Development of Computational Methods to Characterize Transcriptional Precision and Chromatin Dynamics in Yeast

Kunal Poorey
Charlottesville VA

M.S, University of Virginia
M.S, University of Virginia
B.Tech, Indian Institute of Technology Bombay

A Dissertation presented to the Graduate Faculty
of the University of Virginia in Candidacy for the Degree of
Doctor of Philosophy

Department of Biochemistry Molecular Biology and Genetics

University of Virginia
Dec 2013

ABSTRACT

TBP nucleates the assembly of the transcription pre-initiation complex. Although TBP can bind promoters with high stability in vitro, recent results establish that virtually the entire TBP population is highly dynamic in yeast nuclei in vivo. The dynamic behavior of TBP is a consequence of the enzymatic activity of the essential Snf2/Swi2 ATPase Mot1, suggesting that ensuring a highly mobile TBP population is critical for transcriptional regulation on a global scale. We studied the effect of altering TBP binding dynamics on transcription using tiling arrays and a new computational analysis method to determine how perturbed TBP dynamics impact the precision of RNA synthesis in *Saccharomyces cerevisiae*. We found that in *mot1-42* cells RNA length changes were observed for 713 genes with prematurely terminated transcripts representing the largest class of events. Genetic and genomic analyses support the conclusion that these effects on RNA length are mechanistically tied to dynamic TBP occupancies at certain types of promoters. To further capture the TBP binding dynamics we developed a novel method based on modified ChIP assay with sub-second temporal resolution. The time dependence of formaldehyde crosslinking was utilized to extract in vivo on- and off-rates for site-specific chromatin interactions varying over a ~100-fold dynamic range. Using this method, we show that a novel regulatory process shifts weakly bound TATA-binding protein to stable promoter interactions, thereby facilitating transcription complex formation.

Dedicated to my lovely daughter Sophie

PUBLICATIONS

The work presented in Chapter II has been published in

"Poorey K, Sprouse RO, Wells MN, Viswanathan R, Bekiranov S, Auble DT "RNA synthesis precision is regulated by preinitiation complex turnover." Genome Res. 2010 Dec; 20(12):1679-88

The work presented in Chapter III have been published in

"Poorey K, Viswanathan R, Craver MN, Karpova TS, McNelly JG, Bekiranov S, Auble DT. "Measuring Chromatin Interaction Dynamics on the Second Time Scale at Single-Copy Genes." Science, 2013 Oct 18;342(6156):369-72"

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First of all I want to thank my advisors, David Auble and Stefan Bekiranov for giving me this incredible opportunity to pursue graduate studies in their lab. They both have been great mentor, guiding me throughout my graduate carrier and giving me opportunity to work on these incredible projects. Stefan has been an incredible resource for not just science but regarding everything in life. He made the work environment much more warm and so incredibly friendly that I looked forward to my day of work. I have always admired David's patience in dealing with me. I was just an Engineer knowing nothing of the biology but it was his persistence in getting me involved in biology made me learn a lot.

Secondly I would like to thank my committee members Jason Papin, Patrick Grant and Jeff Smith for their great insights, enormous support and their easy availability for meetings.

I would like to thank the past and present members of the Auble and Bekiranov lab members specially Ramya for being a great friend inside and outside the academic world. Her scientific curiosity and dedication for her work has been a great motivation for me. I thank Rebekka & Staton for creating a welcoming environment in the lab and also mentoring me in my first two years. I thank Melissa and Ramya for being awesome colleague in the two major projects during my tenure and our constant discussion of future outside the academic world. I thank Joseph and Jason for taking my advice in their work and utilizing the tools and pipeline I build for their projects

I thank Joel Hockensmith and Debbie Sites for making the academic paperwork feel like a breeze. I would also specially thank Manisha Menon for personally tutoring me for the first year core courses. All those tutoring session were crucial for me to make it through those exams. I would also like to thank Ankit for introducing me to Stefan and for his consistent support, guidance and being an incredible friend for past nine years. He guided me through rough times (which were quite a few) and celebrating vigorously through the good times (which were countless).

These past nine year in Charlottesville I have collected many great friends; they are my family away from my family in India. So Ankit, Manisha, Roshan, Nishant, Nivedita, Lakshmi, Ramya, Budha, Mandovi, Prasad, Pooja, Vipul. Ashu, Tahsin, Priyanka, Upasana, Rahul, and Swati thank you for making this feel like a home. Also special thanks to my friend Sikander for being an awesome friend and also a good motivator in any situation there is.

I am thankful to my family members who have patiently supported me in any decision I took and helped me mentally to reach at this stage. I would like to thank my mother and father for great education right from the start and giving me and my sister's education the first priority no matter whatever the cost. I would like to thank my sister Ketaki, who has always truly believed in me and was my cheerleader since our childhood. I have always appreciated her support that and it has been a great motivation in times of need. I am truly indebt of my mother who not only brought me to this world but has sacrificed her comfort to help me whenever I needed. She took care of Sophie (my daughter) when my career demanded for me. I also thank my mother-in-law who stayed to take care of my daughter Sophie so that me and my wife can make their career.

I would like to thank my nine month old daughter Sophie for just being the joy of my life, and making the arduous process of thesis writing fun. The past nine months have passed like a flash (which could also be because I haven't slept much during that time) and I look forward for the fun parenting that I have to do. I would like to thank my wife for being a true friend and support throughout we have known each other. We have truly complimented each other in qualities and strength. Swati has shown great courage and strength while living alone in Columbus till I was able to complete my graduate studies. I know that she is much smarter than me and having her on my side makes me feel quite secure. Without her I would have never made it anywhere.

Chapter I

Introduction

Transcription is a multistep process of RNA synthesis from DNA sequence. Precise regulation of transcription is crucial at every biological step for the organism such as growth, development, response to environment etc. Hence, transcription is well regulated by different types of mechanisms at every stage, including pre-initiation complex (PIC) formation, transcription initiation, elongation, and termination. Transcription in eukaryotes is carried out by RNA polymerases, which are recruited by general transcription factors which organize at core promoters and interact with coactivators and repressors to regulate transcription. The three Polymerases (Pol I, II, and III) have their own transcription machinery sharing only TATA binding protein (TBP) a general transcription factor (Grünberg et al. 2012). Transcription involving Pol II is responsible for all the messenger RNA (mRNA) synthesis (Hahn 2004; Hahn and Young 2011). In this study we investigated the role played by TBP in the transcription initiation process. We developed computational methods to study how transcription factors influence transcription initiation and termination, as well as their effects on RNA level.

1.1 GTF recruitment, PIC assembly and transcription initiation

The Pol II transcription machinery consists of a collection of general transcription factors (GTFs) which include TFIIA TFIIB TFIID, TFIIE, TFIIF, TFIIH and the multisubunit Pol II enzyme itself (Reese 2003; Hahn 2004; Hahn and Young 2011). Assembly of the Pol II preinitiation complex (PIC) on promoters is highly orchestrated

by transcriptional regulators and co-regulators that influence GTF recruitment by direct interaction with the transcription machinery and by modulating the promoter chromatin template (Hahn 2004). Major efforts in the field have been to understand how activators and co activators affect the basic transcription machinery to regulate gene expression, how the activities of these regulators are regulated, and how the various regulators' pathways are coordinated with signal transduction to control gene expression in response to stress or growth.

PIC assembly is nucleated when TBP is recruited to the TATA box, which is an AT rich sequence motif in the promoter region (Hahn et al. 1989; Spencer and Arndt 2002). Once recruited, TBP alone can initiate the recruitment of other GTFs and PIC assembly at promoters. At TATA-less promoters, TBP is still present, but it requires TAFs (TBP-associated factors) that help stabilize TBP binding to the promoter as a complex called TFIID. TFIID binding to promoters is in turn stabilized by TFIIA and TFIIB (Geiger et al. 1996; Jacobson and Tjian 1996; Hahn and Young 2011) and this complex forms a platform for Pol II and the remaining GTFs to bind. TFIIB directly binds to Pol II along with TFIIF, and both are critical for Pol II recruitment, initiation, and start-site recognition and initiation of RNA synthesis (Chen and Hahn 2003; Kostrewa et al. 2009; Liu et al. 2010; Hahn and Young 2011). TFIIE and H subsequently bind and play an important part in opening the DNA strands to provide a template for RNA synthesis. TFIIH is a multisubunit complex possessing helicase and kinase activities, and also plays a role in stabilizing the promoter DNA during the transition of the PIC into an active open complex (Hahn 2004; Thomas and Chiang 2006; Hahn and Young 2011). This involves TFIIH helicase activity which separates DNA strands around

the transcription start site (TSS) and insertion of single stranded DNA template into the active site of Pol II. This active open complex is highly unstable (Wang et al. 1992; Hahn and Young 2011) (See Figure 1-2).

TFIIB has an important role in the assembly of the PIC as it contacts both TBP-DNA and Pol II. In the open complex, Pol II is reads the DNA sequence and recognizes the start site (Kostrewa et al. 2009). The N-terminus of TFIIB binds to a Pol II and positions the TBP-DNA complex over the Pol II active site cleft (Miller and Hahn 2006; Chen and Hahn 2003; Chen et al. 2007). At this stage, TFIIF stabilizes Pol II in the PIC by interaction with DNA upstream of TATA box, setting the transcription start site and also stabilizing the PIC open complex(Eichner et al. 2010). In vitro studies suggest that TFIIF also stabilizes the RNA-DNA hybrid in the elongation complex at the beginning of elongation (Hahn 2004; Thomas and Chiang 2006; Kostrewa et al. 2009; Hahn and Young 2011). After Pol II leaves the promoter to continue transcription elongation, it is proposed that some of the GTFs dissociate and a scaffold complex remains to facilitate transcription reinitiation (Figure 1.1) (Hahn 2004) By this mechanism the PIC can continue to promote active transcription for some time.

1.1.1 Transcription start site scanning

After binding to the promoter, Pol II scans downstream sequence for a suitable transcription start site (TSS) (Giardina and Lis 1995; Kuehner and Brow 2008; David et al. 2006; Steinmetz et al. 2006a). Promoter DNA is unwound from 20 bp to 90 bp downstream of TATA element. Scanning does not involve transcription but does

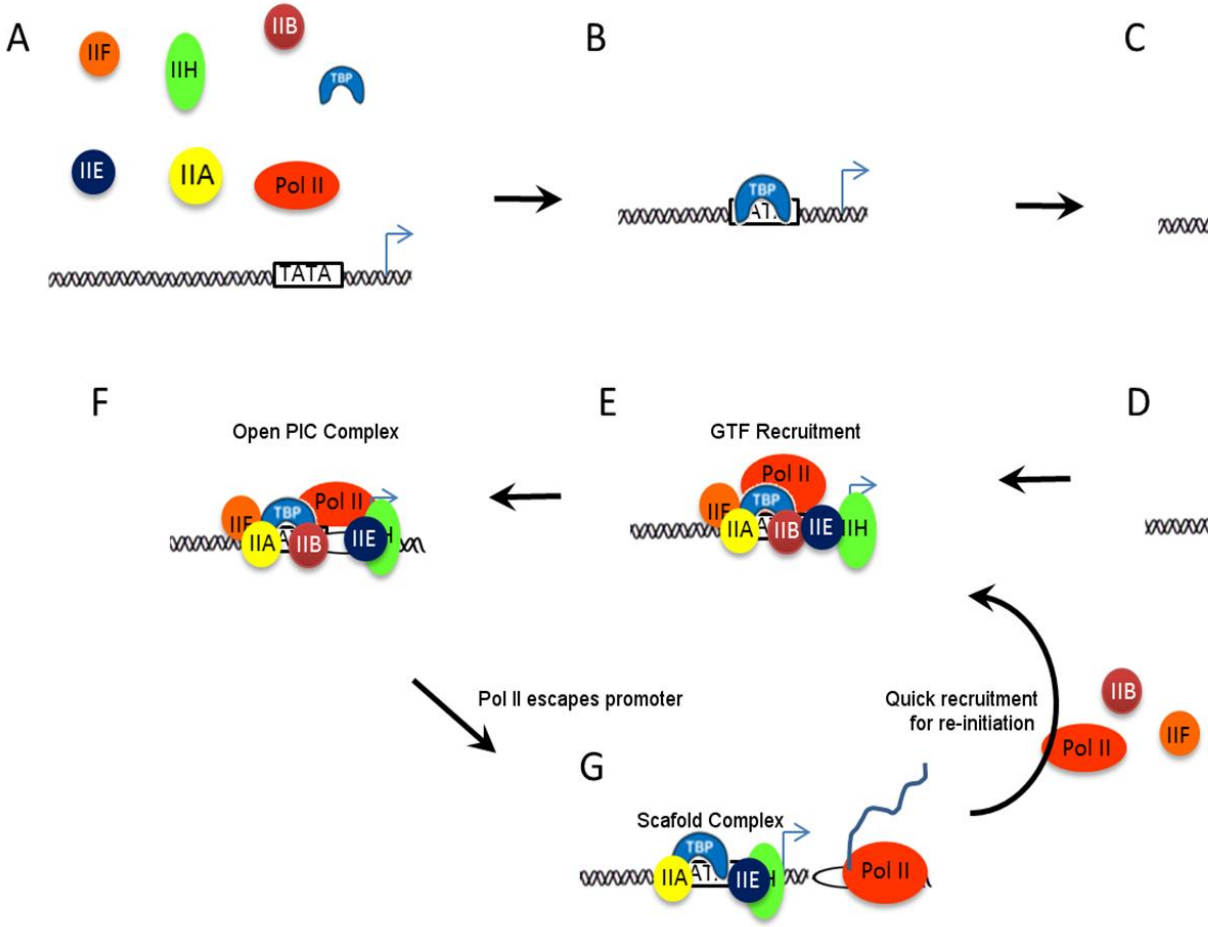

A
IIB
IIF
IIH
TBP
IIE
IIA
Pol II
TATA
B
C
D
E
GTF Recruitment
F
Open PIC Complex
Pol II escapes promoter
G
Scafold Complex
Quick recruitment
for re-initiation

Figure 1- 1: Transcription initiation and GTF assembly at Pol II genes

A) Representation of GTFs, Pol II and promoter DNA. B-E) Steps in the recruitment of GTFs. F) Transition to the open complex for start site scanning/selection and starting elongation. G) Pol II promoter escape. A subset of the GTFs dissociate from the promoter, leaving behind a scaffold complex. This complex facilitates subsequent rounds of transcription via rapid recruitment of the remaining GTFs and Pol II.

involve DNA unwinding and consumes considerable energy in this process. TSS scanning requires TFIIB, TFIIF and Pol II and mutations in any of these factors lead to disruption of proper start site selection (Hampsey 1998; Faitar et al. 2001). Failing to select a proper start site will lead to transcription precision defects. Mutations in TFIIB and Rpb1, a subunit of Pol II, cause the reduced efficiency in recognizing the initiator sequence causing the start of transcription at an alternate start site in the body of the gene (Faitar et al. 2001; Kostrewa et al. 2009). Mutation in TFIIF or in Pol II where it interacts and binds with TFIIF cause a shift in the start of transcription away from the TSS and towards the TATA box (Khaperskyy et al. 2008). In yeast there are multiple TSSs in a single promoter and the mechanism involved in selection of TSS is still not understood.

The Pol II C terminal domain (Pol II CTD) plays unique and important role in PIC assembly. It functionally interacts with co-activators (Mediator) for initiation, interacts with transcription termination factors, and is phosphorylated at Ser2, Ser5 and Ser7 by kinases. Pol II is assembled at the PIC with a non-phosphorylated CTD. During early initiation, the CTD is phosphorylated at Ser5 and Ser7 by a Kin28/Cdk7 which is a subunit of TFIIH (Feaver et al. 1994; Akhtar et al. 2009; Grünberg et al. 2012). Cdk9 activity increases and plays a dominant role of phosphorylation after the transition to the elongation. (Liu et al. 2009b; Kanin et al. 2007) and phosphorylates Ser2 (Buratowski 2009; Liu et al. 2009a; Hahn and Young 2011). The level of these CTD phosphorylations is very closely regulated in initiation, elongation and termination. These markers serve an important function in recruiting many factors, which associate with elongating polymerase and include mRNA capping factors, chromatin modifying enzymes, and termination and mRNA export factors.

1.1.2 Coactivators

Activation is a critical step in regulation of gene expression. Coactivators are large protein complexes that enhance the action of activators by direct contact with GTFs and Pol II or by modifying chromatin. Mediator, SAGA and NuA4 are general coactivators**.** Mediator is a 25 subunit complex in yeast although human homologs exits in multiple forms depending on tissue type and the function of the cell (Conaway et al. 2005). Also, 17 out of the 25 subunits comprise a core complex and other subunits are organism specific (Bourbon 2008; Hahn and Young 2011). Mediator binds to the transcription activation domain and Pol II allowing activator-dependent Pol II activation (Björklund and Gustafsson 2005; Malik and Roeder 2005), although more knowledge of the mechanism is needed to completely understand mediator's role in initiation. Mediator also stimulates basal transcription by stabilizing the PIC and by stimulating TFIIH-dependent phosphorylation of the Pol II CTD (Nemet et al. 2013; Esnault et al. 2008). The effect of mediator on transcription is positive or negative because of its diverse roles at all stages of the process. It is a potential target for promoter-specific regulation in various regulatory pathways (Hahn and Young 2011; Malik and Roeder 2005; Taatjes 2010).

TBP and TAFs which comprise TFIID are targeted by activators to stimulate transcription activation. Although TBP is sufficient to promote basal transcription from TATA containing promoters with purified Pol II and other GTFs, transcription from TATA less promoters as well as response to activators require the TFIID complex. In yeast, Taf1, a subunit of the TFIID complex, has histone acetyl transferase (HAT) activity (Mizzen et al. 1996). SAGA modulates the expression of about 10% of genes that

have TATA boxes (Lee et al. 2000; Huisinga and Pugh 2004). TATA box containing genes are usually genes which are stress regulated and highly inducible. SAGA influences transcription by modifying chromatin and recruiting TBP to the promoters to initiate PIC assembly (Baker and Grant 2007). SAGA is a large multisubunit complex. There are multiple functions of SAGA. Two of the subunits function to suppress TY element expression (Winston et al. 1984). Three other subunits, when mutated, suppress the toxic effect of overexpression of transcription activator (Berger et al. 1992). SAGA was found in a search for co-factors with HAT activity (Grant et al. 1997) in yeast along with NuA4. SAGA preferentially acetylates Histone H3 and NuA4 acetylates Histone H4 proteins and these modifications are in general correlated to activation of genes (Baker and Grant 2007; Zhang et al. 2004). Similar to mediator, SAGA can activate or suppress gene expression. For example Gcn5 can act as coactivator and corepressor for genes in yeast (Xue-Franzén et al. 2010). Gene expression studies have suggested that Gcn5 and SAGA-TBP binding modules have opposite roles in gene expression. Similarly *SPT3* and *GCN5* mutations also have opposite effects on gene expression regulation (Yu et al. 2003; Helmlinger et al. 2008). TFIID and SAGA share some of the same TAFs but they have very different biochemical properties in TBP binding.

1.1.3 Transcription Activation Mechanisms

Controlling transcription initiation is an effective way of controlling transcription itself. This can be achieved by many mechanism (Hahn 1998; Keaveney and Struhl 1998). What important mechanisms result in transcription stimulations?

Some general mechanisms for stimulation of transcription include. 1) Activation by recruitment, 2) Activator induced conformational changes in the transcriptional

machinery, 3) activation through chromatin remodeling 4) activator-mediated regulation post initiation (Hahn and Young 2011).

Activation by recruitment

Activation by recruitment of basal TFs is best studied and mechanistic recruitment studies have been carried out by use of artificial fusion constructs (Chatterjee and Struhl 1995; Klages and Strubin 1995; Xiao et al. 1995) where a specific coactivator or a GTF is targeted to a promoter by fusing it to a DNA binding domain and studying the coactivator or GTF's effect on transcription. There are numerous examples of coactivators, GTFs and chromatin remodelers targeted to promoters and occupancy levels of these factors correlate very well with gene expression.

Other activation mechanisms

Activator-induced conformational changes comprise another class of mechanisms, but to date these have been more prevalent in mammalian systems where activator binding to Mediator causes dramatic changes in conformation of Mediator (Taatjes and Tjian 2004; Taatjes et al. 2002). These conformation changes in the Mediator complex are presumed to result in functional transcription complexes. Chromatin modifications and remodeling directed by transcription factors are also widespread and there are many examples of chromatin remodeling leading to transcription initiation (Narlikar et al. 2002; Li et al. 2007a; Weake and Workman 2010). Many genome-wide studies correlate certain types of chromatin modifications to transcription (Xu et al. 2010; Jung and Kim 2012), but this is not the only type of mechanism used in the processes. This type of mechanism alone cannot lead to transcription initiation without the recruitment of GTFs to promoters (Green 2005). Post-initiation mechanisms of regulation are more prevalent in higher

eukaryotes and one of the best studied mechanism is Pol II pausing shortly after initiation which was first observed in Drosophila (Rasmussen and Lis 1993). Pol II pausing has been observed as Pol II accumulation at the 5' end of the genes in genome-wide studies and is not observed in yeast (Steinmetz et al. 2006b). In Chapter II (and (Poorey et al. 2010; Chu et al. 2007)) we find that deletion of the histone methyltransferase *SET2*, which methylates H3K36, results in cryptic initiation within the body of genes, which was known from previous study Li, Gogol, et al., 2007 from Jerry Workman Lab. Hence chromatin modifications regulate transcription and precision of transcription.

1.2 Transcription Termination

Transcription termination occurs when the RNA Pol II complex is releases the nasent RNA-DNA template and 3' end formation of nascent RNA occurs. This process is important to prevent interference with downstream genes, Pol II turnover and successful 3' end formation of mRNA. Termination of Pol II transcription occurs when cis-acting elements are recognized by RNA binding proteins that associate with the Pol II CTD. Pol II transcription termination is coupled with 3' end formation in which the 3' end of the nascent RNA undergoes cleavage and polyadenylation (Zhao 1999; Rosonina 2006). Termination of Pol II transcription takes place in two steps. 1) transcription of the poly A signal triggers RNA cleavage, and 2) the upstream cleavage product is polyadenylated and the downstream product is degraded. Recognition of the poly (A) signal and cleavage is performed by three factors, CPF, CF1A and CF1B. Mutations in any one of these lead to termination defects.

Termination of most of Pol II genes with a poly (A) signal is carried out by an exonuclease called Rat1(Kuehner 2008; Connelly and Manley 1988; Kim et al. 2004).

Rat1 is involved in RNA processing and termination of rRNA, snRNA and poly (A) containing transcripts. Another pathway, referred to as the Sen1-dependent pathway, is responsible for poly (A) independent termination of transcripts for most of the noncoding Pol II genes (Steinmetz et al. 2006; Steinmetz et al. 2001). The required components for Sen1-dependent termination are Sen1, a presumed helicase, RNA binding proteins Nrd1 and and Nab3, and the CTD phosphatase Ssu72. These factors interact with Pol II and nasent RNA (Kuehner 2008) for termination and 3 prime end formation. In Chapter IV, we use various computational tools to characterize and study the qualitative and quantitative effects of the components of the Sen1 pathway on transcription genome-wide. .

1.3 TBP

TBP is the subunit of the TFIID complex and itserves as a central component of the transcriptional apparatus. Binding of TBP to DNA is the first step in transcription complex assembly. (Auble 2009; Buratowski et al. 1989; Burley and Roeder 1996; Thomas and Chiang 2006; Tora and Timmers 2010). TBP acts as a lynchpin of the Pol II machinery. TBP has important roles in many of the transcription regulatory mechanisms. TBP alone can initiate PIC assembly by binding to an AT rich sequence motif called the TATA box, which is generally represented as TATA(A/T)A(A/T)(A/G) (Basehoar et al. 2004). Previously it was believed that only 20% of promoters in the yeast genome have strong and readily recognizable TATA boxes and hence the promoters were classified into TATA-less and TATA-containing promoters (Basehoar et al. 2004). But a recent study from the Pugh Lab, which used ChIP-exo to precisely map the binding sites of TBP genome-wide, revealed that even so called “TATA less” promoters had some TATA-like

sequence at TBP binding sites (Rhee and Pugh 2012). These two classes have different mechanisms for initiation and recruitment of GTFs. At TATA-containing promoters, TBP recruitment is facilitated by the SAGA (Spt-Ada-Gcn5-acetyltransferace) complex (Bhaumik and Green 2002; Mohibullah and Hahn 2008; Hahn and Young 2011). TBP dynamics are also thought to correlate with the strength of TATA box at the promoter (Tora and Timmers 2010). Genomic studies have revealed that the promoters containing a strong TATA box belong to focused core promoters which control stress response, tissue specific expression, or viral genes in higher eukaryotes (Juven-Gershon et al. 2008; Sandelin et al. 2007; Tora and Timmers 2010). In vitro TBP can form stable complexes with DNA and it directs PIC formation at TATA-containing promoters based on on TATA box strength (Burley and Roeder 1996). The binding of TBP results in two sharp kinks in the DNA at either end of the TATA box which induce an ~90° bend in DNA towards the major groove (Burley and Roeder 1996). Thermodynamic parameters of TBP-DNA binding have been well studied but much remains to be learned about TBP binding at promoters with weak TATA box. The TBP-DNA interaction is stabilized by an extensive hydrophobic interface, and the complex has a half-life of 15 min to 1 hr (Sprouse et al. 2008). This highly stable interaction suggests that TBP dynamics is regulated in vivo by dedicated mechanisms to maintain the soluble pool of TBP as well as to enable PIC turnover in order for transcription to be sensitive to extracellular signals on a rapid time-scale. TBP binds to a variety of sites with high affinity as well as to a variety of non-specific sites (Patikoglou et al. 1999). In principle, this could deplete the soluble TBP pool. In addition, binding of TBP to non-specific promoter sites would

sterically hinder PIC formation. Hence in order for the cell to function properly, mechanisms should be in place to regulate TBP-DNA turnover and dynamics.

1.3.1 Regulation of TBP dynamics

Many factors regulate TBP recruitment to promoters, including Mot1, NC2, Taf1, TFIIA and SAGA (Pereira et al. 2003; Auble 2009). Mot1 and NC2 have global effects on TBP activity. Mot1 is a member of SNF2/SWI2 ATPase family and it displaces TBP from DNA in an ATP-dependent reaction (Auble and Hahn 1993; Auble 2009). In yeast, the robust catalytic activity of Mot1 is responsible for the dynamic nature of TBP and the maintenance of a soluble TBP pool (Sprouse et al. 2008). NC2 is a TBP binding heterodimer which was originally believed to function by blocking PIC formation. However, in addition,NC2 interacts with TBP to form an encircling clamp that allows TBP to diffuse along the DNA contour (Kamada et al. 2001; Schluesche et al. 2007). In this way, NC2 may function to redistribute TBP among different DNA sites in cis. Both NC2 and Mot1 were believed to be repressors of transcriptions (Prelich 1997; Dasgupta et al. 2002; Geisberg et al. 2001, 2002) but they can also act as activators (Collart 1996; Dasgupta et al. 2005; Madison and Winston 1996; Geisberg et al. 2002, 2001). Models by which Mot1 and NC2 can act as global activators and repressors are discussed below.

Activation Mechanisms

- Mot1 can remove a kinetically trapped, inactive form of TBP to facilitate assembly of a functional PIC (Auble 2009; Figure 1-2 D).
- Mot1 maintains a free TBP pool to allow binding to new locations in DNA to form functional PICs (Auble 2009; Sprouse et al. 2008; Figure 1-2 C).

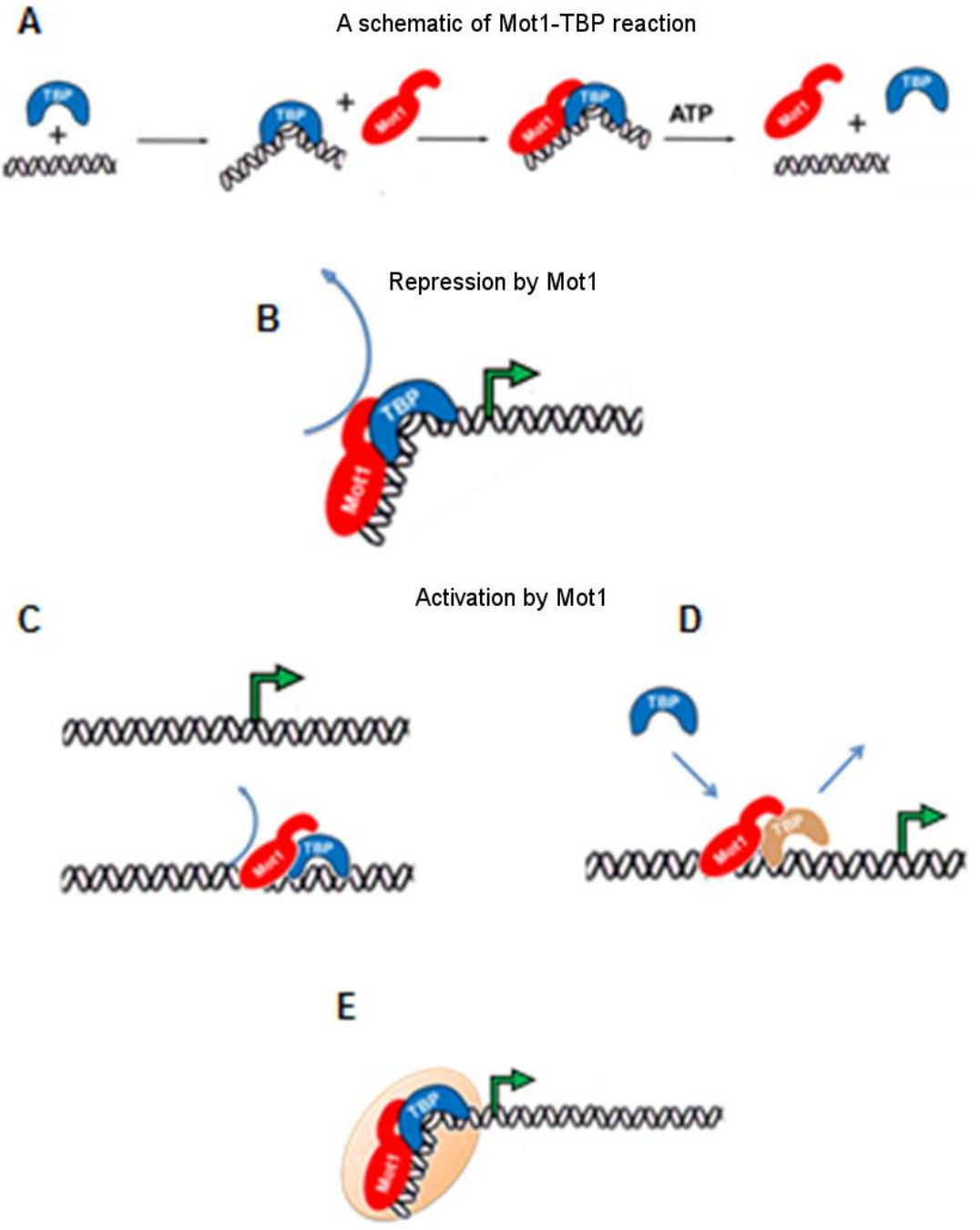

A
A schematic of Mot1-TBP reaction
TBP
Mot1
ATP
B
Repression by Mot1
TBP
Mot1
C
Activation by Mot1
D
TBP
Mot1
E
TBP
Mot1

Figure 1- 2: Regulation of TBP binding and regulation of transcription by Mot1

A) Schematic representation of Mot1 removing TBP from DNA. B) Repression by Mot1. Mot1 removes TBP from active promoters, resulting in repression. C-E) Activation by Mot1. In (C), Mot1 removes TBP from non-specific sites, thereby replenishing the TBP pool. In (D), Mot1 relieves kinetically-trapped and functionally inactive TBP from promoters to allow a functional PIC to form in a subsequent reaction. In (E), Mot1 forms a hypothetical novel complex to activate transcription.

- Mot1 is member of transcriptionally complex active for stress response genes and also believed to have chromatin remodeling properties which are required for gene activation (Auble, 2009; Figure 1-2 E)`.
- NC2 activates transcription by stabilizing weak TBP-DNA or by moving TBP from a non-specific site to the core promoter (Tora and Timmers 2010).

Repression Mechanisms

- Mot1 displaces TBP from core promoters and interrupts the formation of PIC or disassembling the PIC from the promoter (Figure 1-2 B)
- NC2 binds to the TBP DNA complex and sterically hinder the binding of other GTFs to form a functional PIC.

Mot1 has a direct role in regulation of TBP dynamics. Recent measurements of TBP mobility in living yeast cells demonstrate that all detectable TBP is highly mobile, displaying Mot1-dependent FRAP recovery times of < 15 seconds (Sprouse et al. 2008). Importantly, while the recovery times are rapid, they are markedly slower than can be explained by diffusion, and are instead consistent with transient interactions with chromatin. This suggests that nearly the entire TBP pool is rapidly recycled, leading to rapid redistribution of TBP among chromatin binding sites. Our kinetic analysis of TBP binding undertaken by a new method that we developed (Chapter III) shows that there are longer-lived TBP complexes at promoters in vivo, however they represent a tiny fraction of the TBP interactions in the nucleoplasm and escaped detection by live cell imaging.

Several fundamental questions are raised by the observed high mobility of TBP in vivo. If TBP is rapidly recycled from sites on chromatin, what is the nature of these

sites? Are they specific or non-specific interactions? In Chapter II of this dissertation we report the results of our investigation of Mot1's effect on TBP and TFIIB recruitment and on RNA synthesis genome-wide. We found that modulating TBP dynamics by mutations in Mot1 caused RNA synthesis precision defects in 713 genes. Given the pervasive RNA synthesis in yeast cells under these conditions, it would appear that there may be active promoters for which PICs are rapidly recycled. If this is true, how and why are such dynamics important for promoter function? When TBP dynamics are compromised, we also investigate whether new or different types of RNA made, or if simply the quantity at annotated genes is changed.

1.4 Transcription factor dynamics

The many processes of transcription including PIC assembly, initiation, Pol II pausing, reinitiation, elongation, dynamic rearrangements of nucleosomes, histone modifications, termination, and polyadenylation involve dynamic interaction between transcription factors and regulators with chromatin and Pol II and RNA. Most of the things we know about how the transcription machinery is assembled, how it initiates and elongates are deduced from snapshots obtained from biochemical methods including chromatin immunoprecipitation (ChIP) which do not assay dynamics. These methods have been invaluable in showing us the key players involved but not have been insightful in providing the kinetics of these factors in these processes (Hager et al. 2009). The dynamic interaction of proteins with DNA is key to understand PIC assembly, transcription regulation and many other related regulatory processes.

1.4.1 Existing methods to study transcription factor dynamics:

Two general methods exist to measure protein-DNA interaction dynamics in vivo: competition ChIP and fluorescence recovery after photobleaching (FRAP),which we describe below.

Competition ChIP

Competition ChIP is used to measure protein-DNA dynamics at a specific site or across the whole genome (Lickwar et al. 2012; van Werven et al. 2009; Rufiange et al. 2007; Schermer et al. 2005). In this method, the cell contains two copies of the transcription factor or protein with each copy tagged with a different epitope. One of the copies is expressed constitutively and the other is induced as a competitor. Chromatin occupancies of each copy of the protein can be quantified by ChIP, ChIP-Chip or ChIP-seq over time as they compete for the same binding sites. The rate of exchange of the two isoforms enables an estimate of the relative turnover rate and the residence time of the protein on the DNA for each site (Lickwar et al. 2013).

The advantage of using this technique is that site-specific estimates of the relativel turnover and residence time can be can be directly correlated with other genomic features at each promoter across the genome. This method is also readily applicable to any ChIP-based technology and in any organism. Apart from these advantages, the method is only applicable to transcription factors with fairly long chromatin residence times (>500 s) (Lickwar et al. 2013) due to the relatively long induction time (~20 min) of the competitor. Events that are much faster than the rate of induction will appear to have the same kinetic profile driven purely by competitor induction (Lickwar et al. 2013; Deal et al. 2010).

Single cell fluorescence techniques

Fluorescence recovery after photobleaching (FRAP), fluorescence correlation spectroscopy (FCS) (Michelman-Ribeiro et al. 2009) and single-molecule tracking (SMT) are the other common techniques used to study protein-DNA interactions, and all are based on fluorescence imaging. These methods have the advantage that they are applied to single cells and they have high temporal resolution (i.e., seconds). The disadvantage of FRAP is the relatively poor spatial resolution. The extracted binding behavior is averaged over an area which likely includes many promoters and binding sites, and such methods are unable to provide any site specific dynamic information, except for a few special cases where the transcription factor being studied binds to one locus which contains multiple copies of the binding site (Karpova et al. 2008; van Royen et al. 2011).

In Chapter III, I discuss the development of a new method to measure chromatin dynamics which is based on a modified ChIP assay and which provides kinetic parameters on the second time scale with site specificity.

1.4.2 How do transcription factors find targets?

As discussed above, recruitment is a major regulatory step in transcriptional control. Binding sites are in general sparsely located in the genome and comprise a very small percentage of the total genome. Proteins find their binding sites using diffusion which is well supported by FRAP studies (Hager et al. 2009; Gorski and Misteli 2005). TFs diffuse rapidly in the nucleus traveling the length of nucleus within a few seconds and the whole nuclear volume within a few minutes(Hager et al. 2009). FRAP studies revealed that the diffusion rate of TFs was about 2-3 order of magnitude slower than expected (Sprague et al. 2004; Mueller et al. 2008). The loss of mobility could be the

result of TFs integrating into larger complexes, or could be due to transient binding to chromatin (Hager et al. 2009). TFs use of transient binding to find their target sites has been called the "3D scanning model" (Misteli 2001). Most of the TF scanning-time is spent interacting with chromatin non-specifically with an estimate of 50-250 ms between TF-chromatin interactions. This 3D hopping is supplemented by local scanning via 1D diffusion along the DNA has been called facilitated diffusion(Winter et al. 1981). It has been argued that facilitated diffusion makes it possible for TFs to find their targets relatively quickly and efficiently (Phair et al. 2004; Elf et al. 2007). In this model, TFs search the local vicinity using 1D diffusion, scanning about 500 bp before dissociation from DNA (Halford and Marko 2004; Gorman and Greene 2008; Hager et al. 2009)

1.4.3 Modeling of transcription factor binding

Chemical kinetics of various biophysical processes in an organism like TF-DNA binding Protein-protein binding, protein transport, transcription variation/noise, have been modeled for quite some time. Many biological processes can be defined or approximated by a set of reactions known as reaction networks (Wilkinson 2009). These reactions are usually described by nonlinear ordinary differential equations (ODE) or partial differential equations. Solutions of these equations provide valuable insights and dynamics of the studied process. (Resat et al. 2009) For example in Samorodnitsky and Pugh 2010 study they developed a computational program PathCom which under the assumption that occupancy levels can be related to binding duration models in vivo protein-DNA occupancy data as biochemical mechanisms. They model the assembly of the general transcription factors (TBP, TFIIB, TFIIE, TFIIF, TFIIH, and RNA polymerase II) at the genes yeast. They found that TBP occupancy at promoters is rather

transient compared to other general factors. Another example in the study Lickwar et al. 2012 better computational model lead to better understanding of competition ChIP dataset with more accurate parameter estimation.

We aimed to develop a mathematical model to understand TF-DNA interactions and their dynamics measured by the ChIP assay. The traditional way to model a time-evolved molecular population in a reacting system is to use a set of coupled first order ODEs called reaction rate equations (RREs). Usually, RREs are valid in spatially homogeneous systems and are nonlinear for bio-molecular reactions. However, in special cases when time scales are widely different or under simplifying conditions or assumptions, the RRE can be reduced to a linear set of differential equations . Assumption we made were 1) TF concentration in the cell is constant 2) all binding events are independent events. Under these assumptions the the ODEs representing our system was simplified to linear set of equations (Chapter III).

Typical steps taken in developing RREs to model a dynamic process:

1.) Identify key species and properties to be simulated including dependent (output) and independent (input) variables.
2.) Model formulation: where we define the changes in the concentration as a function of time and location, constraints are defined and applied in RRE. In the case of models of average dynamics over a population (as is the case for the experiments in this chapter), the RRE is formulated in terms of the average of the relevant variables and kinetic parameters. In the case of stochastic dynamics, the behavior of individual events in a relatively small, molecular-scale volume are

modeled and the RRE takes the form of a master equation in terms of the probability of events occurring.

3.) Solving the deterministic or stochastic RRE analytically or numerically (Resat et al. 2009).

Deterministic approach to model chemical kinetics in general assume that the reactants are abundant and have a level measured on a continuous scale. The state of the system at any particular instant is stated in concentration and the changes in concentration are assumed to occur in a continuous and deterministic process. The rate of the reaction can be represented by simple mass action kinetics or enzyme kinetic law (Wilkinson 2009) these equations can be solved analytically to give an explicit formula but if the mathematical solution is not tractable, numerical methods can be used to determine the solution In contrast, stochastic approaches are used in model the systems which have heterogeneity in them for example in modeling TF-DNA dynamics the concentration of TF might change over time or modelling TF occupancy in a single molecular system. To model these stochastic systems probability theory is used to account of inherent unpredictability and determine RREs using one or multiple stochastic variables. Single cell systems have more. We applied deterministic methods to solve the TF-DNA binding equations and the relatively few underlying assumptions allowed us to solve the system analytically (Chapter III).

1.5 Scope of the Study

With the help of various computational techniques (Chapter II) we studied the effect of Mot1 on TBP and TFIIB binding genome wide and correlated these effects with gene expression. We found that while a mutation in Mot1 elicited only modest changes in

gene expression, it caused significant changes in RNA synthesis precision. To do this, qualitative and quantitative changes in the transcriptome we analyzed using a computational pipeline that I developed called TraPP (Transcription Precision Pipeline) which was designed to quantify and classify transcription precision defects in a mutant strain compared to a control. Also the effects observed in *mot1* cells appeared to be caused by changes in TBP dynamics. To capture these changes in dynamics we developed a novel method called CLK ChIP (Chapter III) to capture in vivo DNA-protein interaction using a modified ChIP assay. We used deterministic methods to model TF-DNA binding and crosslinking. Non-linear regression was used to fit the mathematical model to the experimental data, yielding kinetic parameters describing protein-DNA interactions in vivo for a number of TFs. Specifically, we derive the on and off rates, the molecular crosslinking rate and in vivo occupancy at steady state for these TFs. In Chapter IV, we expand the application of TraPP to study factors involved in the Sen1 transcription termination pathway.

Chapter II

Defining Transcription Precision Defects and Correlation with PIC Dynamics

The work presented in this Chapter was published in Poorey et al. 2010

Transcription regulation plays an important role in development, differentiation, and disease models. TBP nucleates the assembly of the PIC and although TBP can bind very stably to promoters in vitro, it has been found to be very dynamic in vivo. Mot1 is a SWI2/SNF2 ATPase responsible for dissociation of TBP from DNA. By this reaction, Mot1 can serve as a transcription activator by removing TBP from unproductive promoters and as a repressor by removing TBP from active promoters. We studied the effect of altering TBP binding dynamics on transcription using tiling arrays and a new computational analysis method to determine how perturbed TBP dynamics impact the precision of RNA synthesis in *Saccharomyces cerevisiae*. We found that Mot1 plays a broad role in establishing the precision and efficiency of RNA synthesis. In *mot1-42* cells, RNA length changes were observed for 713 genes, about twice the number observed in *set2*Δ cells, which display a previously reported propensity for spurious initiation within open reading frames (Li et al. 2007b). Loss of Mot1 led to both aberrant transcription initiation and termination, with prematurely terminated transcripts representing the largest class of events. Genetic and genomic analyses support the conclusion that these effects on RNA length are mechanistically tied to dynamic TBP

occupancies at certain types of promoters. These results suggest a new model whereby dynamic disassembly of the PIC can influence productive RNA synthesis.

2.1 Introduction

As described in Chapter I, the Pol II transcription machinery consists of a collection of GTFs and the multisubunit Pol II enzyme itself (Reese 2003; Hahn 2004). Assembly of the PIC is highly orchestrated by transcriptional regulators and co-regulators that influence GTF recruitment by direct interaction with the transcription machinery and by modulating the promoter chromatin template (Hahn 2004). PIC assembly is nucleated by TBP, which physically interacts with multiple GTFs and DNA. TBP recruitment to promoters is often rate limiting for transcription in vivo (Pugh 2000).

Interaction of the TBP saddle with the TATA Box results in severe bending and unwinding of the DNA (Burley and Roeder 1996). In vitro, the resultant complex forms a specialized, long-lived substrate for accrual of the other GTFs. Biochemical evidence indicates that a TBP-containing subcomplex remains on promoter DNA following departure of Pol II (Hahn 2004). This complex, termed the scaffold, can facilitate transcription reinitiation in vitro (Hahn 2004). Although the in vitro evidence in support of a stable reinitiation intermediate is strong, PIC dynamics may be influenced by other factors in vivo. For example, stable TBP-DNA binding is antagonized by Mot1, a Snf2/Swi2-related ATPase that dissociates the TBP-DNA complex (Auble 2009). As another example, the NC2 heterodimer interacts with TBP to form an encircling clamp that allows TBP to diffuse along the DNA contour (Kamada et al. 2001; Schluesche et al. 2007). In fact, recent measurements of TBP mobility in living yeast cells demonstrate that all detectable TBP is highly mobile, displaying Mot1-dependent FRAP recovery

times of < 15 seconds (Sprouse et al. 2008). Importantly, while the recovery times are rapid, they are markedly slower than can be explained by diffusion, and are instead consistent with transient interaction with chromatin. This suggests that the entire (or nearly entire) TBP pool is rapidly recycled, leading to rapid redistribution of TBP among chromatin binding sites.

Several fundamental questions are raised by the observed high mobility of TBP in vivo. If TBP is rapidly recycled from sites on chromatin, what is the nature of these sites? Given the pervasive RNA synthesis in yeast cells under these conditions, it would appear that there may be active promoters for which PICs are rapidly recycled. If this is true, how and why are such dynamics important for promoter function? When TBP dynamics are compromised, are new or different types of RNA made, or is simply the quantity at annotated genes changed? To begin to address these questions, we developed a general genomic strategy to identify aberrant RNA species in mutant strains of interest. Surprisingly, we find that compromising TBP dynamics via conditional mutation of Mot1 gave rise to many hundreds of changes in RNA length, the largest category of which includes transcripts that were apparently initiated properly but failed to reach the end of the gene. In parallel, we determined how Mot1 affects TBP occupancy genome-wide for comparison with the RNA effects. The results support a model in which Mot1-mediated TBP dynamics at the promoter influence transcription elongation efficiency. These results argue that in contrast to prevailing views, at many promoters PIC dynamics can play an important role in conferring efficiency and accuracy of transcription elongation.

2.2 Materials and Methods

All wet-bench procedures were performed by Melissa Wells Carver and Rebekka Sprouse. I was responsible for all computational analyses of the resulting data.

2.2.1 Yeast strains and growth conditions

The strains and the growth conditions used for the study of RNA and ChIP in WT, *mot1-42* and *set2Δ* are described in the material and method section of the publication described in detail in (Poorey et al. 2010).

2.2.2 Expression Analysis

Total RNA was isolated using hot-acid phenol extraction (Schmitt et al. 1990). For the tiling arrays, cDNA was synthesized from 7 µg of total RNA using the Affymetrix GeneChip WT Double-Stranded cDNA Synthesis Kit as recommended by the manufacturer, but with the addition 0.4 mM dUTP for subsequent fragmentation and biotin end-labeling. dsDNA was purified using the GeneChip Sample Cleanup Module (Affymetrix, Inc.). Fragmentation and labeling were performed using the GeneChip WT Double Stranded DNA Terminal Labeling Kit. Fragmented DNA was confirmed to be between 25-100 bp using the Agilent RNA 6000 Nano Kit and an Agilent 2100 Bioanalyzer. Efficient labeling was confirmed by gel shift assay using NeutrAvidin. Samples were hybridized to S. cerevisiae Tiling 1.0R Arrays (Affymetrix, Inc.) and raw data were generated by the Microarray Core Facility at UVA. WT, *mot1-42*, and *set2*Δ RNA analyses were performed using two independent biological replicates for each. For the experiment in Figures 2-7 A and B, WT (TBP), *mot1-42* (TBP-Y185C), and *mot1-42* (TBP-F207L) RNA was analyzed as a single replicate. For real-time qPCR validation,

RT-PCR was first performed using the iScript Select cDNA Synthesis Kit (BioRad) according to manufacturer's instructions. cDNA was then quantitated by real-time PCR using iQ SYBR Green Supermix (BioRad) and the BioRad MyiQ Single Color Real-time PCR detection system. Each experiment includes two independent biological replicates. Northern blotting was performed as previously described (Dasgupta et al. 2005) using strand-specific probes obtained by single-primer PCR in reactions that included □-32P-dATP. In each case, the template for probe synthesis was a gel-purified PCR fragment spanning the differentially affected transcribed region. The Tiling Array hybridization for total RNA in WT, *mot1-42*, and *set2*Δ was performed by Rebekka O Sprouse, and the hybridization for total RNA in *mot1-42* (TBP-Y185C), and *mot1-42* (TBP-F207L) was performed by Melissa Wells Carver with the validation experiments involving PCR and Northern Blotting.

