromatin Immuno-precipitation by RNA polymerase antibodies (RNAP-ChIP) is the most widely used[2]. RNAP-ChIP uses the immuno-purification of RNA polymerase and extracts DNA fragments cross-linked to the polymerase molecules. It can be used in tissues, but ChIP has low genomic resolution, high background, and lacks the information about the direction of transcription on DNA[2].

Nuclear run-on has been the method that uses the endogenous enzymatic activity of RNA polymerases to label the nascent RNA for genome-wide mapping. Coupling this to next-generation sequencing, Global Run-On sequencing (GRO-seq) is a powerful method that uses isolated nuclei from cells for the run-on reaction[3]. It shows the direction of transcription, has very low background, and only detects the polymerases that are actively engaged. However, the resolution is limited by difficulty in controlling the length of labeled-NTP incorporation. Precision Run-On sequencing (PRO-seq) is a variation of GRO-seq that improves the resolution problem by using only the biotin-NTPs for the run-on substrate[4]. Under this condition, the polymerase only incorporates one or at most a few biotin-NTPs, and the 3’ ends of these labeled nascent RNA represents the precise position of RNA polymerase. Since this method analyzes the portion of nascent RNA near the 3’ end that is protected by the polymerase molecule, the result is relatively less affected by RNA degradation than simple RNA-seq. These methods can in some cases be applied to tissues, but the nuclei isolation procedures require laborious optimization for each tissue and time consuming nuclei preparation.

Here we demonstrate the first use of the nuclear run-on method on mammalian solid tissues using a rapid efficient processing procedure. Analysis of the data shows a high-resolution tissue specific transcriptional landscape that defines the tissue identity as well as some novel potential regulatory elements.

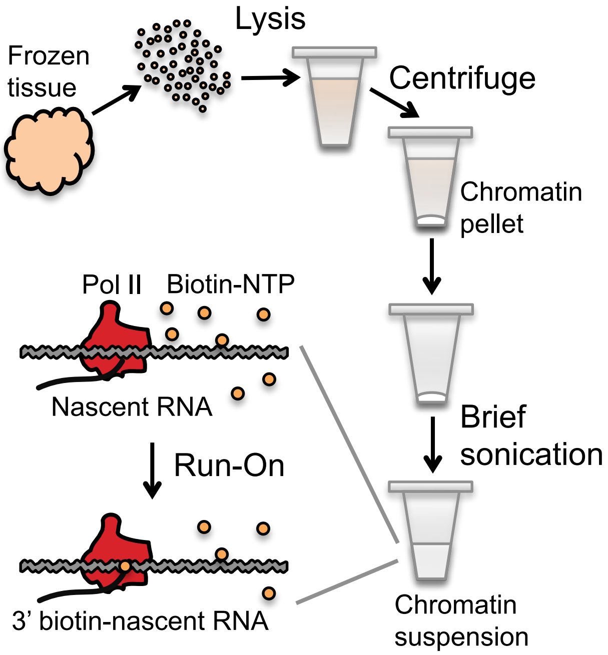


Figure 5.1. Scheme of the Chromatin Run-On method

The Chromatin Run-On Method

To establish a general cell or tissue preparation for nuclear run-on that does not include laborious nuclei isolation, we developed the Chromatin Run-On method. By coupling Chromatin Run-On with sequencing (ChRO-seq) in mouse liver and brain, we generated tissue specific nascent transcriptome data. Briefly, cells or frozen tissues are lysed in buffer containing urea and detergent followed by centrifugation, and pelleted chromatin is mechanically disrupted into suspension for biotin nucleotide incorporation (Fig. 5.1). This provides a method to label nascent transcripts engaged with active RNA polymerases for high specificity mapping in cultured cells or tissues without isolating the cell nuclei.

Table 5.1. Sequencing and alignment statistics

	Liver #1	Liver #2	Brain #1	Brain #2
Total reads	24,510,462	25,212,384	27,703,805	24,587,514
After adaptor clipping	22,623,570	16,997,599	21,394,140	27,876,076
Mapped to rDNA	3,618,882 (16.0%)	4,124,860 (24.3%)	1,539,163 (7.2%)	2,697,042 (9.7%)
Uniquely mapped	12,480,668 (55.2%)	6,856,166 (53.3%)	11,019,407 (51.5%)	13,458,703 (53.5%)

