Studies of the Mechanisms of TFIIH

and Noncoding RNAs in Eukaryotic Transcription

by

Jessica Marie Hattle Thrall

B.S., University of Kansas, 2002

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirement for the degree of

Doctor of Philosophy

Department of Chemistry and Biochemistry

2012

This thesis entitled:
Studies of the Mechanisms of TFIIH and Noncoding RNAs in Eukaryotic Transcription

written by Jessica Marie Hattle Thrall
has been approved for the Department of Chemistry and Biochemistry

James A. Goodrich

Jennifer F. Kugel

Dylan J. Taatjes

Robert D. Kuchta

Robin D. Dowell

January 6th, 2012

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

ABSTRACT

Thrall, Jessica Marie Hattle (Ph.D., Biochemistry)

Studies of the Mechanisms of TFIIH and Noncoding RNAs in Eukaryotic Transcription

Thesis directed by Professors James A. Goodrich and Jennifer F. Kugel

The control of eukaryotic transcription is carefully orchestrated and involves many types of regulatory factors. Transcription is the underlying mechanism that controls all cellular processes and when left unchecked results in diseased cell states and cell death. Understanding the detailed mechanisms and processes of eukaryotic transcription is the goal of these studies.

Inspired by our previous eukaryotic transcription kinetic studies, Chapter 2 describes identifying a factor that accelerates the rate of promoter escape. Spiking *in vitro* transcription assays with a nuclear extract resulted in an increase in the rate of *in vitro* transcription from the adenovirus major late promoter. With the understanding that many factors are involved in transcriptional regulation, we hypothesized that a factor could function to enhance the rate of transcription after being recruited to promoters. I set out to purify, identify, and characterize this factor. I developed a rate assay to monitor purification of the factor over several columns. The purified rate-accelerating factor was identified to be the general transcription factor TFIIH. Comparing my purified TFIIH to two standard TFIIH purifications revealed that high concentrations of TFIIH accelerated the rate of early transcription.

Recent studies have identified thousands of long noncoding RNAs (lncRNAs) with the potential to regulate gene expression, some on a single gene level and others potentially

regulating multiple genes through mechanisms controlling chromatin structure. At the time this work was started, there were no genome-wide methods to determine whether these lncRNAs interact directly with chromatin, and if so, where. I developed a method named ChOP-seq to identify the genomic regions with which the lncRNA HOTAIR associates. I was ultimately able to show RNA-dependent enrichment of specific genomic regions using the ChOP technique, identifying a diverse set of genes that may be regulated by HOTAIR. We are positioned to apply our new knowledge of ChOP assays to other ncRNAs. This method has the potential to extend our understanding of the mechanisms that contribute to epigenetic programming.

ACKNOWLEDGEMENTS

I want to thank everyone who encouraged and helped me to achieve my Ph.D. Special thanks to my graduate advisors, Professors Jim Goodrich and Jen Kugel, for their mentorship and allowing me to do my research studies in their lab. Thank you for the career advice and for the careful editing of my dissertation. Thanks also to Dr. Taatjes, Dr. Kuchta, and Dr. Dowell for their time, discussions, and being part of my graduate committee.

I'd like to express my gratitude to my lab-mates Dr. Stacey Wagner, Ryan Walters, Becca Blair, Steve Ponicsan, Ben Gilman, and Dr. Petro Yakovchuk for the discussions, happy hours, and random lab fun. A special shout out to my classmates and friends: Dr. J'aime Manion, Dr. Sarah Altschuler, Dr. AnGayle Vasiliou, Dr. Janet McCombs, Dr. Andrew Olsen, Dr. Amy Gelinas, Dr. Chris Ebmeier, Doug Chapnick, Francis Reyes, and David Wren for their insightful discussions, and support outside of the lab. J'aime thank you for making our first year one of the best imaginable. Special thanks to all of my dear friends inside and out of CU and for all the adventures we have had.

I also want to thank my advisors and friends, Dr. Jeffrey Urbauer and Ramona Urbauer for taking me under their wings as an undergrad, introducing me to the world of research, and for their inspiration and encouragement to pursue my Ph.D. CSU PI's Dr. Orme, Dr. Gonzales-Juarrero, and Dr. Izzo, thank you for my first post-undergraduate job and for your support and endorsement for graduate school.