2.2.3 Chromatin immunoprecipitation (ChIP)

ChIP assays were performed exactly as previously described (Dasgupta et al. 2005) with the following antibodies: for TBP ChIP, anti-myc (9E10) (Dasgupta et al. 2005); for TFIIB ChIP, a TFIIB rabbit polyclonal antibody (Dasgupta et al. 2005); and for RNA pol II ChIP, the Pol II monoclonal antibody 8WG16 (Thompson et al. 1989), (Bhaumik and Green 2001). ChIP material was then used for hybridization to tiling arrays or for quantitation by real-time qPCR. For the tiling arrays, library preparation and amplification of DNA for both ChIP and mock IP samples were performed using the GenomePlex Complete Whole Genome Amplification Kit (Sigma) as described (O'Green et al. 2006) with several modifications: dUTP was added in equimolar concentration to the dNTP mix (0.4 mM), and a second amplification was performed for both the ChIP

and mock samples using 10 ng of the previously amplified material. Samples were purified with the QIAquick PCR purification kit (QIAGEN) prior to re-amplification and fragmentation. Duplicate samples were combined to obtain 7.5 µg of material for fragmentation and labeling. Samples were hybridized to S. cerevisiae Tiling 1.0R arrays (Affymetrix, Inc.) (Experiments for TBP hybridization were done by Rebekka O. Sprouse and for TFIIB hybridization were carried on by Melissa Wells Carver) and raw data were generated by the Microarray Core Facility at UVA. Three independent biological replicates were analyzed for TBP and two independent biological replicates for were analyzed for TFIIB. Real-time qPCR was performed on ChIP, mock IP, and total samples as described for the RNA analysis. ChIP signals were obtained by subtracting mock IP signal from ChIP signal and normalizing against the input. ORF primers were identical to those used in the expression analysis. Three independent biological replicates were performed for each ChIP analysis. P-values were obtained by log transforming the calculated ChIP signals and a one-tailed, paired, t-test was conducted. The locus specific ChIP quantification and the analysis were performed by Melissa Wells Carver.

2.2.4 Tiling array data analysis

We performed both gene biased and unbiased analyses of the tiling array data. For unbiased analysis, we used the Affymetrix Tiling Array Software (TAS) which quantile normalizes replicate arrays, scales their median intensity to a user defined value and calculates the $-10\log_{10}$(p-value) and $\log_2$(pseudo-median) (or signal strength) associated with a one- or two-sample Wilcoxon signed rank test over a sliding window (Cawley et al. 2004). For the ChIP-chip analysis of TBP and TFIIB for *mot1-42* and WT strains, we applied the two-sample test of the ChIP sample compared to mock IP using a

window size of 500bp. For the total RNA analysis, we applied the one-sample test using a window size of 50bp. For the biased analysis, a gene-centric library (CDF) file was generated from yeast gene annotations (Fisk et al. 2006) and the Affymetrix tiling array library (BPMAP) file, which contained a probe set for every annotated yeast gene comprised of all probes whose central position fell within the annotated start and stop of the gene. Normalized gene expression estimates were obtained by quantile normalizing the arrays and applying GCRMA. Lists of differentially expressed genes were obtained using the limma package in Bioconductor and applying a 5% False Discovery Rate cutoff.

2.2.5 ChIP Tiling Array Data

Raw array data (CEL files) or the intensity data of the treatment (ChIP sample) and control (mock sample) were quantile normalized separately within replicate groups and the median intensity was scaled to a common target of 100. To determine the size and the significance of the difference between ChIP and mock sample data (signal estimates and the p values) we applied a Wilcoxon Rank-Sum test to log transformed PM-MM values, log2(max(PMi-MMi,1)), whose genomic coordinate 'i' (i.e., central position of 25mer PM/MM probe sequence) fell within a 500 bp (i.e., the ChIP fragmentation size) sliding window. The Hodges-Leman estimator, which is the estimator associated with the Wilcoxon Rank Sum test, was calculated along with -10log10(p-value). This analysis was applied to the following datasets:

1. WT TBP occupancy: 3 replicates of myc-tagged TBP ChIP as treatment and mock ChIP as control in WT cells.

2. *mot1-42* TBP occupancy: 3 replicates of myc-tagged TBP ChIP as treatment and mock ChIP as control in *mot1-42* cells.

3. WT TFIIB occupancy: 2 replicates of TFIIB ChIP as treatment and mock ChIP as control in WT cells.

4. *mot1-42* TFIIB occupancy: 2 replicates of TFIIB ChIP as treatment and mock ChIP as control in *mot1-42* cells.

5. Differential TBP occupancy: 3 replicates of myc-tagged TBP ChIP in *mot1-42* cells as treatment and 3 replicates of myc-tagged TBP ChIP in WT cells as control.

6. Differential TFIIB occupancy: 2 replicates of TFIIB ChIP in *mot1-42* cells as treatment and 3 replicates of TFIIB ChIP in WT cells as control.

The above analysis was performed using TAS (Tiling Analysis Software version 1.1) to generates graph files (.bar and .txt) files which contain signal and log transformed p-values as a function of genomic coordinates.

2.2.6 RNA Tiling Array Data

Estimates of RNA levels were made from the raw array data (CEL files) by first quantile normalizing all replicate arrays and scaling the data to a target median intensity of 100. We applied the Wilcoxon Signed-Rank test to the normalized log2(max(PMi-MMi,1)) values whose genomic coordinate 'i' fell within a 100 bp (i.e., a length much smaller than the typical ORF) sliding window to calculate the log transformed probability (-10log10(p-value)) that the RNA was detected above noise levels. The associated

Hodges-Leman estimator was used to estimate RNA levels. This analysis was applied to the following samples:

1. Total RNA in WT cells (MOT1 isogenic control, 2 replicates).

2. Total RNA in *mot1-42* cells (2 replicates).

3. Total RNA in WT cells (SET2 isogenic control, 2 replicates)

4. Total RNA in *set2*Δ cells (2 replicates)

Differential RNA levels were estimated using the same procedure described for the ChIP data including a window size of 500 bp with *mot1-42* RNA samples as treatment and WT RNA samples as control.

2.2.7 ChIP-chip Peak Identification

In order to identify the location and height of TBP and TFIIB ChIP-chip peaks, we started by applying a series of cutoffs to the signal data, from 0.3 to 4.3 in intervals of 1 (as illustrated in Figure 2-1), which result in a series of segments or intervals for each cutoff. To avoid characterizing spurious, noisy local peaks, we further joined intervals for every cutoff that are > 50 bp apart and eliminated those that are > 100 bp long. This yielded a series of intervals which encompass TBP and TFIIB peaks which were found by searching for the maximum signal in each interval.

2.2.8 Associating ChIP-chip Peaks to Genes

Using the TBP and TFIIB peaks, we associated a peak with a gene if it was within -300 to +50 bp of the annotated transcription start site (TSS). We allowed multiple genes which satisfy the distance criteria, to be associated with a single peak. Conversely, a

single gene can be associated with multiple peaks, each of which satisfies the distance cutoff separately.

2.2.9 Average Plots

We generated average RNA and ChIP enrichment plots for selected genes in Figures 2-8, 2-10, and 2-11 and 2-9A, C, D by setting the TSS of the genes to a common value of 0 (effectively aligning the TSSs) and averaging the signal over genes as a function of genomic coordinate. For Figure 2-8 B, the gene end was set to 0 and the signal was averaged over the gene upstream and downstream. We applied modest smoothing to the data by calculating the averages in a sliding window ranging from 1 to 50bp in length depending on the plot.

2.2.10 Heat Maps

The Heat maps shown in Figures 2-10 C- and D were generated by discretizing absolute and differential *mot1-42* versus WT TBP and TFIIB signals (i.e., maximum signal value within -300 to +50 bp of the TSS) into a 30x30 grid. The position of the box on the grid represents the median absolute or differential TBP (x-axis) and TFIIB (y-axis) value of the genes in the box. Similarly, absolute or differential RNA levels of genes falling within a box were summarized by their median value and assigned a color based on percentile (e.g. dark brown for upper 20% absolute or differential RNA levels).

2.2.11 Transcription Precision Pipeline (TraPP)

This pipeline was developed to measure and classify the defects in RNA transcript

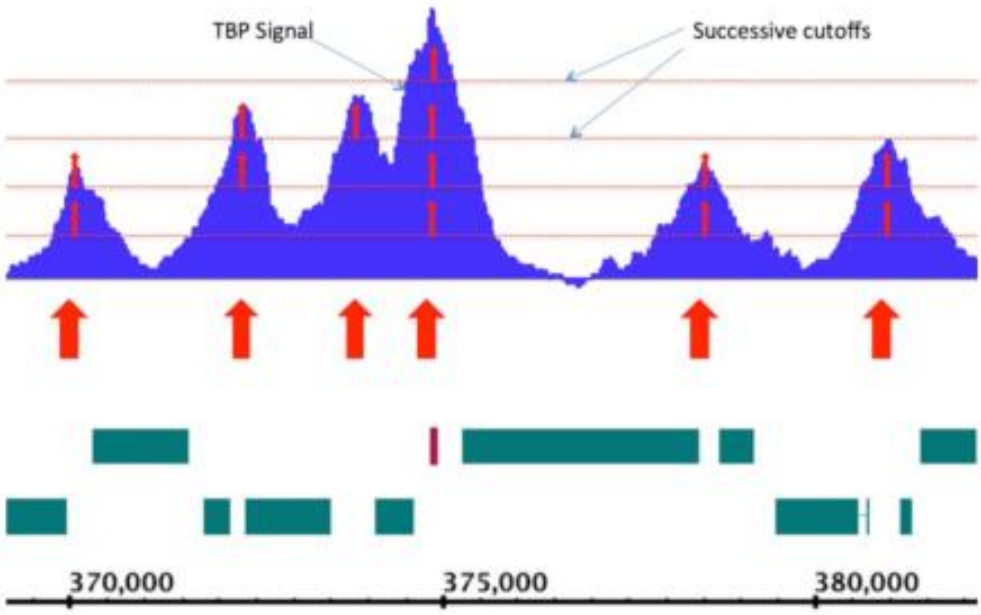


Figure 2-1: peak-finding method.

In each successive segment (horizontal line), the position of maximum signal was computed, thereby estimating the location of the bound factor (red arrows). Annotated Pol II transcribed genes are indicated by green bars, and the red bar denotes a tRNA gene.

lengths. We applied a 0.3 cutoff to differential RNA signal data generated using a 500 bp sliding window. By comparing these segments to annotations, we identified 5' and 3' length changes in *mot1-42* relative to WT transcription. These length changes were either positive or negative and represented a different RNA defect as described in Figure 2-3. Cases where the significant differential expression segment mapped to two or more annotated genes were separately characterized. Putative length changes that did not satisfy all the criteria below were filtered out:

1. The overlap between annotations and the significant differential signal segments was at least 100 bp long.
2. The length of the defect computed was > 150 bp long.
3. Median signal in the defect was significantly different from the baseline differential expression value as described in Figure 2-2. In the case of extensions where the segment fell beyond the boundaries of the annotation, the median signal in the extended region, S_{ext}, was greater than 0.44 to make a length change call. In the case of truncations, we defined two quantities: (1) $S_{overlap}$ as the median of the differential signal in the region where the differential segment and the annotation overlapped and (2) S_{int} as the median signal in the region internal to the gene that did not overlap the differential expression segment. For a truncation to be called, we required $S_{overlap} = S_{int} > 0.44$.
4. Spliced genes (283 total) were excluded because differential expression of spliced genes spans more than one CDS segment and were therefore erroneously called as length changes.

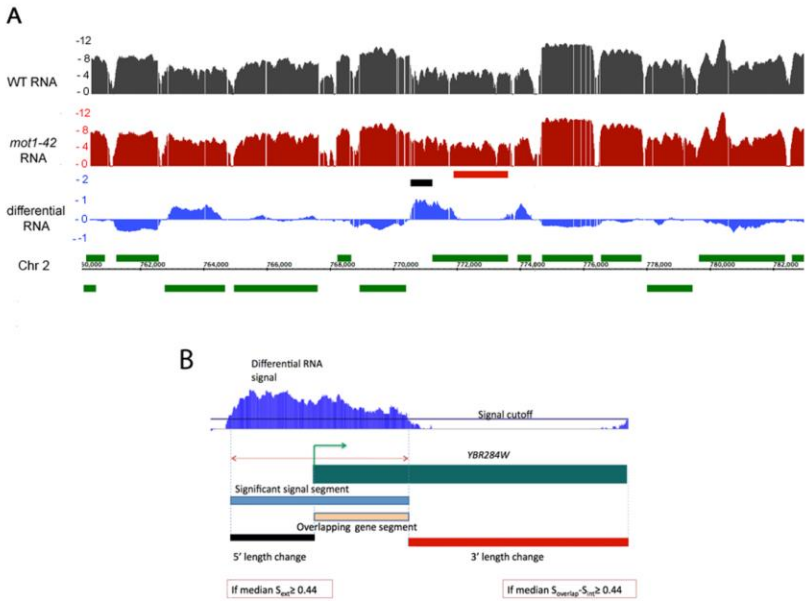


Figure 2-2: Overview of the TraPP method.

illustrated with an enlargement of the chromosomal region in the center of the screen shot in (A). In the case of the gene *YBR284W*, the significant differential RNA segment found using signal $\log_2$(*mot1-42*/WT) $\geq$ 0.3 (blue segment) was compared with the annotation, and aberrant transcript length changes (denoted by the red and the black segments) were calculated. Thresholds requiring a $\geq$ 100 bp overlapping gene segment and length changes $\geq$ 150 bp long were applied. $S_{overlap}$ and S_{ext} refer to the differential RNA signal in the region where the differential segment overlaps (orange segment) and is external (black segment) to the annotation, respectively. S_{int} refers to the differential signal within the portion of the annotation that does not overlap the differential segment (red segment). By provisionally defining the differential RNA with respect to the overlapping gene, the aberrant RNAs were sorted among 4 groups (see text). In this example, a 5' length change is classified as an upstream initiation event whereas the partial overlapping gene segment and consequent change in 3' length define a premature termination event.

2.2.12 SUT/CUT Analysis

We assessed the association between SUTs/CUTs (Neil et al. 2009) and significant changes in Mot1-mediated changes in RNA expression (i.e., differential level and/or location of expression) from the tiling array analysis. Genes were classified according to the location of SUTs/CUTs: (1) overlapping the gene; (2) located in the promoter of the gene and (3) located at the 3' end of the gene. We used BEDTools to perform the overlap analysis (Quinlan and Hall 2010). Based on an analytical formula, we found the overlap of SUTs/CUTs with RNA length changes was the expected number due to random chance. However the associations with RNAs that were differentially up or down regulated were found to be potentially significant from the same analysis. To further test this association, we randomly associated SUTs/CUTs with genes while maintaining the same number of overlapping, 5' and 3' proximal cases. We generated 10,000 random associations of SUTs/CUTs with genes and calculated both p-values and the enrichment of observed over random.

2.2.13 Comparison of Shifts in TBP Localization and RNA Length Changes

A number of analyses were performed to assess the extent to which aberrant TBP localization correlated with changes in RNA lengths. Similar to the RNA length change calculation, we first calculated the differential TBP signal between *mot1-42* and WT cells. We then calculated average plots of the differential signal near the TSS for all genes. We found a general tendency for the differential TBP peak to be localized closer to the TSS or even within the ORF when compared to the WT peak. This is consistent with the average plots shown in Figure 2-10 A where the differences between *mot1-42* and WT tend to be largest near the TSS. To assess if this differential shift in TBP signal

correlated with RNA length changes, we identified the differential peak positions and calculated the difference between the differential peak and the WT peak. For each of the four RNA length change classes as well as those that display no detectable length changes (i.e., nulls), we computed the distribution of □ values. We observed little to no significant differences between each of the RNA length change classes and the null class distributions. This was reflected in insignificant p-values derived from a Kolmogorov-Smirnov test between the null and each of the RNA length change classes separately. We then assessed the extent to which the shifts between the WT and the *mot1-42* TBP peaks correlated with the LI and EI length change classes. We first calculated the difference between *mot1-42* and WT peak positions as well as the difference between the apparent site of initiation of the new RNA and the annotated TSS. We calculated the Pearson correlation coefficient for these two sets of differences and found a correlation coefficient of- 0.51 for the downstream initiation events.

2.3 Results

We first compared RNA from wild-type (WT) and *mot1-42* yeast cells using Affymetrix genomic tiling arrays which interrogate the yeast genome at 5 bp resolution. *mot1-42* is a temperature-sensitive allele that encodes a protein that is biochemically inactive in vitro (Darst et al. 2003), and prior work established that this allele induces changes in gene expression in vivo that closely parallel other severe, conditional *mot1* alleles (Dasgupta et al. 2002). WT and mutant strains were grown at permissive temperature (30^{o} C), then shifted for 45 min to 35^{o} C prior to harvesting RNA. This temperature shift did not impair growth of WT cells, but dramatically inhibited growth of *mot1-42* cells (Darst et al. 2003). We calculated gene expression changes from the tiling

array data using methods similar to those applied to conventional gene expression arrays (see Materials and Methods). These expression changes were well correlated with expression changes defined previously (Sprouse et al. 2009; Figure 2-3), thus validating the use of tiling arrays for quantitative RNA analysis.

Although Mot1 exerts a global effect on transcription, most of these transcriptional effects are modest in magnitude. To better understand why Mot1-catalyzed TBP recycling is essential, we took advantage of the tiling arrays to determine whether Mot1-mediated TBP dynamics affect RNA precision as well as quantity. To test this idea, we developed a method to capture significant RNA hybridization signals that deviate from gene annotations (Figure 2-4). This approach was possible because in the overwhelming majority of cases, the WT RNA signals were closely aligned with annotated genes (an example is shown in Figure 2-2A, and others are discussed below). By comparing the coordinates of gene annotations with differential RNA segments, a methodology was developed that can detect RNA segments that extend outside a gene annotation, or differential signals that overlap with only a portion of an open reading frame. The gene-dense nature of the budding yeast genome presented challenges for the analysis because differential RNA signals could potentially overlap with more than one gene annotation, giving rise to different types of apparent RNA length changes based on the relative orientation of the two genes (Figure 2-4). Nonetheless, it was possible to segregate the differential RNA signals into four categories in which the RNA segment overlap was defined provisionally with respect to the gene annotation (Figure 2-2 B). Thus, “upstream initiation” segments are RNAs that extended upstream of the normal transcription start site, “downstream initiation” events are RNAs apparently initiated

within the open reading frame (a.k.a. cryptic initiation events), "premature termination" segments correspond to RNAs within an open reading frame that do not include the normal 3' end, and "downstream termination" events are those in which RNA extended beyond the termination site in WT cells. Although the arrays do not provide information about the strand specificity of the hybridized RNA, the strand-specific analyses described below confirm that most if not all of the detected events correspond to changes in RNA sense strand abundance.

To validate the array-detected RNA length changes, RNA from *set2*Δ cells was compared to WT using the same methodology. Loss of the H3 K36 methyltransferase activity in *set2*Δ cells is well known to result in cryptic initiation within open reading frames (Workman 2006). Consistent with the published data (Li et al. 2007a), 206 instances of cryptic initiation were detected in *set2*Δ cells, and these "downstream initiation" events comprised the largest class of RNA length changes by far (Table 2-1).

Among the *set2*Δ-induced array detected RNA length changes, we found and validated cryptic initiation in *STE11* and *PCA1*, two genes previously shown to be susceptible to cryptic initiation in *set2*Δ cells (Li et al. 2007; Figure 2-5).

Using the same approach, we characterized RNA length changes from differential *mot1-42* versus WT RNA. Strikingly, twice as many aberrant RNA species were detected in *mot1-42* cells compared to *set2*Δ cells (Table 2-1), and the *mot1-42*-induced length changes had a significantly different distribution among the four length change classes. Notably, a Mot1 defect led to 338 genes showing "premature termination", the most prevalent class of events. The "premature termination" events fell into two

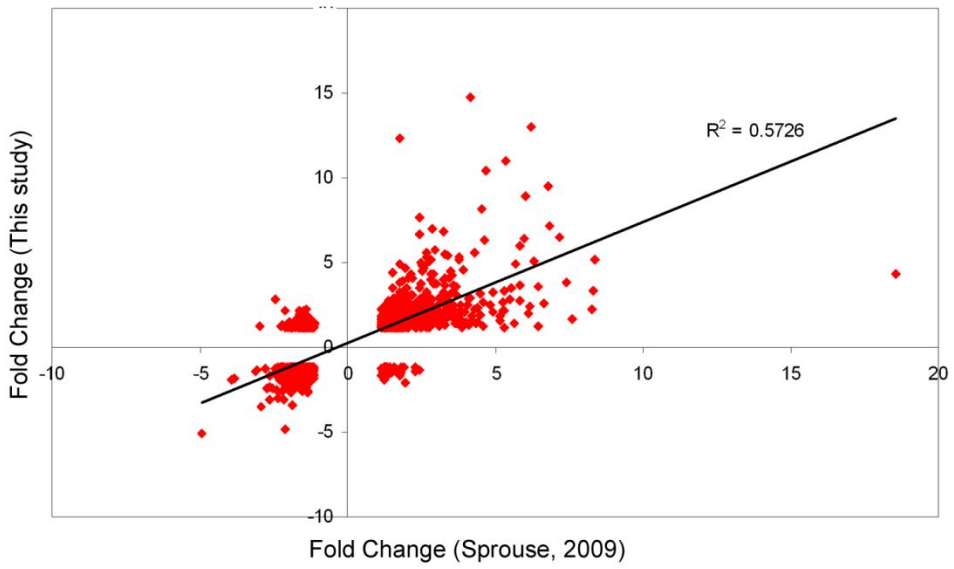


Figure 2-3 Comparison of *mot1* microarray fold changes from Sprouse et al. (2009) and the present study.

Note the linear correlation between Mot1-dependent fold changes derived from Agilent Yeast Oligo arrays (conventional microarrays; Sprouse et al. 2009) and those derived from this study using Affymetrix *S. cerevisiae* 1.0R tiling arrays. The slope of the best-fit line is 0.73. The technical differences in experimental design, array sensitivity, and data analysis account for why the datasets are correlated but the slope of the line is not equal to 1.

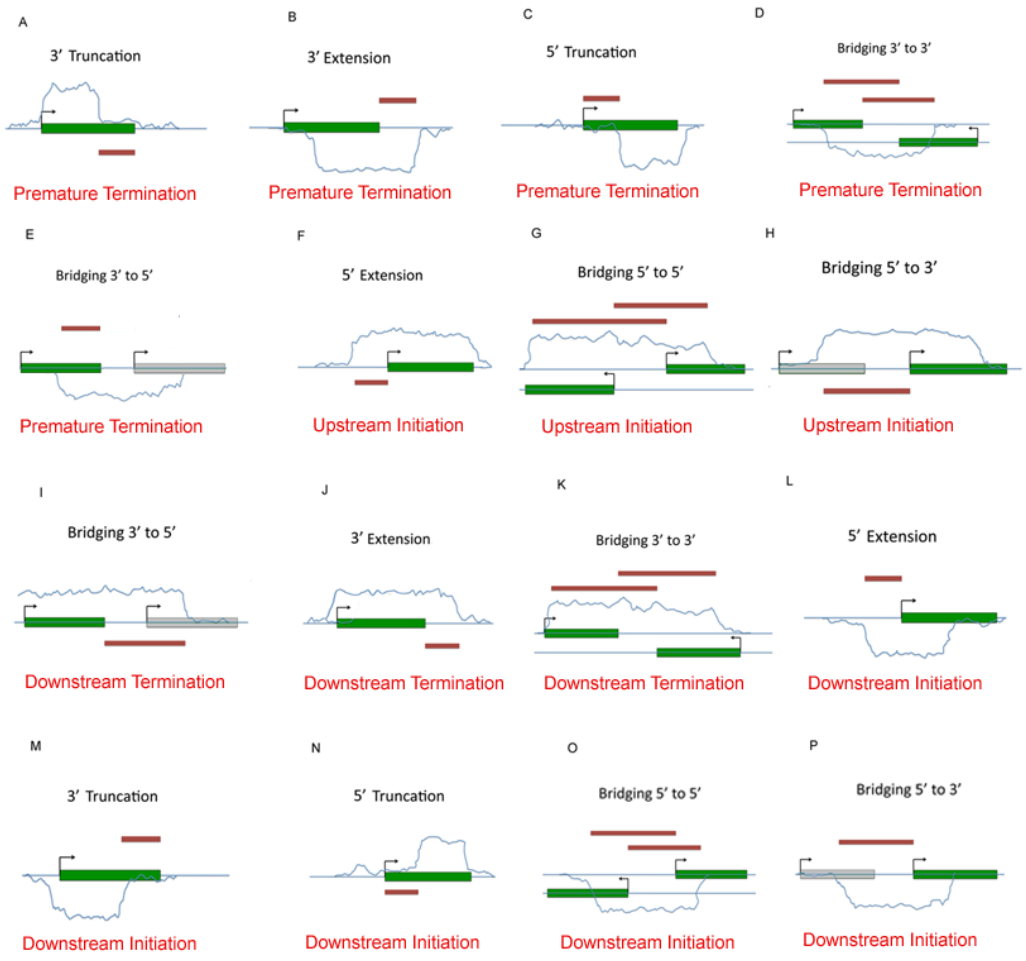


Figure 2-4: Biological complexity in the observed RNA length changes.

(A-P) The gene-dense nature of the *S. cerevisiae* genome, which gives rise to transcription of ~90% of the sequence, leads to complexity in the interpretation of observed RNA length changes, particularly when a differential RNA signal bridges two adjacent annotations. Schematic representations of RNA length changes are shown, with differential RNA shown in the blue curves and the brown segments showing the RNA length change segments called by the algorithm. The length changes were called with respect to a given gene or genes as shown in green.

RNA Length Changes	*set2*Δ	*mot1-42*
Upstream initiation	29	89
Downstream initiation	206	112
Premature termination	82	338
Downstream termination	45	174
Total	362	713

Table 2-1: The table summarizes the number of aberrant RNAs identified from the differential RNA signal *mot1-42* and *set2*Δ cells compared to WT cells. Of particular note is the large number of premature termination events in *mot1-42* cells and the enrichment in downstream (cryptic) initiation events in *set2*Δ cells.

categories: (1) “differential down” instances in which the RNA level was similar in the 5’ end of the open reading frame (ORF) but diminished in the 3’ end of the ORF in *mot1-42* cells compared to WT (77%) and (2) “differential up” instances in which a defect in Mot1 led to increased RNA in the 5’ portion of the ORF but the differential RNA failed to extend to the 3’ end of the ORF (23%). Although the *mot1-42* and *set2*Δ datasets have different numbers of genes and different distributions among the length change categories, there are 12 genes that displayed premature termination in both *mot1-42* and *set2*Δ cells. While few in number, the overlap is statistically significant ($p = 0.02$), suggesting that elongation efficiency in some genes may be under both Mot1 and Set2 control.

Selected RNA length changes were validated by real time PCR, including two genes with premature termination defects (Figure 2-6 A-G). Note that strand-specific real time PCR showed that the differential RNA effects are attributable largely if not entirely to changes in sense strand abundance. The results thus far support a role for

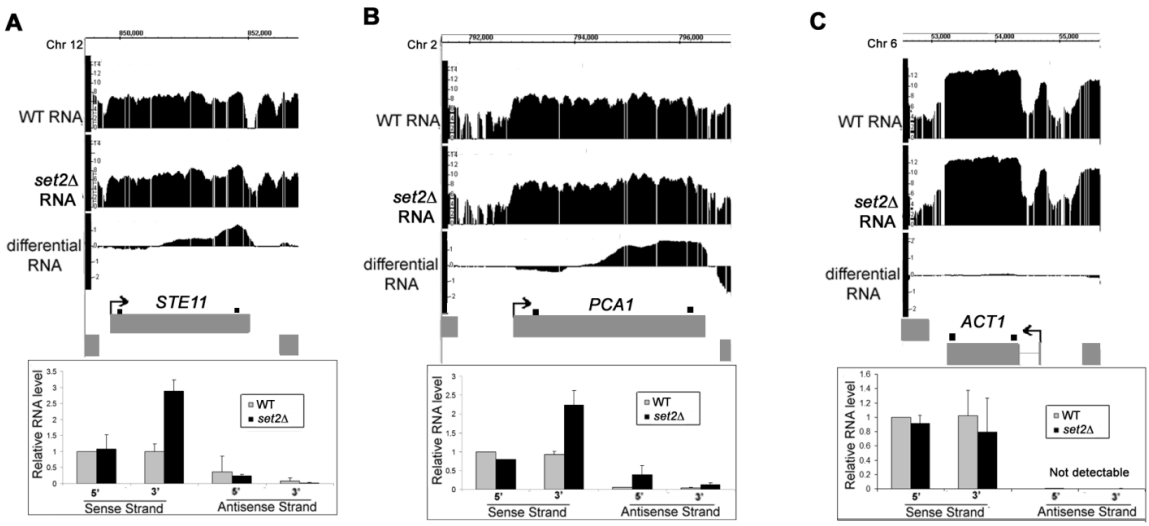


Figure 2-5: Confirmation of RNA length changes in *set2*Δ cells.

(A-C) Integrated Genome Browser screenshots of log2 WT, *set2*Δ, and differential RNA levels (*set2*Δ/WT) for *STE11*, *PCA1* and *ACT1* genes. The bar graphs show the relative RNA levels (arbitrary units) quantified by real time PCR using sense and antisense-specific 5' and 3' primer sets for each gene (depicted by small black boxes above each gene). Average values are shown + the standard error obtained by analysis of two independent RNA samples for each strain. Cryptic initiation was detected in *STE11* and *PCA1*, whereas there was no significant change in *ACT1* expression. The length change confirmation experiments were performed by Melissa Wells Carver.

Mot1 in maintaining RNA profiles that match annotated genes, but do not address whether this effect is mediated through transcription or some other effect on RNA processing. To address this question, ChIP was performed to assess the Pol II density on genes for which the RNA length changes occurred. As shown in Figures 2-6 differential changes in RNA level correlated with changes in Pol II occupancy as expected if the differential RNA signals arose through changes in transcription.

To further confirm the interpretations of the differential RNA tiling array data, we analyzed RNAs by Northern blotting using strand-specific probes. We chose genes associated with premature termination for which the full-length and predicted short RNAs were appropriately sized for detection on the blot, as well as being reasonably well resolved from each other. *THI2* displayed premature termination and was in the "differentially up" class. As shown in Figure 2-7 A, primarily two species of *THI2* RNA were detected in WT cells, a full-length species and a discrete shorter RNA which has not been characterized. In *mot1-42* cells, most of the full-length *THI2* RNA was slightly shifted to a position of smaller size (marked by asterisk), and a heterogeneous smear of shorter RNAs was detected (marked by bracket). This is consistent with the tiling array and RT-PCR data showing an increase in prematurely terminated RNA in *mot1-42* cells. As expected, all of the detectable RNA was sense RNA, confirming the premature termination designation. *PAN1* was also detected as a gene associated with premature termination, but in the "differentially down" class. *PAN1* is a Mot1-activated gene, and as expected, there was less full-length *PAN1* RNA in *mot1-42* cells compared to WT. A smear of shorter *PAN1* RNA was readily detected, and as expected for a gene in the "differentially down" class, *PAN1* short RNA was at least as abundant in *mot1-42* cells as

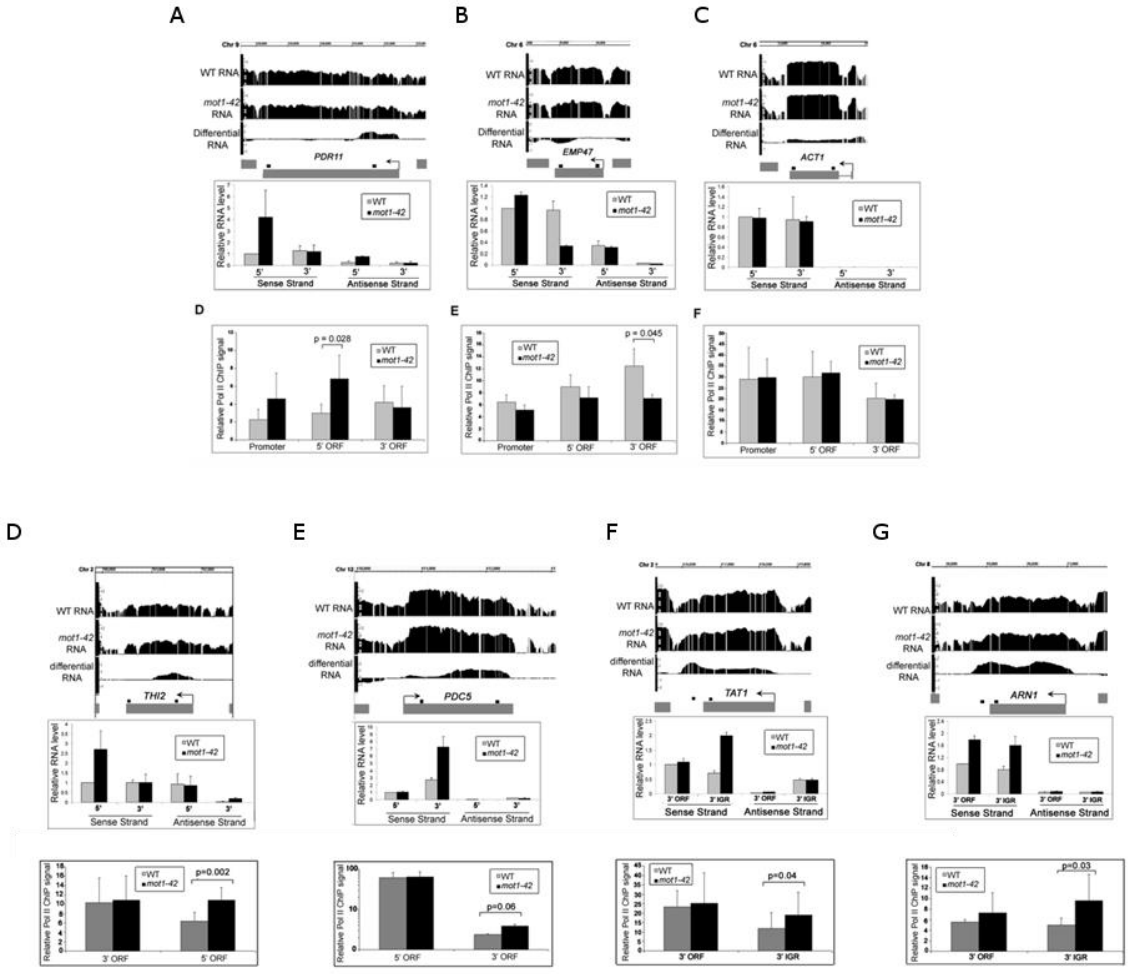


Figure 2-6: Validation of premature termination RNA length changes in *mot1-42* cells and correlation with Pol II density.

(A-G) Upper panel, screen shots showing WT, *mot1-42* and differential RNA signals across *PDR11*, *EMP47*, and *ACT1* genes. Middle panel, relative RNA levels were quantified by real time PCR for both the sense and antisense strands. Primers were specific for either the 5' or 3' end of each gene and the amplified product is represented by the small black square above each gene in the upper panel. Results shown are from the means of two independent RNA samples with associated standard error. *PDR11* is an example of a gene with increased 5' transcription in *mot1-42* cells (classified as

"differentially up") whereas *EMP47* had similar levels of 5' RNA but less 3' RNA in *mot1-42* versus WT cells (classified as "differentially down"). No RNA length change was detected for ACT1, and the change in RNA level across the open reading frame was small (< 5%) in comparison to the total level of total *ACT1* RNA in both strains. Anti-sense *ACT1* RNA was not detectable. Last panel show relative Pol II ChIP signals in the promoter, 5' end, and 3' end of each ORF. The results were obtained using the 8WG16 antibody and are shown as the mean of 3 independent biological replicates + the standard deviation. The indicated p-values were determined using a one-tailed paired t-test of the log-transformed ChIP values. Note the correspondence between the changes in Pol II ChIP and RNA length changes: the 5' ORF of *PDR11* had an increased level of Pol II in *mot1-42* cells that corresponded with increased 5' ORF RNA level. Similarly, Pol II ChIP signal was decreased in the 3' ORF of *EMP47*, consistent with the decrease in 3' RNA in *mot1-42* cells. As expected, there were no significant changes in Pol II ChIP for ACT1. Confirmation of cryptic initiation and 3' transcript length changes in *mot1-42* cells. (D-G) Screenshots of $\log_2$ WT, *mot1-42,* and differential RNA levels for 4 selected genes are shown in the upper part of each panel. Relative RNA levels quantified by real-time PCR using sense- and antisense-specific primers are shown in the bar graphs. The graphs show the average $\pm$ the standard error obtained by analysis of two independent RNA samples for each strain. The small black boxes above each gene depict the locations of each primer set. Bottom panel show relative RNA Pol II ChIP signals were obtained at the 5' end and the 3' end of the ORF (*THI2* and *PDC5*) or the 3' end of the ORF and the 3' intergenic region (*TAT1* and *ARN1).* The results were obtained using the 8WG16 antibody and are shown as the mean of 3 biological replicates ± the standard

deviation. The indicated p-values were determined using a one-tailed paired t-test of the log-transformed ChIP values. The same primer sets were used as the RNA analysis. *THI2* is an example of prematurely terminated RNA and *PDC5* is an example of downstream initiation. In contrast, *TAT1* and *ARN1* are examples of downstream termination RNAs. For all of the genes there is an increase in Pol II ChIP levels that correspond with the RNA length change. Note that the Pol II ChIP data for *PDC5* are shown on a log scale. The strong signal at the 5' end of the *PDC5* ORF is consistent with previously published data (Steinmetz et al. 2006). The experimental confirmation of length changes and Pol II ChIP were performed by Melissa Wells Carver.

it was in WT cells. Moreover, we observed that the shortest RNA species detected were more abundant in *mot1-42* cells compared to WT cells (small bracket). As expected, all of the detectable *PAN1* RNA was sense RNA. In contrast, *ACT1* RNA was a discrete species unaffected by *mot1-42* (Figure 2.7 C). Interestingly, the presence of shorter *THI2* and *PAN1* RNA in WT cells suggests that these genes are prone to premature termination even in WT cells, but that these effects are exaggerated by mutation of *MOT1*.

In a previous study (Sprouse et al. 2009), two TBP alleles were studied that bypass the requirement for Mot1 in vivo. These TBPs restored appropriate expression levels to many Mot1-regulated genes in vivo and allowed cells to grow without Mot1, which is otherwise essential (Sprouse et al. 2009). Biochemically, the bypass TBPs were defective for interaction with other GTFs or DNA, consistent with the critical activity Mot1 provides in destabilizing TBP-containing complexes in vivo (Sprouse et al. 2009). RNA tiling array analysis demonstrated that the two TBP bypass alleles suppressed the premature termination RNA length changes observed in *mot1-42* cells (Figures 2-8 A and B). Suppression of premature termination RNA synthesis from the "differential up" class was essentially complete, whereas the bypass TBPs partially restored efficient RNA synthesis to the "differentially down" gene class. Interestingly, the bypass TBPs were able to suppress each of the other RNA length change classes as well (Figure 2-8). We conclude that the *mot1-42*-mediated effects on RNA length can be explained by a direct effect of Mot1 on TBP dynamics.

Computational approaches were employed to determine if there are promoter features or aspects of local genomic organization that correlate with the RNA length changes observed in *mot1* cells. First, we found that upstream initiation and downstream

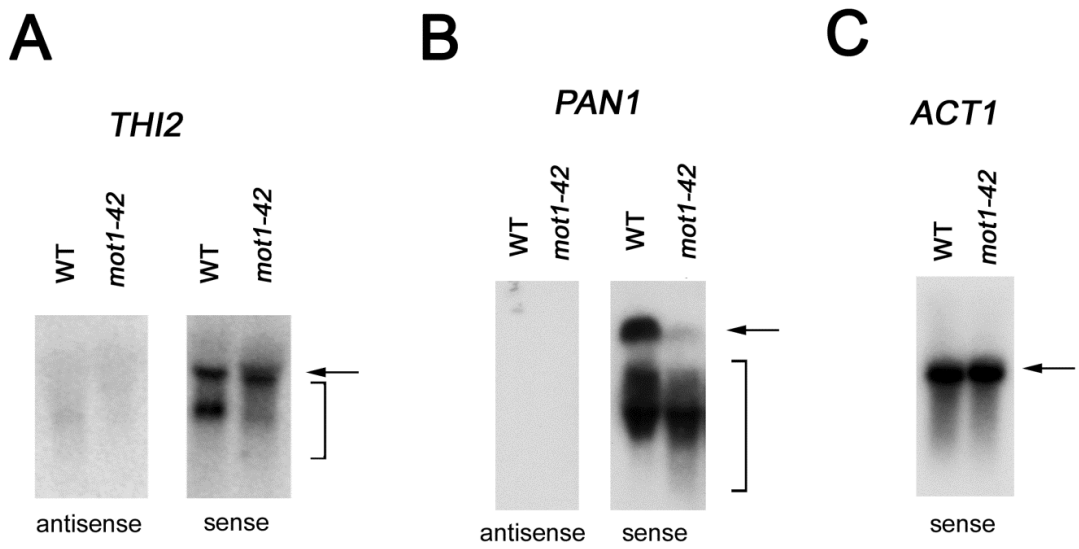


Figure 2-7: Confirmation of RNA length changes by Northern blotting.

Total RNA from WT or *mot1-42* cells was resolved by gel electrophoresis, transferred to nylon membrane and probed with radiolabeled single DNA strands to detect gene-specific sense or antisense RNAs, as indicated. (A), *THI2*. Full-length RNA is indicated by the arrow. Note the slight shift in mobility of full-length RNA (indicated by the asterisk) and the smear of smaller RNA species (bracket) in *mot1-42* versus WT samples. Only sense RNA was detected. The probes spanned +240 to +930 with respect to the start of the open reading frame. (B), *PAN1* was identified as a Mot1-activated gene displaying premature termination in the "differentially down" class. Consistent with this, note the decrease in full-length RNA (arrow) in *mot1-42* cells, which was in contrast to the shorter RNAs (large bracket) whose abundance was comparable in *mot1-42* versus WT cells. However, the small RNA species were distributed such that the shortest RNAs were more prominent in *mot1-42* cells compared to WT (denoted by small bracket).

Only sense RNA was detected. The probes spanned +420 to +864 with respect to the start of the open reading frame (C) *ACT1* control RNA was detected with a sense-strand specific probe. As expected, discrete full-length bands were detected in both WT and *mot1-42* cells, with no quantitative difference between them. The *ACT1* probe spanned +14 to +184 with respect to the start of the open reading frame. These experiments were performed by Melissa Wells Carver.

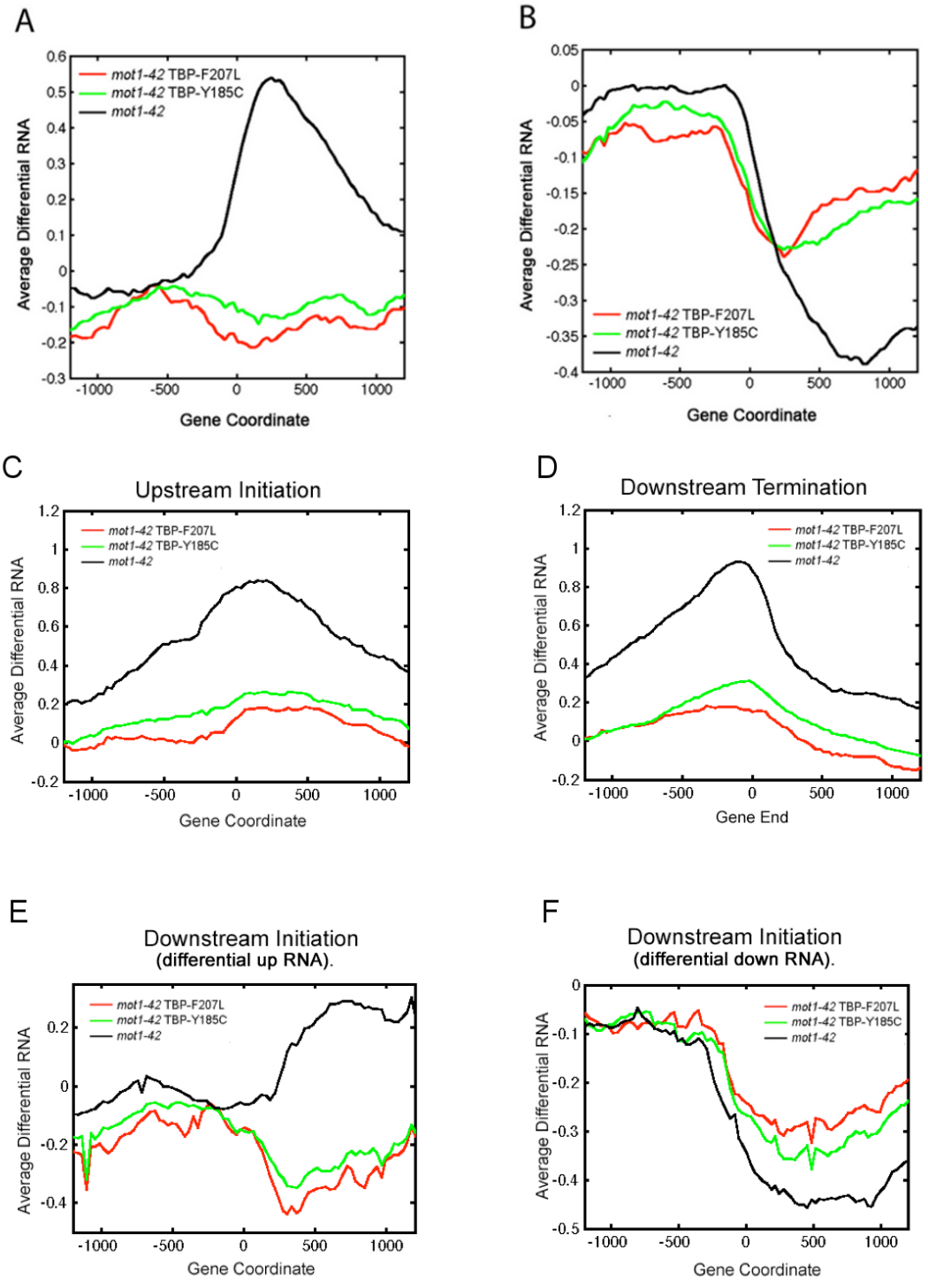

A
mot1-42 TBP-F207L
mot1-42 TBP-Y185C
mot1-42
Average Differential RNA
Gene Coordinate
B
mot1-42 TBP-F207L
mot1-42 TBP-Y185C
mot1-42
Average Differential RNA
Gene Coordinate
C
Upstream Initiation
mot1-42 TBP-F207L
mot1-42 TBP-Y185C
mot1-42
Average Differential RNA
Gene Coordinate
D
Downstream Termination
mot1-42 TBP-F207L
mot1-42 TBP-Y185C
mot1-42
Average Differential RNA
Gene End
E
Downstream Initiation
(differential up RNA).
mot1-42 TBP-F207L
mot1-42 TBP-Y185C
mot1-42
Average Differential RNA
Gene Coordinate
F
Downstream Initiation
(differential down RNA).
mot1-42 TBP-F207L
mot1-42 TBP-Y185C
mot1-42
Average Differential RNA
Gene Coordinate

Figure 2-8: Suppression of the premature termination defects in *mot1-42* cells by mutations in TBP.