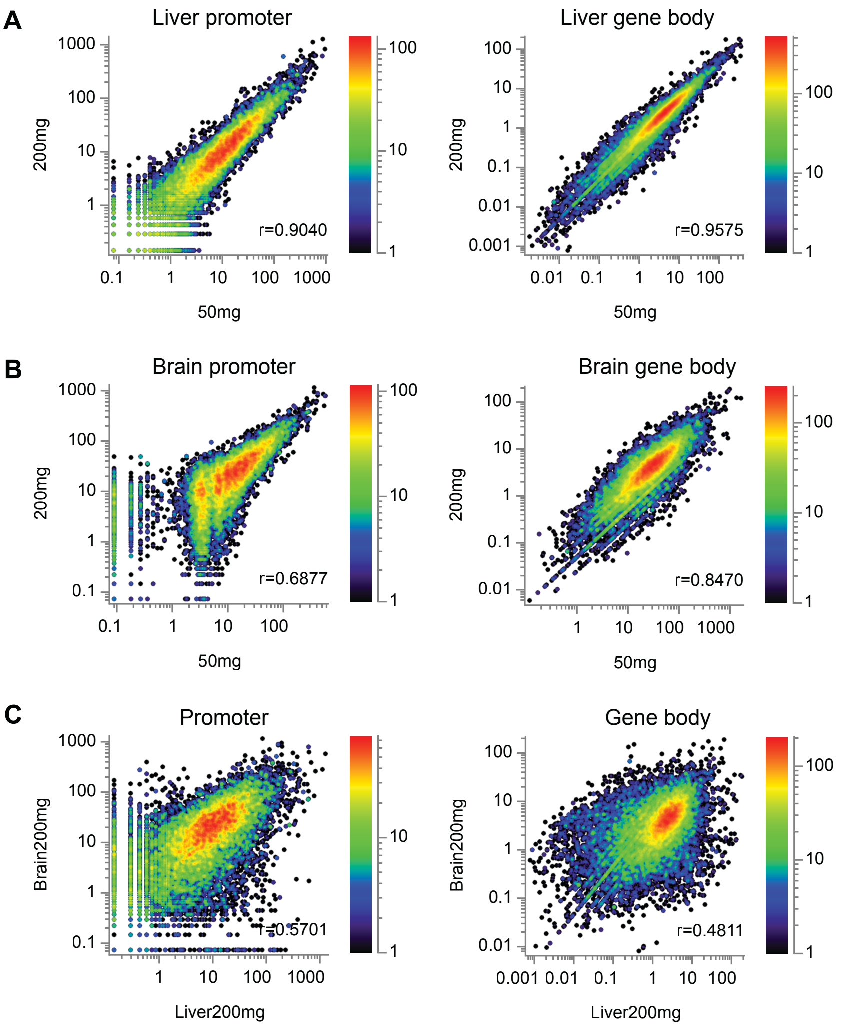


Figure 5.2. Correlation between the ChRO-seq libraries (**A**) Between liver repeats. (**B**) Between brain repeats. (**C**) Between liver and brain libraries.

We obtained sequencing results for two repeats (50mg and 200mg tissues) per each organ, and analyzed them (Table 5.1). To validate whether the repeats correlate well with each other, we generated scatterplots of the promoter proximal read counts and gene body read counts. The liver datasets generated from 50 mg and 200 mg tissues correlated well with each other (Figure 5.2A). However, we saw biased correlation between the promoter read counts of the brain datasets (Figure 5.2B). We speculate that this was due to the lack of sufficient complexity of the library; therefore we only used the 200 mg brain library for further analysis. Distributions between the liver and brain 200mg libraries were evenly dispersed, suggesting that the complexity of the brain 200mg library is adequate (Figure 5.2C).

We show examples of the ChRO-seq profiles at selected positions on the mouse genome. Figure 5.3A shows the UDP glucuronsyltransferase cluster where the genes are specifically expressed in liver tissues[5]. The distribution of Pol II is consistent with the ChIP-seq data in the same tissues[6], demonstrating the validity of the ChRO-seq method. Compared to Pol II ChIP-seq data, ChRO-seq has higher resolution and shows the direction of transcription. The directionality is critical, not only to indicate the direction of the gene, but also to detect the positions where transcription arises. Gene promoters and some active enhancers are known to have characteristic signatures of initiating bidirectional transcriptions[3,7-9]. For example, Figure 5.3B shows a position where two convergent genes are expressed reciprocally between the two tissue types. Compared to Pol II ChIP-seq, ChRO-seq efficiently identifies the TSS of the Arhgap24 gene in the liver tissue by identifying the divergent peak of (-) strand Pol II near the start site (blue arrow). In the brain tissue, Arhgap24 gene is not active and its pairing convergent Mapk10 gene is activated. Interestingly, we saw a pair of strong bidirectional peaks near the 3' end of Arhgap24 gene showing an alternative TSS or an enhancer that might have a regulatory role (red arrow).