Enormous thanks to my family for their love and support and endless understanding that graduation was always "two-to-life" away. Dad, thank you for the lifetime of encouragement, for believing in me, and for all of our science discussions. Mom, thank you

for all of your love and our adventures. And to my sisters, thank you for all the welcome distractions, fun, and amusement.

I'd like to express my highest gratitude to my husband Matt and his endless love, for having faith in me, and his encouragement and understanding that helped me scale to the top of this mountain. To Matt, whom I met and fell in love with during my graduate career; I will always be grateful for the love and the support you have given me during this period of my life. Thank you for sharing these happy, but often stressful years with me. I love you.

Thank you everyone, listed and not listed, for helping me achieve this momentous goal in my life. Now, "time to get a real job". Love and Cheers! Jessica.

CONTENTS

CHAPTER 1 1
Eukaryotic gene expression and its regulation
Introduction 2
Mammalian promoter sequences and structure 4
Transcription factors and the transcription reaction 10
Eukaryotic epigenetic and chromatin regulation 16
Long non-coding RNAs regulate the transcription reaction 20

CHAPTER 2 24
Identification and characterization of a factor that accelerates the rate-limiting step of human RNA polymerase II transcription
Summary 25
Introduction 26
Results 30
Discussion 54
Materials and experimental methods 58

CHAPTER 3 63
Chromatin oligonucleotide precipitation-sequencing (ChOP-seq): Developing a new technique to monitor lncRNA occupancy on chromatin
Summary 64
Introduction 65
Results 69
Discussion 94
Materials and experimental methods 99

BIBLIOGRAPHY 107

TABLES

Table 2.1. Purification table of rate-enhancing activity 37

Table 2.2. Mascot identification of peptides 46

Table 3.1. Oligonucleotide probes used in ChOP assays 77

Table 3.2 Summary of technical differences between ChOP, ChIRP, and CHART 97

FIGURES

Figure 1.1. Eukaryotic Pol II promoter elements 5

Figure 1.2. Eukaryotic basal transcription reaction 11

Figure 1.3. Transcriptional distribution of histone modifications 18

Figure 2.1. Rate constants of eukaryotic transcription on a negatively supercoiled AdMLP template 27

Figure 2.2. Minimal transcription assay on negatively supercoiled DNA 31

Figure 2.3. HeLa nuclear extract accelerates the rate of promoter escape 32

Figure 2.4. Activity assay to identify the factor(s) in HeLa nuclear extracts (HNE) responsible for accelerating the rate constant of promoter escape.....34

Figure 2.5. Purification scheme of activity from HeLa nuclear extracts 36

Figure 2.6. Purification of activity over first five chromatography columns.....39

Figure 2.7. Purification of activity over heparin and DEAE-5PW columns 43

Figure 2.8. Mass spectrometry gel.....45

Figure 2.9. Assessing partially purified activity on a linear DNA template 47

Figure 2.10. Partially purified activity requires TFIIE, but not dATP or TFIIH 49

Figure 2.11. Activity immunodepletes with TFIIH 51

Figure 2.12. Various TFIIH preparations give similar accelerated transcription rates on an AdMLP template.....52

Figure 2.13. Schematic of different TFIIH requirements during early transcription.....56

Figure 3.1. HOTAIR knockdown upregulates transcription in HOX D genes and decreases H3K27me3 66