(A, B) Profiles of average differential RNA—defined as the average of the differential RNA signal as determined in Supplemental Methods (RNA Tiling Array Data)- for genes showing premature termination length changes in *mot1-42* versus WT cells (black lines). The x-axis indicates the position along the chromosome in base pairs relative to the start of the gene annotation (zero). For comparison, the plots show the average differential RNA profiles for these same genes in *mot1-42* cells harboring TBP-F207L versus WT (red lines) and *mot1-42* TBP-Y185C versus WT (green lines). Premature termination length effects were sub-classified depending on whether the differential signal was positive (A, 77 genes) or negative (B, 264 genes). Average signals were obtained by smoothing over a 30 bp window. Suppression of transcription length changes in *mot1-42* cells by TBP bypass alleles: (A-F) Average profiles of differential RNA for genes showing upstream initiation (C), downstream termination (D), and downstream initiation (E, F) length changes in *mot1-42* versus WT cells (black lines). The x-axis indicates the position along the chromosome in base pairs relative to the start of the gene annotation (zero in C, E, and F) or gene end (zero in D). The average differential RNA profiles for these same genes in *mot1-42* cells harboring TBP-F207L versus WT (red lines) and *mot1-42* TBP-Y185C versus WT (green lines) show that the bypass alleles at least partially suppress each of the RNA length change categories.

termination classes of RNA length changes are enriched in genes whose promoters possess TATA boxes (Table 2-2; Basehoar et al. 2004). The TATA box association with premature termination genes was only detected for the "differentially up" class; the "differentially down" premature termination genes are significantly under-represented in TATA-containing genes. In contrast, no significant TATA box enrichment or depletion was seen for the downstream initiation genes. These results argue that the nature of the core promoter is related to the propensity to generate a certain type of aberrant RNA. This observation provides further support for the notion that these RNA length changes result from direct, promoter-mediated effects on transcriptional elongation.

Recent work has revealed two classes of noncoding RNA in yeast, termed SUTS (stable unannotated transcripts) and CUTS (cryptic unstable transcripts) (David et al. 2006; Davis and Ares Jr. 2006; Xu et al. 2009; Neil et al. 2009; Wyers et al. 2005). An extensive and detailed investigation uncovered no statistically significant relationship between the occurrence of a particular type of aberrant RNA in *mot1-42* cells and the occurrence of a SUT or CUT flanking or within the annotated gene in which the RNA length change was detected (Figure 2-9 and data not shown). As a second approach, we classified annotated genes as Mot1-activated, Mot1-repressed or Mot1-unaffected, based on the overall change in RNA level quantified in the gene-centric analysis of the tiling data described above (Figure 2-3). Interestingly, in this case, significant relationships were discovered between Mot1-regulated genes and the presence of a SUT or CUT proximal or overlapping the annotation. As shown in Figure 2-10, Mot1-repressed genes (scored as "differentially up") are enriched in genes with a SUT that overlaps their transcribed regions. These SUTs are transcribed in the opposite sense as the affected

Length Change Class	Number of promoters with TATA boxes	Total number of genes in the length change class	Percentage		p-value
Premature Termination	51	338	15%	-	0.16
Premature Termination Mot1-repressed	24	77	31%	enriched	0.002
Premature Termination Mot1-activated	28	264	11%	depleted	0.002
Upstream Initiation	31	89	35%	enriched	<0.001
Downstream Termination	75	174	43%	enriched	<0.001
Downstream Initiation	20	112	18%	-	0.48

Table 2- 2: Relationship between RNA length change class and occurrence of TATA box

Figure 2-9: Relationship between *mot1*-induced expression changes and CUT and SUT RNAs.

The diagrams illustrate the two significant relationships detected between Mot1-mediated gene expression and noncoding transcription. Mot1-activated genes tend to have an overlapping CUT transcribed in the antisense direction. The CUT transcripts initiate within the transcribed region of the Mot1-activated gene, ~40 bp on average upstream of the Mot1-activated gene termination site. In contrast, Mot1-repressed genes tend to have an overlapping antisense SUT. These SUTs initiate ~40 bp on average downstream from the termination site for the Mot1-repressed gene.

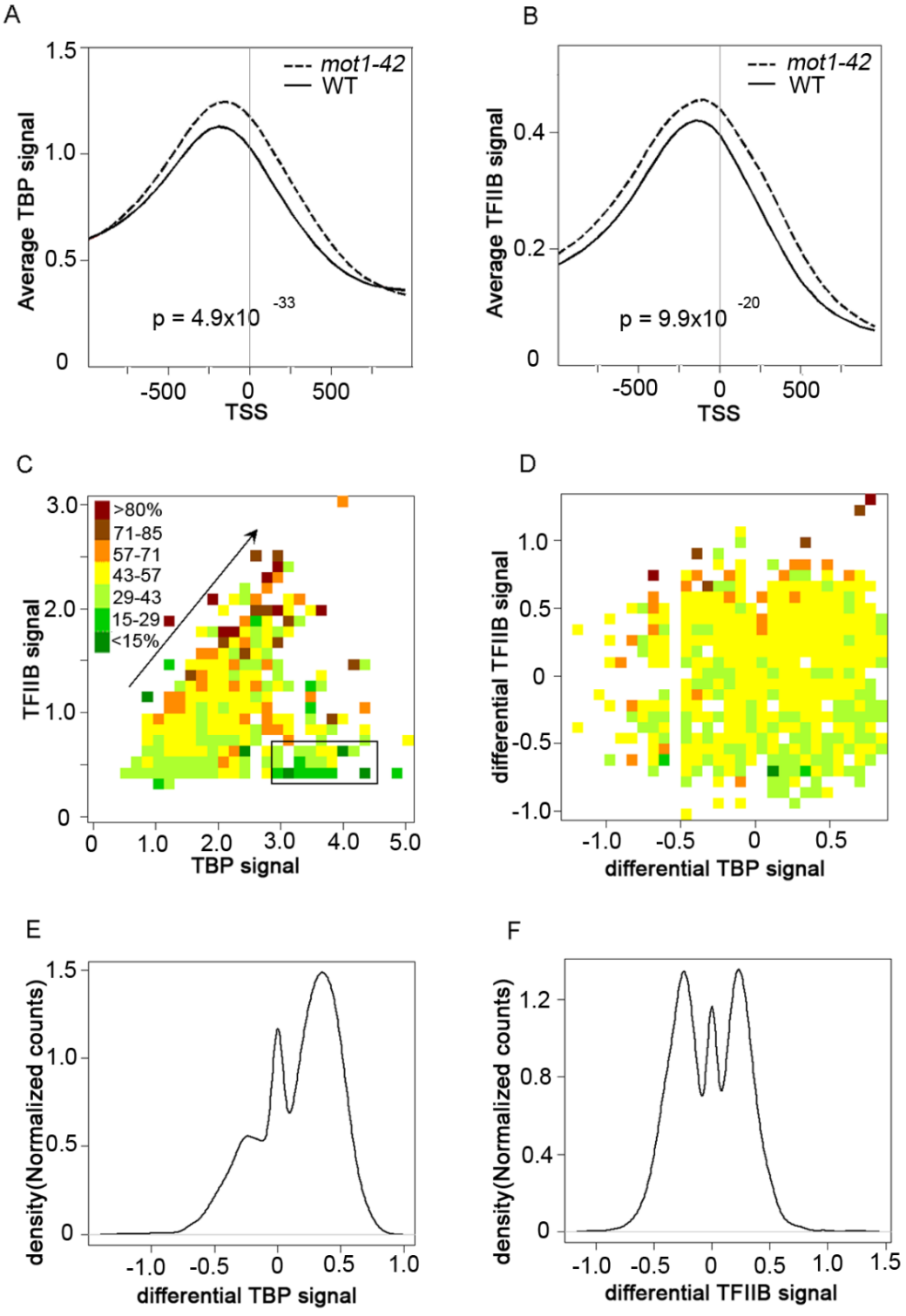

A
Average TBP signal
mot1-42
WT
p = 4.9x10 -33
TSS
B
Average TFIIB signal
mot1-42
WT
p = 9.9x10 -20
TSS
C
TFIIB signal
>80%
71-85
57-71
43-57
29-43
15-29
<15%
TBP signal
D
differential TFIIB signal
differential TBP signal
E
density(Normalized counts)
differential TBP signal
F
density(Normalized counts)
differential TFIIB signal

Figure 2-10: Global effects of Mot1 on TBP and TFIIB genomic distribution and correlations with transcription.

(A) Plots of average TBP and (B) average TFIIB chromatin occupancies obtained by aligning all annotated Pol II genes with respect to the transcription start site (TSS). The signals were smoothed over a 50 bp sliding window (WT, solid curves; *mot1-42*, dashed curves). In both cases, the differences in the distributions are highly significant as determined by the Kolmogorov-Smirnov test (p-values as indicated). (C) The heat map displays the transcriptome-wide relationship between TBP occupancy (x-axis), TFIIB occupancy (y-axis), and RNA level (color) in WT cells. Each box represents one or more genes whose TBP and TFIIB occupancies fall within the box's x-axis/y-axis values. The box color is the relative median expression value on a scale defined by the range of all medians in the dataset. Note the general trend that increasing TBP and TFIIB promoter occupancy is correlated with increasing RNA level (arrow). However, the correlation between TFIIB and RNA levels is better because there are genes with high TBP occupancy but low expression (black rectangle). (D) The heat map is similar to panel C, but shows the transcriptome-wide relationship between the *differential* TBP signal (x-axis), the *differential* TFIIB signal (y-axis) and the *change* in RNA level (color) in *mot1-42* versus WT cells. As in (C), changes in TFIIB occupancy are reasonably well correlated with changes in transcription whereas changes in TBP promoter occupancy do not correlate as well with changes in RNA level. (E, F) Density distributions of the differential TBP and differential TFIIB signals (as indicated), over the promoters for all the genes. The plots show that in *mot1-42* cells compared to WT, TBP occupancies increased at the majority of promoters. In contrast, TFIIB occupancies increased or

decreased at roughly equal numbers of promoters, consistent with changes in gene expression in both positive and negative directions.

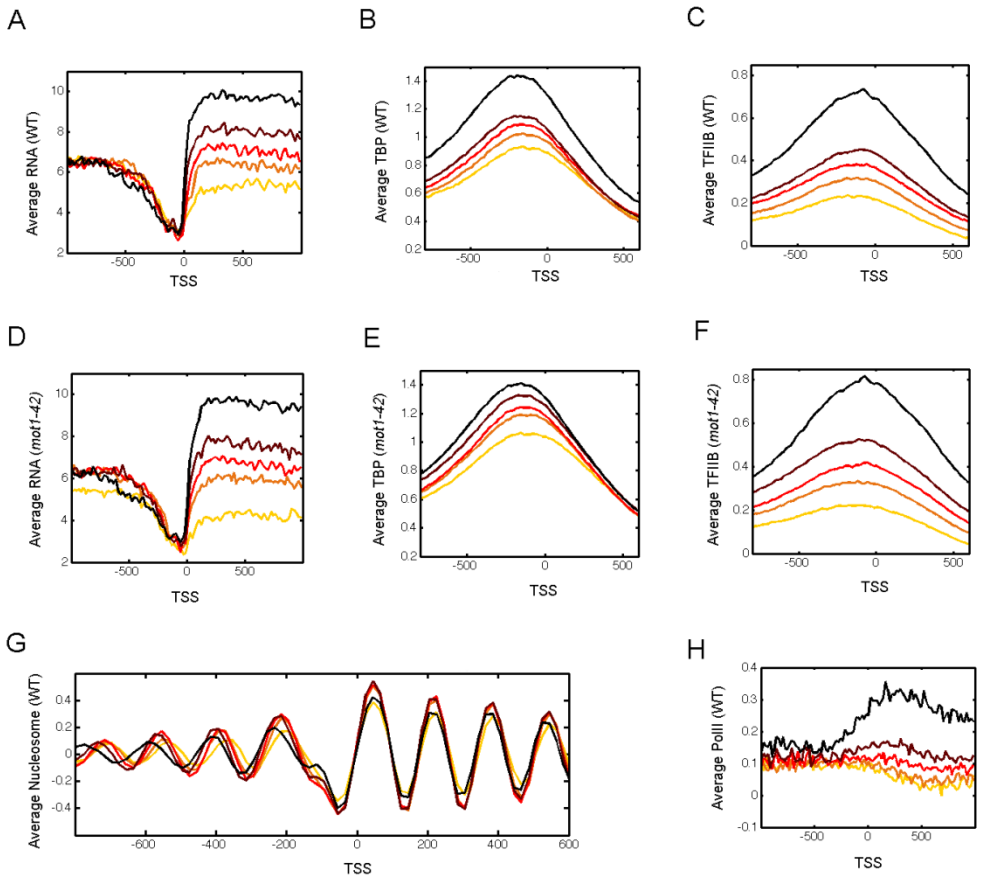


Figure 2-11: Average profiles of TBP, TFIIB, nucleosome positioning, and Pol II stratified by expression level.

(A) Average RNA plots for genes in WT cells stratified by expression level (i.e., quintiles). (B, C) Average plots for TBP and TFIIB (as indicated) in WT cells for genes grouped by RNA level depicted in (A). The curves were generated using a 20 bp sliding window. Note that on average, TBP and TFIIB levels increased with increasing RNA level. (D-F) Average profiles for RNA, TBP and TFIIB as in (A-C) but for genes in *mot1-42* cells. (G) Average nucleosome profile for genes segregated by expression level. Nucleosome data were obtained from Whitehouse et al. 2007. (H) Average Pol II

occupancies in WT cells (Steinmetz et al. 2006) stratified by relative RNA level. The curves were made with 50 bp window smoothing of the signal.

gene. On the other hand, Mot1-activated genes (scored as "differentially down") tend to have an overlapping CUT near the 3' end of the gene. Again, these CUTs are transcribed in the opposite sense as the affected gene. Although the underlying mechanisms are unknown, collectively, these results suggest that Mot1's effects on transcription are influenced by antisense transcription of SUTs or CUTs with a particular proximal relationship to the affected gene.

Next, we sought to correlate TBP chromatin localization with the global changes in transcription observed in *mot1* cells. Published work has documented TBP distribution genome-wide (Kim and Iyer 2004; Zanton and Pugh 2004; van Werven et al. 2008; Venters and Pugh 2009), as well as the Mot1 distribution and its correlation with TBP genome-wide(van Werven et al. 2008; Geisberg and Struhl 2004; Zanton and Pugh 2004). To determine the genome-wide dependence of TBP localization on Mot1 and to compare Mot1-mediated localization to changes in RNA length, chromatin immunoprecipitation was performed using these same tiling arrays (ChIP-on-chip). These results allowed us to define the locations of TBP binding as well as the relative TBP occupancies genome-wide in both WT and *mot1-42* cells. To distinguish productive TBP binding from nonproductive binding to Pol II promoters, ChIP-on-chip was performed in parallel for TFIIB. The results obtained with WT cells are in good agreement with published data (van Werven et al. 2009; Venters and Pugh 2009). For example, TBP and TFIIB were localized near each other and primarily in promoters, 126 and 123 bp, respectively, upstream from the transcription start site (TSS) (Nagalakshmi et al. 2008; Figures 2-10 A and B). As expected, these locations correlate well with nucleosome-free regions (Whitehouse et al. 2007; Figure 2-11 and data not shown). In

addition, although there was a weak positive correlation between TBP promoter occupancy and gene expression, TFIIB promoter occupancy in both WT and *mot1-42* cells tracked much more closely with RNA level (Figures 2-10 C).

Interestingly, in *mot1-42* cells, TBP promoter occupancies were increased genome-wide compared to WT cells (Figures 2-10 A and B). More detailed analyses supported the conclusion that overall, TBP occupancies increased in promoters across the board regardless of whether Mot1 exerted a detectable effect on expression of the gene (Figure 2-10 and data not shown). The implications of these observations are discussed below. Figure 2-10 C shows the global relationship between TBP promoter occupancy (x-axis), TFIIB promoter occupancy (y-axis), and RNA level (color). The diagonal arrow shows that for a significant number of genes, increasing TBP and TFIIB promoter occupancy was correlated with increasing RNA level, as indicated by a transition in the color along the diagonal from green (lower RNA level) to brown/red (higher RNA level). However, the plot also shows that TFIIB promoter occupancy was much better correlated with RNA level than TBP, as indicated by the substantial number of genes with low associated RNA levels (colored green) but with high TBP occupancy (demarcated by the rectangle). The global effects of Mot1 on gene expression, as well as TBP and TFIIB occupancy are shown in Figure 2-10 D. Differential TFIIB promoter occupancy was reasonably well correlated with differential expression mediated by loss of Mot1. Note that an increase in gene expression in *mot1-42* cells (brown/red) was generally associated with an increase in TFIIB promoter occupancy (higher y-axis value), whereas decreases in gene expression (green) generally displayed a decrease in TFIIB promoter occupancy (lower y-axis value). In contrast, there was no obvious correlation between the change in

gene expression (color) and the change in TBP promoter occupancy (x-axis value). Figures 2-10 E and F support these general conclusions. These histograms show how the global distributions of TBP and TFIIB changed in *mot1-42* versus WT cells. Most promoters (60%) had increased TBP occupancies in *mot1-42* versus WT cells, as illustrated by the large peak of positive differential TBP in Figure 2-10 E. The small peak centered over zero indicates that there were some genes (20%) whose TBP occupancies did not change, and the left-hand shoulder reflects promoters whose TBP occupancies decreased in *mot1-42* cells. In contrast, as shown in Figure 2-10 F, the two large peaks indicate that TFIIB occupancies were increased or decreased at roughly equal numbers of promoters (39% increased and 42% decreased). The correlation of differential TFIIB signal with differential RNA (Figure 2-10 D) indicates that the bimodal distribution of differential TFIIB in Figure 2-10 F was a consequence of the nature of the Mot1-mediated transcriptional effect. To further explore the nature of Mot1-mediated effects on TBP, a peak finding algorithm was employed to map more precisely the loci of protein binding (Figure 2-12). Most promoters possessed single TBP peaks (67.4%), about a third (32.7%) possessed TFIIB peaks, and close to half of the detected TBP peaks (47.3%) were associated with a TATA motif (Figure S4). Thus, in many instances the effects of Mot1 on TBP and TFIIB promoter occupancy appear to reflect changes in the occupancies of single, discrete complexes formed on promoters.

Finally, using several computational approaches, we investigated the relationship between TBP occupancy and the RNA length changes that occurred in *mot1-42* cells (see Supplemental Materials and Methods). For the altered initiation events in particular, the relatively small number of affected genes made statistical analysis difficult. Nonetheless,

we observed that the "downstream initiation" events appear to have a relatively straightforward origin: in *mot1-42* cells, the shift in TBP localization parallels the apparent site of initiation of the new RNA (Pearson correlation = 0.51, data not shown). This suggests that changes in initiation occurred because Mot1 failed to clear TBP from cryptic sites that can nucleate the assembly of functional transcription complexes. It is unclear how a defect in Mot1 could give rise to RNAs with extended 3' ends, but the correlation of these different aberrant RNA species with promoter type (Table 2-2) suggests that termination or 3' processing events can somehow be influenced by TBP dynamics, depending on the type of promoter.

2.4 Discussion and Future Directions

The results presented here reveal several new and unanticipated relationships between TBP dynamics and transcription. Notably, we find that impaired turnover of TBP at promoters correlates with the production of a predominant class of "premature termination" RNAs. The Pol II ChIP and TBP allele suppression results argue that Mot1-mediated effects on RNA length distribution are attributable to changes in transcription of the affected genes, and that these occur as a consequence of altered TBP dynamic behavior at promoters. Moreover, the genes that display RNA length changes tend to have certain promoter attributes (TATA versus TATA-less). The stochastic failure of Pol II elongation that ensues when TBP dynamics are perturbed suggests that Mot1-catalyzed clearance of TBP may be important for promoter recruitment or activity of accessory factors that subsequently promote Pol II elongation. Another attractive model is that there may be communication between factors associated with the promoter and the

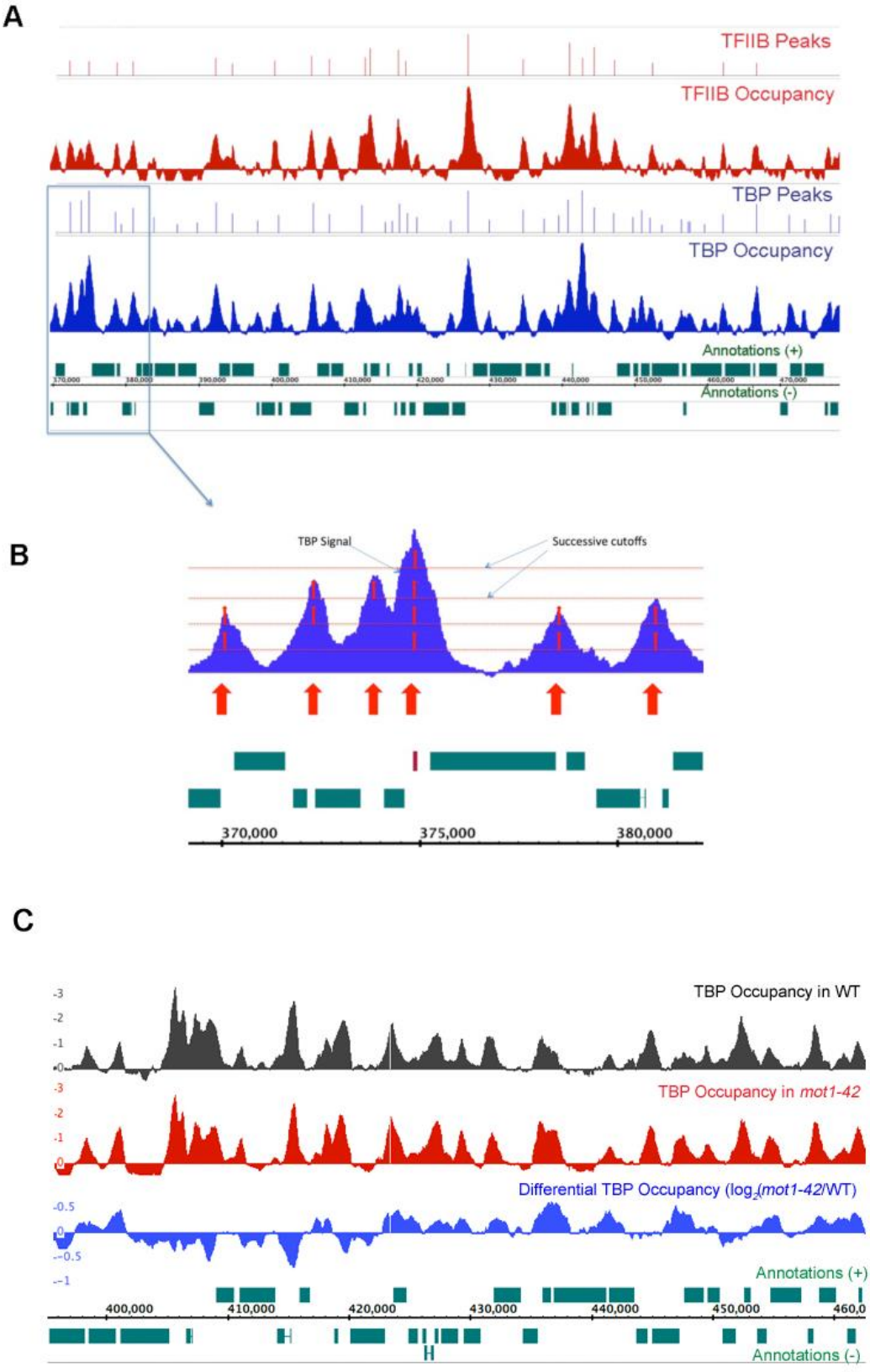

A
TFIIB Peaks
TFIIB Occupancy
TBP Peaks
TBP Occupancy
Annotations (+)
Annotations (-)
B
TBP Signal
Successive cutoffs
370,000
375,000
380,000
C
TBP Occupancy in WT
TBP Occupancy in mot1-42
Differential TBP Occupancy (log2(mot1-42/WT)
Annotations (+)
Annotations (-)
400,000
410,000
420,000
430,000
440,000
450,000

Figure 2-12: Overview of TBP and TFIIB peak-finding method.

(A) Screenshot of a region of chromosome 14 showing TBP occupancies in WT cells (black) and TFIIB (red) ChIP-on-chip signals. The vertical lines shown above each signal track indicate the positions of TBP and TFIIB identified by the algorithm and the annotated genes by green bars (Pol II transcribed) and a red bar (Pol III transcribed gene). (B) Enlargement of a portion of panel (A) shows the principle of the method. In each successive segment (horizontal line), the position of maximum signal was computed, thereby estimating the location of the bound factor (red arrows). Annotated Pol II transcribed genes are indicated by green bars, and the red bar denotes a tRNA gene. (C) Screenshot of a region of chromosome 2 showing TBP occupancy in *mot1-42* (red), TBP occupancy in WT (black), and differential TBP occupancy (*mot1-42*/WT; blue). Annotated genes are shown as green bars.

A

	TBP	TFIIB
Total Peaks - WT	5226	1932
Total Peaks - *mot1-42*	4810	2461

B

Number of TBP Peaks	WT	*mot1-42*
0	2055	1997
1	4697	4772
2	215	194
3	1	5

C

Number of TFIIB Peaks	WT	*mot1-42*
0	4649	4122
1	2279	2756
2	40	87
3	0	3

D

	WT	*mot1-42*
TBP Peaks associated with TATA box	2474	2275
TBP Peaks not associated with TATA Box	2752	2535

Table 2- 3: Summary of counts obtained from the peak finding algorithm.

(A) The total number of TBP and TFIIB peaks computed by the peak finding algorithm in WT and *mot1-42* cells. (B) The table shows the distribution in the number of TBP peaks associated with a given promoter. Note that most promoters (4697) were associated with a single TBP peak in WT cells. (C) The table shows the distribution of the number of TFIIB peaks associated with a given promoter. Note that most promoters (4649) did not have an associated TFIIB peak in WT cells. (D) The table shows the distribution of TBP peaks with and without TATA boxes. TATA boxes were defined as in Basehoar et al. 2004. Note that in contrast to the results shown in Table 2-2, these assignments of TBP peaks to TATA sequences were made without regard for whether the detected TATA box was located within a promoter. Overall, about 20% of Pol II promoters possess a TATA box; the large number of TATA-associated TBP peaks in this analysis indicates that a substantial number of peaks are not in promoters.

elongating Pol II. A possible explanation is that a physical connection persists between the promoter and elongating Pol II which can stall elongation some distance from the promoter. The role of Mot1 then would be to facilitate dissolution of the TBP complex at the promoter to release a restrained Pol II. Alternatively, there may be communication between the promoter and an elongating Pol II to facilitate transit through chromatin or to ensure that downstream RNA processing events are coupled to transcription. Regardless of the mechanism, it is striking that a large number of yeast genes rely to some extent on TBP dynamics to ensure accurate and efficient RNA elongation.

The genome-wide increase in TBP occupancy observed in *mot1-42* cells compared to WT reported here fits well with the rapid mobility of essentially the whole TBP pool assessed by live cell imaging (Sprouse et al. 2008) and shows that TBP occupancy is generally limiting at promoters in vivo. On the other hand, although Mot1 regulates a substantial proportion of the transcriptome, not all yeast genes are affected in *mot1-42* cells. Presumably, Mot1-affected genes are especially sensitive to TBP occupancy whereas other genes are rate-limited in some other critical step of the transcription cycle. A recent study concluded that TBP turnover rates are different at different classes of promoters (van Werven et al. 2009). This study relied on a replacement strategy in which expression of a differentially tagged form of TBP was induced and the rate of its association with promoters was tracked. As the production of the new TBP pool required on the order of 30 min, this approach would not capture the very rapid Mot1-catalyzed dynamics (occurring on the order of seconds) that our previous work has shown accounts for the behavior of the majority of the TBP pool (Sprouse et al. 2008). Although no genomic scale method exists yet for detection of such

rapid, locus specific dynamic behavior, the analysis of the TBP occupancy profile in *mot1-42* cells reported here significantly extends prior work by showing that Mot1 does target virtually the entire pool of promoter bound TBP in vivo. This observation provides support for a recent study that modeled the dynamic behavior of GTFs based on previously published ChIP-chip data, which found that interactions between TBP and chromatin are best described by models in which the interactions are rather transient (Samorodnitsky and Pugh 2010). Our suspicion is that GTF-promoter interactions span a wide range of lifetimes. Because the long-lived interactions detected by kinetic ChIP experiments appear to represent a relatively small (albeit very important) proportion of the total number of chromatin interactions, they may well be below the limit of detection in live cell imaging experiments which capture the behavior of the overall GTF pool.

Recent results show that constitutive yeast genes are expressed by infrequent initiation events clearly separated in time (Zenklusen et al. 2008). These observations also fit well with our measurements of highly dynamic TBP in vivo (Sprouse et al. 2008) and suggest that in contrast to establishing a stable scaffold, many active promoters are subject to occupancy by transcription complexes that undergo rapid cycles of assembly and disassembly. Such dynamic behavior has been speculated to be important for ensuring appropriate start site selection and timely transcriptional regulatory responses (Auble 2009). However, the results presented here support the notion of a fundamental role for PIC dynamics in the process of RNA synthesis itself. More generally, the ability to rapidly characterize the spectrum of aberrant RNAs present in a particular mutant background using the approach outlined here will likely be of use in unraveling the molecular mechanisms responsible for these effects.

The TraPP (Transcription Precision Pipeline) developed in this study has been proven to be a powerful tool to quantify the transcription precision defects in mutants. This tool has also been successfully applied to a study showcasing the role of a new histone modification (Hyland et al. 2011), and to characterize of the Sen1 RNA termination pathway as described in Chapter IV. Currently is also used to characterize precision defects in *spt16* and *mot1* and *spt16* double mutants by Jason True and Joseph Muldoon.

The methodology to measure transcription precision has a lot of scope for improvement. We developed a Multi pass TraPP methodology to be used for datasets with very high variation in differential expression. For these kind of dataset we take multiple cut-offs to define the segments of significant differential expression and use the TraPP for each cut-off. And finally we combine all the captured defects and filter for significant defects. This methodology is applied to capture transcription precision defects in *sen1* dataset in Chapter IV of this dissertation. First a better segmentation method could be applied to the current dataset for capturing the changes in transcription. This will increase our efficiency for defining precision defects as we will get better in defining the transcription boundaries. This would be more efficient from the current method of just taking a single cut-off over the differential RNA signal. One such method is described in (), this method was found to be very computationally intensive. In the current dataset of tiling arrays we only get the information for the sense strand and the use of strand specific arrays would give much more information about the biology and origin of these RNA defects. For example we will be able to know that upstream initiation events for our dataset arise from sense or antisense strand, also weather the

downstream termination events are coming from the antisense transcription from tandem gene or they are real downstream termination. Although all the above improvement could be made by using high throughput sequencing data to measure the transcription precision defects. Although we lack experience in using TraPP on RNA-seq but using sequencing would give us strand specific information and also precise location of where the transcript originate and terminate. Although using sequencing to measure transcription precision is the way to go, further work is needed to develop the analysis pipeline.

Chapter III

CLK ChIP reveals dynamics of TBP invivo

The work presented in this Chapter has been published in Poorey et al. 2013

The chromatin immunoprecipitation (ChIP) assay is widely used to capture interactions between chromatin and regulatory proteins, but it is unknown how stable most native interactions are. No general method thus far can measure short-lived site-specific binding events that live cell imaging suggests are prevalent in vivo. Here we show using a modified ChIP assay with sub-second temporal resolution that the time dependence of formaldehyde crosslinking can be used to extract in vivo on- and off-rates for site-specific chromatin interactions varying over a ~100-fold dynamic range. Using the method, we show that a novel regulatory process shifts weakly bound TATA-binding protein to stable promoter interactions, thereby facilitating transcription complex formation. This assay provides an approach for systematic, quantitative analyses of chromatin binding dynamics in vivo**.**

3.1 Introduction

In the ChIP assay, cellular constituents are crosslinked with formaldehyde, the isolated chromatin is fragmented, and protein-DNA complexes are then recovered by

immunoprecipitation using an antibody that detects a chromatin-associated protein of interest. DNA sequences in the immunoprecipitate are then inventoried by PCR. The assay accurately defines where proteins bind (Rhee and Pugh 2011, 2012), but it provides limited information about how stable the interactions are. For example, a relatively high ChIP signal could reflect high occupancy stable binding, or that a dynamic interaction was efficiently trapped owing to the long formaldehyde incubation period employed in standard assays (Kuo and Allis 1999). In fact, live cell imaging approaches indicate that many chromatin interactions are exceedingly short-lived (Hager et al. 2009; van Royen et al. 2009), although such techniques do not provide high resolution regarding chromatin binding location. Precise chromatin location information can be obtained by competition ChIP, a method that monitors the replacement rate by a differentially tagged factor of interest. However, the method only provides relative turnover rates and the time resolution is limited to ~20 min owing to the delay required to generate the competitor species (e.g., (van Werven et al. 2009; Deal et al. 2010; Lickwar et al. 2012)). A general assay that provides quantitative measures of site-specific on- and off-rates is essential for defining and modeling chromatin regulatory events as they occur in vivo.

3.2 The theory

In this section we provide an overview of the CLK concept, the derivation of the mathematical model, a detailed description of how the method is implemented computationally, and our interpretation of the parameters yielded by the method.

3.2.1 Overview of the CLK Model

The ChIP assay is the most widely used experimental approach for determining where chromatin-binding factors interact with DNA in vivo. However, the standard assay does not provide clearly interpretable quantitative information about chromatin binding. Changes in ChIP signal under different conditions or in comparing different binding sites can be interpreted in many ways, and most importantly, the signal derived from a standard ChIP assay does not provide information about binding kinetics or the fractional occupancy across a cell population. The CLK method capitalizes on the power of the ChIP assay to provide precise location information and extends it so as to provide quantitative kinetic information on a broad time scale.

We applied chemical reaction rate theory to model what happens during a ChIP experiment. The concept is that the dependence of ChIP signal on formaldehyde crosslinking time can be used to extract site-specific kinetic information for a chromatin-binding factor of interest. A simple kinetic model of transcription factor (TF) binding to DNA followed by crosslinking is given by

$$unbound\ TF \underset{r_{1b}}{\overset{r_{1f}}{\rightleftarrows}} bound\ TF \xrightarrow{r_2} crosslinked\ TF$$

where the first reaction represents reversible transcription factor binding to DNA. The overall on rate r_{1f} denotes the forward reaction rate for binding and r_{1b} denotes the dissociation rate. The second reaction represents the overall rate, r_2, at which bound TF-DNA complexes are crosslinked. We assume that the crosslinking reaction is irreversible under our experimental conditions. We obtained the CLK mathematical model by

analytically solving the rate equations derived from this simple scheme. The derivation of the CLK model is presented in Section 3.2.2. Important assumptions are that the concentration of the TF is in excess over the number of binding sites, and that the unbound TFs are not nonspecifically inactivated or crosslinked by formaldehyde. (Figure 3-1; Evidence in support of this is provided in Figure 3-8B and D, and Figure 3-9.) In addition, the method requires that we are able to obtain time-resolved crosslinking data, including on the second time scale and that formaldehyde is not limiting in the reaction. (Figure 3-1; Evidence in support of these assumptions is provided in Figure 3-8A, B and E.)

The CLK model was used to simulate how the ChIP signal varies with formaldehyde crosslinking time (Figure 3-2). The curve shows an initial rapid rise at short crosslinking times (< 5 sec), which corresponds to the formaldehyde fixation of TF-chromatin complexes that were existing at steady-state in the cell population prior to addition of formaldehyde. The steep initial rise is related to the rate constant for the formaldehyde crosslinking reaction. Published work provides support for the suggestion that crosslinking occurs much more rapidly than TF-chromatin dynamics, and our estimates of crosslinking rate obtained with the CLK model are in good agreement with in vitro data (see Section 3.2).

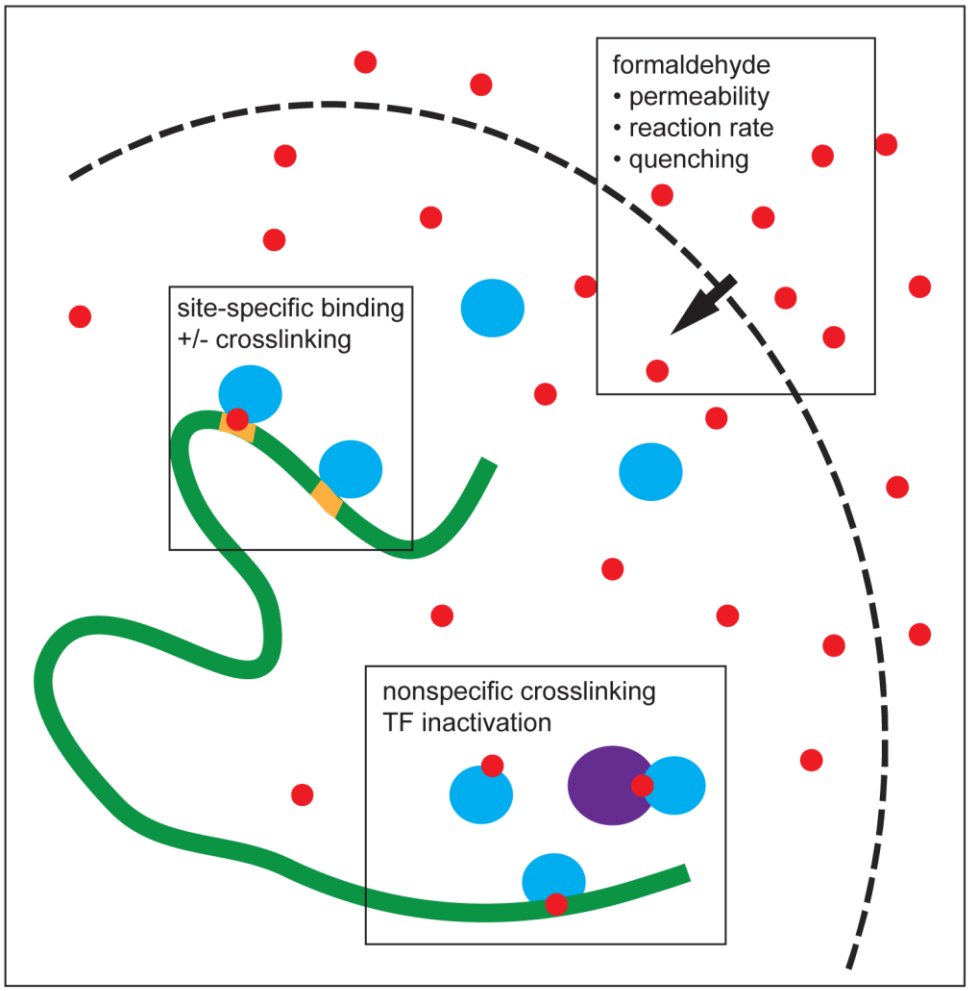


Figure 3-1 : Schematic diagram of a cell

Schematic diagram of a cell with a membrane represented by the dotted arc, a transcription factor (TF) of interest shown as the light blue circle, chromatin shown as the thick green line, and formaldehyde molecules as the small red circles. The small orange chromatin segments represent specific binding sites for the TF. Red circles superimposed on the TF or chromatin represents chemical crosslinking events. The boxes denote the main categories of phenomena occurring in formaldehyde-treated cells that are of importance for understanding how a binding site-specific ChIP signal relates to the time of formaldehyde treatment.

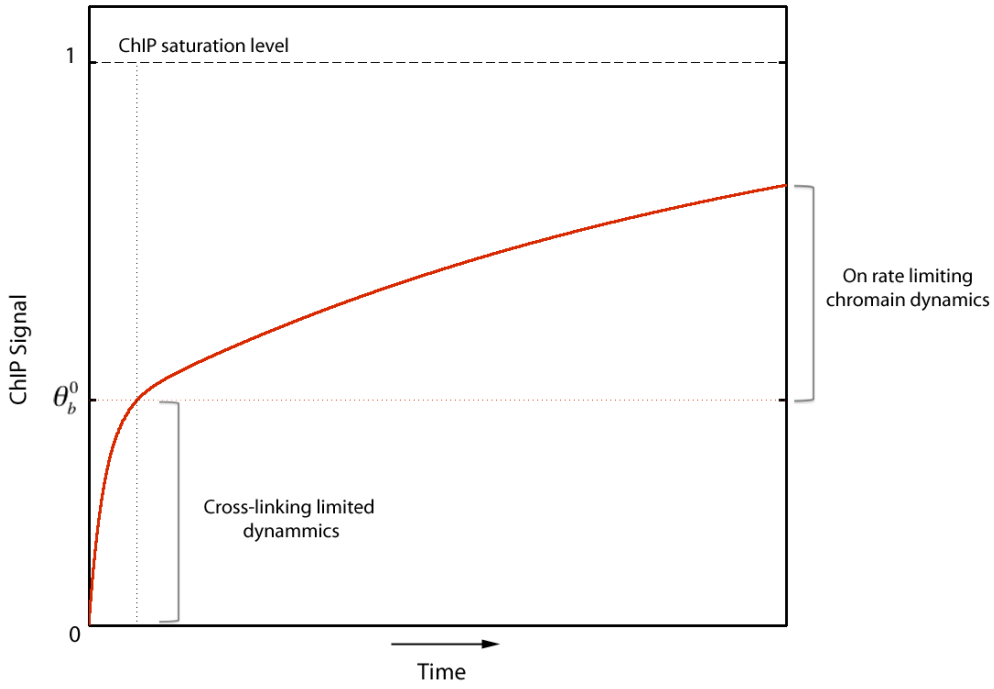


Figure 3-2: Simulation of the CLK model curve

Simulation of the dependence of ChIP signal on formaldehyde incubation time using the CLK model. The early steep rise of the curve shows the formaldehyde crosslinking reaction time dependence of the ChIP signal, which occurs at times much shorter than the formaldehyde incubation time used in traditional ChIP experiments. The later slow rise of the curve shows the increase in the ChIP signal due to the interaction of TF molecules with available sites and their capture by formaldehyde crosslinking. The “knee” at the inflection point indicates the fractional occupancy of the locus for its chromatin binding site at steady-state in the absence of formaldehyde.

The simulation in Figure 3-2 also shows that following the initial rapid increase, the ChIP signal increases more gradually in response to formaldehyde incubation times longer than a few seconds. In the model, this second-phase increase is due to the on-rate driven accumulation of new TF-chromatin interactions, which are fixed by formaldehyde as they form. Eventually, the ChIP signal saturates, reflecting the theoretical state in which all chromatin sites have become occupied and crosslinked. Additional simulations (Figure 3-3) show that the dependence of the ChIP signal on formaldehyde incubation time is expected to be dramatically different for TF-chromatin interactions with different kinetic parameters (e.g. high or low on- or off-rates). For the method to be implemented, it is not necessary that every chromatin binding site is eventually crosslinked to a TF in the sample or that we recover the TF-chromatin complexes with high efficiency. Rather, we assume that regardless of practical limitations of sample handling and recovery that the ChIP signal we measure is proportional to the number of TF-chromatin complexes crosslinked in the population of cells at a particular formaldehyde incubation time.

We make measurements in cells with two different concentrations of the TF. The overall on-rate for chromatin binding contains the TF concentration term, so the rate of increase of the second, slower, phase of the reaction will depend on the TF concentration. Moreover, an increase in the TF concentration will increase by mass action the steady-state fractional occupancy of the chromatin site in the absence of formaldehyde, which is why simulations show that the inflection point or “knee” in the curves moves upward as the TF concentration is increased (see Figure 3-4B). Simultaneous fitting of data sets obtained in cells with two different concentrations of TF thus imposes strong constraints on the mathematical model and reduces the problem of overfitting. In practice, we apply

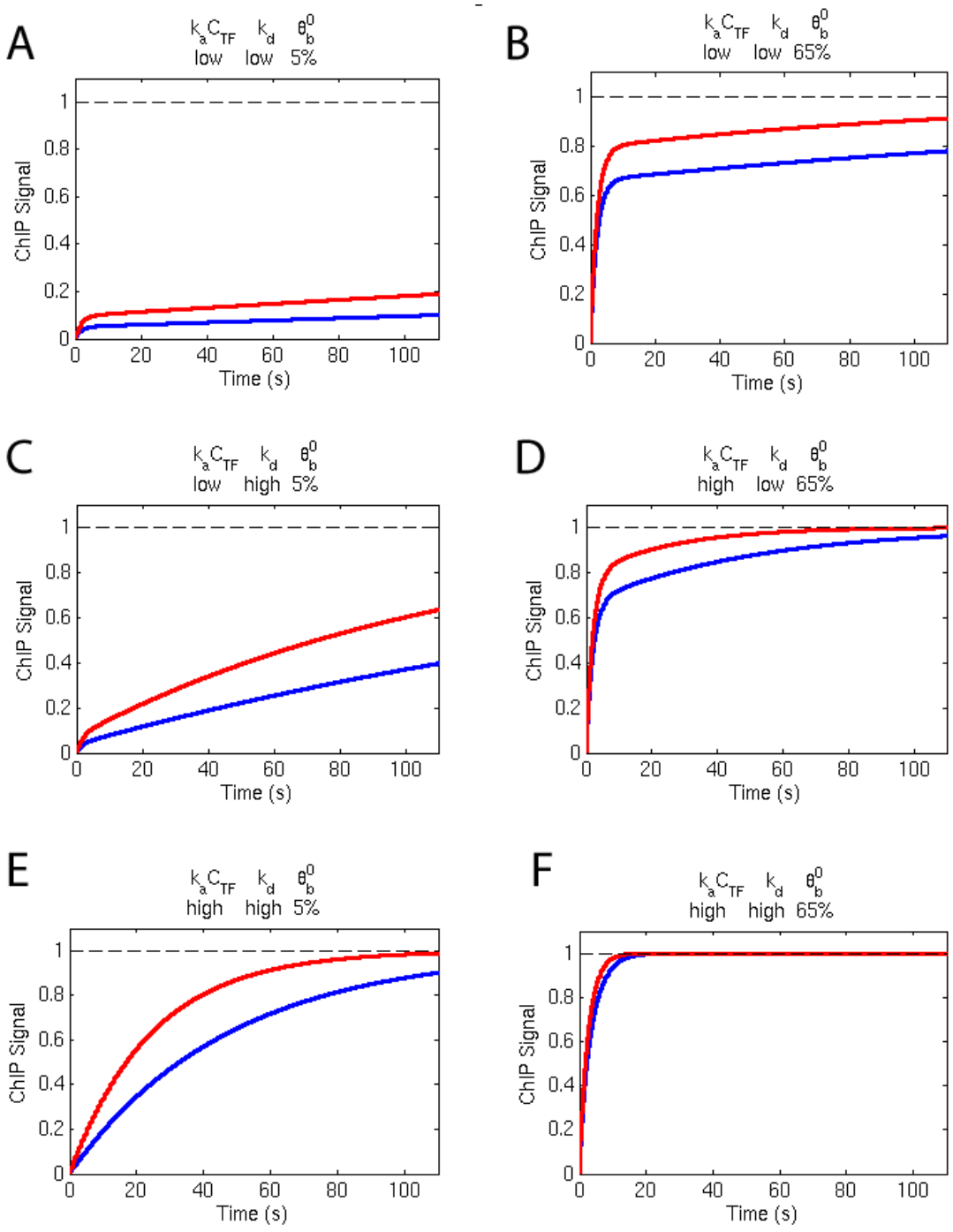

A
$k_a C_{TF}$ k_d θ_b^0
low low 5%
B
$k_a C_{TF}$ k_d θ_b^0
low low 65%
C
$k_a C_{TF}$ k_d θ_b^0
low high 5%
D
$k_a C_{TF}$ k_d θ_b^0
high low 65%
E
$k_a C_{TF}$ k_d θ_b^0
high high 5%
F
$k_a C_{TF}$ k_d θ_b^0
high high 65%
ChIP Signal
Time (s)
0
0.2
0.4
0.6
0.8
1
0
20
40
60
80
100

Figure 3-3: Simulations of the CLK curve for various ranges of kinetic parameters

(A) Simulation of the CLK curve for promoter TF-chromatin interaction with a slow on-rate (k_aC_{TF}), slow off-rate (k_d) and low occupancy (5%; blue lines). The red lines show simulations in which the TF concentration has been increased three-fold compared to the blue line. (B) Simulation of CLK curves as in (A) for a TF-chromatin interaction with slow on-rate, slow off-rate and high occupancy (65%). (C) Simulation of CLK curves for TF-chromatin interaction with slow on-rate, fast off-rate, and low occupancy (5%). (D) Simulation of CLK curves for promoter TF-chromatin interaction with fast on-rate, low off-rate, and high occupancy (65%). (E) Simulation of CLK curves for promoter TF-chromatin interaction with fast on-rate, fast off-rate, and low occupancy (5%). (F) Simulation of CLK curves for promoter TF-chromatin interaction with fast on-rate, fast off-rate, and high occupancy (65%). "Low" on-rate refers to k_aC_{TF} values in the range of 0.4-5 x 10^{-4} s^{-1}, and "high" on-rates varied from 0.2-5 X 10^{-1} s^{-1}. "Low" off-rates ranged from 2-9 x 10^{-3} s^{-1}, and "high" off-rates varied from 0.09 to 1.5 s^{-1}.