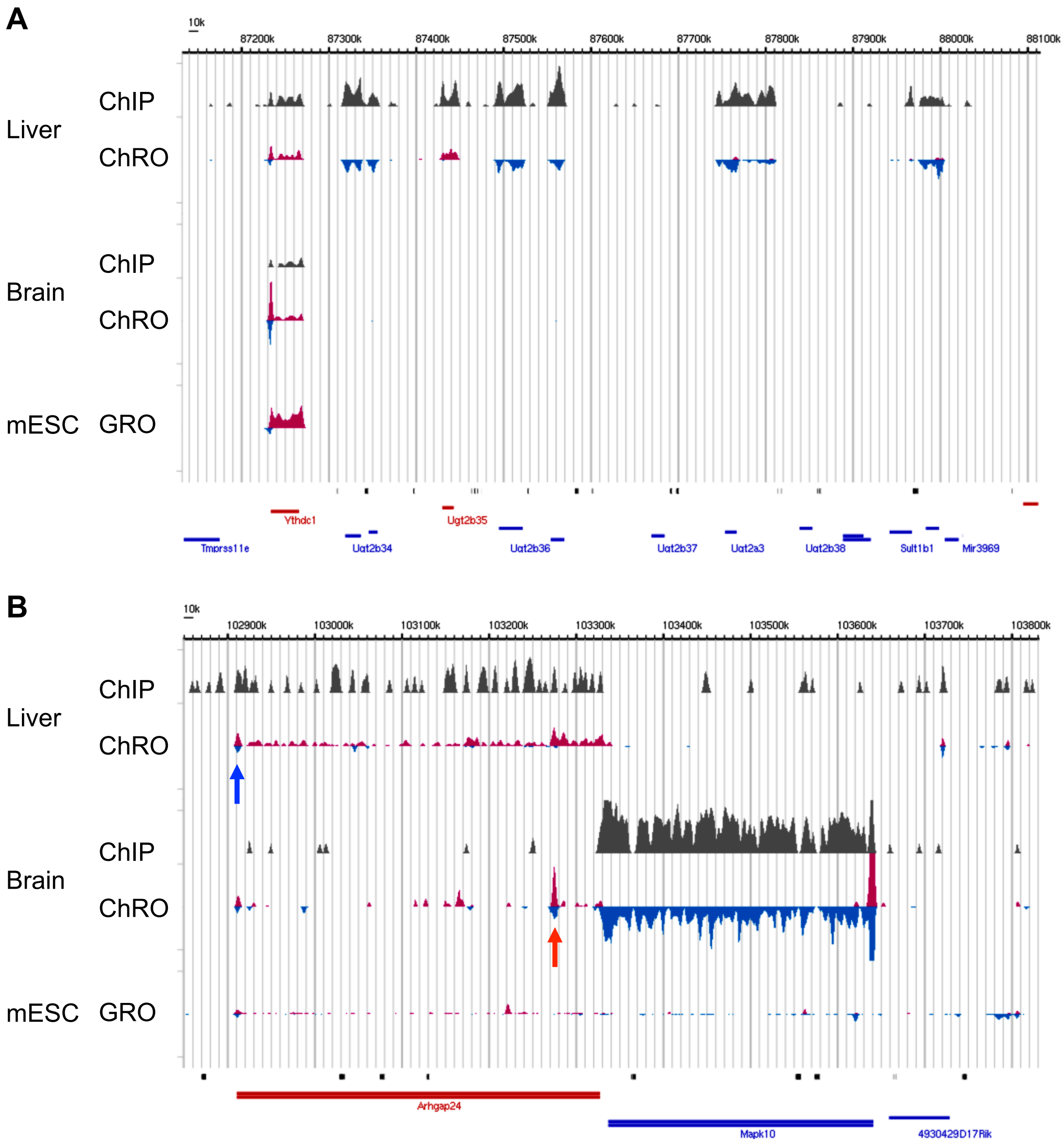


Figure 5.3. Examples of differential ChRO-seq profiles on selected genomic positions. Mouse tissue Pol II ChIP and mESC GRO-seq data are also shown. (**A**) UDP glucuronsyltransferase cluster. (**B**) Arhgap24 and Mapk10 genes.

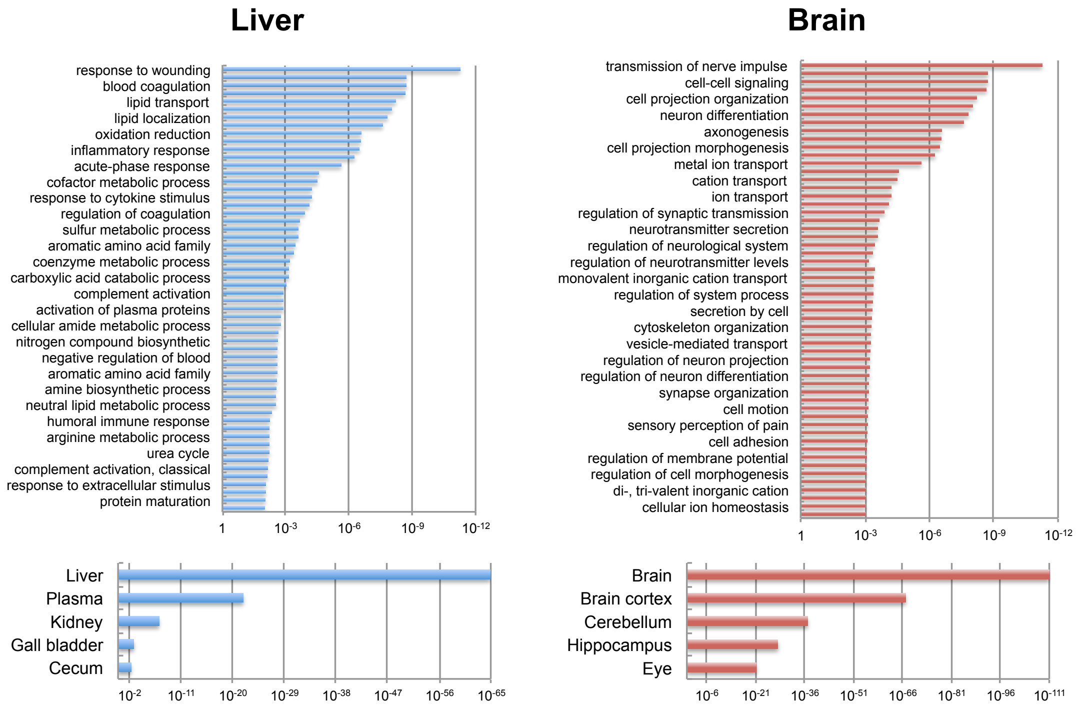


Figure 5.4. Gene ontology analysis based on the expression levels

Analysis of the Tissue Specific Nascent Transcriptome Data

We analyzed the expression level of genes from our ChRO-seq data, and classified the genes that are relatively more expressed in one tissue type versus the other. Using the list of these genes, we carried out a Gene Ontology analysis[10] (Figure 5.4). The genes more expressed in liver are related to wound healing, blood components and metabolism. This agrees well with the function of liver and its regenerative property[11]. On the other hand genes that are more expressed in the brain are involved in the neuronal signaling and differentiation. The tissue type prediction shows with high specificity that the ChRO-seq datasets are from the liver and the brain, respectively.