Figure 3.2. HOTAIR acts as a scaffold for PRC2 and LSD1 complexes 68

Figure 3.3. Schematic of the ChOP-seq assay 70

Figure 3.4. HOTAIR is present in HeLa and Jurkat cells 72

Figure 3.5. HOTAIR cellular localization and recovery after cross-linking.....73

Figure 3.6. Chromatin fragmentation using sonication and DNase I.....75

Figure 3.7. HOTAIR sequence and oligonucleotide probe positions 78

Figure 3.8. ChOP assay testing recovery after formaldehyde cross-linking.....80

Figure 3.9. ChOP assay testing NeutrAvidin versus protein A/G beads.....82

Figure 3.10. ChOP assay testing probe specificity 84

Figure 3.11. Analysis of Illumina peak distribution.....86

Figure 3.12. MEME analysis of promoter-proximal and promoter-distal peaks 88

Figure 3.13. ChOP assay testing sequencing peaks 89

Figure 3.14. ChOP assay of HOTAIR in foreskin fibroblasts.....91

Figure 3.15. ChOP-assay testing genes identified by ChIRP-seq of HOTAIR.....93

CHAPTER 1

Eukaryotic Gene Expression and its Regulation

INTRODUCTION

Eukaryotic transcription is a highly regulated, multi-step process in gene expression that involves RNA polymerase II (Pol II) catalyzing the synthesis of mRNA from a DNA template. Regulation of this step is crucial for maintaining cellular viability and differentiation and is controlled via many mechanisms at the chromatin, protein, and DNA sequence levels. Characterizing new mechanisms of transcriptional regulation is key to developing our understanding of gene expression.

The complex multistep process of transcription involves a broad range of factors specific to the DNA template and cell type. Transcription requires RNA polymerase II (Pol II) and, at most promoters, the general transcription factors (GTFs); TFIIA, TFIIB, TFIID, TFIIE, TFIIF, TFIIH, and the co-activator Mediator *(1)*. For many genes, additional factors are also involved in the regulation of this process such as activators, repressors, co-activators, co-repressors, and chromatin remodeling factors *(1-3)*. Furthermore, the promoter DNA from which transcription initiates contains elements that direct transcription *(4)*. General transcription factors and other factors utilize these sequences to bind and regulate gene expression at specific promoters. The transcribed mRNA or noncoding RNA can also influence the process of transcription *(5-8)*.

Although Pol II is responsible for transcription of protein coding genes, the process is collaborative and dependent on many protein-protein interactions within and between general transcription factors. Each interaction can be regulated and networks of interactions achieve successful transcription. Furthermore, the level of transcription at promoters is controlled by gene and cell specific factors, like activators, repressors, co-activators, and even noncoding

RNAs. Here I will provide an overview of Pol II transcription and the DNA, RNA, and protein factors involved.

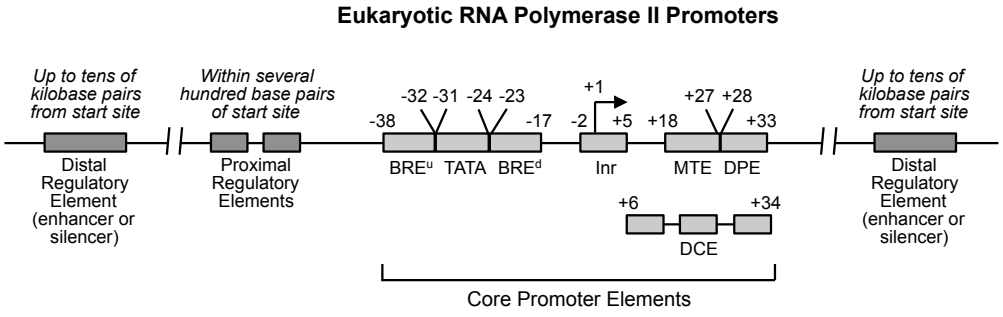
MAMMALIAN PROMOTER SEQUENCES AND STRUCTURE[1]

A promoter is a defined region of DNA that directs the transcription of a gene. Eukaryotic promoters contain two main types of elements: 1) the core promoter elements, which recruit the general transcription machinery and Pol II and set the start site of transcription, and 2) regulatory elements, which recruit sequence specific factors such as activators and repressors that when bound control the level of transcription from the core promoter *(1, 9)*. Each genomic promoter is unique with respect to the elements it contains, the mechanism of regulation, and the amount of transcription it directs. Eukaryotic promoters are discussed below, with an emphasis on well understood promoter elements in eukaryotic protein-encoding genes. The discussion below considers promoters of mammalian and *Drosophila* Pol II transcribed genes, which produce messenger RNA (Figure 1.1)[1]. Eukaryotic promoters contain elements with defined consensus sequences and critical spacing requirements.