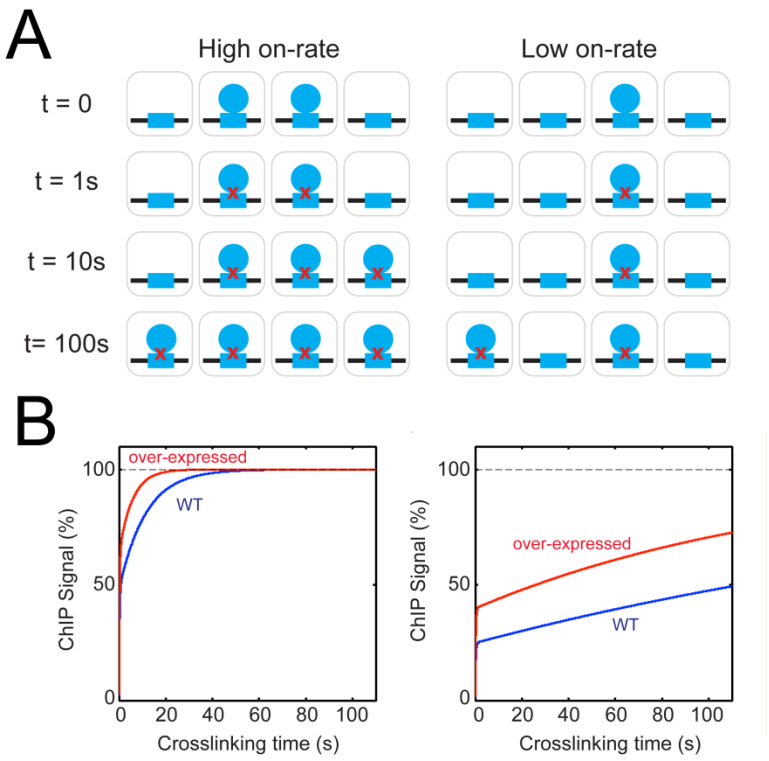


Figure 3-4: Overview of the CLK model.

(A) Schematic showing a chromatin site (blue rectangle) interacting with a transcription factor (blue circle) in a population of four cells in which chromatin binding has a relatively high on-rate (left) or low on-rate (right), but in both cases the off-rate is the same. Rows descending from t=0 show how the site occupancy in the cell population is predicted to change following addition of formaldehyde for 1, 10 or 100 seconds. Red X's indicate crosslinking. (B) Simulations of the two scenarios in (A) using the CLK model (blue lines). The red lines show simulations in which the TF concentration was increased three-fold.

nonlinear regression to fit the CLK model to the experimental CLK dataset to obtain k_a, k_d, k_{xl} and θ_b^0 as parameters (see Section 3.2.4).

3.2.2 – Derivation of the CLK Model

In this model, the transcription factor (TF) can be in one of three states over the course of the crosslinking reaction: unbound, bound to DNA (but not crosslinked) and crosslinked to DNA. In a given ChIP assay only a fraction of chromatin fragments give rise to the ChIP signal. We start by defining n_s as the total number of available binding sites for a given TF at a specific site or array of sites in a population of cells. Denoting $n_b(t)$ as the number of sites bound by the TF as a function of crosslinking time, t, $n_u(t)$ the number of unbound sites as a function of t, and $n_{xl}(t)$ the number of sites with the TF crosslinked to DNA as a function of t, we have $n_s = n_b(t) + n_u(t) + n_{xl}(t)$. Dividing by the total number of binding sites gives

$$\theta_b(t) + \theta_u(t) + \theta_{xl}(t) = 1 \tag{1}$$

where $\theta_b(t) = n_b(t)/n_s$ is the fraction of bound sites; $\theta_u(t) = n_u(t)/n_s$ is the fraction of unbound sites; and $\theta_{xl}(t) = n_{xl}(t)/n_s$ is the fraction of sites with the TF crosslinked to DNA and linearly related to ChIP signals, $Ip(t)$, as $\theta_{xl}(t) = Ip(t)/Ip(\infty)$.

Based on the kinetic model shown above, the rate of change of the fraction of sites bound by the TF is given by

$$\frac{d\theta_b(t)}{dt} = r_{1f} - r_{1b} - r_2 \quad (2)$$

and the rate of change of sites crosslinked to the TF is\

$$\frac{d\theta_{xl}(t)}{dt} = r_2 \quad (3)$$

Assuming first order kinetics, the overall association- or on-rate of TF binding is $r_{1f} = k_a C_{TF} \theta_u$ where C_{TF}, θ_u and k_a are the concentration of the TF in the nucleus, the fraction of unbound sites, and the molecular on-rate, respectively. The overall dissociation- or off-rate is $r_{1b} = k_d \theta_b$ where θ_b and k_d are the fraction of sites bound by the TF, and the molecular off-rate, respectively. In the absence of cross-linking ($r_2 = 0$), equation (2) describes the dynamics of a TF binding to its DNA site in vivo. We assume that the crosslinking reaction is first order with respect to the formaldehyde concentration (C_{FH}) and θ_b, giving $r_2 = k_{xl} C_{FH} \theta_b$ for the overall rate where k_{xl} is the molecular crosslinking rate.

Substituting the overall rates into Eq. (2) and (3) yields

$$\frac{d\theta_b(t)}{dt} = k_a C_{TF} \theta_u(t) - k_d \theta_b(t) - k_{xl} C_{FH} \theta_b(t) \quad (4)$$

$$\frac{d\theta_{xl}(t)}{dt} = k_{xl} C_{FH} \theta_b(t) \quad (5)$$

Boundary Conditions

The boundary conditions can be derived from Eq. (1)-(3) assuming steady-state conditions. Before crosslinking (at *t*=0),

$$\theta_{xl}(0) = 0 \quad (6)$$

by definition. Using this, we can solve for the equilibrium fraction of sites bound by the TF by first setting $r_2 = 0$ and solving Eq. (2) with $d\theta_b/dt = 0$ (i.e., steady-state before addition of crosslinker). This results in the equilibrium fraction of bound sites, θ_b^0, at *t* = 0

$$\theta_b^0 = \frac{k_a C_{TF}}{k_a C_{TF} + k_d} \tag{7}$$

After crosslinker is added and $t \to \infty$, $d\theta_b/dt \to 0$ and $d\theta_{xl}/dt \to 0$ (i.e., steady-state is reached after addition of crosslinker). Use of Eqs. (1)-(3) under steady-state yields

$$\lim_{t\to\infty} \theta_{xl} \to 1 \tag{8}$$

Eqs. (6)-(8) constitute the boundary conditions, which we will use together with Eqs. (1)-(3) to solve for the fraction of sites crosslinked to the TF as a function of time, $\theta_{xl}(t)$.

Solution of the Differential Equations Subject to Boundary Conditions

Differentiating Eq. (4) with respect to t, substituting for θ_u using Eqs. (1), and using (5), we find

$$\frac{d^2\theta_b(t)}{dt^2} = -(k_a C_{TF} + k_d + k_{xl} C_{FH}) \frac{d\theta_b(t)}{dt} + (-k_{xl} C_{FH} k_a C_{TF}) \theta_b(t) \tag{9}$$

which has a general solution of the form

$$\theta_b(t) = Ae^{-t/\tau_+} + Be^{-t/\tau_-} \tag{10}$$

with the inverse of the two time constants or relaxation times—times over which the two dynamic processes shown in Eq. (10) take to reach steady state—$\frac{1}{\tau_+}, \frac{1}{\tau_-}$ given by

$$\frac{1}{\tau_{\pm}} = \frac{(k_a C_{TF} + k_d + k_{xl} C_{FH})}{2} \left[1 \pm \sqrt{1 - \frac{4 k_a C_{TF}}{k_{xl} C_{FH}} \times \frac{1}{\left[1 + \left(\frac{k_a C_{TF} + k_d}{k_{xl} C_{FH}}\right)\right]^2}} \right] \tag{11}$$

Applying the boundary condition at t = 0 shown in Eq. (7), we have

$$\theta_b(0) = \theta_b^0 = \frac{k_a C_{TF}}{k_a C_{TF} + k_d} = A + B \tag{12}$$

Integrating Eq. (5) with respect to *t* and then substituting the general solution of $\theta_b(t)$, Eq. (10), gives

$$\theta_{xl}(t) - \theta_{xl}(0) = k_{xl} C_{FH} \int_0^t \theta_b(t') dt' \tag{13}$$

$$\theta_{xl}(t) = k_{xl} C_{FH} \left[A\tau_+ (1 - e^{-t/\tau_+}) + B\tau_- (1 - e^{-t/\tau_-}) \right] \tag{14}$$

where we have used Eq. (6).

Next, we apply the boundary condition for θ_{xl} as $t \to \infty$, Eq. (8), to Eq. (14), which gives

$$1 = k_{xl} C_{FH} [A\tau_+ + B\tau_-] \tag{15}$$

We use Eqs. (12) and (15) to solve for A and B which when substituted into Eq. (14) yields the fraction of binding sites with crosslinked TF in a population of cells as a function of crosslinking time,

$$\theta_{xl}(t) = 1 - \frac{\tau_+ e^{-t/\tau_+} - \tau_- e^{-t/\tau_-}}{\tau_+ - \tau_-} + \frac{\theta_b^0 \tau_+ \tau_- k_{xl} C_{FH}}{\tau_+ - \tau_-} \left(e^{-t/\tau_+} - e^{-t/\tau_-} \right). \tag{16}$$

3.2.3 – Approximate Forms of the CLK Model

Eqs. (11), (16) describe the relationship of the fractional ChIP signal to chromatin binding dynamics and formaldehyde crosslinking rate. These equations were challenging to understand and implement because of their complexity and the number of parameters involved. We derived simpler approximations to obtain insight into the interpretation of CLK data, and in addition, the approximate models allowed us to obtain accurate initial parameter estimates for subsequent fitting. The experimental results show in general a steep dependence of ChIP signal on time for relatively short incubation times, followed by a more gradual increase in ChIP signal with longer formaldehyde incubation times. This suggested that two processes were occurring that are well separated in time. Thus, we assumed that the two time constants shown in Eq. (11) are well separated (i.e., different orders of magnitude), which led to two simplified approximate models: (1) crosslinking dynamics is much faster than TF-DNA binding dynamics or (2) TF-DNA binding dynamics is much faster than crosslinking dynamics. The detailed derivation of these two approximate models is shown below.

TF-DNA Binding Dynamics-Limited Model

We arrive at the first approximate model by assuming that the crosslinking rate is much greater than transcription factor binding dynamics (i.e., $(k_a C_{TF} + k_d)/k_{xl} C_{FH} \ll 1$). Applying these assumptions, we Taylor expand Eq. (11) in $k_a C_{TF}/k_{xl} C_{FH}$ and $k_d/k_{xl} C_{FH}$ and retain the first order terms

$$\frac{1}{\tau_+} \approx k_{xl} C_{FH} \text{ and } \frac{1}{\tau_-} \approx k_a C_{TF}. \quad (17)$$

We then find the approximate forms for Eq. (16) for relatively short crosslinking times (i.e., $t << \tau_-$) and long crosslinking times (i.e., $t >> \tau_+$). Use of Eq. (17) and $t << \tau_-$, we Taylor expand Eq. (16) in $k_a C_{TF}/k_{xl} C_{FH}$, $k_d/k_{xl} C_{FH}$ and retaining the lowest order terms find

$$\theta_{xl}(t) \approx \theta_b^0 (1 - e^{-k_{xl} C_{FH} t}) \tag{18}$$

which is the approximate form of $\theta_{xl}(t)$ for short crosslinking times (i.e., crosslinking times shorter than or comparable to $\tau_+ \approx 1/k_{xl} C_{FH}$). Use of Eq. (17) and $t \gg \tau_+$, we Taylor expand Eq. (16) in $k_a C_{TF}/k_{xl} C_{FH}$, $k_d/k_{xl} C_{FH}$, and, retaining the lowest order terms, we find the approximate form for $\theta_{xl}(t)$ for relatively long crosslinking times (i.e., crosslinking times much longer than $\tau_+ \approx 1/k_{xl} C_{FH}$),

$$\theta_{xl}(t) \approx 1 - (1 - \theta_b^0) e^{-k_a C_{TF} t}. \tag{19}$$

Equations (18) and (19) have a simple, intuitive interpretation. TFs which are bound to DNA are first rapidly crosslinked at the crosslinking rate as described by Eq. (18)., This continues until the fraction of sites containing crosslinked TF equals the in vivo occupancy (i.e., $\theta_{xl}(t) \approx \theta_b^0$ for $\tau_+ \ll t \ll \tau_-$). Sites are then crosslinked to TFs at the in vivo overall on-rate, $k_a C_{TF}$ of the TF as shown in Eq. (19). This continues until all the sites are crosslinked at crosslinking times much longer than the time-scale associated with the in vivo on-rate (i.e., $\theta_{xl}(t) \approx 1$ for $t \gg \tau_-$).

Crosslinking Dynamics-Limited Model

For the second approximate model, we assume that the crosslinking rate is much slower than transcription factor binding dynamics $(k_a C_{TF} + k_d)/k_{xl} C_{FH} \gg 1$. Taylor expanding Eq. (11) in $k_{xl} C_{FH}/k_a C_{TF}$ and $k_{xl} C_{FH}/k_d$ and retaining the lowest order terms, we find

$$\frac{1}{\tau_+} \approx \frac{1}{\tau_{TF}} = k_a C_{TF} + k_d \quad \text{and} \quad \frac{1}{\tau_-} \approx k_{xl} C_{FH} \theta_b^0 \quad . \tag{20}$$

We then find the approximate forms for Eq. (16) for relatively short crosslinking times (i.e., $t << \tau_-$) and long crosslinking times (i.e., $t >> \tau_+$). Here, we start with deriving the approximate form for long crosslinking times. Substituting Eq. (20) into Eq. (16), assuming $t \gg \tau_+$, expanding in $k_{xl} C_{FH}/k_a C_{TF}$ and $k_{xl} C_{FH}/k_d$ and neglecting higher orders terms we find

$$\theta_{xl}(t) \approx 1 - e^{-k_{xl} C_{FH} \theta_b^0 t} \quad , \tag{21}$$

which is the approximate form for the fraction of sites containing crosslinked TF as a function of time for long crosslinking times (i.e., crosslinking time much longer than the in vivo TF binding relaxation time $\tau_+ \approx \tau_{TF}$). Substituting Eq. (20) into Eq. (16), assuming $t \ll \tau_-$, Taylor expanding in $k_{xl} C_{FH}/k_a C_{TF}$ and $k_{xl} C_{FH}/k_d$ and retaining the lowest order terms yields

$$\theta_{xl}(t) \approx \theta_b^0 k_{xl} C_{FH} t \quad . \tag{22}$$

which is the approximate form for the fraction of sites containing crosslinked TF as a function of time for short crosslinking times (i.e., crosslinking times much shorter than $\tau_- \approx 1/k_{xl} C_{FH} \theta_b^0$). Given that Eq. (22) is simply the first term in a Taylor expansion of Eq. (21) for short crosslinking times (i.e., $k_{xl} C_{FH} \theta_b^0 t << 1$), Eq. (21) represents a good approximation of $\theta_{xl}(t)$ for all crosslinking times assuming the crosslinking rate is much slower than transcription factor binding dynamics.

3.3 Materials and Methods

3.3.1 Experimental Methods

All wet-bench procedures were performed by Melissa Wells Carver , Ramya Viswanathan and Shana Cirimotich. Live cell imaging was done by Tatiana Karpove and Jim McNally. My role in this project was to develop and implement the CLK mathematical model.

3.3.1.1 – Yeast strains and growth conditions

S. cerevisiae strains and plasmids used in this study are listed in Tables 3-1 and 3-2. YPH499 (Sikorski and Hieter 1989) cells used for Gal4 ChIP were grown in YEP plus 2% raffinose to an OD_{600} = 0.8. Cells were then pelleted and resuspended in YEP plus 2% galactose for 1 hour prior to addition of formaldehyde. In order to make measurements with cells overexpressing Gal4, YRV004 cells were used. The strain carries a 2μ plasmid, pSJ4, harboring the *GAL4* open reading frame under control of its own promoter (Johnston and Hopper 1982). For Gal4 ChIP in the Gal4 overexpression strain YRV004 cells were grown overnight at 30° C in SC-URA plus 2% raffinose, pelleted and resuspended in YEP plus 2% galactose for 1 hour prior to treatment with formaldehyde. The level of Gal4 protein in cells is extraordinarily low (Table 3-5), so *GAL4-TAP* (YRV005) cells were used in western blotting experiments to quantify Gal4 levels. The level of Tfa1 protein in cell extracts has been reported (Borggrefe et al. 2001), so *TFA1-TAP* (YRV006) cells were used to obtain a quantitation standard. YRV005 and YRV006 cells were obtained from the Yeast TAP-fusion library (Open

Biosystems, provided by Dan Burke and Frank Pugh). YRV005 was grown in the same way as cells for Gal4 ChIP experiments. YRV006 was grown overnight at 30° C in YPD to OD_{600} ~1.0 and then harvested. To quantify the level of Gal4 in the Gal4 overexpression strain, extracts from YRV012 and YRV014 harboring pRV021 were compared with extracts from YRV005. Plasmid pRV021 was constructed by fusing the TAP coding sequence to the 3' end of the *GAL4* open reading frame in plasmid pSJ4 using the Infusion kit (Clontech). The TAP sequence was obtained by PCR amplification using YRV005 genomic DNA.

YTK539 cells used for Ace1 ChIP were grown overnight in CSM-HIS (MP biomedicals) to an OD_{600} = 0.8. Cultures were then induced with 1 ml of 10 mM $CuSO_4$ for 90 minutes and processed immediately for ChIP. The Ace1 overexpression plasmid (pMW101) was constructed by restriction enzyme digestion of pTSK241 with Not1 and Sac II and cloning the triple GFP tag into pTSK65 to replace the single GFP tag on Ace1 with a triple GFP tag. Strain YSC002 was grown at 30° C overnight in CSM-HIS to an OD_{600} = 0.8. Cultures were induced in the same way as for YTK539 cells.

TBP ChIP was performed using YRV018 cells, which were grown in YPD at 30° C to an OD_{600} = 1.0 prior to addition of formaldehyde. A 2μ plasmid carrying the TBP open reading frame under the control of its own promoter (pSH223, a gift from Steve Hahn) was transformed into the TBP shuffling strain (YAD165). The *URA1*-marked *SPT15* plasmid covering the TBP deletion was shuffled out using FOA selection to generate the TBP overexpression strain YSC003. YSC003 was grown in YPD overnight at 30° C to an OD_{600} = 1 prior to the addition of formaldehyde. YAD154 cells, used for quantitation of the soluble pool of TBP, were grown in YPD to an OD_{600} = 1 prior to

harvesting. AY87 cells (Darst et al. 2003), used for TBP ChIP in the *mot1-42* background, were transformed with either pRS426 (control) or pSH224 (2□ TBP overexpression plasmid; a gift of Steve Hahn) and cells were grown in SC-URA medium. For ChIP, YSC004 and YSC005 cells were grown in SC-URA medium and then diluted into YPD for approximately two population doublings before crosslinking. Note that in order to directly compare TBP ChIP signals in WT and *mot1-42* cells, and at each of two TBP expression levels, YSC004 and YSC005 were grown at 30° C in YPD prior to crosslinking and no heat shock was done.

Strain YTK260, used for LacI-GFP ChIP, was grown in SC-HIS medium overnight at 30° C to an OD_{600} ~1.0 prior to addition of formaldehyde. The LacI-GFP overexpression plasmid (pSC001) was constructed by restriction enzyme digestion of pTSK437 with Kpn1 and Not1 and subcloning of the LacI-GFP cassette into pRS426 (Sikorski and Hieter, 1989). Strain YSC001 harboring pSC001 was grown at 30° C overnight in SC-HIS-URA medium to an OD_{600} ~1.0 prior to addition of formaldehyde.

3.3.1.2 – Chromatin Immunoprecipitation (ChIP)

ChIP was performed as described in Dasgupta et al. (2005), but with varying crosslinking times. Unless otherwise indicated, formaldehyde was added to a final concentration of 1% (360 mM) for various times and quenched by adding glycine to 250 mM (final concentration). The shortest crosslinking times (1.37 s and shorter) were achieved using a quench flow apparatus, which is described below. For longer but still relatively short crosslinking times (5 s to 60 s), formaldehyde and glycine were added to

cell cultures while rapidly mixed using a stir bar. After incubation with glycine for 5 minutes, cells were washed with cold TBS (40 mM Tris-HCl pH 7.5, 300 mM NaCl) with 125 mM glycine and once with cold TBS. Cell pellets were then resuspended in ChIP lysis buffer (50 mM HEPES pH 7.5, 1% Triton-X 100, 0.1% sodium deoxycholate) with 140 mM NaCl and protease inhibitors (Roche Complete Protease Inhibitor Cocktail Tablet) and were lysed using acid-washed glass beads (Sigma) in a FastPrep machine (MP Biomedicals). The whole cell extracts were then sonicated and subsequently quantitated using Bradford Reagent. Immunoprecipitation (IP) was performed overnight at 4°C using 1 mg chromatin protein. For Gal4 ChIP, Gal4-TA C-10 antibody (sc-1663x; Santa Cruz Biotechnology) was used. For Ace1-GFP and LacI-GFP IPs, anti-GFP antibody was used (Invitrogen). TBP immunoprecipitations were performed using anti-TBP antibodies (Sigma, clone 58C9). Following antibody incubation with sonicated chromatin, 40 µL Protein A sepharose beads (Amersham) were added and samples were mixed by rotation for 2 hours at 4° C. Mock IPs were performed by combining 1 mg total chromatin protein with the protein A sepharose beads, without addition of antibody. The beads were then washed twice with 1 ml of each of the following buffers: ChIP lysis buffer (140 mM NaCl), ChIP lysis buffer (500 mM NaCl), LiCl buffer (10 mM Tris pH 8.0, 250 mM LiCl, 0.5% NP-40, 0.5% sodium deoxycholate, 1 mM EDTA), and TE (10 mM Tris-Cl pH 8.0, 1mM EDTA). Protein-DNA complexes were then eluted with 50 mM Tris pH 8.0, 1% SDS, 10 mM EDTA twice for 10 minutes at 65°C, and formaldehyde crosslinks were reversed by incubation overnight at 65°C. DNA was purified using a QIAquick PCR purification kit (QIAGEN) according to the manufacturer's instructions. ChIP DNA was then quantified by real-time qPCR.

3.3.1.3 – KinTek ChIP

A KinTek quench flow instrument (model RQF-3, KinTek corporation) was used for formaldehyde crosslinking reactions too short in duration to be performed by simple hand mixing. The KinTek apparatus is encased in a waterbath whose temperature was set to 30° C. One syringe was filled with 5 ml of yeast cell culture, while the other syringe was filled with 5 ml of 2% formaldehyde. The quench syringe was not used. Instead, different times were obtained by adjusting the length of the exit tube, whose end was placed in 10 ml of the quenching solution (500 mM glycine). The stepping motor speed was set to 200 and there were 60,000 steps per cycle. The mixing time and effectiveness of the quenching arrangement were calibrated according to the manufacturer's instructions using the standard reaction of hydrolysis of benzylidenemalononitrile by NaOH at 20° C. The calibrated mixing times and errors are shown in Table 3-3. Following quenching, the formaldehyde-treated cells were pelleted and washed as described above.

3.3.1.4 – ChIP quantitation

ChIP, mock IP, and total samples were quantitated by real time PCR using iQ SYBR Green Supermix (BioRad) and the BioRad MyiQ Single Color Real Time PCR detection system. Relative ChIP signals were obtained by subtracting the mock IP signal from the ChIP signal and normalizing against the input. Two to three independent biological replicates were averaged for each time point. Oligonucleotides used for PCR are listed in Table 3-4. Oligonucleotides used for the *lacO* array ChIP anneal to a unique region located just outside the array to avoid amplifying the repetitive sequence.

3.3.1.5 – Nuclear protein concentration of factors and the amount overexpressed

The numbers of Ace1-GFP and TBP molecules per cell have been reported previously (See Table 3-5). The nuclear concentration of these factors was estimated based on the nuclear volume reported in Jorgensen et al. (2007). As Gal4 levels are very low, the nuclear concentration of Gal4 for cells grown in galactose-containing media was determined by Western blot analysis of TAP-tagged Gal4 by normalizing the signal with TAP-tagged TFIIE, whose concentration is known(Borggrefe et al. 2001) .

The amount of each factor overexpressed in cells, except for LacI-GFP, was quantified by Western blot analysis. Strains were grown as described above, pelleted and washed with cold TBS. For Western blot analysis of Gal4 and Ace1-GFP, the cells were then resuspended in Benoit's buffer (200 mM Tris-Cl pH 8, 400 mM $(NH_4)_2SO_4$, 10 mM $MgCl_2$, 1 mM EDTA, 10% glycerol) plus protease inhibitors and lysed using acid-washed glass beads as was done for ChIP. After incubating on ice for 30 minutes, the extracts were then clarified by centrifugation at 14000 rpm for 30 minutes in a microcentrifuge. The protein amounts in the supernatant were quantified as for ChIP using Bradford Reagent using bovine serum albumin as the standard. Extracts normalized for total protein were boiled with sample buffer and loaded onto SDS polyacrylamide gels and Western blotted using antibody against the various factors as for ChIP. α-protein A was used for TAP-tagged protein and anti-GFP antibody (Invitrogen) was used for Ace1-GFP. To quantify TBP levels YPH499, YRV018 and YSC003 cells were lysed as described (Borggrefe et al, 2001) and extracts were Western blotted using anti-TBP antibody (Sigma, clone 58C9) and purified recombinant yeast TBP as a standard. Quantification

was done using Image J software (NIH). The overexpression level of TBP was the same in WT and *mot1-42* cells.

To measure the LacI-GFP level in cells and extent of over-expression, fluorescence microscopy was used. The spindle pole body (SPB) and the *lacO* array under conditions of saturated LacI-GFP expression were imaged for calibrating the relationship between average intensity and number of molecules per pixel. Based on the calibration curve, the molecules per pixel for each structure were determined. Then, this number was multiplied by the measured area of the structure to obtain the estimated number of total molecules (Table 3-6).

3.3.1.6 – Quantitation of soluble protein pools with/without formaldehyde treatment

YRV005 and YAD154 cells were used to quantify the soluble Gal4 and TBP prior to and after formaldehyde treatment. At an OD of 1, 250 mM glycine was added to one-fourth of the volume of cells and they were harvested (0 minute sample). To the remaining culture, 1% formaldehyde was added and samples were quenched by adding 250 mM glycine after 5, 10 or 15 min incubation with formaldehyde. The soluble protein fraction was separated from the chromatin-bound fraction for each sample by each of two different methods. In one method, the cells were lysed in Benoit's buffer and the extracts were treated the same way as described for quantification in Section 3.3.1.5 above. In the second method, cells were spheroplasted using the procedure described (Muldrow et al., 1999). After the spheroplasts were allowed to recover in YPD-S media (10 g of yeast extract, 20 g of peptone, 20 g of glucose, and 182.2 g of sorbitol per liter) by shaking

gently at 30° C, they were pelleted at 4000 rpm for 9 minutes in a clinical centrifuge and washed thrice with lysis buffer (0.4 M Sorbitol, 150 mM potassium acetate,, 2 mM magnesium acetate, 20 mM Pipes/KOH, pH 6.8, 1 mM phenylmethylsulfonyl fluoride, 10 µg/mL leupeptin, 1 µg/mL pepstatin A, 10 mM benzamidine) (Donovan et al. 1997). Cells were then resuspended in ~200 µL of lysis buffer to which was added Triton-X100 to a final concentration of 1%. The supernatant and chromatin-enriched fractions were separated by centrifuging the extracts for 15 minutes at 14,000 rpm in a microcentrifuge

Table 3-1 *S. cerevisiae* strains used in this study.

Strain	Genotype	Reference or source
YPH499	*MATa ura3-52 lys2-801a ade2-101o trp1-Δ63 his3-Δ200 leu2-Δ1*	Sikorski and Hieter, 1989 (*19*)
YRV004	*MATa * pSJ4 [GAL4 URA3 2μ]*	This study
YRV005	*MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0 GAL4-TAP*	Ghaemmaghami *et al. 2003* (*35*)
YRV006	*MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0 TFA1-TAP*	Ghaemmaghami *et al. 2003*
YRV012	*MATa * pRV021[GAL4-TAP URA3 2μ]*	This study
YRV014	*MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0 GAL4-TAP pRV021[GAL4-TAP URA3 2μ]*	This study
YTK539	*MAT*a *his3-Δ1 leu2Δ0 met15Δ0 ura3Δ0 ace1Δ* :: *KAN TRP1:: pCap2-ACE1-tripleGFP-HIS3*	Karpova et al, 2008 (*11*)
YTK934	*MATα his3Δ1 leu2Δ0 lys2Δ0 ura3Δ0 ACE1-tripleGFP-URA3*	This study
YRV018	*MAT*a *ade2–1 his3–1115 leu2–112 trp1–1 ura3–1 can1–100 abf1*::*HIS3MX6 pRS415-ABF1-FLAG*	Miyake et al, 2004 (*36*)
AY87	*MATa * mot1Δ::TRP pMOT221 [LEU2 CEN ARS]*	Darst et al, 2003 (*22*)
YSC002	*MATα his3Δ1 leu2Δ0 lys2Δ0 ura3Δ0 ACE1-tripleGFP-URA3 pMW101 [ACE1-triple GFP HIS3 2μ]*	This study
YSC004	*MATa * mot1Δ::TRP pMOT221[LEU2 CEN ARS] pRS426*	This study
YSC005	*MATa * mot1Δ::TRP pMOT221[LEU2 CEN ARS] pSH224 [TBP URA3 2μ]*	This study
YAD154	*MATa * SPT15-myc*	This study
YTK260	*MAT a/α, HIS5/his3Δ1, leu2Δ0/leu2Δ0, ura3Δ0/ura3Δ0, met15Δ0 LYS2::pHIS3-lacI-GFP-NLS-NAT1, CU3::KAN-(LacO)256 , CU1::(LacO)256*	This study
YSC001	*MAT a/α, HIS5/his3Δ1, leu2Δ0/leu2Δ0, ura3Δ0/ura3Δ0, met15Δ0 LYS2::pHIS3-lacI-GFP-NLS-NAT1, CU3::KAN-(LacO)256 , CU1::(LacO)256 pSC001 [pHIS3-GFP-LacI URA3 2μ]*	This study
YSC003	*MATa * spt15::natMX pSH223 [TBP LEU 2μ]*	This study

** ura3-52 lys2-801a ade2-101o trp1-Δ63 his3-Δ200 leu2-Δ1*

Table 3-2 Plasmids used in this study.

Plasmid name	Information	Reference or source
pSJ4	*GAL4 URA3 2µ*	Johnston and Hopper, 1982 (*20*)
pSH223	*TBP LEU 2µ*	Steve Hahn
pSH224	*TBP URA3 2µ*	Steve Hahn
pRV021	*GAL4-TAP URA3 2µ*	This study
pSC001	*pHIS3-GFP-LacI URA3 2µ*	This study
pMW101	*ACE1-triple GFP HIS3 2µ*	This study
pTSK65	*ACE1-GFP HIS3 2µ*	Karpova et al., 2004 (*37*)
pTSK241	*pCap2-ACE1-tripleGFP HIS3*	Karpova et al., 2008 (*11*)
pTSK437	*pHIS3-GFP-LacI HIS3*	This study

Table 3-3 KinTek calibrated times with errors.

Mixing time point	Standard deviation
142 ms	± 52 ms
264 ms	± 16 ms
441 ms	± 30 ms
814 ms	± 170 ms
1.37 s	± .12 s

Table 3-4 Oligonucleotides Used for Real-Time PCR (5'-3').

Name	Sequence
CUP1-F	AGA AGC AAA AAG AGC GAT GC
CUP1-R	GAC AAT CCA TAT TGC GTT GG
LOS1-F	TTT GAG AAG TTG TCG GTA AGC A
LOS1-R	GCA TTC CTC GAT TTG ACT GG
ACT1-F	CAG CTT TTA GAT TTT TCA CGC TTA
ACT1-R	TTT TCG ATC TTG GGA AGA AAA A
HSC82-F	TCT TGA AAC GCT ACA GAA CCA A
HSC82-R	CAC CAG CCA TAT TTC AGA ATG A
URA1-F	AAG ATG CCC ATC ACC AAA AA
URA1-R	AAG AAT ACC GGT TCC CGA TG
NTS2-F	GCA CCT GTC ACT TTG GAA AAA
NTS2-R	TCG CCG AGA AAA ACT TCA AT
U6-F	TTC GTC CAC TAT TTT CGG CTA
U6-R	GGG TTA CTT CGC GAA CAC AT
INO1-F	GTT GGC GGC AAT GTT AAT TT
INO1-R	CGA CAA CAG AAC AAG CCA AA
GAL3 UAS-F	CCG AAC ATG CTC CTT CAC TA
GAL3 UAS-R	GCA TGG CGA TTT CAT TCT TT
GAL3 ORF-F	GCC AAA ACT AAA GGC CAC AC
GAL3 ORF-R	GGC GAT GAC GAA ACT GAT TT
CU3-F	TCT CGG CCT AGC TCA TCA GT
CU3-R	AAG ACA GAT CCA CGT CTT TGG

Table 3-5 Estimate of nuclear protein concentrations, based on nuclear volume from Jorgensen et al, 2007.

Factor	Concentration in the nucleus* (µM)	*Reference	Overexpression concentration (µM) (This study)
Ace1-GFP	1	Ghaemmaghami S et al, 2003 (*35*), Karpova et al, 2004 (*37*)	10
TBP	12	Borggrefe et al, 2001 (*21*)	38
Gal4	0.18	This study	0.45
LacI-GFP	1	This study	3.6

Table 3-6 Measurement of the total number of LacI-GFP molecules per cell using fluorescence microscopy.

	Total molecules	Molecules/pixel	Intensity/pixel
Overexpression strain	6150	3.99	227
Basal expression strain	1691	0.91	25

(Donovan et al., 1997). Bradford assays were used to quantitate the total protein levels in the soluble pools, and 100 µg of total protein were mixed with equal volumes of 2X sample buffer, boiled, and loaded onto denaturing gels and Western blotted using α-protein A (to detect Gal4-TAP) or α-myc (to detect TBP-myc) (Dasgupta et al., 2005). The blots were also probed for G6PDH using α-G6PDH antibody, which served as a control. Quantification was done using Image J software (NIH) and both methods yielded similar results. The reported soluble protein levels were obtained by averaging the data from three independent sets of biological replicates.

3.3.1.7 – Experiments to test the efficiency of formaldehyde crosslinking reaction

This procedure refers to the results presented in Figure 3-8B. YPH499 cells were grown in YEP plus 2% raffinose as described above. At an OD_{600} = 0.8, cells were split into three samples. To one sample, formaldehyde was added for 1 minute and then quenched with glycine. The cells were pelleted, washed and processed as described above for ChIP. To the second sample, formaldehyde was added for 1 minute and the reaction was quenched as for the first sample. Then these formaldehyde-treated cells were resuspended in YEP plus 2% galactose and incubated for 20 minutes at 30 degrees with shaking. Then formaldehyde was added again for 5 minutes and the reaction was then quenched with glycine. The cells were pelleted, washed, and processed for ChIP as described above. To the third sample, *GAL* gene expression was induced by resuspending the cells in YEP plus 2% galactose, and then cells were crosslinked by incubating with 1% formaldehyde for 5 minutes. The reaction was then quenched and processed for ChIP as described above. Note that in a separate experiment we confirmed that Gal4 bound to the *GAL3* promoter within 20 minutes post induction with galactose (data not shown).

3.3.1.8 – Imaging of live cells

Live yeast cells were imaged in LabTek II coverglass chambers (Nalge Nunc Intl., Rochester, NY). Before an experiment, 500 ml of the mid-log phase yeast culture was concentrated by centrifugation, and then 5 ml of the concentrated suspension was placed into a Lab Tek II chamber and subsequently covered by a 10 mm x 10 mm agarose slab cut from the solid NF-His/agarose medium.

3.3.1.9 – FRAP

FRAP experiments were carried out on a Zeiss 510 confocal microscope with a 100X/1.3 NA oil immersion objective. To reduce bleaching due to imaging, cells were imaged with a 488 nm laser line from a 30 mW argon laser operating at low laser intensity (0.75%). One of the two *CUP1* loci or lacI/lacO markers in a diploid cell was photobleached using a short (17 msec) laser pulse with the laser operating at 75% of full power. Fluorescent recovery for LacI was monitored at 30 sec time intervals for 240 sec (24 cells). Fluorescent recovery for CUP1 was monitored at 10 sec time intervals for 235 sec (30 cells). 3D image stacks (11 focal planes at 250 nm z step size) were collected, and intensities of both the bleached and unbleached locus were measured, and image background was subtracted from each measurement. To correct for bleaching due to imaging in each cell, intensities from the bleached locus were divided by those from the unbleached locus. The resulting curve was normalized to the prebleach level of array intensity, and these normalized curves were then averaged. The curves were fit with the

reaction-dominant model (Sprague et al. 2004) $FRAP(t) = 1 - Ae^{-k_{off}t}$, with time t, and A and k_{off} free parameters determined by the fit. The half-time, $t_{1/2}$, is equal to $ln2/k_{off}$.

3.3.2 Computational Methods

3.3.2.4 – Non-linear regression analysis using the approximate CLK models

We took an agnostic view regarding which approximate equation (Section 3.2.2) would yield the best fit, and hence, best explanation of the CLK data. This yielded four Cases (Figure 3-5). We fit each case to determine which gave the best initial and final parameters as determined by the full model with the lowest RMSE. This in turn selected the best performing case. While we exhaustively tested each Case (see Figure 3-6), we found that Case 1 (illustrated in Figure 3-2) yielded the best RMSE between the model estimates and the experimental data, and moreover, all of the CLK model-fitted curves shown in this study are approximated by Case 1.

We arrived at the initial estimates of the parameters by fitting approximate generalized linear equations shown in Eqs. (18), (19), (21) and (22) using linear regression. Indeed, taking the natural log of Eqs. (18), (19) and (21) gives the following expressions

$$ln(1 - \theta_{xl}(t)/\theta_b^0) \approx -k_{xl}C_{FH}t, \tag{24}$$

$$ln(1 - \theta_{xl}(t)) \approx ln(1 - \theta_b^0) - k_a C_{TF} t, \tag{25}$$

$$ln(1 - \theta_{xl}(t)) \approx -k_{xl}C_{FH}\theta_b^0 t \tag{26}$$

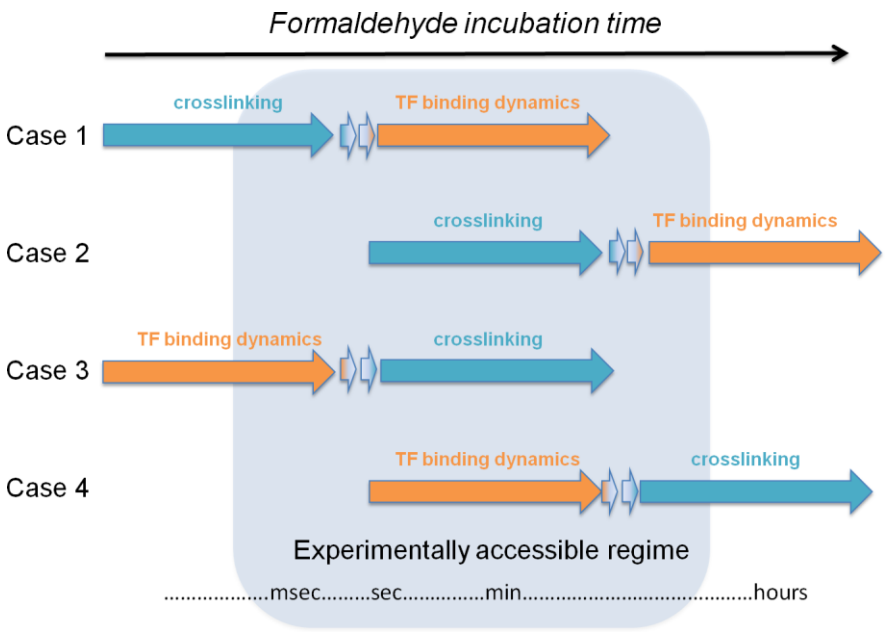


Figure 3-5: Schematic illustrating four possible cases of regimes considered for deriving CLK model

Schematic illustrating four possible cases in which crosslinking kinetics (blue arrows) and TF binding dynamics (orange arrows) contribute to the increase in ChIP signal with increasing formaldehyde incubation time (black arrow at top of figure). The four cases arise as a result of our experimental observation that in general, the ChIP signal increases dramatically in response to relatively short formaldehyde incubation times, and then more gradually in response to longer incubation times. This suggests that the processes of formaldehyde crosslinking and chromatin binding dynamics are themselves well separated in time. Reactions too fast (less than ~100 ms) and too slow (>40 min) are outside the experimentally accessible regime (shown by the central light blue shaded area). In Cases 1 and 2, crosslinking kinetics is assumed to be much faster than TF-chromatin binding dynamics. In Case 1, crosslinking occurs very rapidly (seconds time scale), followed by TF-chromatin binding which is on-rate limited and occurs on the

order of seconds to ~30 minutes. Case 2 is similar to Case 1 in that crosslinking occurs on a faster time scale than binding dynamics, but in this case crosslinking occurs more slowly and TF-DNA binding dynamics is even slower still. In Case 2, on-rate limited chromatin binding dynamics occurs on the minutes to hours time scale (i.e., much of it is beyond our experimentally accessible time range). In Cases 3 and 4, TF-chromatin binding dynamics is much faster than crosslinking kinetics. In Case 3, TF-chromatin binding dynamics happens over the first few seconds while most of the measured ChIP signal increase would be explained by the crosslinking reaction rate. In Case 4, the overall ChIP reaction is limited by the crosslinking reaction rate. The experimentally accessible increase in ChIP signal is linearly dependent on the crosslinking rate. As the formaldehyde incubation time increases (~30 minutes to hours), the crosslinking-limited reaction drives the ChIP signal to saturation by an exponential relationship with the crosslinking rate. As discussed in the text, the CLK data reported here are best described by Case 1.

$$\theta_{xl}(t) = 1 - \frac{\tau_+ e^{-t/\tau_+} - \tau_- e^{-t/\tau_-}}{\tau_+ - \tau_-} + \frac{\theta_b^0 \tau_+ \tau_- k_{xl} C_{FH}}{\tau_+ - \tau_-}\left(e^{-t/\tau_+} - e^{-t/\tau_-}\right)$$

where

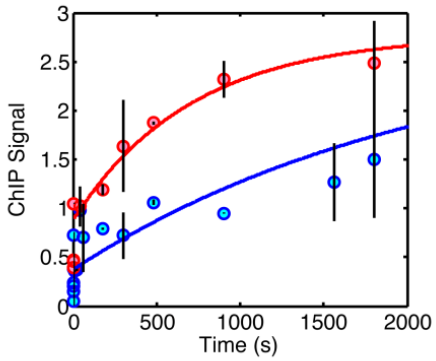
$$\frac{1}{\tau_+} = \frac{(k_a C_{TF} + k_d + k_{xl} C_{FH})}{2}\left[1 + \sqrt{1 - \frac{4k_a C_{TF}}{k_{xl} C_{FH}} \times \frac{1}{\left[1 + \left(\frac{k_a C_{TF} + k_d}{k_{xl} C_{FH}}\right)\right]^2}}\right]$$

$$\frac{1}{\tau_-} = \frac{(k_a C_{TF} + k_d + k_{xl} C_{FH})}{2}\left[1 - \sqrt{1 - \frac{4k_a C_{TF}}{k_{xl} C_{FH}} \times \frac{1}{\left[1 + \left(\frac{k_a C_{TF} + k_d}{k_{xl} C_{FH}}\right)\right]^2}}\right]$$

A

$t \sim \tau_-$ **and** $k_{xl} C_{FH} \gg k_a C_{TF}, k_d$

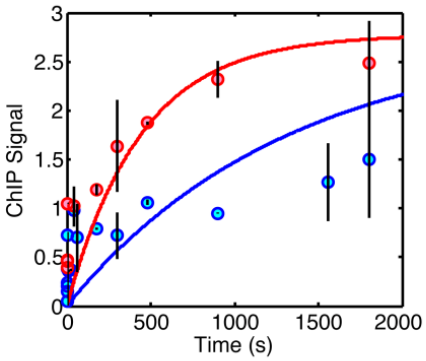
$$\theta_{xl}(t) \approx 1 - \left(1 - \theta_b^0\right) e^{-k_a C_{TF} t}$$



C

$t \sim \tau_-$ **and** $k_{xl} C_{FH} \ll k_a C_{TF}, k_d$

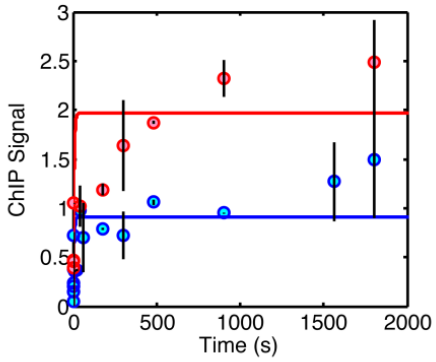
$$\theta_{xl}(t) \approx 1 - e^{-k_{xl} C_{FH} \theta_b^0 t}$$



B

$t \sim \tau_+$ **and** $k_{xl} C_{FH} \gg k_a C_{TF}, k_d$

$$\theta_{xl}(t) \approx \theta_b^0 (1 - e^{-k_{xl} C_{FH} t})$$



D

$t \sim \tau_+$ **and** $k_{xl} C_{FH} \ll k_a C_{TF}, k_d$

$$\theta_{xl}(t) \approx \theta_b^0 k_{xl} C_{FH} t$$

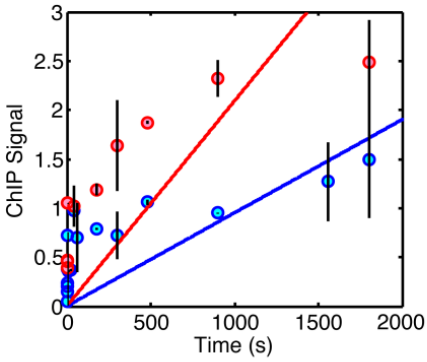


Figure 3-6: Example of CLK data fitting to the four approximate models

An example of CLK data fitting to the four approximate models, using data obtained for Gal4 binding to the *GAL3* promoter. See Section 2.3 for derivations and description of these limiting cases. In each graph, the blue circles correspond to data obtained from cells with WT levels of Gal4, and the red circles to data obtained from cells with overexpressed Gal4. Red and blue curves correspond to the fits obtained in each case. The full CLK model (Eq (16)) is shown at the top. (A) CLK model fits for Case 1 (described by Eq (19)). (B) CLK model fits for Case 2 (described by Eq (18)). (C) CLK model fits for Case 3 (described by Eq (21)). (D) CLK model fits for Case 4 (described by Eq (22)). The approximate model equation for each case is shown above the graph and the assumptions giving rise to each approximate case are shown in red text.

which are linear in the crosslinking time, *t*. We note that the last approximate model, Eq. (22), is linear in crosslinking time.