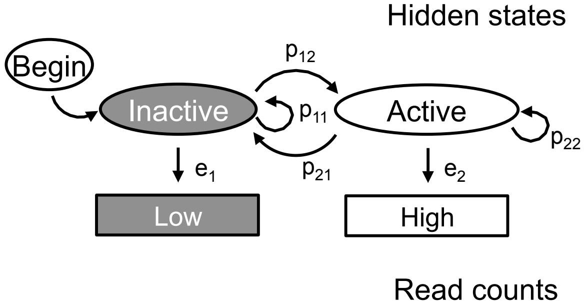


Figure 5.5. Scheme of the de-novo transcription unit calling using HMM

We further performed an analysis attempting to define tissue specific un-annotated transcription units using a computational algorithm. Hidden Markov Model (HMM) is a widely used method to predict the unidentified statuses along a linear chain from probabilistic observations[12,13] (Figure 5.5). We applied this method for the genome-wide de-novo identification of un-annotated potential regulatory transcription units.

For example, Srrm4 gene is a splicing regulator in neuronal cells that is not expressed in the liver[14] (Figure 5.6A). It has an antisense non-coding transcript (2410137F16Rik) arising from the 3' end of the gene. In the liver cells, there is a characteristic bidirectional peak indicative of a TSS or an enhancer located about 10 kb downstream of the annotated non-coding transcript start site (Figure 5.6, black arrow). This was detected by the HMM algorithm as two bidirectional transcription units (tar#10229 and tar#11065; red and blue arrows) that arise at a similar location. This transcription activity is not observed in the brain tissue where Srrm4 transcription is dominant. Therefore, we suggest that the current annotation does not accurately reflect the transcriptional landscape in the liver tissues, and the novel ChRO-seq based transcription unit may have be associated with regulating the expression of the Srrm4 gene.

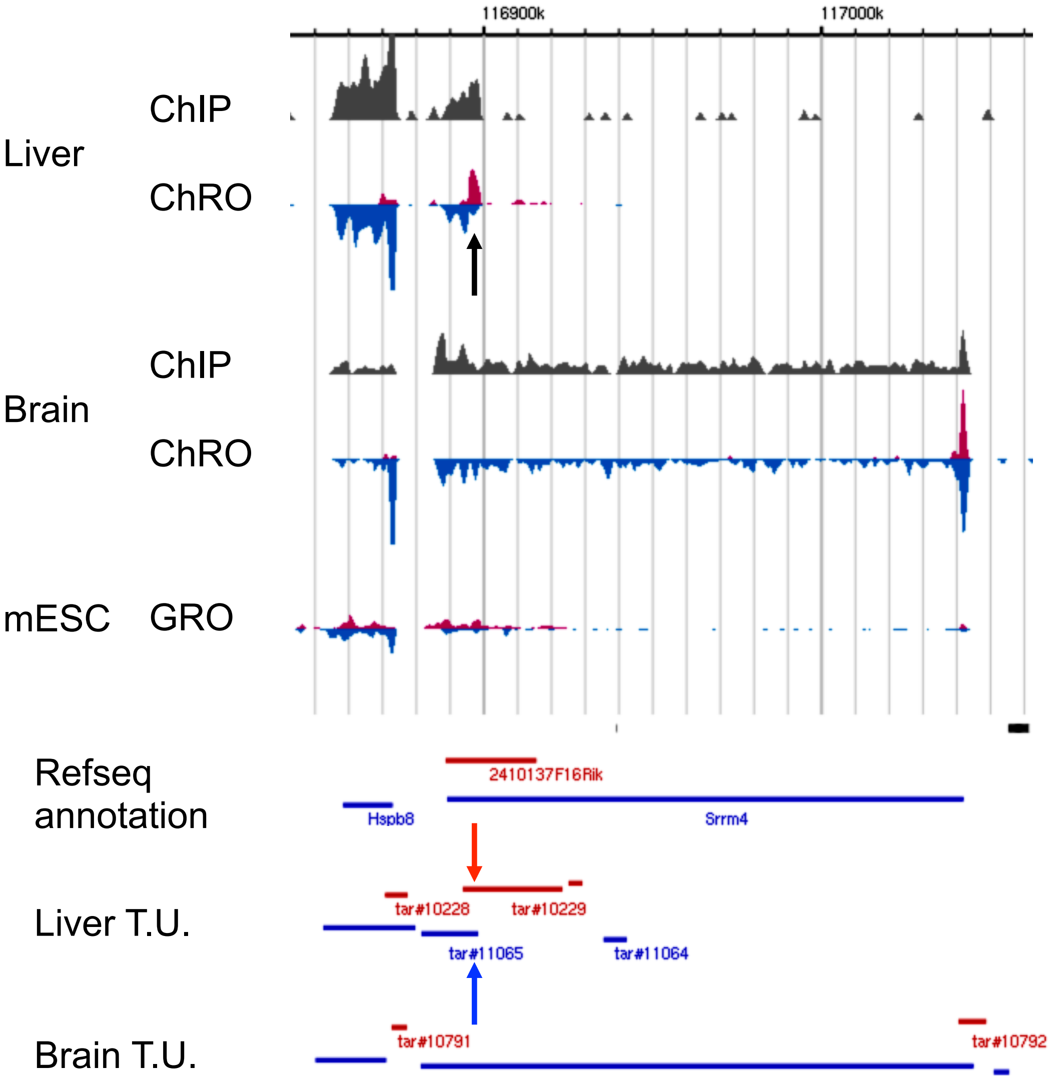


Figure 5.6. Identification of tissue specific re-annotated de-novo transcription units. A pair of liver specific transcription units are identified, of which the expression is reciprocal to Srrm4 gene expression. Black arrow is the start of the bidirectional transcription is ChRO-seq profile. Red and blue arrows point to the newly annotated transcription starts using HMM on the plus and minus strands of the reference genome respectively. (T.U.: transcription unit)

Another example is the Slc2a13, which is a membrane transport protein expressed in brain[15] (Figure 5.7). In the brain ChRO-seq data, a bidirectional peak is located about 40kb upstream of the TSS, which could potentially be an active enhancer (black arrow). With ChIP-seq, it would be difficult to assess this since the bidirectionality cannot be determined. This un-annotated transcription unit was also detected by the HMM algorithm in brain but not in liver (red arrow), demonstrating that the genome-wide transcription unit annotation can serve as a useful tool to detect and evaluate potential regulatory elements on the genome.

Discussion

We have demonstrated that the genome-wide nuclear run-on can be performed in solid tissues using our new ChRO-seq variation of the nuclear run-on protocol. The results demonstrate the differences in the transcriptional landscape between tissues. This method has several unique features compared to conventional genome-wide transcriptional analyses.

ChRO-seq provides a direct instantaneous processing of cells without nuclei isolation. Unlike other methods, the cells are lysed instantly in a mild denaturing solution where RNA polymerase would not be active. Therefore, there is little possibility of secondary effects altering the polymerase distribution during the preparation step. For tissues, instantaneous freezing in liquid nitrogen can ensure that RNA polymerase distributions are not altered. The frozen tissues can be processed by grinding and lysed within a few seconds. This is a major advantage over other methods, because it is extremely simple and efficient.

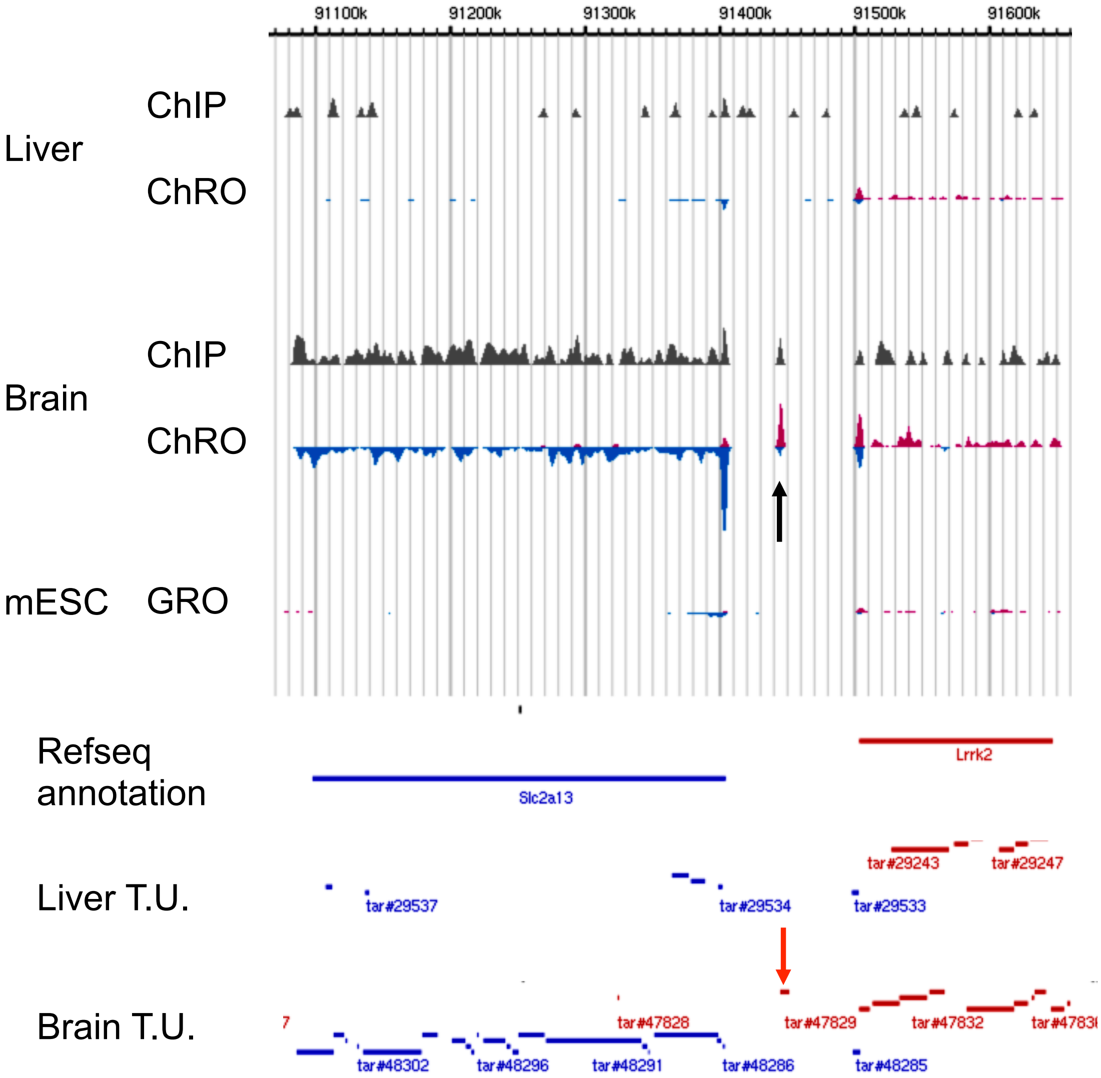


Figure 5.7. Identification of tissue specific un-annotated de-novo transcription unit. A potential enhancer site of Slc2a13 gene expression in brain is identified. Black arrow is the start of the bidirectional transcription is ChRO-seq profile and the red arrow point to the newly annotated transcription start using HMM on the plus strand. (T.U.: transcription unit)