The specific approach for fitting experimental data to the CLK model is outlined in Figure 3-7. With Eqs. (18), (19) and (21), as starting points, the step wise procedure involves robust linear regression to fit these log-linear equations (detailed in Figure 3-7-B), followed by nonlinear regression to fit Eqs. (18), (19) and (21) (detailed steps shown in Figure 3-7-C), followed by nonlinear regression to fit the full CLK model, Eqs. (11), (16), by a Multi-Pass Parameter Estimation Procedure (MPPEP; Figure 3-7-D) to the experimental data. For Eq. (22), the stepwise procedure is similar. Each step uses the previous steps estimated parameters as initial guesses. We apply this overall procedure starting with the different approximate equations derived above to arrive at final fits and determine the best case, hence dynamic model, based on the lowest RMSE.

Case 1: For the first case we use Eq. (25) to fit the CLK model with robust linear regression to arrive at initial estimates of k_a, θ_b^0 and k_d. Using the parameters we gather from the linear regression as initial guesses we apply nonlinear regression to fit *Ip(t)* (ChIP signals) to Eq. (19). The parameters obtained after this step are θ_b and $k_a C_{TF}$ and *Ip(∞)*).

In order to fit the full CLK model, Eqs. (11), (16), to experimental CLK data, we use the parameter estimates from fitting Eq. (19) together with a series of initial guesses

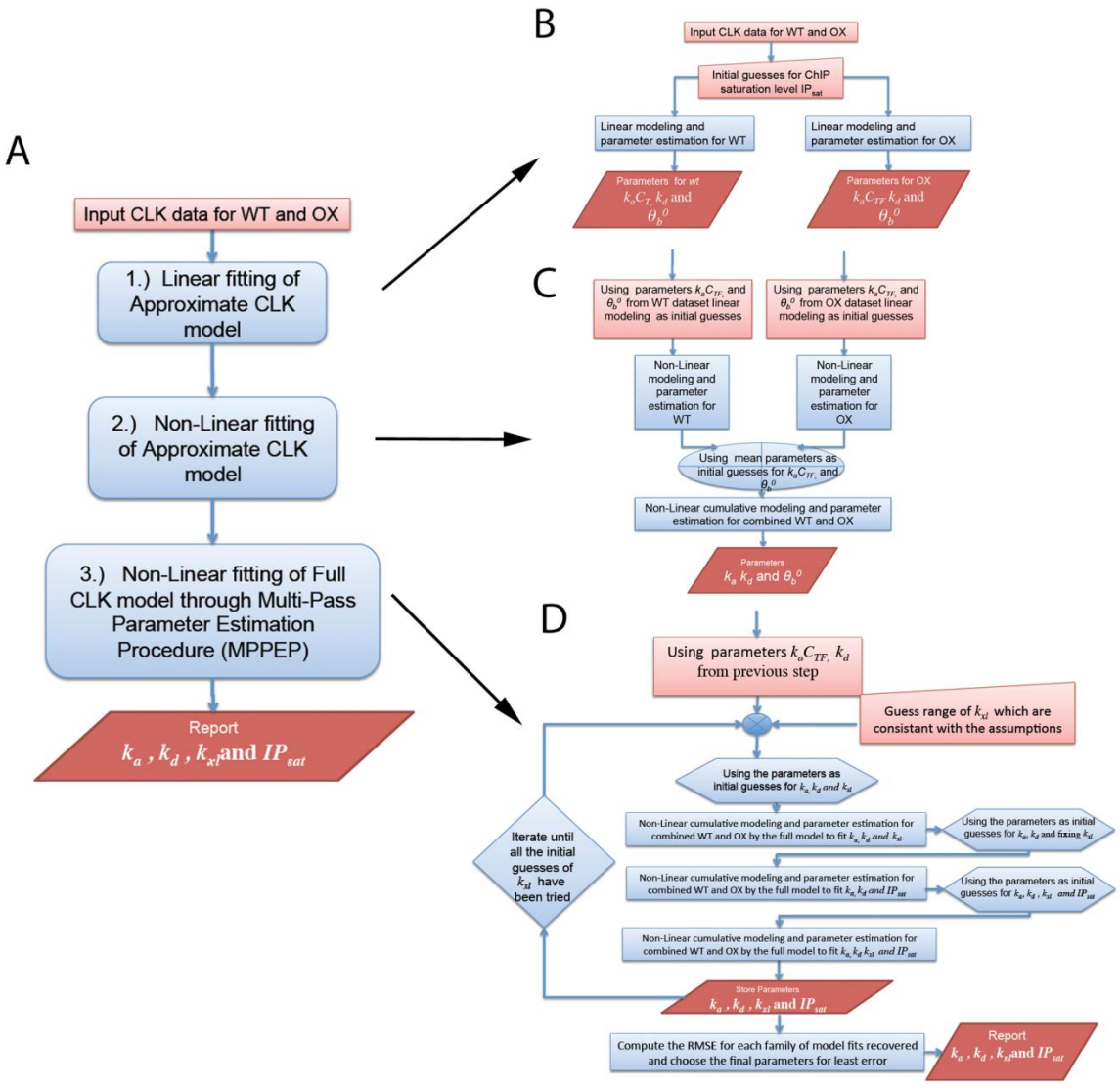


Figure 3-7: Flow charts describing the nonlinear regression fitting procedure

(A) Overview of the nonlinear regression procedure. (B) Flow chart of the linear regression procedure. (C) Flow chart of the nonlinear regression procedure using the approximate models. (D) Flow chart for the nonlinear regression fitting procedure using MPPS (Multi-Pass Parameter Estimation Procedure).

for $k_{xl}C_{FH}$ which are consistant with approximations used to derive this equation, $k_{xl}C_{FH} > k_aC_{TF}$ and k_d. Specifically, we generated an array of initial $k_{xl}C_{FH}$ values ranging from 2× to $10^6 \times \max(k_aC_{TF}, k_d)$ in 1.5 to 2-fold steps. We arrive at the final fitted values for k_aC_{TF}, k_d, $k_{xl}C_{FH}$, andand $Ip(\infty)$ by simultaneously fitting wt and ox CLK data to Eqs. (11), (16) using the Matlab function 'nlinfit' within a Multi-Pass Parameter Estimation Procedure (MPPEP) described below.

Case 2: Similar to the method described in Case 1 we used Eq. (24) to apply robust linear regression to obtain $k_{xl}C_{FH}$. Next we apply nonlinear regression to fit Eq (20) to CLK data using initial guesses for $Ip(\infty)\cdot\theta_b^0$ followed by robust linear regression to arrive at refined initial estimates of θ_b^0 and k_{xl}. Use of these parameters from fits to wt and ox data allowed us to derive estimates of $Ip(\infty)$, $K = k_aC_{TF}/k_d$, and $k_{xl}C_{FH}$ where we took the mean of $k_{xl}C_{FH}$ from the wt and ox fits. Using these parameters as initial guess next we apply nonlinear regression analysis, we fit Eq. (18) simultaneously to wt and ox CLK data from which we derived refined estimates of $Ip(\infty)$, $K = k_aC_{TF}/k_d$, and $k_{xl}C_{FH}$.

Assuming that the crosslinking reaction rate is much faster than TF-chromatin binding dynamics ($k_{xl}C_{FH} >> k_aC_{TF}$ and k_d), we select an array of initial guesses for k_aC_{TF} and k_d which are at least an order of magnitude smaller than $k_{xl}C_{FH}$ and satisfy $K = k_aC_{TF}/k_d$ where K is the TF-DNA equilibrium binding constant obtained from previous step. Using these as initial estimates of the parameters, we fit Eqs. (11), (16) simultaneously to wt and ox CLK data using 'nlinfit' within our MPPEP (described below) to derive the final values for k_aC_{TF}, k_d, $k_{xl}C_{FH}$, and $Ip(\infty)$.

Case 3: Similar to the earlier described two cases we use Eq (26) to estimate $k_{xl}C_{FH}\theta_b^0$. Using this as initial guess we apply nonlinear regression to fit Eq (24) using 'nlinfit' to obtain estimates of θ_b^0, $k_{xl}C_{FH}$.and $K = k_a C_{TF}/k_d$.

Using the assumption that the crosslinking reaction rate is much slower than TF-chromatin binding dynamics, we select an array of initial guesses for k_aC_{TF} and k_d that also satisfy the estimated values from previous steps. Using these as initial estimates of the parameters, we fit Eqs. (11), (16) simultaneously to wt and ox CLK data using 'nlinfit' within our MPPEP (described below) to derive the final values for k_aC_{TF}, k_d, $k_{xl}C_{FH}$, and $Ip(\infty)$.

Case 4: We fit Eq. (22) to CLK data using the robust linear regression function 'regress' to

arrive at estimates of $\theta_b \cdot k_{xl}C_{FH}$ from the slope of the line. Using these to arrive at initial guesses for k_aC_{TF}, k_d, $k_{xl}C_{FH}$, and $Ip(\infty)$ which are compatible with the assumptions made to derive Eq (22), we apply non-linear regression fit of CLK data to Eq. (22) to arrive at estimates of k_aC_{TF}, k_d, $k_{xl}C_{FH}$, and $Ip(\infty)$.

Using the estimates of k_aC_{TF}, k_d, $k_{xl}C_{FH}$, and $Ip(\infty)$ from the previous step as initial guesses to the MPPEP (described below), we fit the full model, Eqs. (11), (16), to CLK data to arrive at final estimates of k_aC_{TF}, k_d, $k_{xl}C_{FH}$, and $Ip(\infty)$.

MPPEP (Multi-Pass Parameter Estimation Procedure)

Step 1:We fit ChIP signal as a function of crosslinking time to the full CLK model shown in Eqs. (11), (16) with k_aC_{TF}, k_d and $k_{xl}C_{FH}$ as free parameters and $Ip(\infty)$ fixed using the

Matlab 'nlinfit' function. Initial values for all these parameters are obtained from the fits to the approximate equation associated with each case.

Step 2:Using the values obtained for k_aC_{TF}, k_d, $k_{xl}C_{FH}$ from step 1 and $Ip(\infty)$ from the fit to the approximate equation associated with each case as initial guesses, we fit ChIP signal as a function of crosslinking time to the full CLK model, Eqs. (11), (16), with k_aC_{TF}, k_d, and $Ip(\infty)$ as free parameters and $k_{xl}C_{FH}$ fixed using the Matlab function 'nlinfit'.

Step 3:Using the values for k_aC_{TF}, k_d, $k_{xl}C_{FH}$, and $Ip(\infty)$ obtained from step 2 as initial guesses, we fit ChIP signal as a function of crosslinking time to the full CLK model, Eqs. (11), (16), with k_aC_{TF}, k_d, $k_{xl}C_{FH}$, and $Ip(\infty)$ as free parameters.

We follow steps 1-3 for each value of an array of initial guesses (e.g., array of guesses for $k_{xl}C_{FH}$ in step 4 of case 1) and select the fit and corresponding estimates of k_aC_{TF}, k_d, $k_{xl}C_{FH}$, and $Ip(\infty)$ which yields the smallest MSE.

TF-Specific Fitting Approaches: While the majority of fits to CLK data were executed using the steps detailed in Case 1 above, the procedure for fitting Ace1 at *CUP1* and LacI at *LacO* deviated from this despite the fact that the final fitted curves satisfy case 1 dynamics (Figure 3-2). For Ace1 at *CUP1*, the procedure detailed in case 2 lead to the final fitted curves shown in Figure 3-10A. For LacI at *LacO*, we tuned the overexpression concentration and found $C_{TF} = 9.5\mu M$ produced the best fits. We note that this value differs from the estimated value shown in Table 3-5 by 2.6-fold. Use of the estimated overexpression concentration shown in Table 3-5 yielded un-physically high values for the ChIP saturation signal, $Ip(\infty)$.

3.3.2.5 – Error Estimation

In order to estimate the errors in the estimates of the kinetic parameters, we sampled the error in the measured ChIP signal to generate multiple curves using the fitting procedures described above. Specifically, we calculated the mean and standard deviation of the ChIP signal from biological replicates at each experimental CLK datapoint. We randomly sampled a normal distribution with the estimated mean and standard deviation of the ChIP signal in order to generate error-sampled CLK data. We did this 10000 times. We fit these error-sampled data to Eqs. (11), (16) using the fitting procedures described above to arrive at 10000 values for $k_a C_{TF}$, k_d, k_{xl}, θ_b^0, and $t_{1/2}$. In Figure 3-12-14, we display distributions of $ln(k_a C_{TF})$, $ln(k_d)$, and $ln(k_{xl} C_{FH})$, and, for Figure 3-15, we plotted the same for $ln(k_a C_{TF})$, $ln(k_d)$, $ln(k_{xl} C_{FH})$, $ln(t_{1/2})$,and $ln(\theta_b^0)$. We note that in each case where the distribution of a parameter's estimates was unimodal, the distribution appeared more normal for log-transformed parameter estimates compared to that of the untransformed parameter estimates. We then calculated the left and right tail standard deviations from the log-transformed parameters for $k_a C_{TF}$, k_d, k_{xl}, θ_b^0, and $t_{1/2}$, which correspond to the lower and upper bounds of the parameters, respectively, shown in Tables 3-7 – 3-9.

3.4 Results

To test the CLK method, we analyzed Gal4 binding to the single UAS in the *GAL3* promoter. The Gal4 system has provided a paradigm for transcriptional regulation(Traven et al. 2006), but the in vivo stability of the Gal4-promoter interaction has been the subject of debate (Collins et al. 2009; Nalley et al. 2006). A quench flow apparatus was adapted to acquire formaldehyde-treated samples on the sub-second

timescale, and longer time points were obtained by hand mixing, prior to quenching in glycine (supplementary online text). As predicted by the simulations (Figure 3-4, Figure 3-2 and Figure 3-3), the ChIP signal increased dramatically at short formaldehyde incubation times (< 5 sec), and then gradually following longer incubation times (Figure 3-8A, blue curve). The time dependence of the ChIP signal substantiates several key aspects of the model (Figure 3-1), and other fundamental suppositions were validated experimentally. First, the steep increase in ChIP signal at short crosslinking times demonstrates that crosslinking occurred rapidly and that glycine efficiently quenched the reaction (Figure 3-8A,C), as stipulated in the model. The curve was shifted upward in cells with a 2.5-fold increase in Gal4, (Figure 3-8A, red curve), consistent with the time dependence of the slower phase of the ChIP signal being driven by the overall on-rate for Gal4 chromatin binding and not formaldehyde reaction kinetics. In the model, the ChIP assay rapidly captures specifically bound TFs but does not inactivate or nonspecifically crosslink the remaining TF pool. Remarkably, the Gal4-promoter interaction occurred in cells even when binding was induced after formaldehyde pre-treatment (Figure 3-8B). Thus, Gal4 was not nonspecifically inactivated by formaldehyde. Moreover, the levels of soluble Gal4 and other proteins were reduced less than two-fold in cell extracts following formaldehyde incubation, and their apparent molecular weights were not detectably affected (Figure 3-8D, Figure 3-9). In addition, ChIP signals were indistinguishable over an 8-fold range of formaldehyde concentration (Figure 3-8E), demonstrating that formaldehyde was not limiting in the reaction. CLK analysis revealed that the Gal4-*GAL3* interaction had a $t_{1/2}$ of about 10 min (Figure 3-8A; Table 3-9), suggesting that a single Gal4 complex facilitates multiple rounds of transcription initiation. Combined with

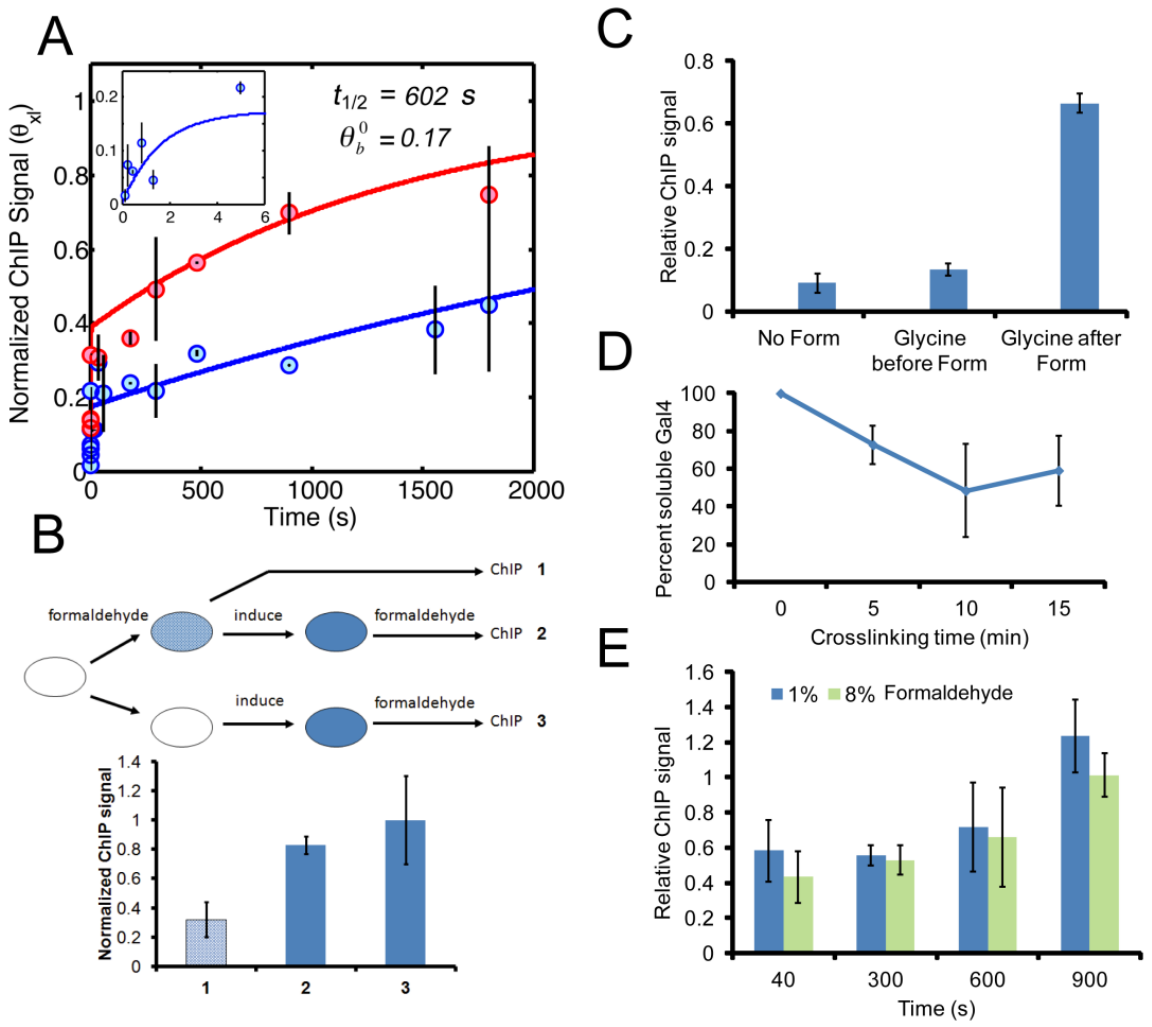


Figure 3-8: CLK analysis of Gal4 and tests of model assumptions.

(A) Model fits of CLK data for Gal4 binding to the *GAL3* promoter in cells with WT Gal4 levels (blue line) and cells with 2.5-fold overexpression of Gal4 (red line). Inset shows first 5 sec of time course from cells with WT Gal4 levels. (B) Gal4 ChIP results obtained with cells treated as shown in the schematic. ChIP signal obtained in formaldehyde treated, uninduced cells (1), formaldehyde treated cells subsequently induced by addition of galactose (2), and cells induced with galactose and subsequently treated with formaldehyde (3). Note that Gal4 chromatin binding was fully inducible in

formaldehyde treated cells. (C) Glycine addition prior to formaldehyde (Form) prevents crosslinking. The graph shows the relative Gal4 ChIP signal obtained when glycine was added prior to formaldehyde or 8 min after formaldehyde treatment, compared to cells in which no formaldehyde was added. (D) Relative soluble Gal4 protein level in extracts from cells treated with formaldehyde for the indicated times. Gal4 was quantified by Western blotting. (E) Gal4 ChIP signals at *GAL3* obtained using cells treated with 1% or 8% formaldehyde for the indicated times. ChIP signals did not depend on formaldehyde concentration.

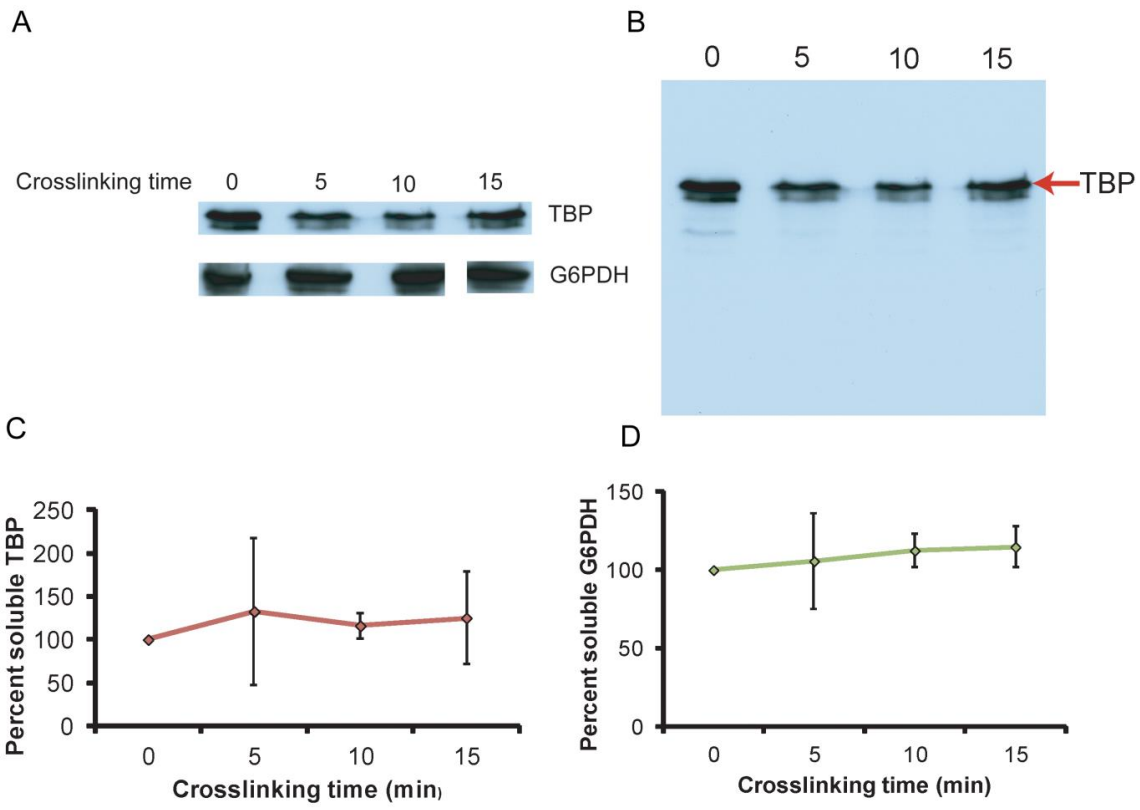


Figure 3-9: Quantification of the soluable TBP in formaldehyde treated cells

Soluble TBP levels are reduced less than two-fold in cell extracts prepared from formaldehyde-treated cells. (A) Western blotting analysis of soluble TBP (top panel) and G6PDH (below) in extracts made from cells treated with 1% formaldehyde for the indicated times in minutes. (B) Western blot of soluble fraction of TBP as shown in A. This image of the entire blot shows that there was no detectable proportion of TBP with an aberrant migration pattern as a consequence of formaldehyde treatment. (C, D) Quantitation of soluble TBP and G6PDH from Western blots such as those shown in (A) and (B).

the low fractional promoter occupancy (~0.17), we conclude that the *GAL3* gene is likely transcribed in infrequent bursts.

To better define the dynamic range within which the CLK method can capture kinetic information in vivo, we analyzed two TFs whose widely divergent dynamic behavior could be independently measured by fluorescence recovery after photobleaching (FRAP). FRAP was possible in these cases because the fluorescently tagged factors interact with tandem arrays of binding sites, making the chromosomal loci visible by microscopy. The CLK measured $t_{1/2}$ for the interaction of the Ace1 TF with the *CUP1* gene array(Karpova et al. 2008) was 11 sec, in excellent agreement with the value of 31 sec obtained by FRAP (Figure 3-10A,B; Table 3-8). The interaction of LacI-GFP with an array of 256 *Lac* operators (Robinett et al. 1996) was far more stable, and the two methods yielded $t_{1/2}$ values again within about a factor of four (Figure 3-10C,D; Table 3-8). Thus, as validated by an independent approach, the CLK method can reveal rank-ordered estimates of TF-chromatin interaction stability over a wide range in vivo, including interactions that persist for mere seconds. Compared to other methods, the CLK method increases the time resolution of chromatin dynamics at single copy loci by two to three orders of magnitude.

To further explore transcription dynamics using this method, we investigated the interaction of the TATA-binding protein (TBP) with each of seven different promoters possessing diverse transcriptional activities and driven by RNA polymerases I, II or III. Consistent with expectation (Roberts et al. 2003), the Pol III-driven *SNR6* (U6) promoter had the highest occupancy, however, interestingly, occupancies of all promoters were well below saturation (Figure 3-11A; supplementary online text). Moreover, TBP-

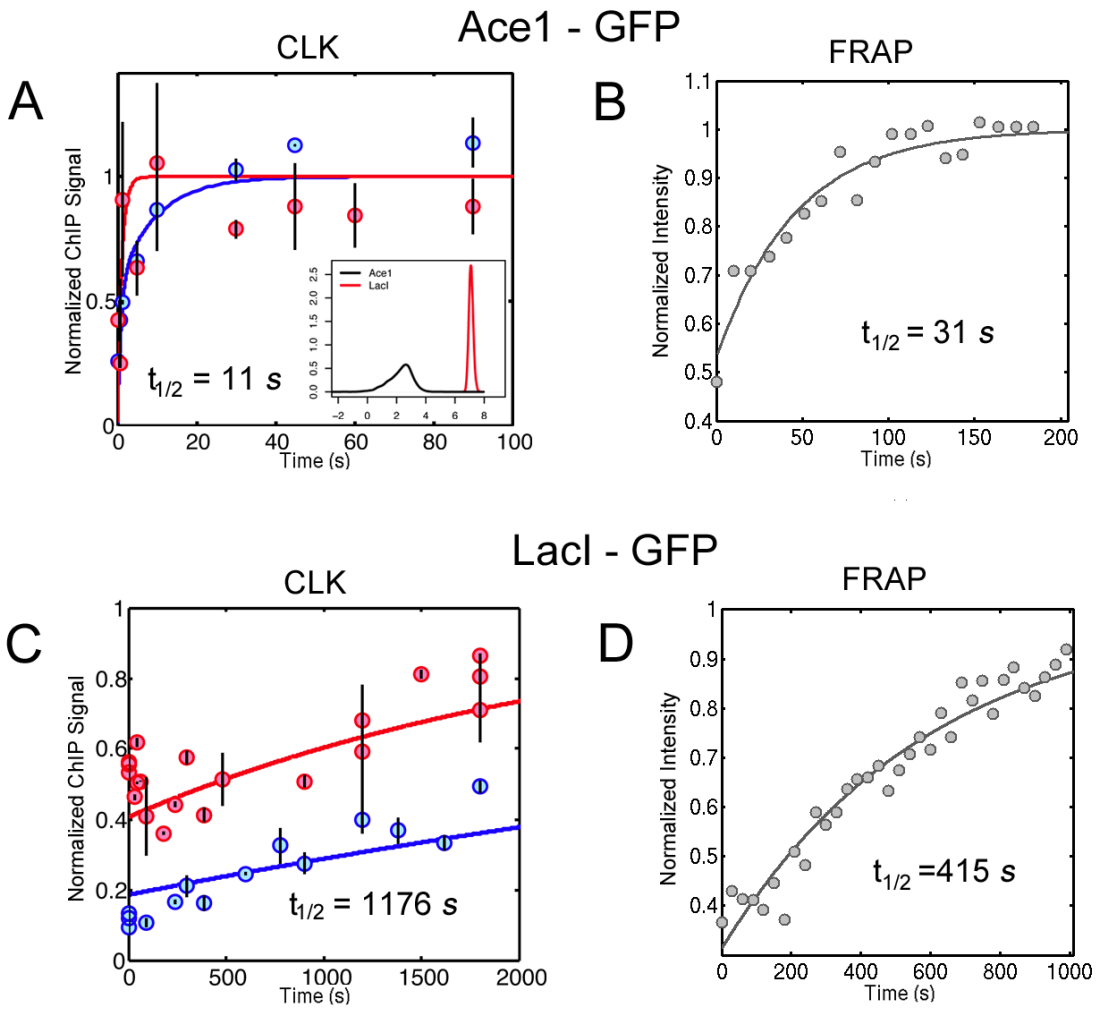


Figure 3-10: Comparison of TF-chromatin dynamics by CLK and FRAP

(A) Model fits of CLK data for Ace1-GFP binding to *CUP1* in cells with two different expression levels of Ace1-GFP (low, blue curve; high, red curve; Table 3-2). Inset: distributions of $t_{1/2}$ values obtained from multiple independent fits of the Ace1-GFP or LacI-GFP CLK data. (B) FRAP of Ace1-GFP in cells with low Ace1-GFP levels. (C) Model fits of CLK data for LacI-GFP binding to the *Lac* array in cells with low (blue curve) or high (red curve) levels of LacI-GFP (Table 3-2). (D) FRAP of LacI-GFP in cells with low LacI-GFP levels.

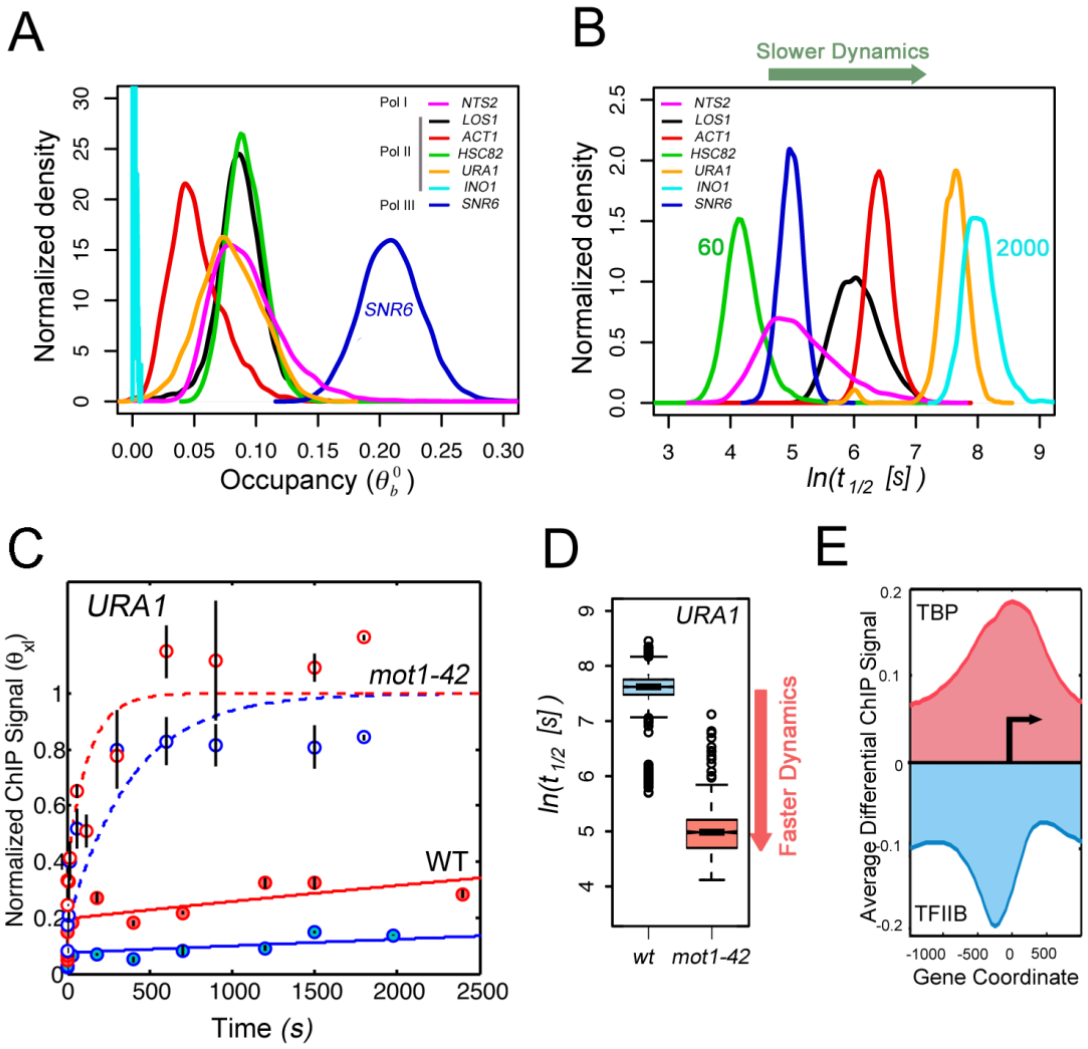


Figure 3-11: TBP dynamics and regulation by Mot1

(A) Distributions of TBP occupancy at different yeast promoters obtained by multiple independent fits of the CLK data . (B) Distributions of TBP-promoter half-lives, whose mean values vary from 60 to about 2000 seconds. (C) Model fits of CLK data for TBP binding to the Mot1-activated *URA1* promoter in WT (sold lines) and *mot1-42* cells (dashed lines). Data and fits from cells expressing WT levels of TBP are shown in blue, results from cells over-expressing TBP are in red. (D) Boxplots for distribution of $t_{1/2}$

values (log scale) for TBP binding to the Mot1-activated *URA1* promoter in WT (blue) and *mot1-42* cells (red). (E) Average genome-wide $\log_2$ differential TBP and TFIIB ChIP-chip signals at promoters in *mot1-42* versus WT cells shown with respect to the transcription start site (arrow) (Poorey et al. 2010).

Table 3-7 Estimated kinetic parameters for TBP binding to the indicated promoters and in the indicated strains.

Promoter	$k_aC_{TF}(s^{-1})$	$k_a(M^{-1}s^{-1})$	$k_d(s^{-1})$	$k_{xl}(M^{-1}s^{-1})$	IP_{sat}	θ_b^0	$t_{1/2}(s)$
LOS1	1.7 (+1.2, -0.8)x10^{-4}	1.4 (+1, -0.6)x10^{1}	1.7 (+0.7, -0.6)x10^{-3}	6.7 (+4.5, -2.8)	1.1 (+0.4, -0.3)	0.09 (+0.03, -0.03)	406 (+198, -118)
ACT1	6.8 (+2.7, -3.3)x10^{-5}	5.66 (+2.2, -2.7)	1.2 (+1.5, -0.7)x10^{-3}	1 (+10^{3}, -1)x10^{2}	1.8 (+1.2, -0.4)	0.05 (+0.07, -0.03)	574 (+867, -320)
NTS2	4.4 (+3, -2)x10^{-4}	3.7 (+2.4, -1.6)x10^{1}	4.5 (+4, -2.1)x10^{-3}	6.0 (+6.6, -3.2)	1.3 (+0.3, -0.2)	0.09 (+0.03, -0.03)	155 (+137, -73)

Promoter	Strain	$k_aC_{TF}(s^{-1})$	$k_a(M^{-1}s^{-1})$	$k_d(s^{-1})$	$k_{xl}(M^{-1}s^{-1})$	IP_{sat}	θ_b^0	$t_{1/2}(s)$
URA1	WT	2.7 (+2.1, -1.0)x10^{-5}	2.2 (+1.7, -0.9)	3.3 (+0.9, -0.5)x10^{-4}	1.0 (+0.3, -0.1)	2.8 (+1.1, -0.9)	0.08 (+0.04, -0.02)	2120 (+389, -440)
	mot1-42	2.6 (+0, -1.7)x10^{-3}	2.2 (+0, -1.4)x10^{2}	1.3 (+0, -1)x10^{-2}	16.3 (+0, -13.7)	0.66 (+0.1, -0)	0.17 (+0.06, -0)	53 (+157, -0)
HSC82	WT	1.1 (+0.5, -0.4)x10^{-3}	9.0 (+4.4, -3.1)x10^{1}	1.2 (+0.6, -0.6)x10^{-2}	1 (+23, -1)x10^{9}	1.1 (+0.1, -0.1)	0.08 (+0.04, -0.01)	57 (+58, -19)
	mot1-42	1.3 (+4, -0.8)x10^{-3}	1.1 (+3.3, -0.6)x10^{2}	6.1 (+37, -4.8)x10^{-3}	8.6 (+58, -7)	0.8 (+0, -0.14)	0.17 (+0.1, -0.05)	114 (+433, -98)
INO1	WT	3.6 (+32, -0)x10^{-8}	3.0 (+26, -0)x10^{-3}	2.0 (+1, -0.2)x10^{-4}	3.0 (+1.2, -0.5)	975.6 (+0, -851)	2 (+10, -0)x10^{-4}	3529 (+387, -1247)
	mot1-42	2.6 (+0.7, -0.5)x10^{-4}	2.2 (+0.6, -0.5)x10^{1}	1.2 (+0.3, -0.2)x10^{-3}	8.5 (+2.1, -1.7)	0.8 (+0.07, -0.07)	0.19 (+0.02, -0.02)	604 (+154, -123)
SNR6	WT	1.3 (+0.1, -0.1)x10^{-3}	1.1 (+0.1, -0.1)x10^{2}	4.9 (+0.8, -1)x10^{-3}	6.3 (+3.7, -2.2)	3.3 (+0.1, -0.1)	0.21 (+0.03, -0.02)	142 (+35, -21)
	mot1-42	2.2 (+1.5, -1.3)x10^{-3}	1.9 (+1.2, -1.1)x10^{2}	1.3 (+3.7, -1.1)x10^{-2}	16.2 (+88, -14)	2.1 (+0.2, -0.1)	0.15 (+0.1, -0.06)	53 (+256, -39)

Table 3-8 Estimated kinetic parameters for Ace1-GFP and LacI-GFP chromatin binding.

Transcription factor	Promoter	$k_a C_{TF}\,(s^{-1})$	$k_a\,(M^{-1}s^{-1})$	$k_d\,(s^{-1})$	$k_{xl}\,(M^{-1}s^{-1})$	IP_{sat}	θ_b^0	$t_{1/2}\,(s)$
Ace1	*CUP1*	1.1 (+0.8, -0.5)x10^{-1}	1.1 (+0.8, -0.5)x10^{5}	6.1 (+10, -3)x10^{-2}	2.9 (+1.6, -0.7)	1.1 (+0.05, -0.02)	0.64 (+0.1, -0.1)	11 (+17, -7)
LacI	*LacO*	1.3 (+0.1, -0.3)x10^{-4}	1.3 (+0.1, -0.3)x10^{2}	5.9 (+0.8, -1)x10^{-4}	1.6 (+0.3, -1.5)x10^{10}	4.8 (+0.4, -0.2)	0.19 (+0.01, -0.02)	1176 (+228, -135)

Table 3-9 Estimated kinetic parameters for Gal4 binding to the *GAL3* promoter

Transcription factor	$k_a C_{TF}\,(s^{-1})$	$k_a\,(M^{-1}s^{-1})$	$k_d\,(s^{-1})$	$k_{xl}\,(M^{-1}s^{-1})$	IP_{sat}	θ_b^0	$t_{1/2}\,(s)$
Gal4	2.4 (+5.4, -1.7)x10^{-4}	1.4 (+3, -1)x10^{3}	1.2 (+1, -0.4)x10^{-3}	1.7 (+0.8, -0.7)	3.3 (+4.7, -1.8)	0.17 (+0.18, -0.1)	602 (+358, -227)

promoter interactions varied dramatically, with $t_{1/2}$ values ranging from one to about thirty minutes (Figure 3-11B; Figure 3-16 and Table 3-7), and in many cases half-lives were much shorter than distinguishable by any other current technique. To test whether the method can quantify a dynamic difference associated with a perturbation in cellular transcription, we next compared TBP dynamics in WT and *mot1-42* cells. Mot1 is an essential regulator of TBP, which can use its ATPase activity to dissociate TBP from DNA in vitro (Viswanathan and Auble 2011). Evidence supports a direct role for Mot1 in gene activation, but how it accomplishes this is unknown. Using *URA1* as a model Mot1-activated gene (Sprouse et al. 2008), we observed dramatically different CLK curves for TBP binding to the *URA1* promoter in WT and *mot1-42* cells (Figure 3-11C; Table 3-7). Surprisingly, mutation of Mot1 gave rise to TBP binding that was far more dynamic than in WT cells (Figure 3-11D. Similar results were observed at *INO1*, another Mot1-regulated promoter, but not control promoters (Figure 3-16; Table 3-7). Rather than catalyzing dissociation of stable TBP-chromatin interactions, these results reveal a Mot1-mediated mechanism responsible for dissociating weakly bound TBPs in promoter regions, thereby facilitating much more stable binding of TBP in functional transcription complexes. This enzyme-catalyzed change in TBP dynamics appears essential for proper gene expression; analogous processes may operate to facilitate functional high affinity chromatin binding at the expense of weak binding by other TFs as well.

Time-dependent formaldehyde crosslinking ChIP data and CLK model fits for TBP binding to promoters referred to are shown in Figure 3-16. The full set of parameters obtained by CLK model fitting of all of the data sets in this study are shown in Tables 3.7-3.9, and errors in the parameters are presented below. Box plots showing

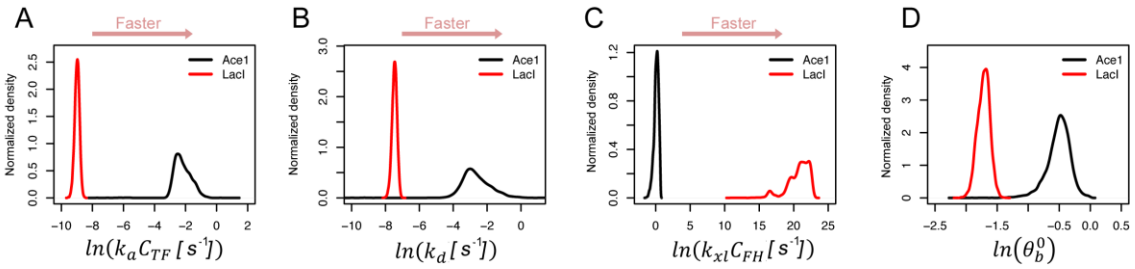


Figure 3-12: Density distributions of kinetic parameters and occupancy for LacI and Ace1 binding

Density distributions of kinetic parameters and occupancy (as indicated) obtained by multiple independent fits of the Ace1-GFP (black lines) and LacI-GFP (red lines) CLK data. The red arrow at the top of the figure shows the direction for a faster parameter set.

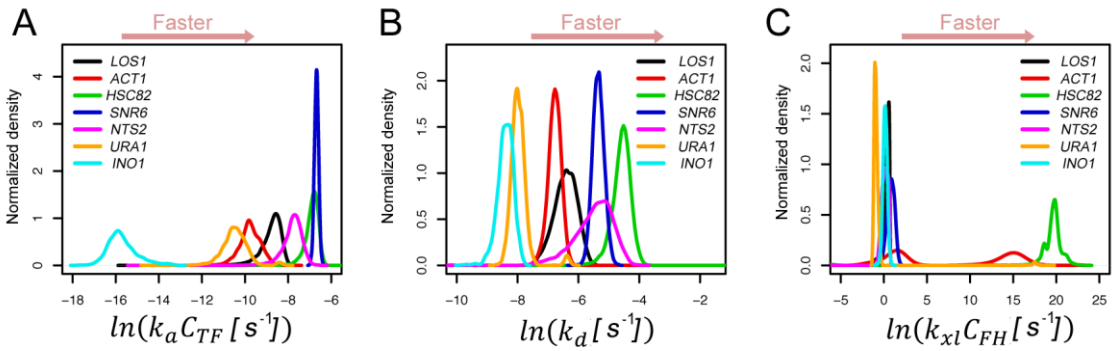


Figure 3-13: Density distributions of kinetic parameters for TBP binding

Density distributions of kinetic parameters for TBP binding to the indicated promoters obtained by multiple independent fits of the CLK data.

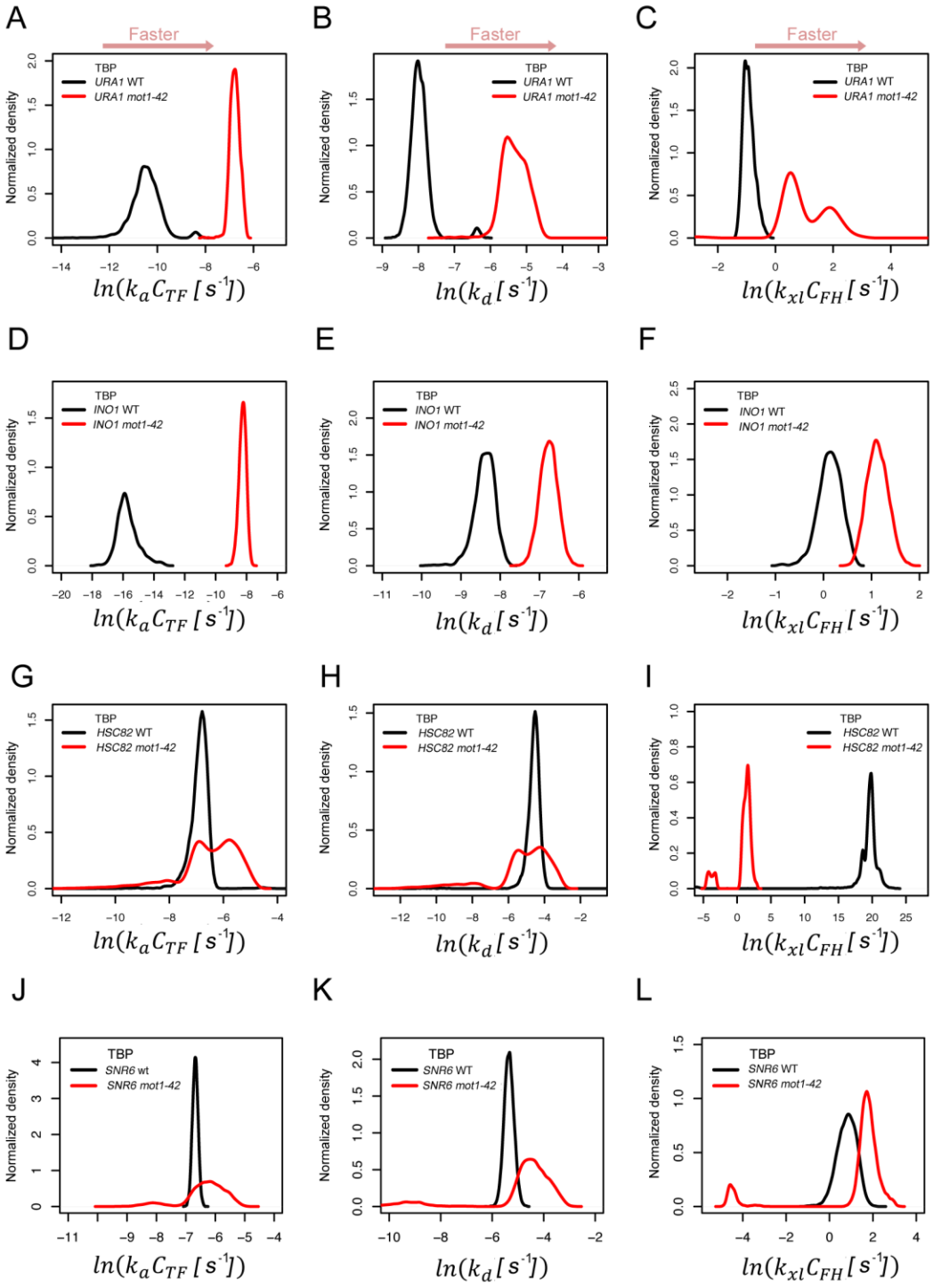


Figure 3-14: Density distribution of kinetic parameters for TBP binding in WT and *mot1-42*

Density distribution of kinetic parameters for TBP binding to the indicated promoters in WT (black lines) or *mot1-42* cells (red lines) obtained by multiple independent fits of the CLK data.