ChRO-seq detects endogenous enzymatic activity of RNA polymerase from pelleted chromatin. Chromatin fractionation method originally described by Wuarin and Schilber[16] has been previously used to isolate nascent RNA associated with chromatin. However, it has never been shown that the RNA polymerase in the chromatin pellet contains the enzymatic activity after the pellet is re-suspended. Using this enzymatic activity for labeling nascent transcript is critical for the three reasons. First, it provides affinity handle to highly enrich for nascent RNA engaged with active polymerase and no other forms of RNA[3]. Second, identifying the sequences of the labeled nascent RNA from the 3' end makes it possible to map where the RNA polymerase was positioned to a base-pair resolution[4]. Finally, it is possible to selectively ablate the undesired cells or parts of the tissues by inactivating the RNA polymerase molecule, for instance with UV laser.

By using the PRO-seq scheme for making the library for high-throughput sequencing, ChRO-seq can still obtain the genomic resolution up to a single base-pair and has all the other advantages of PRO-seq[4]. Biotin affinity purification system is established from many available commercial sources with well-controlled nuclease contamination. Also, it is less affected by deteriorated sample quality and RNA degradation. RNA polymerase active sites are mapped by aligning the 3' end of nascent RNA sequence to the reference genome. Typically in mammalian cells, this requires about 30 bases of RNA sequence from the 3' end. Most of this length of nascent RNA is structurally protected by RNA polymerase complexes from the degradation by RNases[17,18]. Therefore, ChRO-seq can be less affected by the quality of the sample or RNA.

We expect that the nascent transcriptome analysis using ChRO-seq can be easily extended to the diagnostics. Previous methods to measure the genome-wide distribution of RNA polymerases are practically limited to cultured cells or blood cells. But this method can be used as a diagnostic nascent transcriptome analysis in pathologic specimen from solid tissues. Excess surgical specimen after a pathologic diagnosis, as little as 100mg of fresh frozen tissue, can be used. Coupled to the clinical studies that associate the disease outcome and therapeutic responses to drugs, ultimately this method can serve as a predictive tool. Currently, possible diagnostic genomic analyses are DNA sequencing of the exomes (exome-sequencing)[19], and mRNA sequencing (RNA-seq). Because of the impractical depth of sequencing required to identify mutations across the whole genome by DNA sequencing, DNA sequencing at this stage is limited to the exons of all annotated genes (exome)[19]. However, critical mutation can take place at other regions of the genome such as regulatory regions of introns. Exome-sequencing only detects the mutations that may have the potential for a mis-regulation, but not directly at the gene expression products. RNA-seq measures the gene expression products, but also is limited to mature mRNAs, and cannot detect transcriptional activity at regulatory regions. Therefore, measurement of the transcriptional activity across the whole genome by ChRO-seq can make it serve as an efficient tool for the disease diagnostics as well as for the studies of the transcriptional mechanism itself.

Materials and Methods

Chromatin Run-On sequencing (ChRO-seq)

We used NUN (NaCl-Urea-NP40) buffer which is 20mM HEPES pH 7.5, 1mM DTT, 7.5mM $MgCl_2$, 0.2mM EDTA, 1M Urea, 0.3M NaCl, 1% NP-40, RNase inhibitor, Protease inhibitor cocktail (modified from Wuarin & Schibler[16]). We lysed cells or finely homogenized frozen tissues in NUN at 4°C by immediate re-suspension and vortexing. The volume of NUN buffer was at least 20 × w/v of the cell or tissue mass. Centrifuge for 20min at >15,000g, 4°C. A thin white pellet formed at the bottom of the vessel. For tissues with high extracellular matrix content (brain), an additional layer of solid pellet formed over the chromatin pellet. However, that layer could be easily disrupted with pipetting while the chromatin pellet remained solid. Then we wash the chromatin pellet at least twice with cold 50mM Tris-HCl pH 7.5 buffer, using caution not to break the pellet since the pellet sometimes adhered to the pipette tips. We removed all wash buffer and add 100 μl of storage buffer, typically per 200mg of tissue or 2×10^7 cells. The storage buffer is 50mM Tris-Cl pH 8.0, 25% glycerol, 5mM $MgAc_2$, 0.1mM EDTA, and 5mM DTT. Then the chromatin pellet was disrupted with tip sonicator with short bursts of < 2 sec each. Repeat sonication until the chromatin pellet is disrupted to white cloudy suspension. Then we instantaneously froze the chromatin suspension in liquid nitrogen. The material was stored at -80°C for at least a few months until run-on processing. For the run-on, we mixed the chromatin suspension with the equal volume of 2 × run-on solution (10mM Tris-HCl pH 8.0, 5mM $MgCl_2$, 1mM DTT, 300mM KCl, 1% sarkosyl, 40μM biotin-11-ATP, 40μM biotin-11-CTP, 40μM biotin-11-GTP, 40μM biotin-11-UTP, and RNase inhibitor). The reaction mixture was incubated at 37°C (mammalian cells) or 30°C (insect cells) for 3 min. Then the labeled RNA was extracted. The extracted RNA was processed in the same manner as described previously for PRO-seq.