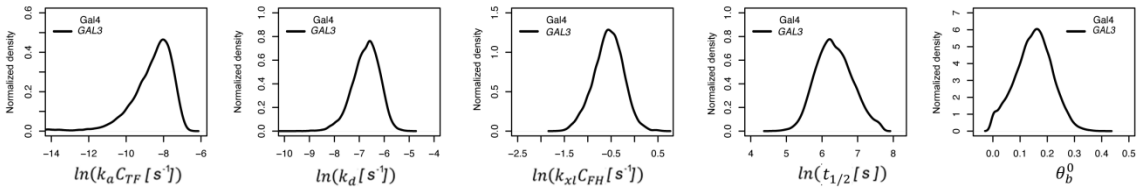


Figure 3-15: Density distributions of kinetic parameters for Gal4 binding

Density distributions of kinetic parameters estimated by CLK model nonlinear regression fits obtained by randomizing data points within the error range of the replicates for Gal4 binding to the *GAL3* promoter.

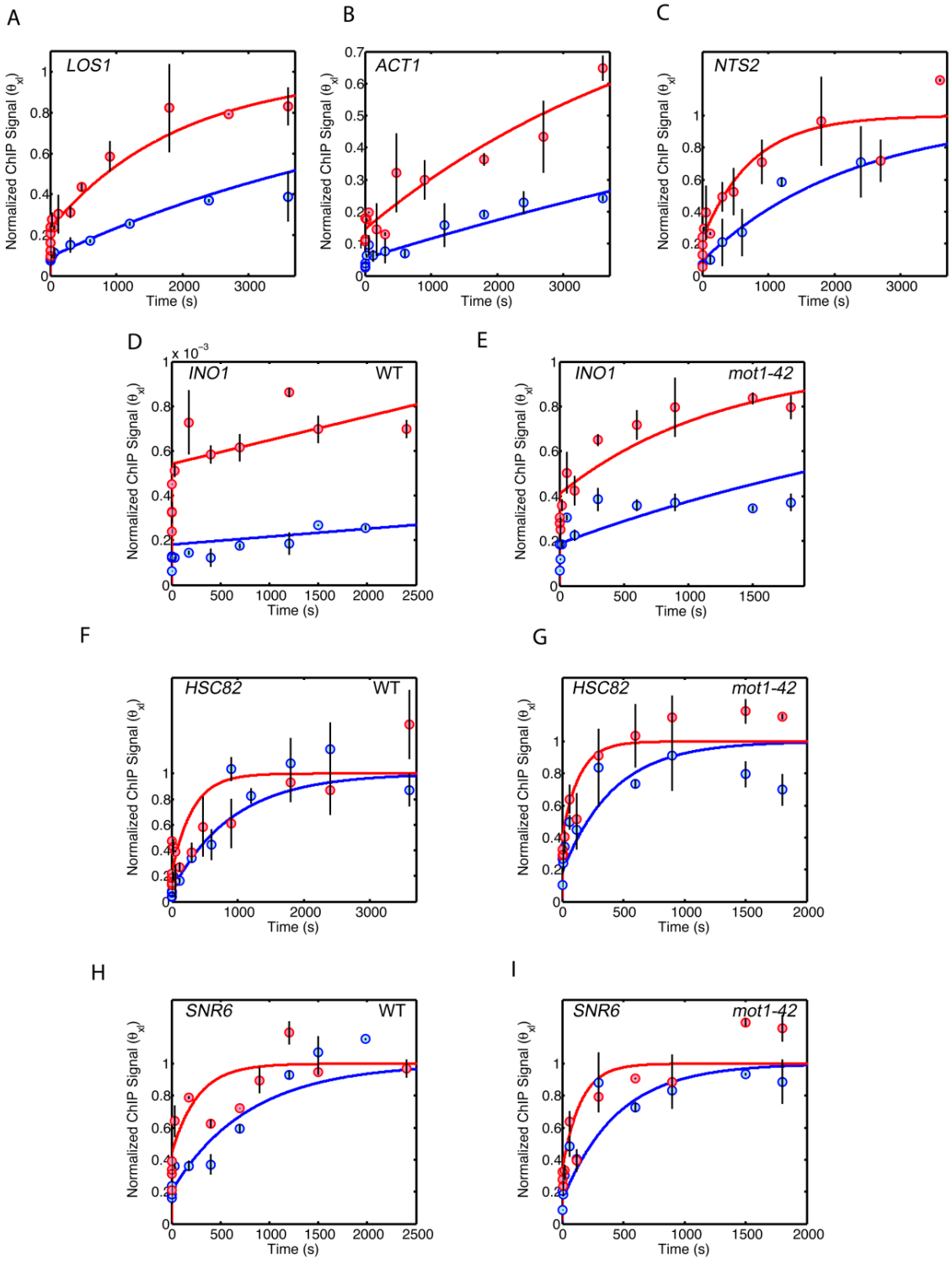


Figure 3-16: CLK model fits for TBP binding

CLK model fits for TBP binding to the promoters of the *LOS1* (A), *ACT1* (B), *NTS2* (RNA Pol I promoter in the ribosomal DNA repeat) (C), *INO1* (D), *HSC82* (F) and *SNR6*, a Pol III-transcribed gene (H). Panels E, G and I show the CLK model fits for TBP binding to the indicated promoters in *mot1-42* cells. The red curve fits the data obtained in cells with WT levels of TBP, the blue curve describes the data obtained in cells in which TBP was over-expressed three-fold over the WT level.

the distributions of complex half-life and fractional occupancy in WT versus *mot1-42* cells are shown in Figure 3-16 and Figure 3-17. Note that in addition to the shorter chromatin complex half-life and higher TBP occupancy observed at the *URA1* promoter in *mot1-42* versus WT cells as reported in the main text, a similar trend in dynamics was observed at the Mot1-repressed *INO1* promoter whereas there was less of a kinetic effect on TBP binding to the *HSC82* or the U6 promoter. In addition, TBP occupancies increased in *mot1-42* cells compared to WT cells at each of the three Pol II-driven promoters but not at the RNA Pol III-driven U6 promoter. Density distribution plots shown in Figures. 3-12-3.14 show how the model parameters obtained from multiple independent fits of each data set (see Section 3.3.2.5) vary for chromatin interactions at different sites and in different cells.

The most notable differences in chromatin interaction behavior evident from the density distribution plots include:

1. On-rate, off-rate, and occupancies are different for Ace1 and LacI binding to their respective sites (Figure 3-12). This is consistent with their dramatically different kinetic behavior measured by both the CLK method and by FRAP.
2. On-rates and off-rates for TBP binding to each of seven different promoters span a broad range (Figure 3-13 A, B). In contrast, with but one exception, the formaldehyde crosslinking rates in WT cells are tightly clustered (Figure 3-13C). See summary section (below) for an interpretation of the crosslinking rates.

On-rates and off-rates for TBP binding are distinctly different in WT versus *mot1-42* cells for interactions at the Mot1-regulated *URA1* and *INO1* promoters but similar in both cell types at the *HSC82* and U6 promoters (Figure 3-14). The crosslinking rates

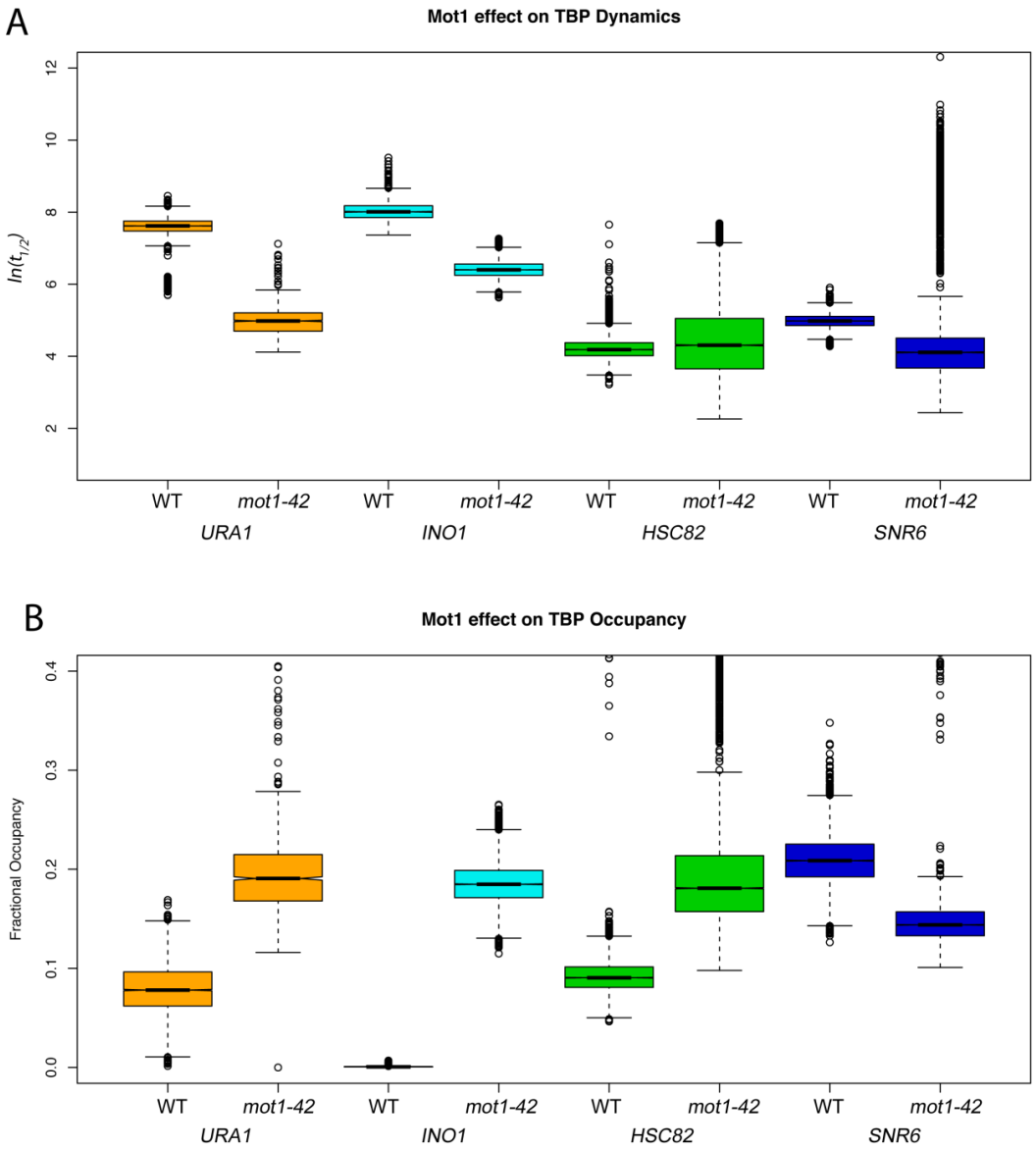


Figure 3-17: Box plot of distribution of parameters for TBP in WT and *mot1-42*

(A) Boxplots for distribution of $t_{1/2}$ values (log scale) for TBP binding to various promoters in WT and *mot1-42* cells. (B) Boxplots for distribution of fractional occupancy levels for TBP binding to various promoters in WT and *mot1-42* cells. Note that the TBP occupancy increases in *mot1-42* cells at each of three Pol II promoters (*URA1*, *INO1* and *HSC82*) but not at the Pol III-driven *SNR6* promoter.

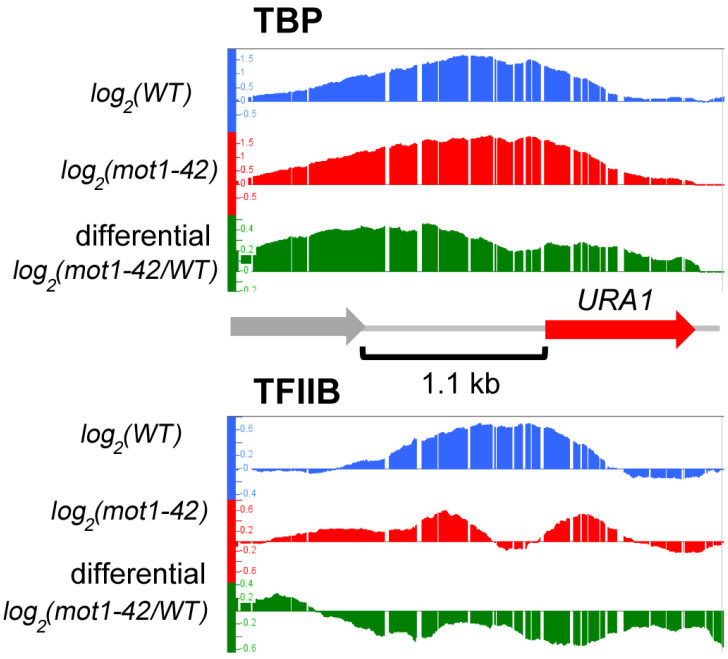


Figure 3-18: Genome wide distribution of TBP and TFIIB at the *URA1* gene

TBP (top panel) and TFIIB (bottom panel) binding to the ~1.1 kb *URA1* promoter in WT and *mot1-42* cells. The *URA1* gene is shown as the red arrow, and is transcribed from left to right. TBP and TFIIB $\log_2$ ChIP-chip signals (*29*) in WT cells are shown in blue; signals in *mot1-42* cells are shown in red. The $\log_2$ fold change differential signals for each factor are shown in green. Note that the TBP signal increased and the peak broadened in *mot1-42* cells compared to WT cells. In contrast, the TFIIB signal decreased, suggesting that the TBP that accumulates in *mot1-42* cells is nonfunctional.

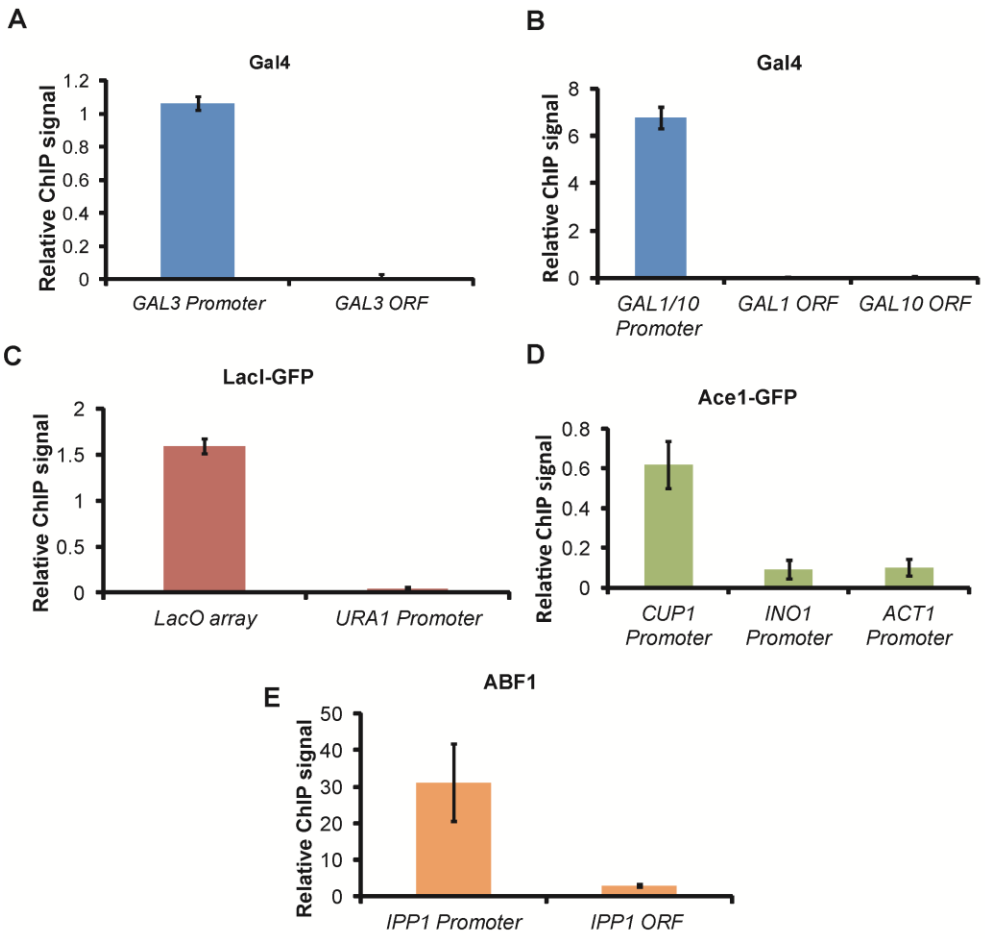


Figure 3-19 :Control for ChIP at non specific sites.

Relative ChIP signals for four TFs at specific versus control loci. Each bar graph shows the relative ChIP signal obtained from two biological replicate cultures for interaction of the indicated TF with known chromosomal binding site regions (leftmost bar in each graph) versus control chromosomal locations either in the open reading frame (ORF) adjacent to the targeted promoter (A, B and E) or at other promoters lacking the sequence recognized by the TF (C, D). Cells were fixed with formaldehyde for 8 min (A, B), 27 min (C), or 20 min (D, E). For Gal4, LacI, and Abf1, nonspecific ChIP signals were barely detectable. The Ace1 ChIP signal at *CUP1* was about six-fold greater than the signals obtained at *INO1* and *ACT1*.

(Table 3-7) also vary in these two cell types, but with one exception, the differences are in general rather modest in magnitude.

3.5 Summary

The CLK assay yields estimates of physical kinetic parameters as opposed to relative rates, and it is applicable over a much broader time scale than competition ChIP since it is not limited by the time required to synthesize or activate a competitor molecule. This will permit rapid chromatin interaction dynamics for a factor to be compared directly to kinetic parameters for functionally related factors or processes. The CLK methodology is in principle not limited to yeast, and it is based on ChIP, one of the most widely used assays in chromatin research. Our data suggest an explanation for why there is no detectable chromatin-bound TBP as judged by live cell imaging (Sprouse et al. 2008), but there are stable TBP complexes as judged by competition ChIP(van Werven et al. 2009). The CLK results show that TBP fractional occupancies are low. Thus, while there are stable TBP-promoter complexes in vivo, most promoters are not occupied at steady state. The surprisingly low occupancies are consistent with results showing that transcription in vivo occurs via uncoordinated stochastic cycles separated in time (Larson et al. 2011; Suter et al. 2011) . CLK results also illustrate the danger of inferring relative occupancies or dynamics from ChIP assays employing single, long formaldehyde incubation times. TBP ChIP signals are much greater in *mot1-42* cells than in WT cells, but the higher ChIP signals result from highly dynamic TBP molecules being trapped during the formaldehyde incubation period, rather than reflecting stable TBP binding.

An added advantage is that the approach is based on ChIP, one of the most widely used assays in chromatin research. The ability to obtain binding kinetic parameters will

permit direct comparison with rates for other steps in the transcription cycle. Moreover, the approach could be useful for analysis of other chromatin-based processes

TBP ChIP signals are much greater in *mot1-42* cells than in WT cells, but CLK analysis shows that the higher ChIP signals result from highly dynamic TBP molecules being trapped during the formaldehyde incubation period, rather than reflecting stable TBP binding. Formaldehyde-mediated trapping of dynamic binding may be a general ChIP phenomenon as the crosslinking reaction rates we calculate are in the same range as the rates of spontaneous DNA base flipping, the postulated rate limiting step in formaldehyde reaction with DNA.

3.6 Future Directions

3.6.1 Study of PIC dynamics

The CLK method can be used to measure the dynamics of binding for PIC factors at promoters of interest and to study recruitment dynamics and regulation at relatively high temporal resolution. Some progress along these lines has already been made by Ramya Viswanathan and Savera Shetty, who are using the CLK assay to measure TBP, TFIIB and Pol II binding dynamics at the *GAL3* promoter. Additional measurements at different promoters would be useful to study the recruitment dynamics of TFIIB, TFIIA, TFIIH, TFIIE, and Pol II will further our understanding of PIC assembly. Although locus-specific studies are informative and will provide insight into mechanism of PIC assembly dynamics, genome wide CLK–ChIP using high throughput sequencing technology will be required to relate the range of PIC assembly/disassembly dynamics to gene expression levels, RNA synthesis precision, response to stress, TATA-containing and TATA-less sequences, histone modifications, and other regulatory elements across

genes.. In order to see the scope of application of CLK method a dynamic system should also be studied in mammalian system to prove the applicability of the method for more complex organisms such as humans and mouse.

3.6.2 Genome-wide application of CLK-ChIP

Another advantage of using a ChIP based approach method is the versatility of the application and its wide application in genome-wide studies. We have used the CLK method to study TBP binding dynamics through ChIP-Chip. We have collected 7 crosslinking time points and are in the process of developing a computational pipeline (discussed in Appendix A) to analyze the dataset. Although we carried out experiments with genomic tiling arrays for practical reasons at the time, the best way to conduct genome wide CLK experiments would be to use ChIP-seq which should yield more accurate and precise genome-wide CLK data (See Appendix A).

3.6.3 Simplified CLK

Although proven successful in providing insights about chromatin interactions, much effort goes into collecting the data. Each CLK curve involves fitting 15-18 time-points including measurements using the KinTek apparatus and strains with two different TF concentrations in biological duplicates requiring 50-60 ChIP measurements. Accounting for sample preparation time, generation of a CLK data set for one TF at a given locus required 5-6 weeks of hard work. Since the CLK method was published (Poorey et al. 2013) interest in application of the CLK method by other groups is driving the development of simpler, less labor intensive approach than that detailed in this chapter. The approach is to use an approximate CLK model (described in Eq(19)) with a subset of time points to determine a minimal number of data-points required to obtain

reasonable estimates of the kinetic parameters, which agree with our original findings. Our results show that crosslinking rates are much faster than chromatin-binding rates and the crosslinking rates did not vary much for different factors. Using measurements for time points longer than the crosslinking-limited range (typically 1-2 s), the approximate model described in Eq(19) is sufficient to estimate the dynamics for TF-DNA interactions that are relatively slow, as was observed for TBP (for example *LOS1*, *ACT1 INO1* wt, *URA1* wt,). If ChIP measurements for sub-second crosslinking times are not required, then collection of the data using KinTek apparatus will not be needed, reducing the labor required for data collection.

3.6.4 Improvement in the CLK Method

Apart from making the CLK method simpler, there is additional room for improvement to make the method more accurate.

Modeling approaches

The current CLK model is based on the assumption that the TF concentration remains constant. The experimental measurements in Figure 3-8 and 3-9 show that the concentration of free Gal4 and TBP did not change significantly overtime to introduce significant (i.e., order of magnitude) bias in estimation of kinetic parameters But there may be other TFs in different systems which could be dramatically affected. A more comprehensive CLK model could be developed for a system affected by TF concentration depletion.

Another assumption was that all of the TF binding interactions were independent events and were not dependent on any other factors. However, recruitment of TFs to

promoters can be complicated. As detailed in the introduction, recruitment of TFs to a given promoter is much more complicated. Some factors bind cooperatively including those that bind as dimers. Cooperative binding can affect the derived on-rates, off-rates and occupancy relative to binding of the same factor as a monomer. Moreover, incorporating cooperative effects in generalized CLK mathematical models would lead to different interpretations and values for the kinetic parameters. For initial study, modeling two-factor cooperative binding in a system where factors are known to cooperatively bind, for example, TBP and TFIIB binding to a promoter as TFIIB binding is dependent on TBP binding, would be an important start. One could then expand to a multicomponent system like Gal4 binding on a *GAL 1-10* promoter which requires cooperative binding of 4 Gal4 molecules.

Chapter IV

Analysis of the Sen1 Termination Pathway Using the Transcription Precision Pipeline (TraPP)

4.1 Introduction

Transcription termination is a tightly regulated process (Kuehner et al. 2011; Brow 2011). The same core complex can precisely transcribe a short ncRNA transcript and also a long Pol II transcript. The termination signal for most of the cases is carried by the nascent RNA which is recognized by termination factors, cleavage factors and also poly Adenylation (Poly (A)) factors (Zhao et al. 1999; Rosonina et al. 2006). For yeast Pol II two termination pathways have been identified, Poly (A)-dependent where most of the mRNA transcripts are terminated by the action of exonuclease Rat1 and the other is a Poly (A)-independent transcription termination pathway involving Sen1 (Kuehner 2008). Sen1, Nrd1 and Nab3 were identified as termination factors for snoRNA and snRNA in yeast by Steinmetz et al. 2001. It was found that Sen1, Nrd1 and Nab3 associate with Pol II with each other and with the Pol II CTD. Binding of Nrd1 and/or Nab3 to specific sequences in the nascent transcript results in Sen1 dependent termination. (Brow 2011; Steinmetz et al. 2006b; Ursic et al. 1997). As mentioned in Chapter I, the essential components for the Sen1-dependent termination pathway include the Sen1 helicase, Nrd1 and Nab3 RNA binding proteins, the CTD phosphatase Ssu72, and Pol II (Kuehner 2008). We studied the effects of mutations in Sen1, Nrd1, Nab3, Ssu72, Rpb11, and Hrp1 (subunit of Cleavage factor 1) on transcription both qualitatively and quantitatively.

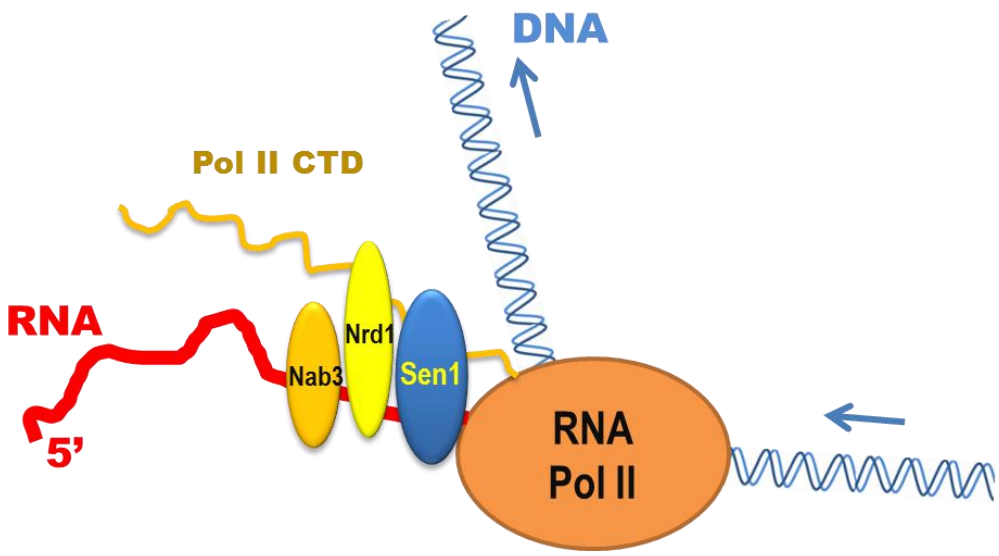


Figure 4-1: Schematic of the Sen1-Nrd1-Nab3 termination machinery scanning nascent transcript shown in red. Figure is adapted from Brow, 2011

Sen1 is presumed to be a helicase based on studies of, Sen1 homolog in *S. Pombe* which has 5'-3' helicase activity (Ursic et al. 2004). Sen1 is essential and conditional loss-of-function mutations in Sen1 result in accumulation of tRNA and rRNA precursors, as well as mis-localization of snoRNA. Such mutations are also known to cause 3' extension and transcriptional read through of some snoRNAs and some short Pol II genes (Ursic et al. 1997, 2004). The human homolog of Sen1 is Sentaxin and mutations in Sentaxin are known to cause the progressive neurological disease Ataxia Oculomotor Apraxia 2 as well as Juvenile Amjotrophic Lateral Sclerosis ALS4 (Chen et al. 2006, 2004). Nrd1 interacts with the large subunit of Pol II at phosphorylated Ser 5 to direct termination of the transcript for non poly(A) transcripts (Steinmetz et al. 2001; Vasiljeva et al. 2008). Hrp1 is a subunit of cleavage factor 1 which is a five subunit complex required for the cleavage and polyadenylation of the pre-mRNA 3' end. It binds to the poly (A) signal sequence (Kessler et al. 1997). Ssu72 is a phosphatase and transcription/RNA processing factor; it associates with Pta1p and removes Ser 5 and Ser 7 phosphorylation of the RNA Pol II CTD (Krishnamurthy et al. 2004). It also associates with TFIIB to ensure accurate start site selection (Sun and Hampsey 1996). Mutation of Ssu72 affects start site selection and causes transcription read through (Sun and Hampsey 1996; Krishnamurthy et al. 2004). Along with this, mutations in the Pol II subunit Rbp11 cause read-through transcription in vivo.(Steinmetz et al. 2006b). It is evident from the wide spread roles played by these factors that their role is not limited to termination of ncRNAs. And study of transcription when these factors are mutated would reveal their involvement in other fundamental pathways to modulate transcription. The study

presented in this chapter would assess the effect of mutation of these factors on ncRNA coding genes and also genes transcribed by Pol II.

4.2 Materials and Methods

4.2.1 Yeast strains

Strains for RNA processing analysis were obtained from David Brow and were previously described in: (*sen1-E1597K*, Steinmetz and Brow 1996; *nrd1-V368G*, Steinmetz and Brow 1996; *nab3-11(F371L, P374T)*, Kuehner 2008; *ssu72-G338*, Steinmetz and Brow 2003; *hrp1-L205S*, Kuehner 2008; *rpb11-E108G*, Eric J Steinmetz et al. 2006.

4.2.2 Expression Analysis

Total RNA was isolated and hybridized to S. cerevisiae Tiling 1.0R Arrays (Affymetrix, Inc.) and raw data were generated by the Microarray Core Facility at UVA using the method described in chapter II for WT, *sen1-E1597K*, *nrd1-V368G*, *nab3-11*, *ssu72-G338*, *hrp1-L205S* and *rpb11-E108G* RNA analyses were performed using two independent biological replicates for each. Cell growth, RNA isolation and library preparation were performed by Melissa Wells Carver . Estimates of total RNA levels were made from the raw array data (CEL files) by first quantile normalizing all replicate arrays and scaling the data to a target median intensity of 100. We applied the Wilcoxon Signed-Rank test to the normalized log2(max(PMi-MMi,1)) values whose genomic coordinate ‘i’ fell within a 100 bp (i.e., a length much smaller than the typical ORF) sliding window to calculate the log transformed probability (-10log10(p-value)) that the

RNA was detected above noise levels. The associated Hodges-Leman estimator was used to estimate RNA levels. Differential RNA levels were estimated using the same procedure described in Chapter II including a window size of 500 bp with mutant RNA samples as treatment and WT RNA samples as control using Tiling Assesment Software TAS version 1.1.

4.2.3 Multi Pass TraPP Pipeline

We applied the TraPP pipeline described in Chapter II with 0.3, 1.3, 2.3, 3.3 and 4.3 cutoffs of the differential RNA signal ($\log_2$(Mut/WT)) data generated using a 500 bp sliding window. By comparing these segments to annotations, we identified 5' and 3' transcriptional length changes in each mutant relative to WT RNA (Figure 4-2). We used the annotations provided by Saccharomyces Genome Database from the reference genome released in 2004. These length changes were either positive or negative and represent different transcriptional defects as described in Chapter II. Cases where the significant differential expression segment ran across to two or more annotated genes were characterized separately on gene-by-gene basis. Putative length changes that did not satisfy all the criteria below were filtered out:

- The overlap between annotations and the significant differential signal segments was at least 100 bp long.
- The length of the defect computed was > 150 bp long.
- Median signal in the defect was significantly different from the baseline differential expression value as described in Figure 4-2. In the case of extensions where the segment fell beyond the boundaries of the annotation, the median signal in the extended region, S_{ext}, was greater than 0.6 to make a length change call. In

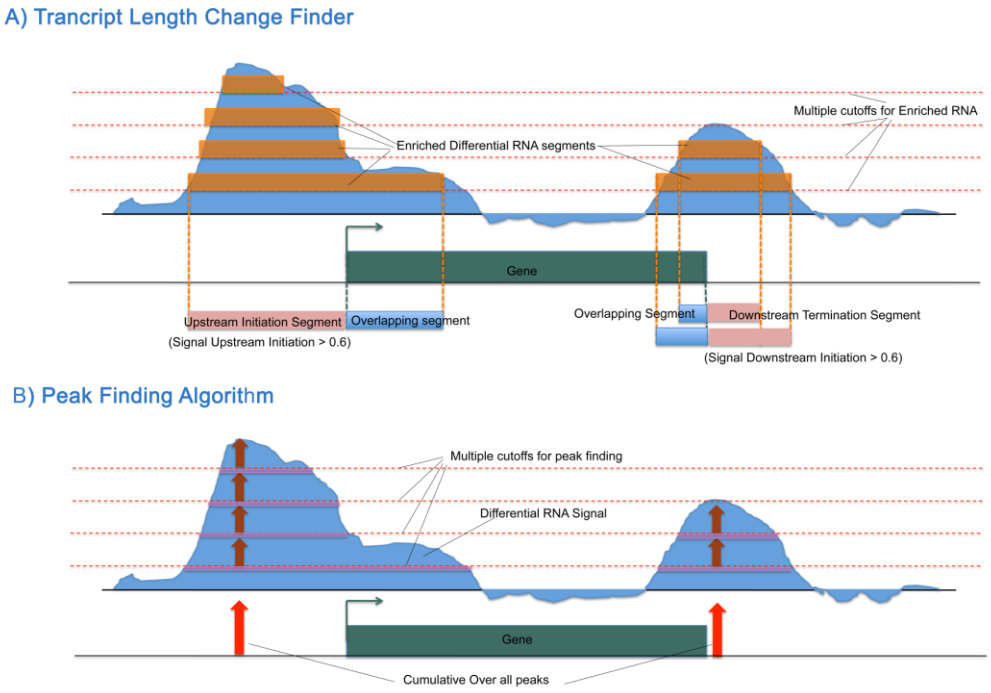


Figure 4-2 Analysis methods used for characterizing the RNA defects as described in the text.

the case of truncations, we defined two quantities: (1) $S_{overlap}$ as the median of the differential signal in the region where the differential segment and the annotation overlapped and (2) S_{int} as the median signal in the region internal to the gene that did not overlap the differential expression segment. For a truncation to be called, we required $S_{overlap} = S_{int} > 0.6$.

4.2.4 Cross Correlation Analysis

This analysis was used to measure the similarities of the differential RNA profiles of two datasets which was sensitive to transcriptional precision defects for all the genes or a given gene list. The first step of this analysis was to select an annotation class for comparing the differential signal. Genomic coordinates of the selected annotations are supplied as input to the program. For this current study, two sets of coordinates were used: (1) gene boundaries and (2) extended gene boundaries which included 150 bp upstream of the TSS and 100 bp downstream of the transcription stop site for each gene in the yeast genome. This was done to ensure we took into account the 5' and 3' RNA precision defects. We used these coordinates to select differential RNA data from two datasets (e.g. Nrd1 and Nab3) and calculated the cross-correlation coefficient using the Matlab function "xcorr" over the normalized differential signal of these datasets. Using the normalized differential signal was crucial so that the overall differential signals did not influence the cross-correlation coefficient.

4.2.5 Clustering Analysis

We used the Matlab function "clustergram" to perform clustering analysis for gene expression and used Pearson, Mahalonobis and Spearman correlation coefficients as metrics to calculate the distances between genes and mutants in the data. We also used

AutoSOME which uses a powerful unsupervised computational method for identifying discrete and fuzzy clusters of diverse geometries from large datasets without requiring an estimate of the number of clusters as input. We divided the gene body into five equal lengths and calculated the median differential signal within these boundaries. Which allowed clustering and visualization of shared transcriptional defects across genes and mutants. We applied AutoSOME to cluster this dataset in order to produce lists of genes with similar transcriptional defects across mutants. The resulting clustered profiles were viewed using the java-Tree View software.

4.3 Results

We compared overall RNA expression levels from *sen1-E1597K, nrd1-V368G, nab3-11, ssu72-G338, hrp1-L205S,* and *rpb11-E108G* with respect to their WT strains using Affymetrix yeast genomic tiling arrays at 5bp resolution as described in the Chapter II Materials and Methods section. Two different WT datasets were used to account for different mating types in the mutant strains. Consistent with the method used in Chapter II, the total RNA profile was calculated by averaging the signal over 100 bp windows and the differential profiles were calculated using a 500 bp window. Figure 4-5 shows a screenshot of genomic region of chromosome 7 showing total RNA from WT and *sen1* as well as the differential RNA signal which effictively is $\log_2$(Mut/WT). Although the Sen1 pathway is believed to target genes involving poly (A)-independent termination and 3'end formation, we found that a large portion of Pol II genes show aberrant differential RNA precision defects at both 5' and 3' regions We estimated overall gene expression changes from the tiling array data by calculating the median $\log_2$(Mut/WT) signal within the gene body. In the *sen1* strain, it was observed that the

Table 4- 1: GO term and pathway results from DAVID

GO Terms - Sen1 significant differentially expressed genes

GO Term	Count	%	PValue
maltose metabolic process	12	1.31291	6.07E-11
sporulation	53	5.798687	2.35E-10
sporulation resulting in formation of a cellular spore	53	5.798687	2.35E-10
ascospore formation	32	3.501094	3.20E-08
sexual sporulation	32	3.501094	3.20E-08
sexual sporulation resulting in formation of a cellular sp	32	3.501094	3.20E-08
ascospore wall biogenesis	19	2.078775	3.20E-08
spore wall assembly	19	2.078775	3.20E-08
ascospore wall assembly	19	2.078775	3.20E-08
spore wall biogenesis	19	2.078775	3.20E-08
cell wall assembly	19	2.078775	4.87E-08
fungal-type cell wall biogenesis	20	2.188184	1.64E-07
disaccharide metabolic process	13	1.422319	4.75E-07
M phase of meiotic cell cycle	40	4.376368	1.03E-06
meiosis	40	4.376368	1.03E-06
reproductive developmental process	35	3.829322	1.44E-06
meiosis I	24	2.625821	1.65E-06
meiotic cell cycle	40	4.376368	2.41E-06
response to toxin	17	1.859956	2.52E-06
reproductive process in single-celled organism	36	3.938731	1.68E-05
hexose metabolic process	28	3.063457	2.20E-05
monosaccharide metabolic process	30	3.282276	2.32E-05
siderophore transport	7	0.765864	4.86E-05
cell wall biogenesis	22	2.407002	6.76E-05
oligosaccharide metabolic process	14	1.531729	6.83E-05
vacuolar protein catabolic process	26	2.844639	8.22E-05
reproductive cellular process	42	4.595186	1.77E-04
synapsis	8	0.875274	2.28E-04
chromosome organization involved in meiosis	8	0.875274	2.28E-04
response to temperature stimulus	38	4.157549	2.97E-04
energy reserve metabolic process	13	1.422319	3.89E-04
reciprocal meiotic recombination	14	1.531729	4.00E-04
regulation of glucose metabolic process	10	1.094092	8.18E-04
cellular amide metabolic process	16	1.750547	9.43E-04
sulfur metabolic process	21	2.297593	0.001025
glucose metabolic process	21	2.297593	0.001025
reproduction of a single-celled organism	37	4.04814	0.00103
iron assimilation by chelation and transport	5	0.547046	0.001062
siderophore-iron transport	5	0.547046	0.001062
iron chelate transport	5	0.547046	0.001062
glycogen metabolic process	10	1.094092	0.001424
sexual reproduction	43	4.704595	0.001455
carbohydrate transport	12	1.31291	0.001502
cellular response to heat	31	3.391685	0.001754
regulation of cellular carbohydrate metabolic process	10	1.094092	0.00184
autophagy	26	2.844639	0.00197

Kegg Path way analysis - Sen1 significant differentially expressed genes

Term	Count	%	PValue
Starch and sucrose metabolism	15	1.641138	3.77E-06
Galactose metabolism	9	0.984683	4.78E-04
Meiosis	23	2.516411	0.001657
Fructose and mannose metabolism	7	0.765864	0.028111
Phenylalanine metabolism	4	0.437637	0.074739
Glutathione metabolism	5	0.547046	0.092725

Sen1 very high differential expression (> 3 fold change)

Term	Count	%	PValue
polyol biosynthetic process	2	3.174603	0.054138
meiosis I	3	4.761905	0.091861

expression changes were relatively large compared to other mutants studied. Table 4.1 shows the GO term and KEGG pathway analysis and the enriched terms and pathways highlighted by DAVID. Most of the enriched terms are related to meiosis and reproduction with very significant p-values. This strongly suggests that Sen1 has a function in suppressing meiosis related genes in vegetative state. Furthermore, when we focus on the genes which are highly differentially expressed, they show enriched for meiosis pathway (Table 4.1) as well. And 21 out of the top 26 higly expressed genes were caused by read through from snoRNA genes. This behavior of read-through transcription was also observed from the average plot of total RNA in WT and *sen1* cells, as well as from average profile of differential RNA for snoRNA genes shown in Figure 4-3. To capture these readthrough events TAS was used to calculate the differential RNA signal using a 50 bp smoothing window so that these relatively shorter genes could be studied. The average profile shows clear signs of read-through transcription in the mutant (Figure 4-3 A).

In order to capture and quantify these read-through events from snoRNA genes we modified the Multi-Pass TraPP algorithm to be applied for snoRNAs, which are relatively short. First, we used the a smaller smoothing window of 50 bp to estimate differential RNA signal. The RNA precision defects for snoRNAs are shown in Table 4-2. The defects were classified into upstream initiation (UI), premature termination (PT), downstream termination (DT) and downstream initiation (DI). As expected, the dominant class of RNA precision defects was downstream termination (DT) with more than 60% of snoRNA genes affected. Further analysis showed that all 41 cases exhibitread-through transcription and led to RNAs that apparently ran into 77 downstream Pol II genes.

Figure 4-4 shows two examples of snoRNA read-through transcripts which runs through multiple Pol II genes. Surprisingly ~40% of the snoRNA genes were not affected by the mutation in Sen1, which suggests the possibility of another termination pathway for these genes.

Average profiles of total RNA and differential RNA in *sen1* versus WT cells at Pol II genes suggest that both quantity and precision of RNA at Pol II genes is widely affected (Figure 4.3 B). Both 5' and 3' regions of Pol II genes show in general an accumulation of differential RNA, which suggests prominent upstream initiation and downstream termination defects. To further investigate this phenomenon we used the Multi-pass TraPP method for all the Pol II genes. Table 4.3 shows the number of significant transcript precision defects detected. We also applied the peak calling algorithm to the differential RNA signal data to classify the 5' and 3' accumulation of differential RNA captured in the average plots in Figure 4-3. Figure 4-5 shows an example of the accumulation of the differential RNA signal and how well these effects can be captured and classified by peak calling. Table 4.3 B shows the number of genes associated with 5' or 3' differential RNA peaks. Associations were made if the peak was within 150 bp of the TSS and 100 bp downstream of the transcription stop site. Figure 4-6 shows an example of all the methods used to classify RNA length defects for the *sen1*data. There were many orphan differential RNA peaks which were not associated with any genes. This is surprising given the gene dense nature of yeast genome. However, the Sen1 pathway is involved in degrading SUTs (Kuehner 2008)and further investigation is needed to determine if there is an association between

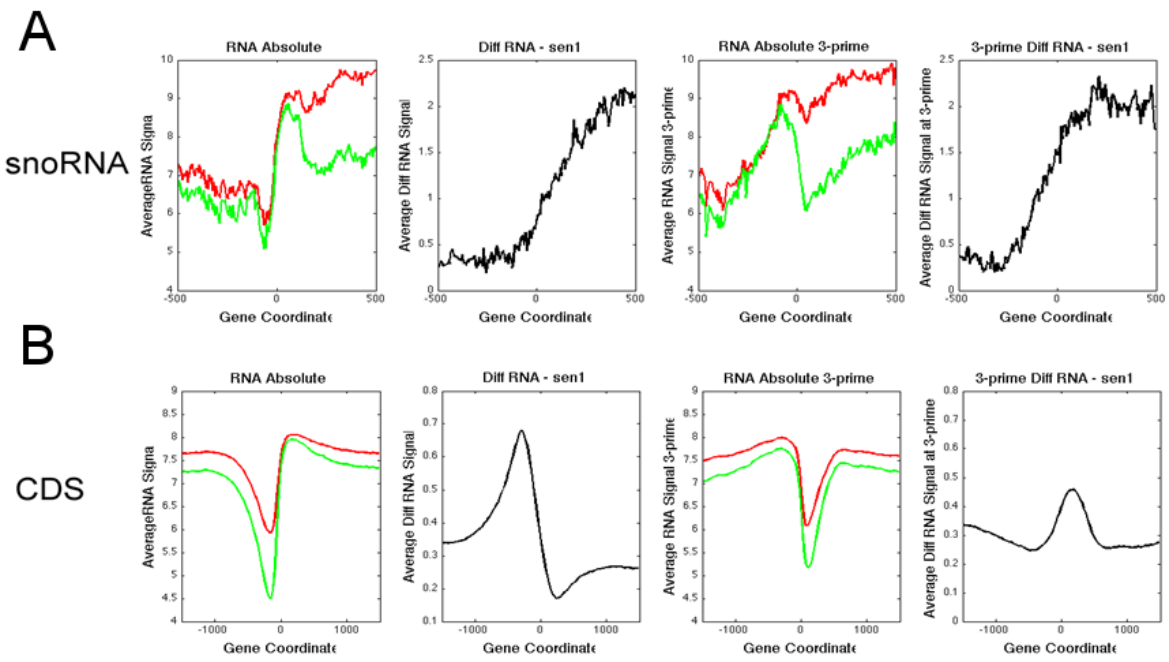


Figure 4-3: Average profiles of total RNA and defferential RNA signal centered over gene start and Gene end for snoRNA and CDSs.

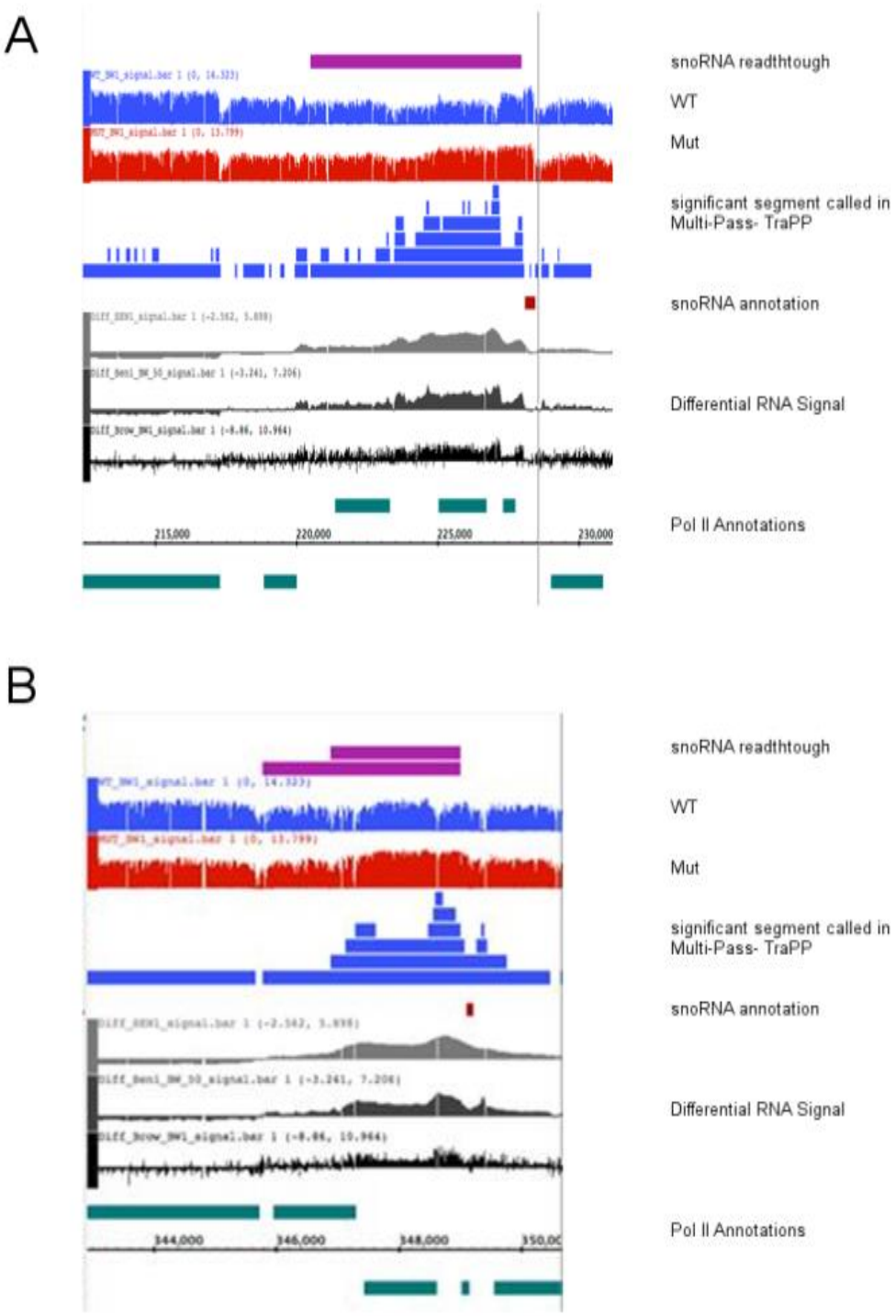


Figure 4-4: Example of snoRNA read through downstream Pol II genes.

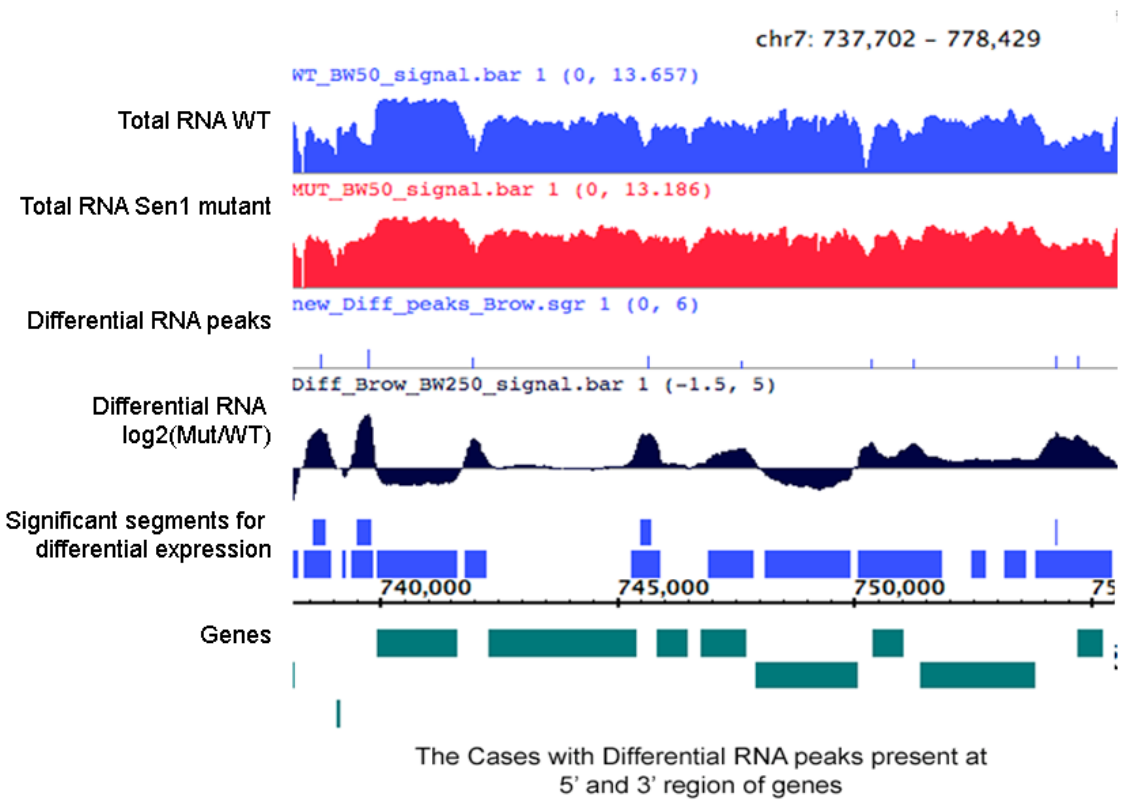


Figure 4-5: Example of differential RNA signal (differential RNA peaks) at 5' and 3' end of the genes in *sen1-E1597K*.

Table 4- 2: Classification of RNA precision defects in snoRNA.

Sen1 - snoRNA	**Total Genes considered 68**
Upstream initiation (UI)	13
Downstream termination (DT)	41
Late Initiation (LI)	10
Premature termination (PT)	2

Table 4- 3: Classification of RNA precision defects in Pol II genes for the *sen1-E1597K* dataset

A

Type: Sen1 Pol II Genes	Count
Upstream Initiation	1236
Premature Termination	196
Downstream Initiation	184
Downstream Termination	1185
Total Length Changes	3084
Total number of genes involved in Length Changes	1931

B

Type: Sen1 Pol II Genes	Count
5' Associated Peaks (Promoters)	1921
3' Associated Peaks	1738
Overlapping Promoter Peaks(1921) and Upstream Initiation(1236)	647
Overlapping Promoter Peaks (1738) and Downstream termination (1185)	635

these events and the occurrence of SUTs. Further wet-bench experimentation is also needed of the 5’ and 3’ peaks to understand the origins of the signal. Table 4-6 shows the collection of enriched GO terms and pathways for the genes which have significant length changes. Similar to the results of the snoRNA-focused analysis described above, meiosis is highlighted in a number of significant terms.

The TraPP pipeline as described in Chapter II was also used to analyze and classify the RNA precision defects in other components of the Sen1 termination pathway (Nrd1, Nab3, Ssu72, Rpb11 and Hrp1). The significant RNA length changes detected are summarized in Table 4-4 A. Nab3, Hrp1 and Nrd1 mutants show a premature termination phenotype. A mutation in Hrp1 gave rise to a large number of premature termination defects, as expected, and these defects were mostly differentially down premature termination events, whereas Nab3 and Nrd1 show a mixture of both differentially down and differentially up premature terminations. Ssu72 and Rpb11 mutants show downstream termination and upstream initiation effects to be the dominant phenotypic effects, similar to what was observed in *sen1* cells. We also calculated the overlap of the genes sets involved in significant length changes and the overlap numbers are enriched over what was expected from random chance, which is expected since all these factors are part of the same pathway (Table 4-4 B).

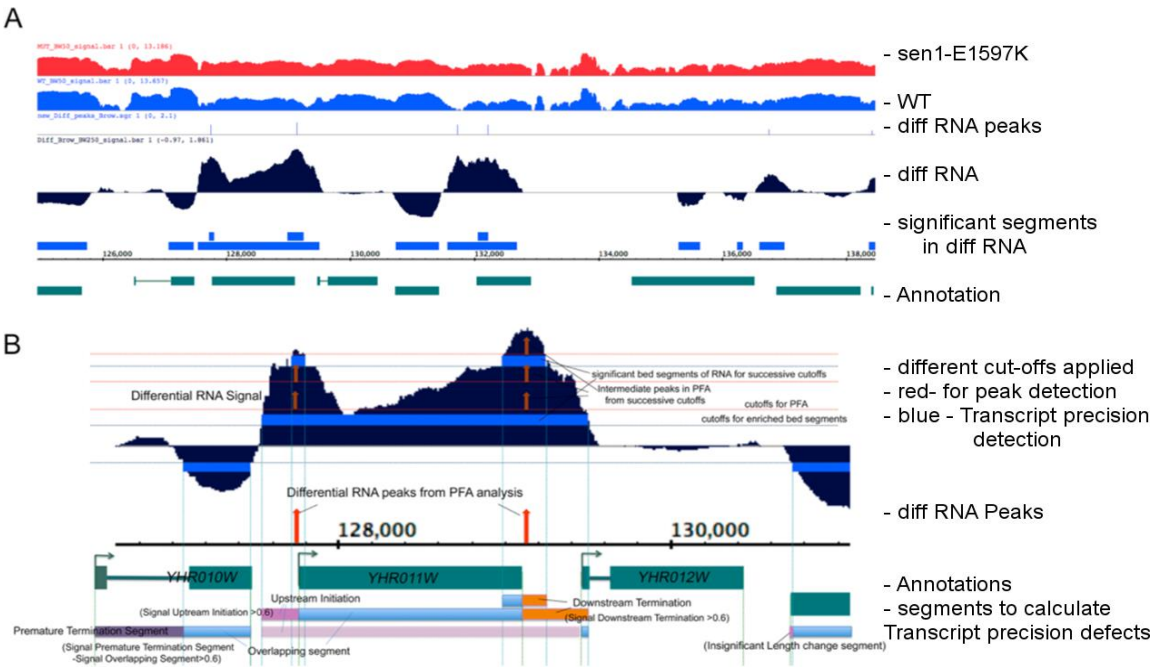


Figure 4-6: Example of computational methods applied to *sen1* mutant dataset. A) Integrated Genome Browser view of a segment of chromosome 2. B) Example of Multi-pass TraPP and peak finding algorithm applicatied to *YHR011W* , showing red peaks as final location calls and transcription precision defects called by pink and orange bands at the bottom for 5' and 3' changes respectively.

Table 4- 4: Summary of transcript length changes for Pol II genes

A) Number of instances of different classes of transcription precision defects

B) Overlap of genes involved in different datasets

A Classification of the transcript length changes

Type:	Nrd1 -α	Sen1	Nab3 -α	Hrp1 -α	Rpb11	Ssu72 -α
Upstream Initiation	258	1236	316	150	698	529
Premature Termination	286	196	463	1619	163	104
Downstream Initiation	248	184	206	272	169	501
Downstream Termination	175	1185	192	99	700	1087

B Overlap Analysis for Length Changes

	Nrd1	Nab3	Sen1	Rpb11	Ssu72	Hrp1	
Nrd1	803	347	470	293	387	371	Nrd1
Nab3		980	558	346	437	423	Nab3
Sen1			1931	634	928	707	Sen1
Rpb11				1302	524	488	Rpb11
Ssu72					1751	747	Ssu72
Hrp1						1953	Hrp1

Table 4- 5: Cross-correlation coefficients A) for Pol II genes B) for ncRNA genes.

For Pol II genes the cross-correlation was computed using extended gene boundaries which included 150 bp upstream of the TSS and 100 bp downstream of the transcription stop site for each gene. And for the ncRNA genes only gene boundaries are used to compute the cross-correlation coefficient.

A

	Nrd1	Nab3	Sen1	Rpb11	Ssu72	Hrp11	Set2	
Nrd1	1	**0.44**	**0.29**	-0.04	**0.33**	**0.26**	0	Nrd1
Nab3		1	**0.29**	0	0.2	**0.38**	0.06	Nab3
Sen1			1	**0.29**	**0.32**	0.19	0.11	Sen1
Rpb11				1	0.09	0.06	0.16	Rpb11
Ssu72					1	-0.04	0.06	Ssu72
Hrp11						1	0.06	Hrp11
Set2							1	Set2

B

	Nrd1	Nab3	Sen1	Rpb11	Ssu72	Hrp11	Set2	
Nrd1	1	**0.49**	**0.26**	-0.2	**0.4**	0.08	0.2	Nrd1
Nab3		1	0.06	-0.48	0.62	**0.32**	0.47	Nab3
Sen1			1	**0.23**	-0.02	0.05	-0.09	Sen1
Rpb11				1	-0.34	-0.16	-0.38	Rpb11
Ssu72					1	0.15	0.49	Ssu72
Hrp11						1	0.29	Hrp11
Set2							1	Set2

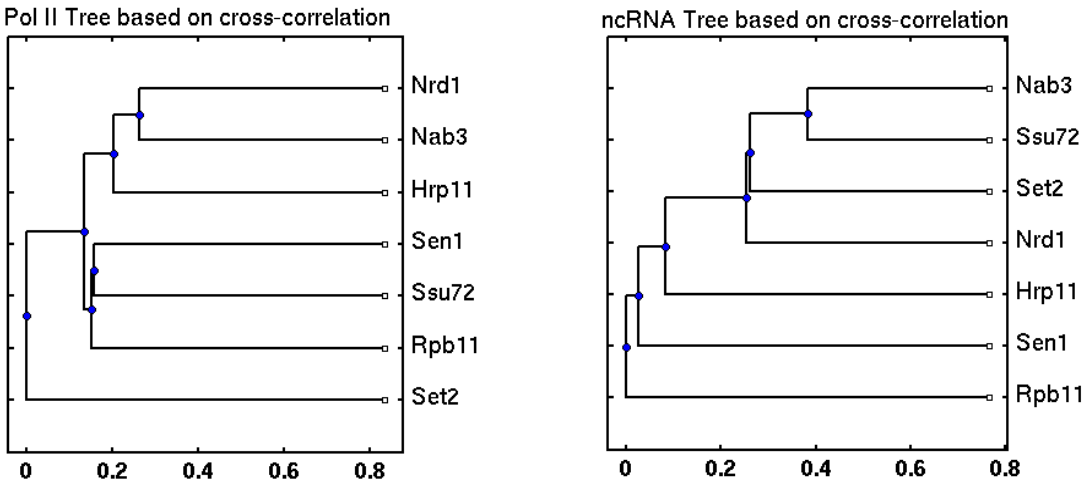


Figure 4-7 Tree based cross-correlation coefficient computed in the Table 4-5 as a distance matrix. For the Pol II genes the cross-correlation coefficient was computed by comparing the differential signal within the extended gene boundaries which included 150 bp upstream of the TSS and 100 bp downstream of the transcription stop site for each gene. The median cross-correlation coefficients were used as a distance matrix to form a tree based on similarities in transcription defects. Similar method was applied for ncRNA genes but only gene boundaries were considered for comparison

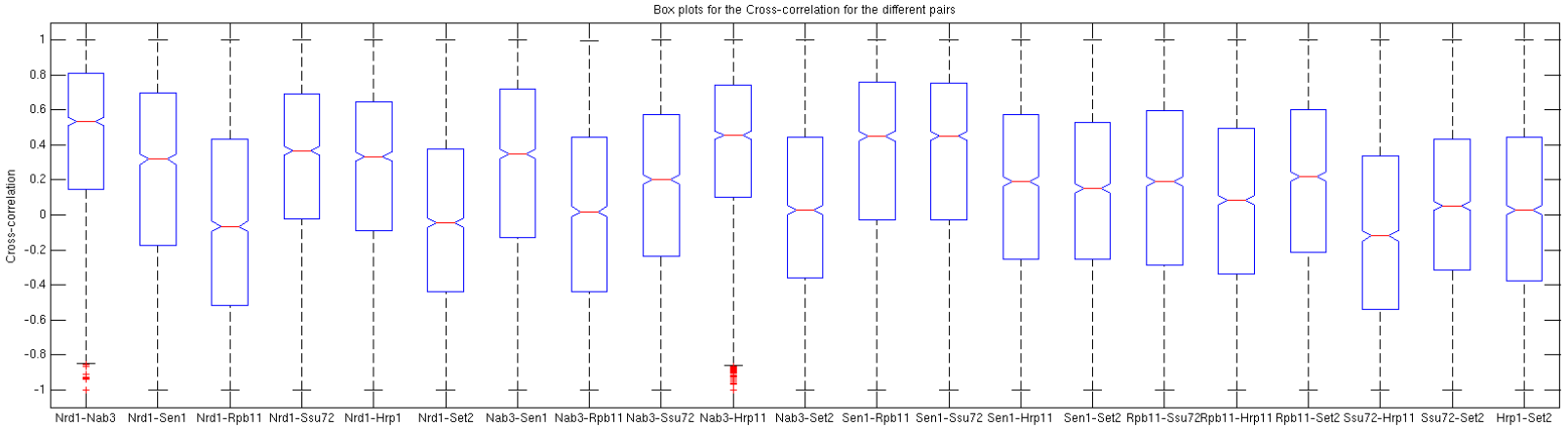


Figure 4-8: Distribution of Cross-correlation-coefficients for all the pairs of mutants in Sen1 termination pathway

Table 4- 6: GO term and pathway enriched terms for genes involved in transcription length changes in *sen1* cells.

GO Term - Genes with significant Sen1 Length changes

GO-Term	Count	%	PValue
cellular response to heat	86	4.479167	2.73E-08
response to temperature stimulus	99	5.15625	2.80E-08
meiosis I	48	2.5	3.47E-08
M phase of meiotic cell cycle	85	4.427083	1.60E-07
meiosis	85	4.427083	1.60E-07
meiotic cell cycle	86	4.479167	4.03E-07
reciprocal meiotic recombination	31	1.614583	5.99E-07
sporulation resulting in formation of a cellular spore	95	4.947917	1.26E-06
sporulation	95	4.947917	1.26E-06
response to heat	89	4.635417	1.34E-06
cellular response to stress	207	10.78125	1.46E-06
autophagy	64	3.333333	2.38E-05
response to toxin	27	1.40625	4.37E-05
DNA recombination	75	3.90625	5.39E-05
ascospore formation	52	2.708333	6.56E-05
sexual sporulation resulting in formation of a cellular s	52	2.708333	6.56E-05
sexual sporulation	52	2.708333	6.56E-05
reproductive developmental process	65	3.385417	1.10E-04
M phase	125	6.510417	1.38E-04
response to abiotic stimulus	126	6.5625	3.21E-04
vacuolar protein catabolic process	51	2.65625	4.00E-04
disaccharide metabolic process	16	0.833333	4.33E-04
regulation of cellular catabolic process	23	1.197917	4.61E-04
maltose metabolic process	10	0.520833	9.77E-04
iron ion transport	19	0.989583	0.001172
oligosaccharide metabolic process	22	1.145833	0.001372
regulation of catabolic process	29	1.510417	0.002092
reproductive process in single-celled organism	68	3.541667	0.002146
regulation of glucose metabolic process	16	0.833333	0.003786
spore wall biogenesis	22	1.145833	0.004156
spore wall assembly	22	1.145833	0.004156
ascospore wall assembly	22	1.145833	0.004156
ascospore wall biogenesis	22	1.145833	0.004156
reproductive cellular process	88	4.583333	0.004202
transition metal ion transport	27	1.40625	0.004443
cell wall assembly	22	1.145833	0.005771
recombinational repair	18	0.9375	0.007704
ascospore-type prospore membrane formation	7	0.364583	0.007837
membrane biogenesis	7	0.364583	0.007837
sexual reproduction	97	5.052083	0.007917

Kegg Pathway analysis - Genes with significant Sen1 Length changes

Term	Count	%	PValue
Meiosis	48	2.5	2.34E-04
Starch and sucrose metabolism	19	0.989583	0.001697
Butanoate metabolism	10	0.520833	0.056803
Galactose metabolism	10	0.520833	0.056803
Regulation of autophagy	8	0.416667	0.086534

In the six datasets we analyzed for RNA lengths, change there were two broad classes of factors based on the precision defects that we identified. 1.) Mutations in Nrd1, Nab3, and Hrp1 caused premature terminations and 2.) mutations in Sen1, Rpb11 and Ssu72 caused 5' and 3' RNA extension defects (UI and DT). As discussed above, we already knew that the genes with associated RNA length changes in each mutant significantly overlap, However do they share similar specific precision defects? To investigate this, we calculated the cross-correlation coefficients using differential RNA profiles for all the pairs of mutants at every gene locus. In two different analyses, we used differential RNA profiles from (1) the gene body and (2) the gene body, promoter and 3' region to compute the cross-correlation coefficient for Pol II genes and ncRNA genes. Table 4-5 shows the median cross-correlation coefficient between mutants where we analyzed 5' and 3' extended boundaries for Pol II genes and only the gene body for ncRNA genes. Because ncRNAs are in general short and the value of cross-correlation coefficient would be heavily biased for the signal in the regulatory region rather than the whole body we used only the gene body for analysis of ncRNA genes. These coefficients quantify the similarity of the differential RNA profile between each pair of mutants (cross-correlation coefficient nearing 1 corresponds to a highly correlated profile and cross-correlation coefficient nearing -1 corresponds to a highly anti-correlated profile). We can use these coefficients across mutants and form a distance matrix in order to generate a tree based on similarities of transcriptional defects. Figure 4-7 shows trees made using the median coefficients calculated for Pol II genes. We also normalized the differential expression profile to vary from -1 to 1 for this analysis so the overall expression levels do not drive the relationship between mutants. This was necessary due

to higher levels of expression levels in certain datasets (e.g. *sen1*). F). For this analysis we also only included the union of genes, which were involved in RNA precision defects across mutants. By doing this we ensured that differential expression alone does not play a dominant role in determining the overall relationships. Figure 4-8 show boxplots of the cross-correlation coefficients calculated for Pol II genes. In this case we found that the termination factors can be subdivided into two classes. Nrd1 and Nab3 mutant defects are quite closely related across genes, which is consistent with their known biochemical relationship. We also included the *set2Δ* dataset described earlier as a control, and its relationship with the other datasets was observed to be the weakest, as expected.

Next we applied clustering techniques over gene expression and differential gene expression profiles to determine if overall levels of expression yielded similar associations across mutants as those that resulted from the analysis of defects in transcriptional precision. Two-way clustering was done using Matlab for the gene expression data and the resulting dendrogram is shown in Figure 4-9. The pattern of the tree emerging from the two-way clustering of the Pol II genes is different than that observed for the cross-correlation analysis. This is because the large gene expression profiles of the factors drive the clustering approach. Sen1 seems to be more closely related to Nrd1 and Nab3 based on overall expression but it seems to play a different role in modulating termination, hence yielding a different RNA profile. This defect was better characterized by cross-correlation analysis above and beyond the results of the gene expression.

We next computed the median differential RNA signal in seven regions for every gene as described in the material and method section. We used Autosome to cluster these

seven regions of genes in all the datasets including *set2Δ* as a control, resulting in 49 columns to cluster. This was done to ensure that clustering approach can capture the structure of genes associated with precision defects and not just gene expression patterns. Sixty clusters were obtained from this analysis with thirty clusters containing more than 50 genes. Overall trends across the clusters were difficult to determine but some of the clusters showed very distinct patterns. For example, Cluster 5 (Figure 4-10) contains genes that predominantly show premature termination defects. The 5' and 3' differential RNA accumulation in *sen1*data is evident in this cluster. *nrd1* and *nab3* show mixture of both premature termination and downstream initiations, *rpb11*is differentially down-regulated and *hrp1*shows premature termination in all the displayed genes. Furthermore the control dataset, *set2Δ,* show no consistency in the pattern.

Figure 4-11 shows an Autosome clustering plot of the genes showing premature termination in *hrp1* cells. It has been observed that Hrp1 functions independent of the Sen1 pathway (Kuehner 2008). To confirm this, we made average plots of differential RNA signal over the 5' end of genes involved in length changes where we included (Figure 4-12 black curves) and excluded Sen1-affected genes (Figure 4-12 red curves) for the purpose of comparison. All the datasets except Hrp1 show a difference in the average differential profile (Figure 4-12). This suggest that Hrp1 is part of a mutually exclusive termination pathway then Sen1.

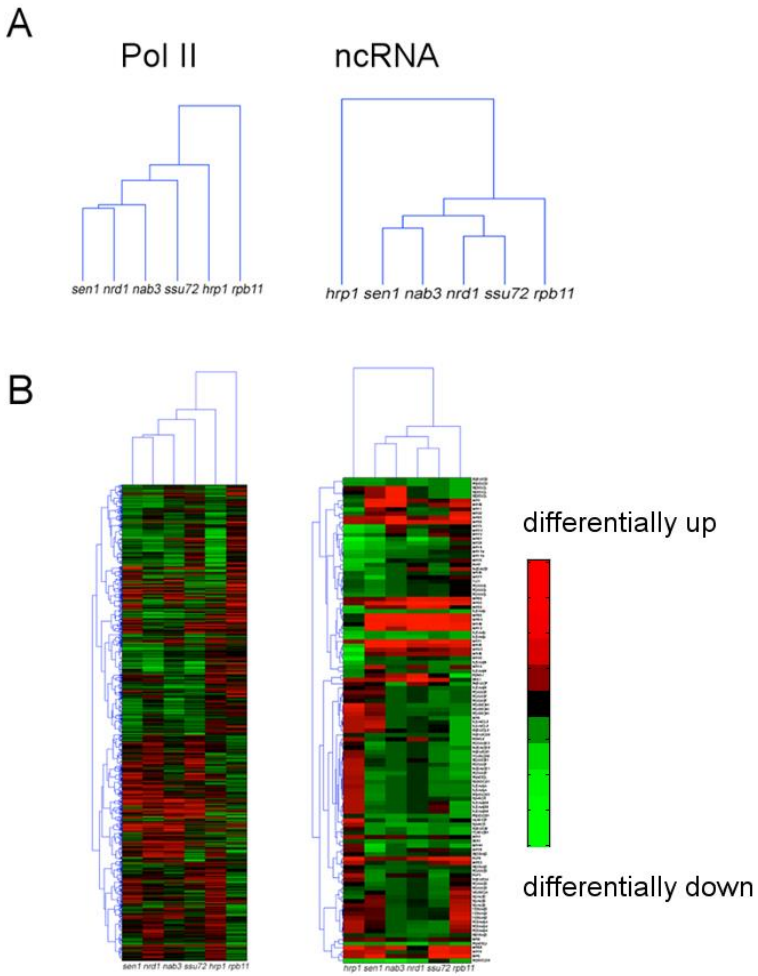


Figure 4-9: Dendrogram obtained from two-way clustering of gene expression for Sen1, Nrd1, Nab3, Ssu72, Rpb11 and Hrp1 datasets .

A) The tree showsthe overall relationships between the datasets with distances calculated by clustering. B) Heatmaps of gene expression with red – downregulation and green – upregulation.

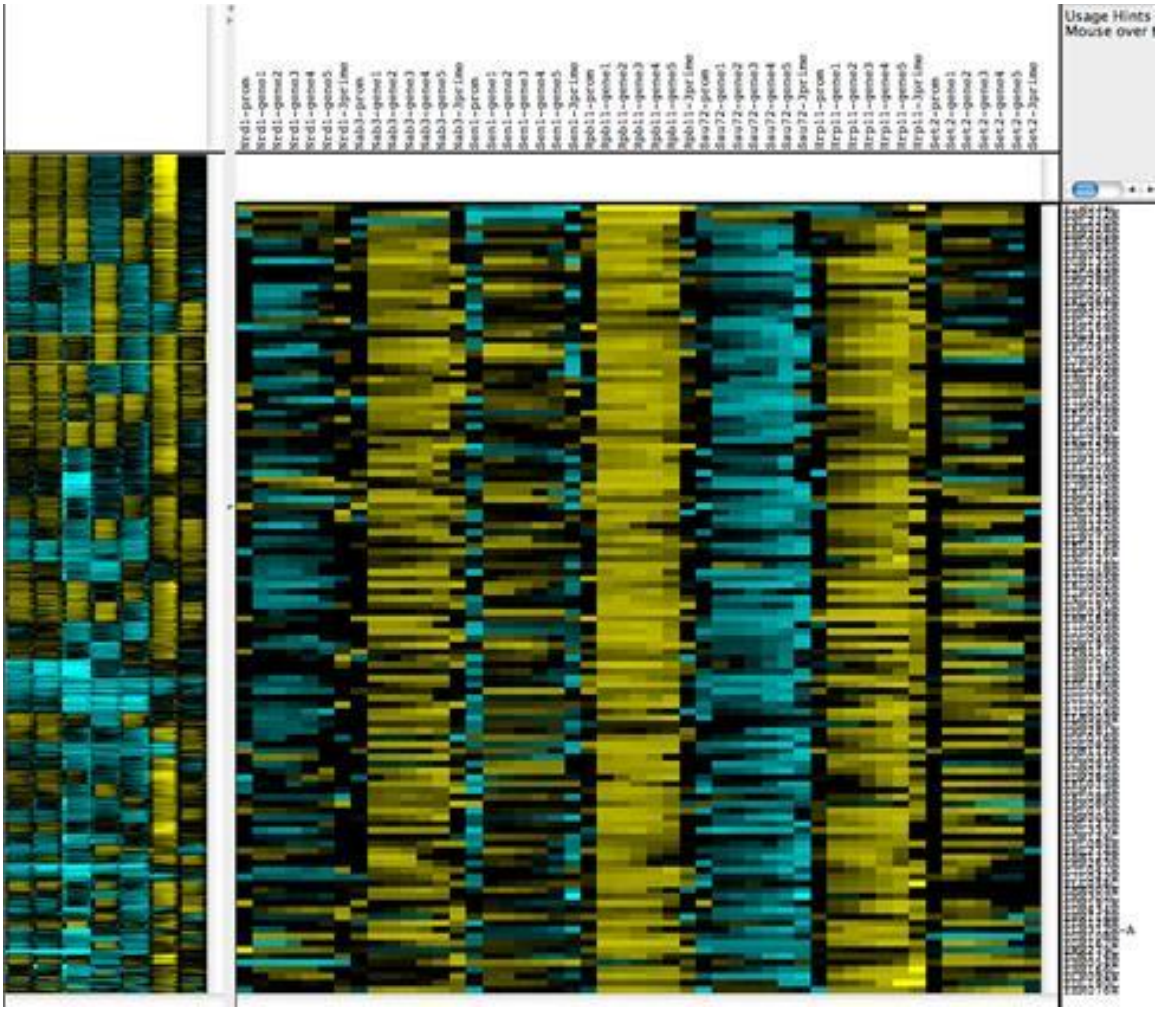


Figure 4-10: Example of AutoSOME output.

The left panel shows a segment of the overall AutoSOME clustering results. The right panel shows Cluster 5 from the AutoSOME output with blue representing differentially up signal and yellow showing differentially down signal. The order of columns from left to right is Nrd1, Nab3, Sen1, Rpb11, Ssu72, Hrp1, and Set2 dataset.

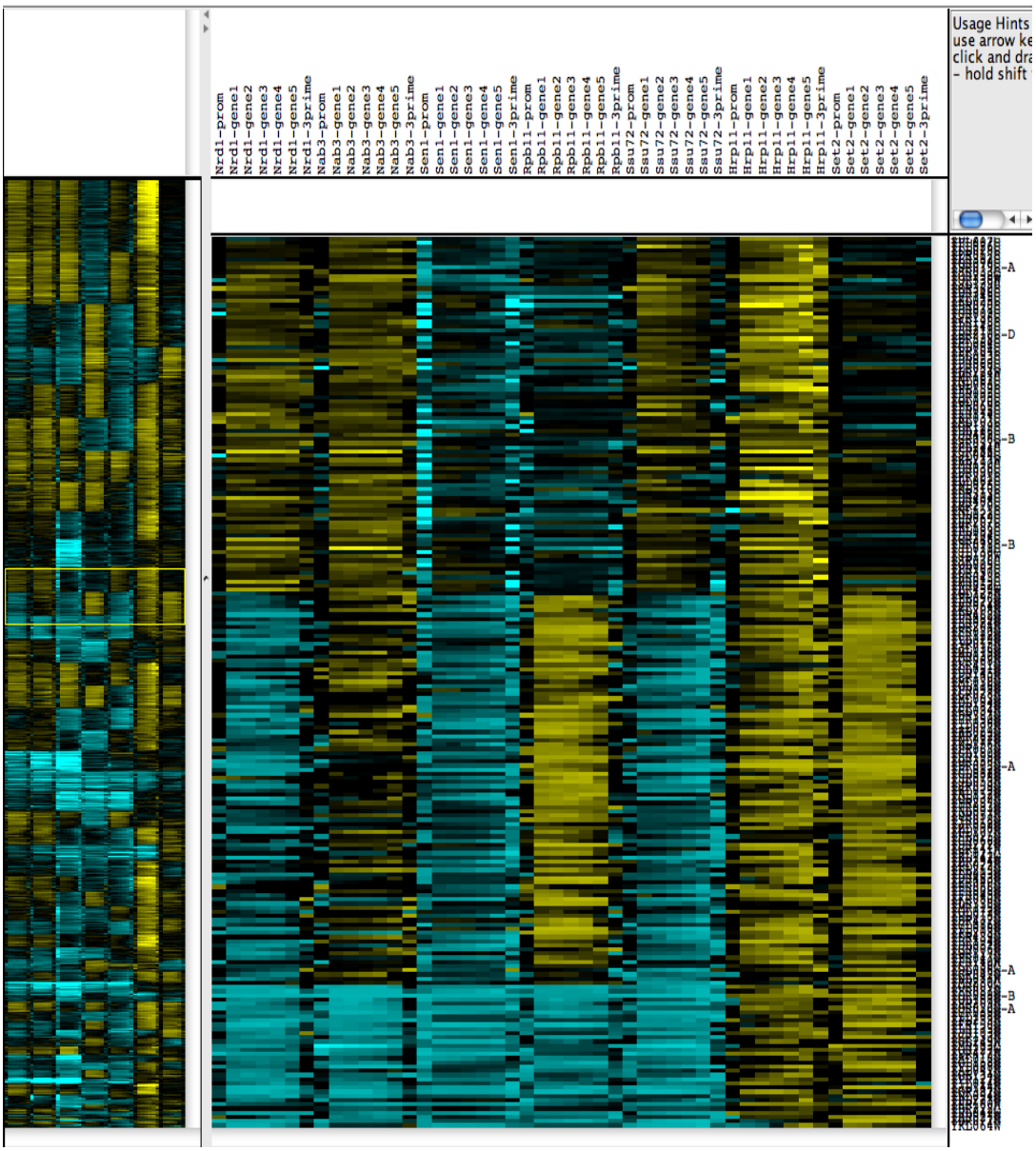


Figure 4-11: Example clusters with Hrp1 showing premature termination.

Java tree view of Clusters 11-13 which show premature termination to be the predominant class of defects in *hrp1* cells. Blue color represents positive differential RNA signal and yellow representing differential negative signal. Panel on the left showing collection of all the clusters and panel on the right showing the genes corresponding to the selected view.

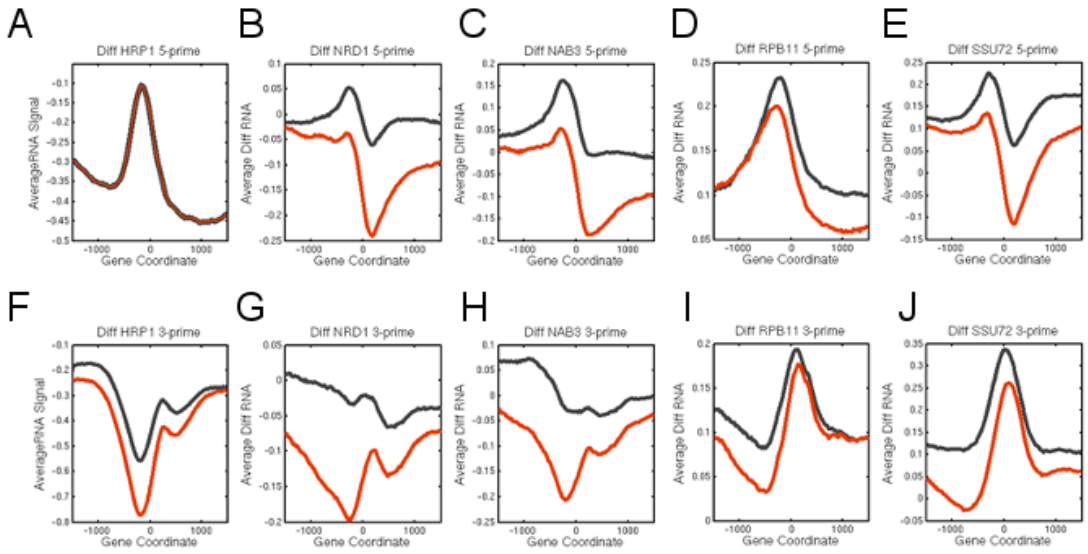


Figure 4-12: Differential RNA profile for genes showing length changes in various factors

Average differential RNA profiles for Hrp1, Nrd1, Nab3, Rpb11, and Ssu72 data sets at the 5' regions (in A-E) and at 3' regions (F-J) of the genes with transcription precision defects in black and with genes containing transcription precision defects with no Sen1 affect. An interesting thing to note is that in (A), the Hrp1 profile is not affected from black to red (the black and red curves overlap each other perfectly) showing independence from Sen1 pathway.

4.4 Summary and Further Analysis

We analyzed the effect of mutations in factors in the Sen1 termination pathway on transcription using gene expression and analyzing the data using the TraPP. Mutation in Sen1 had a very dramatic effect on transcription, showing extensive differential expression and a very distinct signature of transcription precision defects. Mutation of Sen1 resulted in read-through transcripts in 60% of snoRNA genes, with remaining 40% of snoRNA unaffected by mutations in any of the factors studied, suggesting presence of another termination pathway or exisitance of other unknown factors in the termination pathway for these genes. The gene expression and transcription precision results obtained strongly suggest that Sen1 is involved in suppression of meiosis-related genes in the vegetative state. Though thought to be involved only in poly (A)-independent termination, mutations in all these factors show termination defects in many Pol II genes. Sen1 pathway is responsible for degradation of the SUTs and CUTs and further analysis of SUT/CUT annotations using TraPP would shed light on whether these effects are direct or indirect. Also other classes of ncRNA should be analyzed by TraPP to test how these factors are involved in their processing.

APPENDIX A

Genome wide analysis of TBP binding dynamics by CLK ChIP-Chip

This includes preliminary results from analysis of genome-wide CLK ChIP Chip data for TBP.

Methods

Experimental data was collected using *S. cerevisiae* strain YAD155 (*SPT15-MYC*) which was derived from YPH499 (Sikorski and Hieter 1989) and is described in Chapter II and III. ChIP experiments were conducted by Melissa Wells Carver as previously described in Chapter II and with modifications described in Chapter III. Cells were crosslinked for different times: 15 min (3 biological replicates), 8 min (2 biological replicates), 5 min (2 biological replicates), 3 min (2 biological replicates), 1 min (2 biological replicates), 5 sec (2 biological replicates), 0.5 sec (2 biological replicates) along with native ChIP where no formaldehyde was added. Mock ChIP samples were also processed for different time points of crosslinking : 15 min (2 biological replicates), 8 min (1 sample), 5 min (1 sample), 3 min (2 biological replicates), 1 min (2 biological replicates), and 5 sec (2 biological replicates).

Data Analysis

Tiling Array Analysis

Tiling Analysis Software (TAS) version 1.1 was used to processes the raw files in the same manner as described in Chapter II with a slight variation in which we used all the Mock samples as a control to compute the signal intensities. This was done to reduce the noise variation or

apparent noise in the data. The output of TAS are graph files (.bar and .txt) files which contain signal and log-transformed p-values as a function of genomic coordinate.

ChIP-chip Peak Identification

In order to compare the ChIP levels we applied the method described in Chapter II and Figure 2-1 to identify TBP peak positions. Notably we only searched peaks within the promoter region (-300 to +50 bp region of the gene start site).

Consensus Peak finding Algorithm (CoPFA)

For CLK analysis, we need ChIP signals across crosslinking times spanning seconds to minutes. The following algorithm was developed to link the peaks in different time points to a promoter region. The resolution of ChIP-Chip is relatively poor and the TBP occupancy peaks are very broad, and when we calculate peak locations, these locations can vary hundreds of base pairs for different time points. We found the nearest neighbor of the peaks found in the datasets in other time-points within 350 bp region of the peak in question. Because new neighboring peaks can appear within 350 bp of a newly calculated centroid, we iterated this procedure until a centroid is stably identified (i.e., the centroid does not change between two successive iterations). The peaks used to calculate the centroids were used for the CLK analysis. Due to the broad resolution of tiling arrays, not all of the peaks were successfully clustered, especially the low peaks obtained at short time points (0.5 and 5 sec datasets). For the missing dataset points we computed the median signal within a window of 50 bp flanking the centroid. An illustration of the calculation of these centroids is shown in the Figure A-2. And Figure A-3 shows a screen shot showing centroid locations calculated with respect to each time-point.

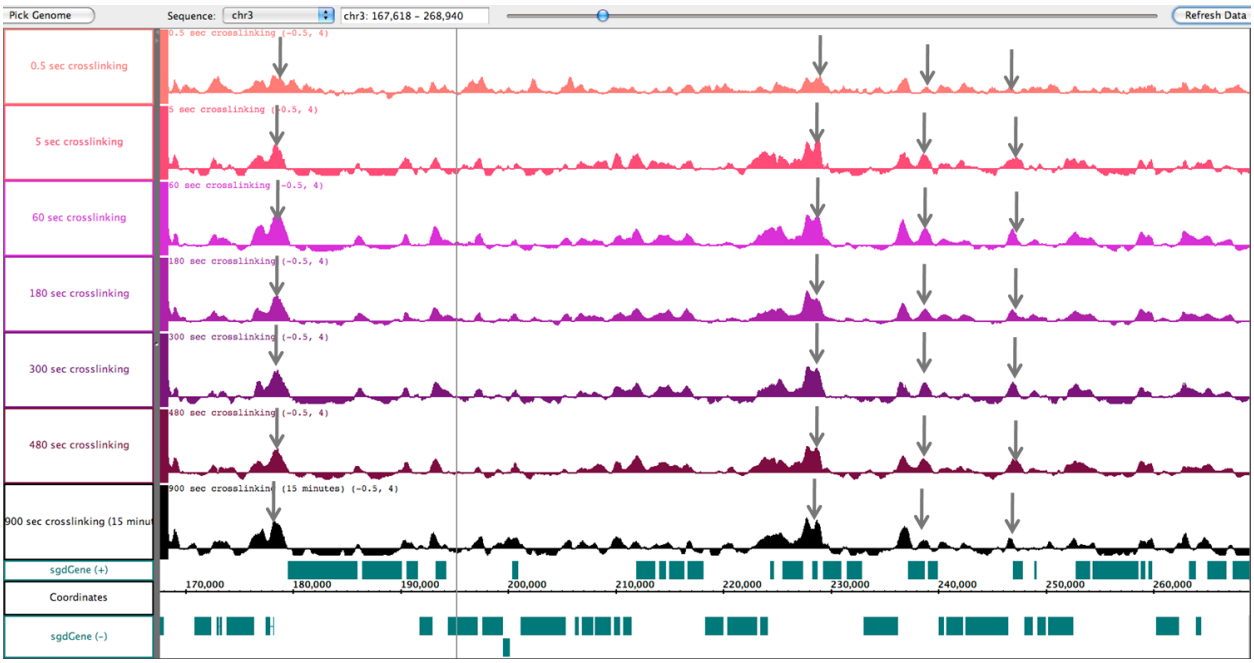


Figure A- 1: A Screen shot of a region of chromosome 3 showing the genome-wide CLK data for TBP. The arrows show peaks examples of TBP peaks which with increasing crosslinking time.

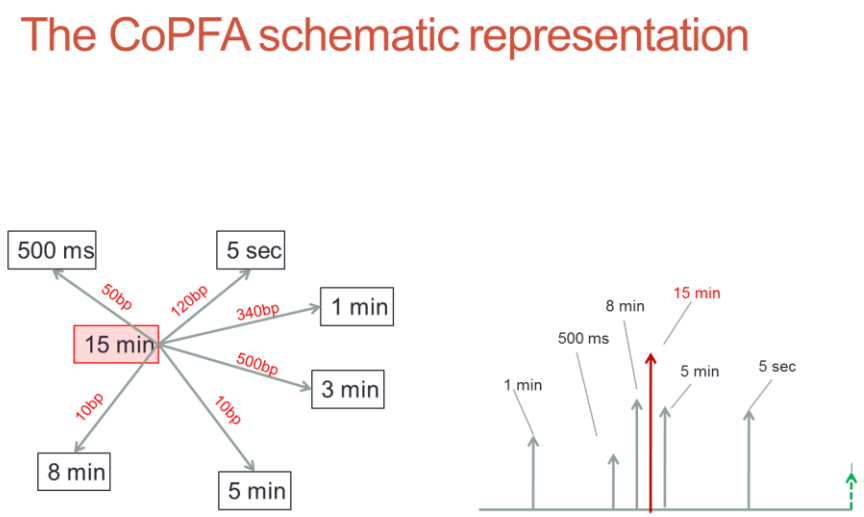


Figure A- 2: Schematic representation of how CoPFA computes centroids of peak clusters.

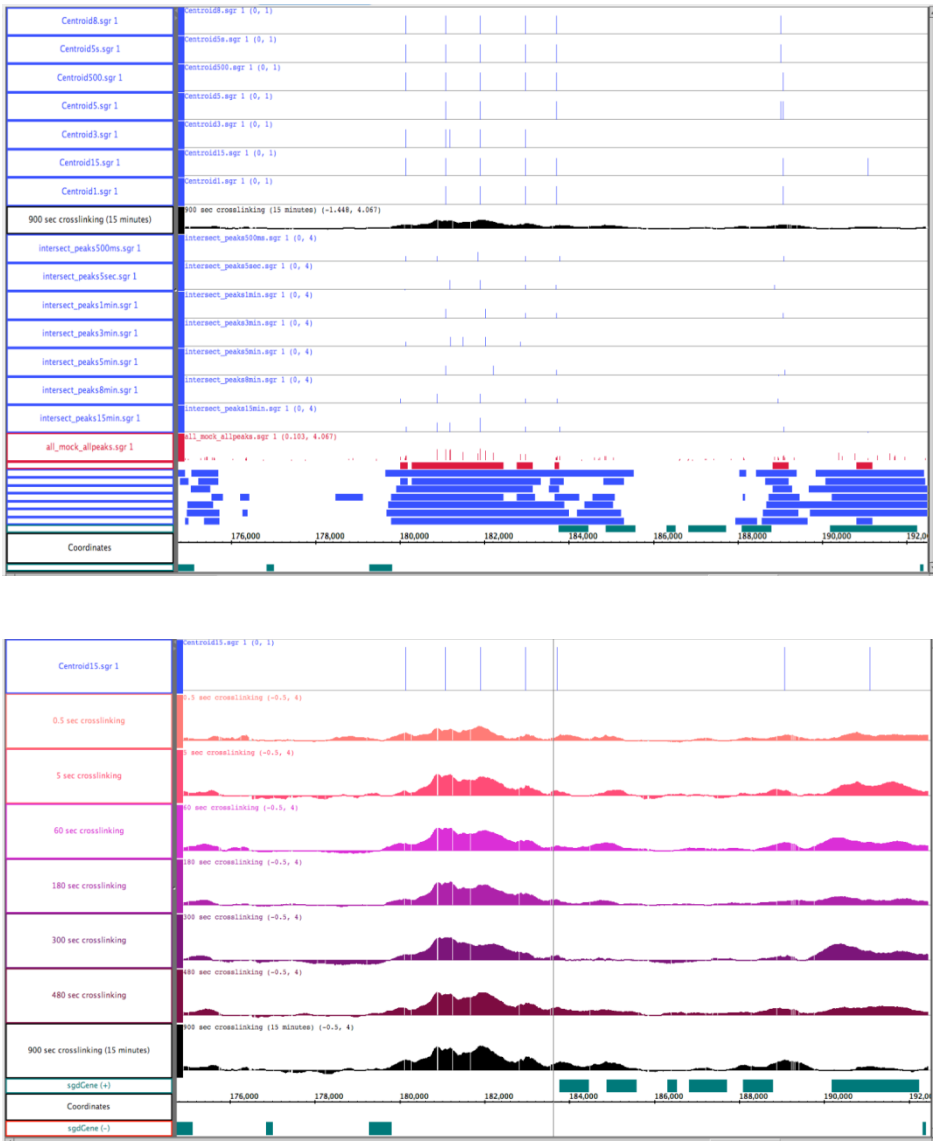


Figure A- 3: Example of CoPFA in the top panel showing all the different centroids computed by CoPFA for every time point, and the bottom panel shows the final centroids used to compute the signal used for CLK analysis.

Associating ChIP-chip Peaks to Genes

Using the TBP peaks, we associated a peak with a gene if it was within -300 to +50 bp of the annotated transcription start site (TSS). We allowed multiple genes which satisfy the distance criteria, to be associated with a single peak. Conversely, a single gene can be associated with multiple peaks, each of which satisfies the distance cutoff separately.

CLK model fitting

An approximate CLK model described by Eq. (19) in Chapter III was used to fit the datasets. Because of the sparsity of the time course data, only generalized linear regression could be effectively applied using the Eq. (25) described in Chapter III

$$\theta_{xl}(t) \approx 1 - (1 - \theta_b^0)e^{-k_a C_{TF} t}. \tag{19}$$

$$ln(1 - \theta_{xl}(t)) \approx ln(1 - \theta_b^0) - k_a C_{TF} t, \tag{25}$$

we used Eq. (25) to fit the CLK model with robust linear regression to arrive at initial estimates of k_a, θ_b^0 and k_d where

$$\theta_{xl}(t) = \frac{I_p(t)}{I_p(\infty)}$$

and $I_p(\infty)$ is defined by the maximum measured ChIP signal. These steps were sequentially applied to all the promoters which resulted data from CoPFA.