De-novo annotation of transcription units using Hidden Markov Model

For the full length of each chromosome, ChRO-seq read coverage was used to make binary observation values for a bin of the Markov process, 1 if there is a ChRO-seq read within the bin and 0 if there are no read. Therefore, each binned position contains can have either active or inactive states, and the probability of the each state giving ChRO-seq read coverage is the emission probability e. We assumed 2 hidden states ‘transcriptionally inactive’ and ‘transcriptionally active’, with two emission probabilities e_1 and e_2. The transition probabilities between the ‘affected’ and ‘unaffected’ states are p_{11}, p_{12}, p_{21}, and p_{22}. The Baum-Welch algorithm is used to estimate the transition and emission probabilities, by iterative calculations until the probabilities converge. Then the Viterbi path was calculated to define transcriptionally active regions for each chromosome.

REFERENCES

1. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10,** 57–63 (2009).

2. Nechaev, S. & Adelman, K. Pol II waiting in the starting gates: Regulating the transition from transcription initiation into productive elongation. *Biochim. Biophys. Acta* **1809,** 34–45 (2011).

3. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. *Science* **322,** 1845–1848 (2008).

4. Kwak, H., Fuda, N. J., Core, L. J. & Lis, J. T. Precise Maps of RNA Polymerase Reveal How Promoters Direct Initiation and Pausing. *Science* **339,** 950–953 (2013).

5. Tephly, T. R. & Burchell, B. UDP-glucuronosyltransferases: a family of detoxifying enzymes. *Trends Pharmacol. Sci.* **11,** 276–279 (1990).

6. Sun, H. *et al.* Genome-wide mapping of RNA Pol-II promoter usage in mouse tissues by ChIP-seq. *Nucleic Acids Res.* **39,** 190–201 (2011).

7. Seila, A. C. *et al.* Divergent transcription from active promoters. *Science* **322,** 1849–1851 (2008).

8. Sigova, A. A. *et al.* Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells. *Proc. Natl. Acad. Sci. U.S.A.* **110,** 2876–2881 (2013).

9. Wang, D. *et al.* Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature* **474,** 390–394 (2011).

10. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4,** 44–57 (2009).

11. Fausto, N., Campbell, J. S. & Riehle, K. J. Liver regeneration. *J. Hepatol.* **57,** 692–694 (2012).

12. modENCODE Consortium *et al.* Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science* **330,** 1787–1797 (2010).

13. Danko, C. G. *et al.* Signaling Pathways Differentially Affect RNA Polymerase II Initiation, Pausing, and Elongation Rate in Cells. *Mol Cell* **50,** 212–222 (2013).

14. Raj, B. *et al.* Cross-regulation between an alternative splicing activator and a transcription repressor controls neurogenesis. *Mol Cell* **43,** 843–850 (2011).

15. Uldry, M. *et al.* Identification of a mammalian H(+)-myo-inositol symporter expressed predominantly in the brain. *EMBO J.* **20,** 4467–4477 (2001).

16. Wuarin, J. & Schibler, U. Physical isolation of nascent RNA chains transcribed by RNA polymerase II: evidence for cotranscriptional splicing. *Mol Cell Biol* **14,** 7219–7225 (1994).

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

18. Kostek, S. A. *et al.* Molecular architecture and conformational flexibility of human RNA polymerase II. *Structure* **14,** 1691–1700 (2006).

19. Meyerson, M., Gabriel, S. & Getz, G. Advances in understanding cancer genomes through second-generation sequencing. *Nat. Rev. Genet.* **11,** 685–696 (2010).

CHAPTER 6

CONCLUSION

An integrative model of Pol II elongation

RNA polymerase II (Pol II) is the key molecule of eukaryotic transcription. In this dissertation, I have described some aspects of the mechanisms of Pol II during early and productive elongation. Although these studies were carried out in different systems such as *Drosophila* and mouse cells or tissues, the findings could be integrated into a single model that gives a general picture of how Pol II elongates.

First, after the initiation and promoter clearance, Pol II starts the synthesis of nascent RNA chains[1]. At this stage, a part of the initiating complex on the promoter may still remain associated with the promoter DNA elements and retain contact with very early elongating Pol II[2]. As shown in the figures in chapter 1, the 3 dimensional space within the promoter and Pol II is quite restricted, and the complexes are within proximity to each other for such potential interactions to occur. The Pol II at this stage has only some of its C-terminal domain (CTD) phosphorylated at Ser5 position[3]. This very early elongating Pol II less than ~20-60 nt experience backtracks and pauses due to the energy landscapes of the transcription bubbles[4] and potential interaction with large promoter protein complexes such as TFIID or TFIIH. GAGA factor and appropriately positioned GAGA element might have a role in stabilizing this promoter protein complex. This stabilizing of the promoter complex may be closely related to the positioning of core promoter DNA elements[5] as described in Chapter 2.

Second stage is the consolidation of the transient pausing through the recruitment of pausing factors, NELF and DSIF. The precise mode of their recruitment is still not very well known, but is speculated to be through the cooperative effect involving the binding to nascent RNA chains[6], promoter protein complexes such as

GAGA factor[7], and Pol II itself[8]. Binding of NELF and DSIF may lock the nascent RNA chain from exiting and modify the structure of Pol II active site through allosteric interaction with clamp regions of Pol II[9]. At the same time, TFIIS may get recruited to the other side of Pol II near the secondary channel[10], and rescue backtracked Pol II by cleaving protruded nascent RNA, putting the polymerase molecule back to the paused position[4].