Results

The ChIP-chip data was analyzed with TAS as described in the Method section. Centroids and peaks were estimated for each promoter that containe significant ChIP signal

yielding a time series of ChIP signal data which allow fitting of the CLK model shown above in order to derive kinetic parameters.. We found ~1900 clusters of peaks which could be associated with genes and centroids with more than 3 peaks from individual time points from which the centroids were derived using CoPFA. Figure A-4 shows a heat map of the ChIP signal across time points derived using CoPFA. For each promoter with an associated centroid, we fit the CLK signal data derived using CoPFA to the CLK model using robust linear regression. Figure A-5 shows some of the model fits of selected promoters. Figure A-6 shows the distribution of parameters calculated across promoters. With this approach we found that the CLK model was able to fit about 900 promoter sites yielding ~300 physically plausible parameters. However, comparisons between the locus specific parameters and the ChIP-chip derived parameters showed relatively large discrepancies. The main source of the discrepancy was the estimation of saturation. In the Genomic CLK ChIP due to sparcity in the time course and also lack genomic data for over expression of TBP in the cell the fitting procedure only attempted to fit two parameters through linear regression. The ChIP saturation level was set to the maximum measured ChIP signal from the measured ChIP signal. This creates bias for datasets to be interpreted for high occupancy of TBP. This is an artifact of the analysis which can easily be rectified by including overexpression of TBP dataset in parameter estimation method.

Potential Pitfalls and Caveats in this approach

Although we derive kinetic parameters from some promoters, this approach could be significantly improved. We find reasonable fits for only 10% of the sites with peaks/centroids. ChIP-chip is a powerful approach, but the peaks are quite broad and the loss of peak-resolution contributes significantly to the effective noise in the data used for fitting the CLK model. Figure A-7 shows the DNA quantification data after the first amplification of DNA for different time

points. The x-axis represents the crosslinking time in seconds and y-axis represents the amount of DNA acquired after first amplification (ng) normalized to the number of cells used.

It is evident that the total DNA content increases with the increase in crosslinking time; however, ChIP-chip requires two rounds of amplification to generate the minimum amount of DNA for array hybridization. Although this amplification step is necessary to ensure detectable signal intensities, it creates an amplification bias for the lower time points which display lower levels of originally collected DNA. Notably, there are traces of this amplification-based bias in the background levels (i.e., background levels in lower time points are higher), however, they are challenging to quantify. This amplification bias can be easily avoided if ChIP-seq method is applied to perform CLK analysis because only one amplification step is required. Another major is problem in the dataset is lack of CLK-ChIP data-points for overexpressed TBP. As it has discussed above lack of TBP overexpression data-points for fitting the CLK model leaves us with only two parameters and also a poor estimate of ChIP saturation levels. And it is evident from comparing the parameters computed in the genomic study and the locus specific study (Chapter III) that the parameters computed in in genomic CLK-ChIP is biased towards saturated ChIP signal. Hence in order to get more meaningful parameters there should be atleast one datapoint of ChIP for over-expressed TBP to have a better estimate of ChIP saturation levels. One other problem with using Tiling arrays is that CLK method has been optimized to be applied to ChIP non log transformed signal and TAS produces log transformed ChIP signal, and conversion to non-logtransformed data increases the error exponentially. Again this error can be reduced by using mapped ChIP-seq reads for doing the CLK analysis, but further study and optimization is needed to see the applicability of CLK method using ChIP-seq.

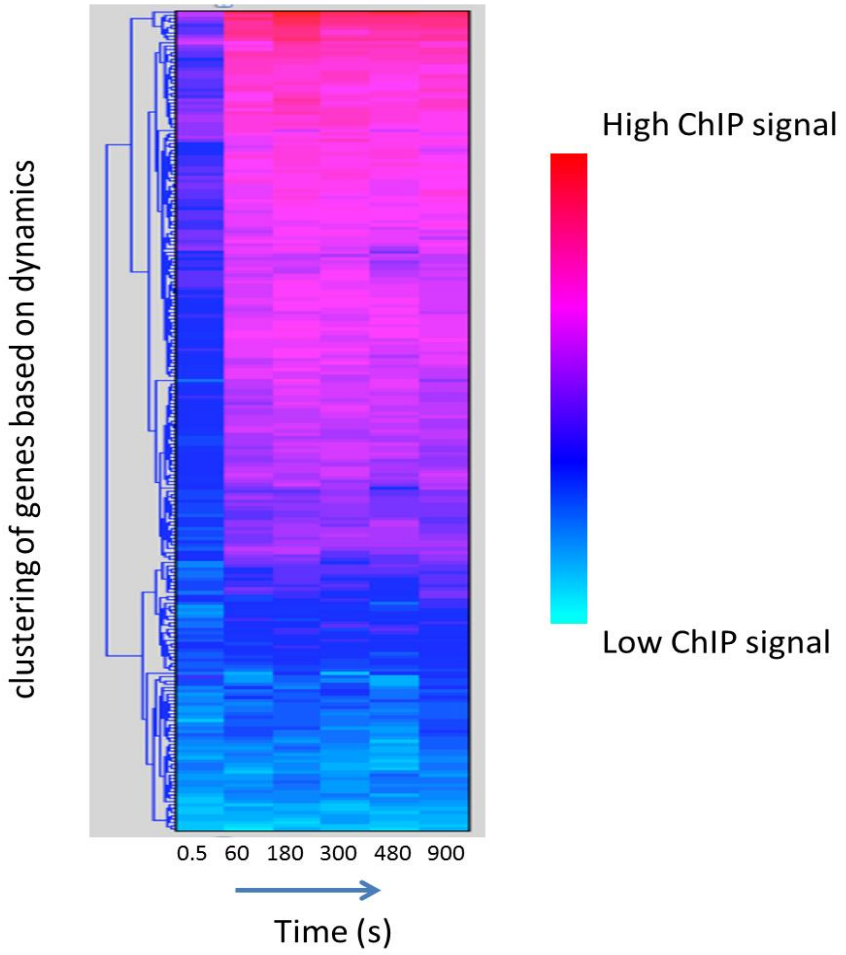


Figure A- 4: A heatmap of ChIP Signals computed from the clusters of the peaks and signal associated with promoters

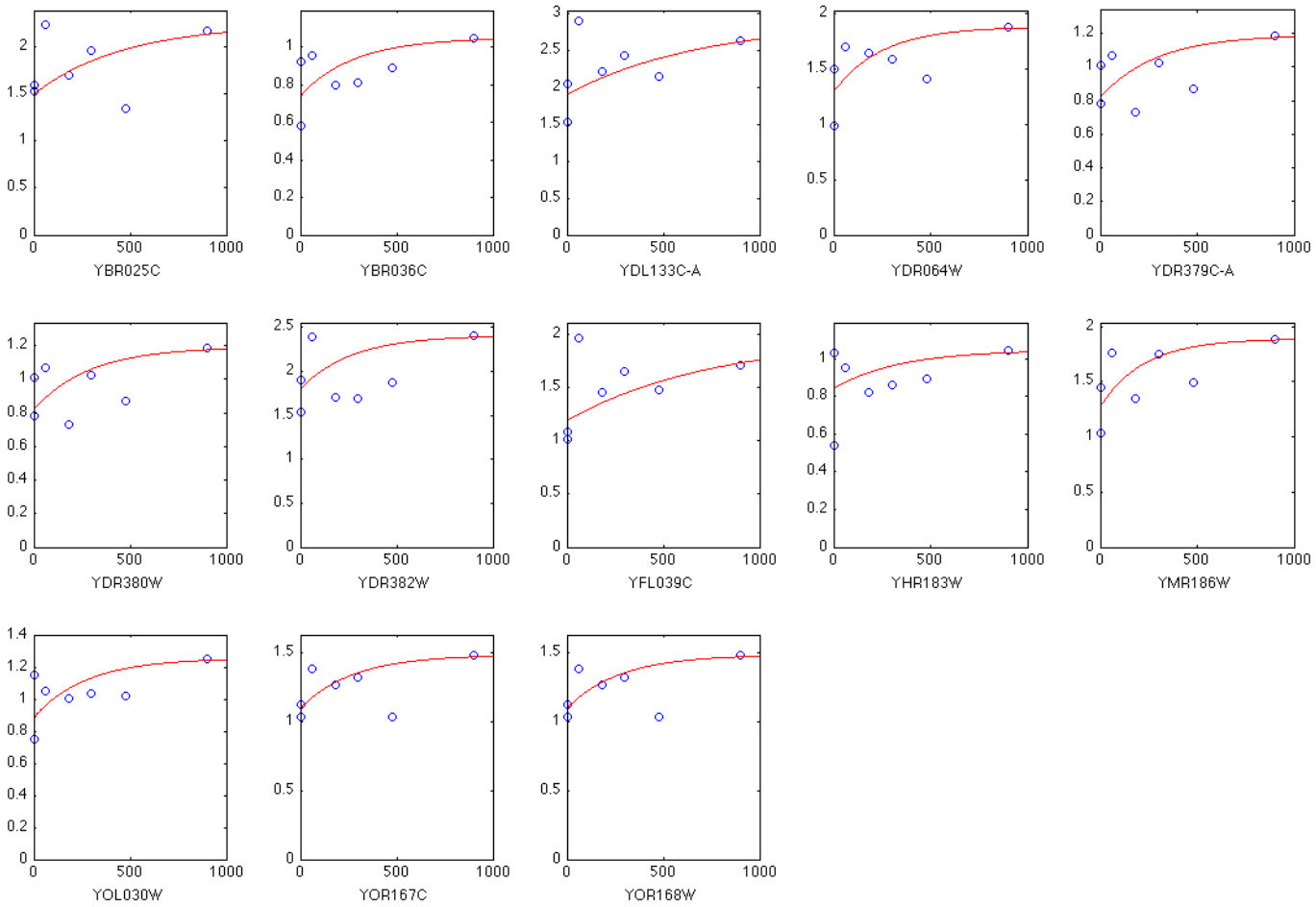


Figure A- 5: Examples of CLK model fits for TBP binding to various promoters.

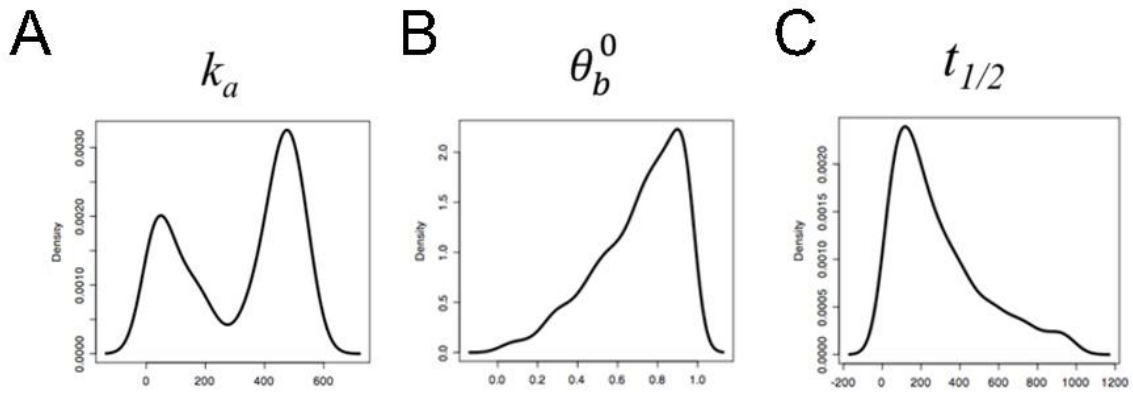


Figure A- 6: Distribution of parameters collected by fitting the approximate CLK model using robust linear regression to the datasets obtained using microarrays.

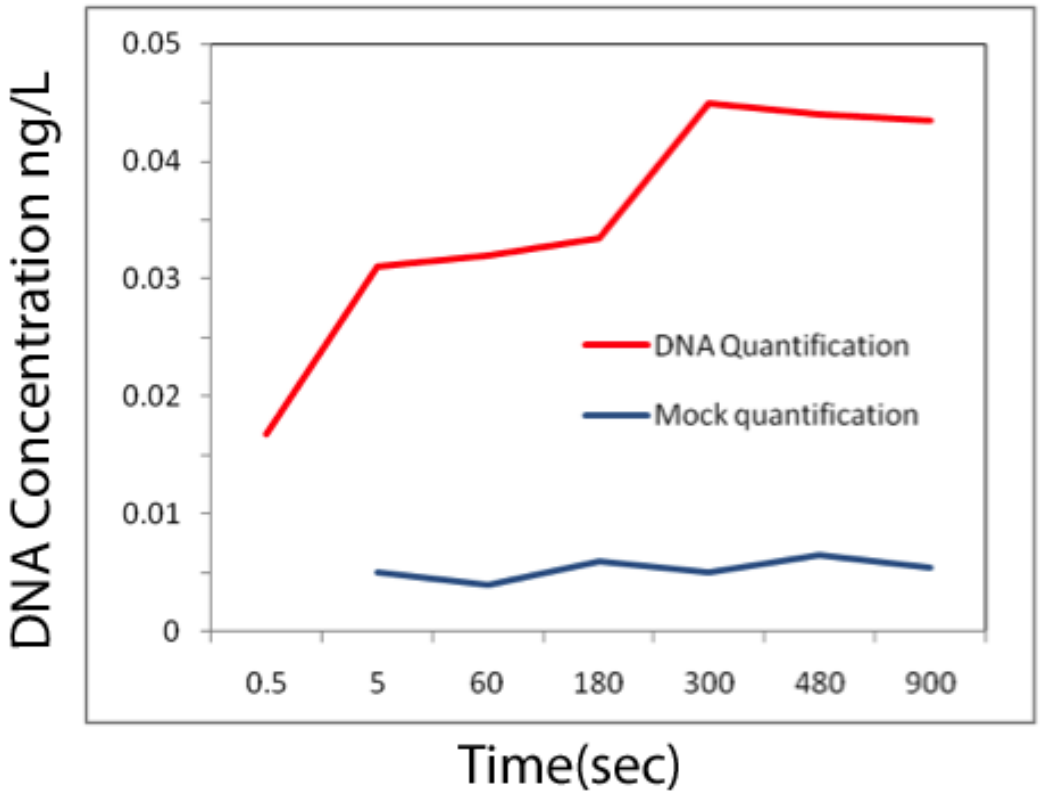


Figure A- 7: DNA quantification data after the first round of DNA amplification.

References

Akhtar MS, Heidemann M, Tietjen JR, Zhang DW, Chapman RD, Eick D, Ansari AZ. 2009. TFIIH kinase places bivalent marks on the carboxy-terminal domain of RNA polymerase II. *Mol Cell* **34**: 387–93.

Auble DT. 2009. The Dynamic Personality of TATA-Binding Protein. *Trends Biochem Sci* **34**: 49–52.

Auble DT, Hahn S. 1993. An ATP-dependent Inhibitor of TBP Binding to DNA. *Genes Dev* **7**: 844–856.

Baker SP, Grant PA. 2007. The SAGA continues: expanding the cellular role of a transcriptional co-activator complex. *Oncogene* **26**: 5329–40.

Basehoar AD, Zanton SJ, Pugh BF. 2004. Identification and Distinct Regulation of Yeast TATA Box-Containing Genes. *Cell* **116**: 699–709.

Berger SL, Piña B, Silverman N, Marcus GA, Agapite J, Regier JL, Triezenberg SJ, Guarente L. 1992. Genetic isolation of ADA2: A potential transcriptional adaptor required for function of certain acidic activation domains. *Cell* **70**: 251–265.

Bhaumik SR, Green MR. 2002. Differential Requirement of SAGA Components for Recruitment of TATA-Box-Binding Protein to Promoters In Vivo. *Mol Cell Biol* **22**: 7365–7371.

Bhaumik SR, Green MR. 2001. SAGA Is an Essential In Vivo Target of the Yeast Acidic Activator Gal4p. *Genes Dev* **15**: 1935–1945.

Björklund S, Gustafsson CM. 2005. The yeast Mediator complex and its regulation. *Trends Biochem Sci* **30**: 240–4.

Borggrefe T, Davis R, Bareket-Samish A, Kornberg RD. 2001. Quantitation of the RNA Polymerase II Transcription Machinery in Yeast. *J Biol Chem* **276**: 47150–47153.

Bourbon H-M. 2008. Comparative genomics supports a deep evolutionary origin for the large, four-module transcriptional mediator complex. *Nucleic Acids Res* **36**: 3993–4008.

Brow DA. 2011. Sen-sing RNA Terminators. *Mol Cell* **42**: 717–718.

Buratowski S. 2009. Progression through the RNA Polymerase II CTD Cycle. *Mol Cell* **36**: 541–546.

Buratowski S, Hahn S, Guarente L, Sharp PA. 1989. Five Intermediate Complexes in Transcription Initiation by RNA Polymerase II. *Cell* **56**: 549–561.

Burley SK, Roeder RG. 1996. Biochemistry and Structural Biology of Transcription Factor IID (TFIID). *Ann rev Biochem* **65**: 769–799.

Cawley S, Bekiranov S, Ng HH, Kapranov P, Sekinger EA, Kampa D, Piccolboni A, Sementchenko V, Cheng J, Williams AJ, et al. 2004. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* **116**: 499–509.

Chatterjee S, Struhl K. 1995. Connecting a Promoter-Bound Protein to TBP Bypasses the Need for a Transcriptional Activation Domain. *Nature* **374**: 820–822.

Chen HT, Hahn S. 2003. Binding of TFIIB to RNA Polymerase II: Mapping the Binding Site for the TFIIB Zinc Ribbon Domain within the Preinitiation Complex. *Mol Cell* **12**: 437–447.

Chen H-T, Warfield L, Hahn S. 2007. The positions of TFIIF and TFIIE in the RNA polymerase II transcription preinitiation complex. *Nat Struct Mol Biol* **14**: 696–703.

Chen Y-Z, Bennett CL, Huynh HM, Blair IP, Puls I, Irobi J, Dierick I, Abel A, Kennerson ML, Rabin BA, et al. 2004. DNA/RNA helicase gene mutations in a form of juvenile amyotrophic lateral sclerosis (ALS4). *Am J Hum Genet* **74**: 1128–35.

Chen Y-Z, Hashemi SH, Anderson SK, Huang Y, Moreira M-C, Lynch DR, Glass IA, Chance PF, Bennett CL. 2006. Senataxin, the yeast Sen1p orthologue: characterization of a unique protein in which recessive mutations cause ataxia and dominant mutations cause motor neuron disease. *Neurobiol Dis* **23**: 97–108.

Chu Y, Simic R, Warner MH, Arndt KM, Prelich G. 2007. Regulation of histone modification and cryptic transcription by the Bur1 and Paf1 complexes. *EMBO J* **26**: 4646–4656.

Collart MA. 1996. The NOT, SPT3, and MOT1 Genes Functionally Interact to Regulate Transcription at Core Promoters. *Mol Cell Biol* **16**: 6668–6676.

Collins GA, Lipford JR, Deshaies RJ, Tansey WP. 2009. Gal4 Turnover and Transcription Activation. . *Nature* **461**: E7.

Conaway RC, Sato S, Tomomori-Sato C, Yao T, Conaway JW. 2005. The mammalian Mediator complex and its role in transcriptional regulation. *Trends Biochem Sci* **30**: 250–5.

Connelly S, Manley JL. 1988. A functional mRNA polyadenylation signal is required for transcription termination by RNA polymerase II. *Genes Dev* **2**: 440–52.

Darst RP, Dasgupta A, Zhu C, Hsu J-Y, Vroom A, Muldrow TA, Auble DT. 2003. Mot1 Regulates the DNA Binding Activity of Free TATA-Binding Protein in an ATP-Dependent Manner. *J Biol Chem* **278**: 13216–13226.

Dasgupta A, Darst RP, Martin KJ, Afshari CA, Auble DT. 2002. Mot1 Activates and Represses Transcription by Direct, ATPase-Dependent Mechanisms. *PNAS* **99**: 2666–2671.

Dasgupta A, Juedes SA, Sprouse RO, Auble DT. 2005. Mot1-Mediated Control of Transcription Complex Assembly and Activity. *EMBO J* **24**: 1717–1729.

David L, Huber W, Granovskaia M, Toedling J, Palm CJ, Bofkin L, Jones T, Davis RW, Steinmetz LM. 2006. A High-Resolution Map of Transcription in the Yeast Genome. *PNAS* **103**: 5320–5325.

Davis CA, Ares Jr. M. 2006. Accumulation of Unstable Promoter-Associated Transcripts Upon Loss of the Nuclear Exosome Subunit Rrp6p in Saccharomyces cerevisiae. *PNAS* **103**: 3262–3267.

Deal RB, Henikoff JG, Henikoff S. 2010. Genome-Wide Kinetics of Nucleosome Turnover Determined by Metabolic Labeling of Histones. *Science (80-)* **328**: 1161–1164.

Donovan S, Harwood J, Drury LS, Diffley JFX. 1997. Cdc6p-dependent loading of Mcm proteins onto pre-replicative chromatin in budding yeast. *Proc Natl Acad Sci* **94**: 5611–5616.

Eichner J, Chen H-T, Warfield L, Hahn S. 2010. Position of the general transcription factor TFIIF within the RNA polymerase II transcription preinitiation complex. *EMBO J* **29**: 706–16.

Elf J, Li GW, Xie XS. 2007. Probing transcription factor dynamics at the single-molecule level in a living cell. *Science (80-)* **316**: 1191–1194.

Esnault C, Ghavi-Helm Y, Brun S, Soutourina J, Van Berkum N, Boschiero C, Holstege F, Werne M. 2008. Mediator-Dependent Recruitment of TFIIH Modules in Preinitiation Complex. *Mol Cell* **31**: 337–346.

Faitar SL, Brodie SA, Ponticelli AS. 2001. Promoter-specific shifts in transcription initiation conferred by yeast TFIIB mutations are determined by the sequence in the immediate vicinity of the start sites. *Mol Cell Biol* **21**: 4427–40.

Feaver WJ, Svejstrup JQ, Henry NL, Kornberg RD. 1994. Relationship of CDK-activating kinase and RNA polymerase II CTD kinase TFIIH/TFIIK. *Cell* **79**: 1103–9.

Fisk DG, Ball CA, Dolinski K, Engel SR, Hong EL, Issel-Tarver L, Schwartz K, Sethuraman A, Botstein D, Cherry JM, et al. 2006. Saccharomyces cerevisiae S288C genome annotation: a working hypothesis. *Yeast* **23**: 857–865.

Geiger JH, Hahn S, Lee S, Sigler PB. 1996. Crystal Structure of the Yeast TFIIA/TBP/DNA Complex. *Science (80-)* **272**: 830–836.

Geisberg J V, Holstege FC, Young RA, Struhl K. 2001. Yeast NC2 Associates with the RNA Polymerase II Preinitiation Complex and Selectively Affects Transcription In Vivo. *Mol Cell Biol* **21**: 2736–2742.

Geisberg J V, Moqtaderi Z, Kuras L, Struhl K. 2002. Mot1 Associates with Transcriptionally Active Promoters and Inhibits Association of NC2 in Saccharomyces cerevisiae. *Mol Cell Biol* **22**: 8122–8134.

Geisberg J V, Struhl K. 2004. Cellular Stress Alters the Transcriptional Properties of Promoter-Bound Mot1-TBP Complexes. *Mol Cell* **14**: 479–489.

Giardina C, Lis JT. 1995. Dynamic Protein-DNA Architecture of a Yeast Heat Shock Promoter. *Mol Cell Biol* **15**: 2737–2744.

Gorman J, Greene EC. 2008. Visualizing One-Dimensional Diffusion of Proteins Along DNA. *Nat Struct Mol Biol* **15**: 768–774.

Gorski S, Misteli T. 2005. Systems biology in the cell nucleus. *J Cell Sci* **118**: 4083–92.

Grant PA, Duggan L, Cote J, Roberts SM, Brownell JE, Candau R, Ohba R, Owen-Hughes T, Allis CD, Winston F, et al. 1997. Yeast Gcn5 Functions in Two Multisubunit Complexes to Acetylate Nucleosomal histones: Characterization of an Ada Complex and the SAGA (Spt/Ada) Complex. *Genes Dev* **11**: 1640–1650.

Green MR. 2005. Eukaryotic Transcription Activation: Right on Target. *Mol Cell* **18**: 399–402.

Grünberg S, Warfield L, Hahn S. 2012. Architecture of the RNA polymerase II preinitiation complex and mechanism of ATP-dependent promoter opening. *Nat Struct Mol Biol* **19**: 788–96.

Hager GL, McNally JG, Mistelli T. 2009. Transcription Dynamics. *Mol Cell* **35**: 741–753.

Hahn S. 2004. Structure and mechanism of the RNA polymerase II transcription machinery. *Nat Struct Mol Biol* **11**: 394–403.

Hahn S. 1998. The Role of TAFs in RNA Polymerase II Transcription. *Cell* **95**: 579–582.

Hahn S, Buratowski S, Sharp PA, Guarente L. 1989. Yeast TATA-Binding Protein TFIID Binds to TATA Elements with Both Consensus and Nonconsensus DNA Sequences. *Cell* **58**: 5718–5722.

Hahn S, Young ET. 2011. Transcriptional regulation in Saccharomyces cerevisiae: transcription factor regulation and function, mechanisms of initiation, and roles of activators and coactivators. *Genetics* **189**: 705–36.

Halford SE, Marko JF. 2004. How do site-specific DNA-binding proteins find their targets? *Nucleic Acids Res* **32**: 3040–3052.

Hampsey M. 1998. Molecular genetics of the RNA polymerase II general transcriptional machinery. *Microbiol Mol Biol Rev* **62**: 465–503.

Helmlinger D, Marguerat S, Villén J, Gygi SP, Bähler J, Winston F. 2008. The S. pombe SAGA complex controls the switch from proliferation to sexual differentiation through the opposing roles of its subunits Gcn5 and Spt8. *Genes Dev* **22**: 3184–95.

Huisinga KL, Pugh BF. 2004. A Genome-Wide Housekeeping Role for TFIID and a Highly Regulated Stress-Related Role for SAGA in Saccharomyces cerevisiae. *Mol Cell* **13**: 573–585.

Hyland EM, Molina H, Poorey K, Jie C, Xie Z, Dai J, Qian J, Bekiranov S, Auble DT, Pandey A, et al. 2011. An evolutionarily “young” lysine residue in histone H3 attenuates transcriptional output in Saccharomyces cerevisiae. *Genes Dev* **25**: 1306–1319.

Jacobson RH, Tjian R. 1996. Transcription Factor IIA: A Structure with Multiple Functions. *Science (80-)* **272**: 827–828.

Johnston SA, Hopper JE. 1982. Isolation of the yeast regulatory gene GAL4 and analysis of its dosage effects on the galactose/melibiose regulon. *Proc Natl Acad Sci U S A* **79**: 6971–5.

Jorgensen P, Edgington NP, Schneider BL, Rupes I, Tyers M, Futcher B. 2007. The Size of the Nucleus Increases as Yeast Cells Grow. *Mol Biol Cell* **18**: 3523–3532.

Jung I, Kim D. 2012. Histone modification profiles characterize function-specific gene regulation. *J Theor Biol* **310**: 132–42.

Juven-Gershon T, Hsu J-Y, Theisen JW, Kadonaga JT. 2008. The RNA polymerase II core promoter - the gateway to transcription. *Curr Opin Cell Biol* **20**: 253–9.

Kamada K, Shu F, Chen H, Malik S, Stelzer G, Roeder RG, Meisterernst M, Burley SK. 2001. Crystal Structure of Negative Cofactor 2 Recognizing the TBP-DNA Transcription Complex. *Cell* **106**: 71–81.

Kanin EI, Kipp RT, Kung C, Slattery M, Viale A, Hahn S, Shokat KM, Ansari AZ. 2007. Chemical inhibition of the TFIIH-associated kinase Cdk7/Kin28 does not impair global mRNA synthesis. *Proc Natl Acad Sci U S A* **104**: 5812–7.

Karpova TS, Kim MJ, Spriet C, Nalley K, Stasevich TJ, Kherrouche Z, Heliot L, McNally JG. 2008. Concurrent Fast and Slow Cycling of a Transcriptional Activator at an Endogenous Promoter. *Science (80-)* **319**: 466–469.

Keaveney M, Struhl K. 1998. Activator-Mediated Recruitment of the RNA Polymerase II Machinery Is the Predominant Mechanism for Transcriptional Activation in Yeast. *Mol Cell* **1**: 917–924.

Kessler MM, Henry MF, Shen E, Zhao J, Gross S, Silver PA, Moore CL. 1997. Hrp1, a sequence-specific RNA-binding protein that shuttles between the nucleus and the cytoplasm, is required for mRNA 3'-end formation in yeast. *Genes Dev* **11**: 2545–56.

Khaperskyy DA, Ammerman ML, Majovski RC, Ponticelli AS. 2008. Functions of Saccharomyces cerevisiae TFIIF during transcription start site utilization. *Mol Cell Biol* **28**: 3757–66.

Kim J, Iyer VR. 2004. Global Role of TATA Box-Binding Protein Recruitment to Promoters in Mediating Gene Expression Profiles. *Mol Cell Biol* **24**: 8104–8112.

Kim M, Krogan NJ, Vasiljeva L, Rando OJ, Nedea E, Greenblatt JF, Buratowski S. 2004. The yeast Rat1 exonuclease promotes transcription termination by RNA polymerase II. *Nature* **432**: 517–22.

Klages N, Strubin M. 1995. Stimulation of RNA Polymerase II Transcription Initiation by Recruitment of TBP In Vivo. *Nature* **374**: 822–823.

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

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

Kuehner JN. 2008. *Mechanism and Regulation of Yeast RNA Polymerase II Transcription Initiation and Termination*. ProQuest.

Kuehner JN, Brow DA. 2008. Regulation of a Eukaryotic Gene by GTP-Dependent Start Site Selection and Transcription Attenuation. *Mol Cell* **31**: 201–211.

Kuehner JN, Pearson EL, Moore C. 2011. Unravelling the means to an end: RNA polymerase II transcription termination. *Nat Rev Mol Cell Biol* **12**: 283–94.

Kuo MH, Allis CD. 1999. In Vivo Cross-Linking and Immunoprecipitation for Studying Dynamic Protein:DNA Associations in a Chromatin Environment. *Methods* **19**: 425–433.

Larson DR, Zenklusen D, Wu B, Chao JA, Singer RH. 2011. Real-Time Observation of Transcription Initiation and Elongation on an Endogenous Yeast Gene. *Science (80-)* **332**: 475–478.

Lee TI, Causton HC, Holstege FCP, Shen W-C, Hannett N, Jennings EG, Winston F, Green MR, Young RA. 2000. Redundant Roles for the TFIID and SAGA Complexes in Global Transcription. *Nature* **405**: 701–704.

Li B, Carey M, Workman JL. 2007a. The Role of Chromatin during Transcription. *Cell* **128**: 707–719.

Li B, Gogol M, Carey M, Pattenden SG, Seidel C, Workman JL. 2007b. Infrequently transcribed long genes depend on the Set2/Rpd3S pathway for accurate transcription. *Genes Dev* **21**: 1422–1430.

Lickwar CR, Mueller F, Hanlon SE, McNally JG, Lieb JD. 2012. Genome-wide protein-DNA binding dynamics suggest a molecular clutch for transcription factor function. *Nature* **484**: 251–5.

Lickwar CR, Mueller F, Lieb JD. 2013. Genome-wide measurement of protein-DNA binding dynamics using competition ChIP. *Nat Protoc* **8**: 1337–53.

Liu C, van Dyk D, Li Y, Andrews B, Rao H. 2009a. A genome-wide synthetic dosage lethality screen reveals multiple pathways that require the functioning of ubiquitin-binding proteins Rad23 and Dsk2. *BMC Biol* **7**: 75.

Liu W-L, Coleman RA, Ma E, Grob P, Yang JL, Zhang Y, Dailey G, Nogales E, Tjian R. 2009b. Structures of three distinct activator–TFIID complexes. *Genes Dev* **23**: 1510–1521.

Liu X, Bushnell DA, Wang D, Calero G, Kornberg RD. 2010. Structure of an RNA Polymerase II–TFIIB Complex and the Transcription Initiation Mechanism. *Science (80-)* **327**: 206–209.

Madison JM, Winston F. 1996. Evidence that Spt3 Functionally Interacts with Mot1, TFIIA, and TBP to Confer Promoter-Specific Transcriptional Control in Saccharomyces cerevisiae. *Mol Cell Biol* **17**: 287–295.

Malik S, Roeder RG. 2005. Dynamic regulation of pol II transcription by the mammalian Mediator complex. *Trends Biochem Sci* **30**: 256–63.

Michelman-Ribeiro A, Mazza D, Rosales T, Stasevich TJ, Boukari H, Rishi V, Vinson C, Knutson JR, McNally JG. 2009. Direct Measurement of Association and Dissociation Rates of DNA Binding in Live Cells by Fluorescence Correlation Spectroscopy. *Biophys J* **97**: 337–346.

Miller G, Hahn S. 2006. A DNA-Tethered Cleavage Probe Reveals the Path for Promoter DNA in the Yeast Preinitiation Complex. *Nat Struct Mol Biol* **13**: 603–610.

Misteli T. 2001. Protein Dynamics: Implications for Nuclear Architecture and Gene Expression. *Science (80-)* **291**: 843–847.

Mizzen CA, Yang X-J, Kokubo T, Brownell JE, Bannister AJ, Owen-Hughes T, Workman J, Wang L, Berger SL, Kouzarides T, et al. 1996. The TAF250 Subunit of TFIID has Histone Acetyltransferase Activity. *Cell* **87**: 1261–1270.

Mohibullah N, Hahn S. 2008. Site-specific cross-linking of TBP in vivo and in vitro reveals a direct functional interaction with the SAGA subunit Spt3. *Genes Dev* **22**: 2994–3006.

Mueller F, Wach P, McNally JG. 2008. Evidence for a common mode of transcription factor interaction with chromatin as revealed by improved quantitative fluorescence recovery after photobleaching. *Biophys J* **94**: 3323–39.

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. 2008. The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. *Science (80-)* **320**: 1344–1349.

Nalley K, Johnston SA, Kodadek T. 2006. Proteolytic turnover of the Gal4 transcription factor is not required for function in vivo. *Nature* **442**: 1054–1057.

Narlikar GJ, Fan H-Y, Kingston RE. 2002. Cooperation Between Complexes that Regulate Chromatin Structure and Transcription. *Cell* **108**: 475–487.

Neil H, Malabat C, d'Aubenton-Carafa Y, Xu Z, Steinmetz LM, Jacquier A. 2009. Widespread Bidirectional Promoters are the Major Source of Cryptic Transcripts in Yeast. *Nature* **457**: 1038–1042.

Nemet J, Jelicic B, Rubelj I, Sopta M. 2013. The two faces of Cdk8, a positive/negative regulator of transcription. *Biochimie*.

O'Green H, Nicolet CM, Blahnik K, Green RD, Farnham PJ. 2006. Comparison of Sample Preparation Methods for ChIP-chip Assays. *Biotechniques* **41**: 577–580.

Patikoglou GA, Kim JL, Sun L, Yang S-H, Kodadek T, Burley SK. 1999. TATA Element Recognition by the TATA Box-Binding Protein has been Conserved Throughout Evolution. *Genes Dev* **13**: 3217–3230.

Pereira LA, Klejman MP, Timmers HTM. 2003. Roles for BTAF1 and Mot1p in Dynamics of TATA-Binding Protein and Regulation of RNA Polymerase II Transcription. *Gene* **315**: 1–13.

Phair RD, Scaffidi P, Elbi C, Vecerova J, Dey A, Ozato K, Brown DT, Hager G, Bustin M, Misteli T. 2004. Global Nature of Dynamic Protein-Chromatin Interactions In Vivo: Three-Dimensional Genome Scanning and Dynamic Interaction Networks of Chromatin Proteins. *Mol Cell Biol* **24**: 6393–6402.

Poorey K, Sprouse RO, Wells MN, Viswanathan R, Bekiranov S, Auble DT. 2010a. RNA synthesis precision is regulated by preinitiation complex turnover. *Genome Res* **20**: 1679–88.

Poorey K, Sprouse RO, Wells MN, Viswanathan R, Bekiranov S, Auble DT. 2010b. RNA Synthesis Precision Is Regulated by Preinitiation Complex Turnover. *Genome Res* **in press**.

Poorey K, Viswanathan R, Carver MN, Karpova TS, Cirimotich SM, McNally JG, Bekiranov S, Auble DT. 2013. Measuring Chromatin Interaction Dynamics on the Second Time Scale at Single-Copy Genes. *Science (80-)* **342**: 369–72.

Prelich G. 1997. Saccharomyces cerevisiae BUR6 Encodes a DRAP1/NC2alpha Homolog that has both Positive and Negative Roles in Transcription in Vivo. *Mol Cell Biol* **17**: 2057–2065.

Quinlan AR, Hall IM. 2010. The BEDTools manual.

Rasmussen EB, Lis JT. 1993. In vivo transcriptional pausing and cap formation on three Drosophila heat shock genes. *Proc Natl Acad Sci U S A* **90**: 7923–7.

Reese JC. 2003. Basal Transcription Factors. *Curr Opin Genet Dev* **13**: 114–118.

Resat H, Petzold L, Pettigrew MF. 2009. Kinetic Modeling of Biological Systems eds. R. Ireton, K. Montgomery, R. Bumgarner, R. Samudrala, and J. McDermott. *Methods Mol Biol* **541**: 1–19.

Rhee HS, Pugh BF. 2011. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* **147**: 1408–19.

Rhee HS, Pugh BF. 2012. Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature* **483**: 295–301.

Roberts DN, Stewart AJ, Huff JT, Cairns BR. 2003. The RNA polymerase III transcriptome revealed by genome-wide localization and activity-occupancy relationships. *Proc Natl Acad Sci U S A* **100**: 14695–700.

Robinett CC, Straight A, Li G, Willhelm C, Sudlow G, Murray A, Belmont AS. 1996. In vivo localization of DNA sequences and visualization of large-scale chromatin organization using lac operator/repressor recognition. *J Cell Biol* **135**: 1685–700.

Rosonina E, Kaneko S, Manley JL. 2006. Terminating the transcript: breaking up is hard to do. *Genes Dev* **20**: 1050–6.

Van Royen ME, Farla P, Mattern KA, Geverts B, Trapman J, Houtsmuller AB. 2009. Fluorescence recovery after photobleaching (FRAP) to study nuclear protein dynamics in living cells. *Methods Mol Biol* **464**: 363–385.

Van Royen ME, Zotter A, IbraHIM SM, Geverts B, Houtsmuller AB. 2011. Nuclear proteins: finding and binding target sites in chromatin. *Chromosom Res* **19**: 83–98.

Rufiange A, Jacques P-E, Bhat W, Robert F, Nourani A. 2007. Genome-Wide Replication-Independent Histone H3 Exchange Occurs Predominantly at Promoters and Implicates H3 K56 Acetylation and Asf1. *Mol Cell* **27**: 393–405.

Samorodnitsky E, Pugh BF. 2010. Genome-wide modeling of transcription preinitiation complex disassembly mechanisms using ChIP-chip data. *PLoS Comput Biol* **6**: e1000733.

Sandelin A, Carninci P, Lenhard B, Ponjavic J, Hayashizaki Y, Hume DA. 2007. Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nat Rev Genet* **8**: 424–36.

Schermer UJ, Korber P, Hörz W. 2005. Histones are incorporated in trans during reassembly of the yeast PHO5 promoter. *Mol Cell* **19**: 279–85.

Schluesche P, Stelzer G, Piaia E, Lamb DC, Meisterernst M. 2007. NC2 Mobilizes TBP on Core Promoter TATA Boxes. *Nat Struct Mol Biol* **14**: 1196–1201.

Schmitt ME, Brown TA, Trumpower BL. 1990. A rapid and simple method for preparation of RNA from Saccharomyces cerevisiae. *Nucleic Acids Res* **18**: 3091–3092.

Sikorski RS, Hieter P. 1989. A System of Shuttle Vectors and Yeast Host Strains Designed for Efficient Manipulation of DNA in Saccharomyces cerevisiae. *Genetics* **122**: 19–27.

Spencer J V, Arndt KM. 2002. A TATA Binding Protein Mutant with Increased Affinity for DNA Directs Transcription from a Reversed TATA Sequence In Vivo. *Mol Cell Biol* **22**: 8744–8755.

Sprague BL, Pego RL, Stavreva DA, McNally JG. 2004. Analysis of Binding Reactions by Fluorescence Recovery After Photobleaching. *Biophys J* **86**: 3473–3495.

Sprouse RO, Karpova TS, Mueller F, Dasgupta A, McNally JG, Auble DT. 2008. Regulation of TATA binding protein dynamics in living yeast cells. *PNAS* **105**: 13304–13308.

Sprouse RO, Wells MN, Auble DT. 2009. TATA-Binding Protein Variants that Bypass the Requirement for Mot1 In Vivo. *J Biol Chem* **284**: 4525–4535.

Steinmetz EJ, Conrad NK, Brow DA, Corden JL. 2001. RNA-binding protein Nrd1 directs poly(A)-independent 3'-end formation of RNA polymerase II transcripts. *Nature* **413**: 327–331.

Steinmetz EJ, Ng SBH, Cloute JP, Brow DA. 2006a. cis- and trans-Acting determinants of transcription termination by yeast RNA polymerase II. *Mol Cell Biol* **26**: 2688–96.

Steinmetz EJ, Warren CL, Kuehner JN, Panbehi B, Ansari AZ, Brow DA. 2006b. Genome-Wide Distribution of Yeast RNA Polymerase II and Its Control by Sen1 Helicase. *Mol Cell* **24**: 735–746.

Sun ZW, Hampsey M. 1996. Synthetic enhancement of a TFIIB defect by a mutation in SSU72, an essential yeast gene encoding a novel protein that affects transcription start site selection in vivo. *Mol Cell Biol* **16**: 1557–66.

Suter DM, Molina N, Gatfield D, Schneider K, Schibler U, Naef F. 2011. Mammalian Genes Are Transcribed with Widely Different Bursting Kinetics. *Science (80-)* **332**: 472–474.

Taatjes DJ. 2010. The human Mediator complex: a versatile, genome-wide regulator of transcription. *Trends Biochem Sci* **35**: 315–322.

Taatjes DJ, Näär AM, Andel F, Nogales E, Tjian R. 2002. Structure, function, and activator-induced conformations of the CRSP coactivator. *Science* **295**: 1058–62.

Taatjes DJ, Tjian R. 2004. Structure and Function of CRSP/Med2: A Promoter-Selective Transcriptional Coactivator Complex. *Mol Cell* **14**: 675–683.

Thomas MC, Chiang C-M. 2006. The general transcription machinery and general cofactors. *Crit Rev Biochem Mol Biol* **41**: 105–78.

Thompson NE, Steinberg TH, Aronson DB, Burgess RR. 1989. Inhibition of In Vivo and In Vitro Transcription by Monoclonal Antibodies Prepared Against Wheat Germ RNA Polymerase II That React with the Heptapeptide Repeat of Eukaryotic RNA Polymerase II. *J Biol Chem* **264**: 11511–11520.

Tora L, Timmers HTM. 2010. The TATA Box Regulates TATA-Binding Protein (TBP) Dynamics In Vivo. *Trends Biochem Sci* **35**: 309–314.

Traven A, Jelicic B, Sopta M. 2006. Yeast Gal4: a Transcriptional Paradigm Revisited. *EMBO Rep* **7**: 496–499.

Ursic D, Chinchilla K, Finkel JS, Culbertson MR. 2004. Multiple protein/protein and protein/RNA interactions suggest roles for yeast DNA/RNA helicase Sen1p in transcription, transcription-coupled DNA repair and RNA processing. *Nucleic Acids Res* **32**: 2441–52.

Ursic D, Himmel KL, Gurley K a, Webb F, Culbertson MR. 1997. The yeast SEN1 gene is required for the processing of diverse RNA classes. *Nucleic Acids Res* **25**: 4778–85.

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

Venters BJ, Pugh BF. 2009. A Canonical Promoter Organization of the Transcription Machinery and Its Regulators in the Saccharomyces Genome. *Genome Res* **19**: 360–371.

Viswanathan R, Auble DT. 2011. One small step for Mot1; one giant leap for other Swi2/Snf2 enzymes? *Biochim Biophys Acta* **1809**: 488–96.

Wang W, Carey M, Gralla JD. 1992. Polymerase II Promoter Activation: Closed Complex Formation and ATP-Driven Start Site Opening. *Science (80-)* **255**: 450–453.

Weake VM, Workman JL. 2010. Inducible gene expression: diverse regulatory mechanisms. *Nat Rev Genet* **11**: 426–37.

Van Werven FJ, van Bakel H, van Teeffelen HAAM, Altelaar AFM, Koerkamp MG, Heck AJR, Holstege FCP, Timmers HTM. 2008. Cooperative Action of NC2 and Mot1p to Regulate TATA-Binding Protein Function Across the Genome. *Genes Dev* **22**: 2359–2369.

Van Werven FJ, van Teeffelen HAAM, Holstege FCP, Timmers HTM. 2009. Distinct promoter dynamics of the basal transcription factor TBP across the yeast genome. *Nat Struct Mol Biol* **16**: 1043–1048.

Whitehouse I, Rando OJ, Delrow J, Tsukiyama T. 2007. Chromatin remodelling at promoters suppresses antisense transcription. *Nature* **450**: 1031–1035.

Wilkinson DJ. 2009. Stochastic modelling for quantitative description of heterogeneous biological systems. *Nat Rev Genet* **10**: 122–33.

Winston F, Chaleff DT, Valent B, Fink GR. 1984. Mutations Affecting Ty-Mediated Expression of the HIS4 Gene of Saccharomyces cerevisiae. *Genetics* **107**: 179–197.

Winter RB, Berg OG, von Hippel PH. 1981. Diffusion-Driven Mechanisms of Protein Translocation on Nucleic Acids. 3. The Escherichia coli lac Repressor-Operator Interaction: Kinetic Measurements and Conclusions. *Biochemistry* **20**: 6961–6977.

Workman JL. 2006. Nucleosome Displacement in Transcription. *Genes Dev* **20**: 2009–2017.

Wyers F, Rougemaille M, Badis G, Rousselle J-C, Dufour M-E, Boulay J, Regnault B, Devaux F, Namane A, Seraphin B, et al. 2005. Cryptic Pol II Transcripts Are Degraded by a Nuclear Quality Control Pathway Involving a New Poly(A) Polymerase. *Cell* **121**: 725–737.

Xiao H, Friesen JD, Lis JT. 1995. Recruiting TATA-Binding Protein to a Promoter: Transcriptional Activation without an Upstream Activator. *Mol Cell Biol* **15**: 5757–5761.

Xu C, Hoang S, Mayo MW, Bekiranov S. 2010. Application of machine learning methods to histone methylation ChIP-Seq data reveals H4R3me2 globally represses gene expression. . *BMC Bioinformatics* **11**: 396.

Xu Z, Wei W, Gagneur J, Perocchi F, Clauder-Münster S, Camblong J, Guffanti E, Stutz F, Huber W, Steinmetz LM. 2009. Bidirectional promoters generate pervasive transcription in yeast. *Nature* **457**: 1033–1037.

Xue-Franzén Y, Johnsson A, Brodin D, Henriksson J, Bürglin TR, Wright APH. 2010. Genome-wide characterisation of the Gcn5 histone acetyltransferase in budding yeast during stress adaptation reveals evolutionarily conserved and diverged roles. *BMC Genomics* **11**: 200.

Yu Y, Eriksson P, Bhoite LT, Stillman DJ. 2003. Regulation of TATA-Binding Protein Binding by the SAGA Complex and Nhp6 High-Mobility Group Protein. *Mol Cell Bio* **23**: 1910–1921.

Zanton SJ, Pugh BF. 2004. Changes in Genomewide Occupancy of Core Transcriptional Regulators During Heat Stress. *PNAS* **101**: 16843–16848.

Zenklusen D, Larson DR, Singer RH. 2008. Single-RNA Counting Reveals Alternative Modes of Gene Expression in Yeast. *Nat Struct Mol Biol* **15**: 1263–1271.

Zhang H, Richardson DO, Roberts DN, Utley R, Erdjument-Bromage H, Tempst P, Côté J, Cairns BR. 2004. The Yaf9 component of the SWR1 and NuA4 complexes is required for proper gene expression, histone H4 acetylation, and Htz1 replacement near telomeres. *Mol Cell Biol* **24**: 9424–36.

Zhao J, Hyman L, Moore C. 1999. Formation of mRNA 3' ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis. *Microbiol Mol Biol Rev* **63**: 405–45.