Third stage is the maturation of paused Pol II complex. 5' RNA capping can take place at this time, which may be a checkpoint for pause escape and productive elongation[11-13]. If this stage is prolonged, some of the paused Pol II molecules may terminate, as described in Chapter 3. At the same time, assembly of splicing component may also take place during this stage[14,15], preparing Pol II for an efficient splicing of the first intron as proposed in Chapter 4. If this stage is short, Pol II will appear less paused, but at the same time, splicing of the first intron would be less efficient, causing more exon skipping.

Fourth stage is the pause escape, triggered by the recruitment of P-TEFb[3]. P-TEFb phosphorylates Pol II CTD at Ser2 and the pausing factors[16]. The modification and release of the pausing factors release the Pol II from pausing. Immediately after the pause escape, Pol II elongates but not at its full speed. Other functions Pol II may still be 'immature' right after the pause escape. For instance, splicing is less efficient and there are more exon skips in regions immediately downstream of the TSS relative the regions more into the gene bodies as shown in Chapter 4. Assembly of additional factors to the elongation complex may still take place at this stage.

Fifth stage is the productive elongation. Pol II accelerates and reaches a constant maximum speed in the body of the gene. Pol II function is 'mature'; splicing is efficient and there is undetectable amount of termination in the gene body unless there are specific regulatory sites that cause alternative splicing or premature termination[17].

Perspective

The results contained in this dissertation have two major implications. First is on the basic mechanisms of transcription. Pol II pausing and elongation is among the major topics in the study of eukaryotic transcriptional regulation. The newly developed technologies, detailed dissection of the mechanisms of the early Pol II kinetics, and in-depth analysis of elongation rates with its linkage to splicing which were presented in Chapters 2 through 4 broadens the perspective of studying how Pol II elongations. For instance, the role of the components of the initiation complex such as TFIID subunits can be extensively studied using high resolution Pol II mapping. Associations of co-transcriptional processing machinery - capping or splicing factors - with pausing, elongation rates, and termination can be another set of questions that remains to be answered.

Second is on the possible application of the method and the basic mechanisms to more translational biomedical research. The methods described in Chapter 5 are expected to open up a new possibility of conveniently analyzing the statuses of Pol II regulation in a large scale. It may not be restricted to studying the mechanisms of Pol II elongation, but may also be applied to approaching the transcriptional identity of cells or tissues from a different perspective. For instance, we could identify novel transcriptional units that may have regulatory roles rather than protein coding function. The collection of these new genomic elements will make the analysis of transcriptional statuses deeper and more complete. Coupled to computational analysis and modeling, accumulation of the database may lead into a potential use of transcriptional analysis technologies for diagnostic purposes.

REFERENCES

1. Liu, X., Bushnell, D. A., Silva, D.-A., Huang, X. & Kornberg, R. D. Initiation complex structure and promoter proofreading. *Science* **333,** 633–637 (2011).

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

3. Saunders, A., Core, L. J. & Lis, J. T. Breaking barriers to transcription elongation. *Nat. Rev. Mol. Cell Biol.* **7,** 557–567 (2006).

4. Nechaev, S. *et al.* Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila. *Science* **327,** 335–338 (2010).

5. Kwak, H., Fuda, N. J., Core, L. J. & Lis, J. T. Precise Maps of RNA Polymerase Reveal How Promoters Direct Initiation and Pausing. *Science* **339,** 950–953 (2013).

6. Missra, A. A. & Gilmour, D. S. D. 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 (2010).

7. Li, J. *et al.* Kinetic Competition between Elongation Rate and Binding of NELF Controls Promoter-Proximal Pausing. *Mol Cell* **50,** 711–722 (2013).

8. Hartzog, G. A. & Fu, J. The Spt4-Spt5 complex: A multi-faceted regulator of transcription elongation. *Biochim. Biophys. Acta* **1829,** 105–115 (2013).

9. Wu, C.-H. *et al.* NELF and DSIF cause promoter proximal pausing on the hsp70 promoter in Drosophila. *Genes Dev* **17,** 1402–1414 (2003).

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

11. Pei, Y. & Shuman, S. Interactions between fission yeast mRNA capping enzymes and elongation factor Spt5. *The Journal of biological chemistry* **277,** 19639–19648 (2002).

12. Mandal, S. S. S. *et al.* 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 (2004).

13. Lewis, J. D., Izaurralde, E., Jarmolowski, A., McGuigan, C. & Mattaj, I. W. A nuclear cap-binding complex facilitates association of U1 snRNP with the cap-proximal 5' splice site. *Genes Dev* **10,** 1683–1698 (1996).

14. Izaurralde, E. *et al.* A nuclear cap binding protein complex involved in pre-mRNA splicing. *Cell* **78,** 657–668 (1994).

15. Lenasi, T., Peterlin, B. M. & Barboric, M. Cap-binding protein complex links pre-mRNA capping to transcription elongation and alternative splicing through positive transcription elongation factor b (P-TEFb). *J Biol Chem* **286,** 22758–22768 (2011).

16. Price, D. H. P-TEFb, a cyclin-dependent kinase controlling elongation by RNA polymerase II. *Mol Cell Biol* **20,** 2629–2634 (2000).

17. Wagschal, A. *et al.* Microprocessor, Setx, Xrn2, and Rrp6 co-operate to induce premature termination of transcription by RNAPII. *Cell* **150,** 1147–1157 (2012).