Core Promoter Elements

Core promoters in higher eukaryotes span from approximately 40 base pairs upstream to 40 base pairs downstream of the transcription start site. The first discovered and most well studied of the core promoter elements is the AT-rich sequence known as the TATA box, which is located from approximately 24 to 31 base pairs upstream of the transcription start site. The consensus TATA box sequence is TATAWAAR (nontemplate strand; degenerate nucleotide symbols are described in the legend to Figure 1.1), which is conserved from archaebacteria to humans *(10-12)*. During transcription, the TATA binding protein (TBP) of the transcription factor

[1] This section is adapted from **Thrall, JMH** and Goodrich, JA, 2012. Promoters. Encyclopedia of Genetics, 2nd edition. Brenner, S and Miller, J. (eds). Elsevier / Academic Press, 2012. (in Press)



Element	Sequence
BRE^{u}	SSRCGCC
TATA box	TATAWAAR
BRE^{d}	RTDKKKK
Inr	YY<u>A</u>NWYY (human) TC<u>A</u>KTY (*Drosophila*)
MTE	CSARCSSAAC
DPE	RGWYVT (*Drosophila*)
DCE	CTTC from +6 to +11 CTGT from +16 to +21 AGC from +30 to +34

KTY

Figure 1.1. Eukaryotic Pol II promoter elements. Degenerate nucleotide symbols: S= G or C; R= G or A; W= A or T; D= G, A, or T; K= G or T; Y= T or C; V= G, C, or A; N= any nt. The transcription start site is underlined in the Inr consensus sequences.

IID (TFIID) multisubunit complex binds this sequence and bends the DNA to commence recruitment of the transcription machinery *(13-15)*. TATA-containing promoters make up only 10-15% of the mammalian promoters *(12, 16)*. A second core promoter element, the Initiator element (Inr), has the consensus sequence YYANWYY (nontemplate strand) in humans (or TCAKTY in *Drosophila)* and encompasses the transcription start site (underlined in the consensus sequences) *(9, 17)*. The TBP associated factors 1 and 2 (TAF1 and TAF2) of the TFIID complex bind the Inr element *(18)*. The transcription factor IIB (TFIIB) response elements, the BRE^u and BRE^d, are positioned immediately upstream or downstream of the TATA box, respectively *(19-21)*. The BRE^u has a consensus sequence of SSRCGCC and BRE^d has a consensus sequence of RTDKKKK (nontemplate strand). When present in a promoter, TFIIB binds these elements to aid in preinitiation complex formation.

There are also core promoter elements found downstream of the transcription start site. These elements include the Downstream Promoter Element (DPE), the Motif Ten Element (MTE), and the Downstream Core Element (DCE). The DPE has a consensus sequence of RGWYVT in *Drosophila* (nontemplate strand) and spans from 28 to 33 base pairs downstream of the transcription start site *(22)*. It is bound by TAF6 and TAF9 to aid in recruiting the TFIID complex to the promoter. Mutation of the DPE in *Drosophila* promoters causes a reduction in transcription *(23)*. The MTE was identified during biochemical and computational comparisons of *Drosophila* promoter sequences *(24).* The MTE is found just upstream of the DPE from +18 to +27 relative to the +1 start site and has a consensus sequence of CSARCSSAAC (nontemplate strand) *(25)*. The MTE contributes to TFIID binding to the core promoter. Both the DPE and MTE have been found in *Drosophila* and human promoters. The DCE overlaps the same DNA

region as the DPE and MTE, however it is a distinct core promoter element *(26)*. It consists of three short sequences: CTTC from +6 to +11, CTGT from +16 to +21, and AGC from +30 to +34 (all nontemplate strand). The TAF1 subunit of TFIID can associate with the DCE.

Eukaryotic core promoters are not universal in the elements that they contain *(4)*. The MTE can act in synergy with either the TATA box or DPE *(21, 25)*. However, the MTE, and frequently the DPE, can independently act with the Inr in TATA-less promoters *(22, 25, 27)*. The modularity of eukaryotic core promoters allows for a high level of promoter variation, which is likely to impart specificity on the regulation of transcription of each gene *(28)*. Many mammalian core promoters are less well-defined, with multiple transcription start sites that are dispersed over a range of 50 to 100 base pairs. These promoters often contain CpG islands, and generally lack TATA, DPE, and MTE motifs *(9)*. The mechanism of recognition and transcription initiation at these promoters is currently an active area of research.

Regulatory Elements

Eukaryotic regulatory elements can be separated into two groups: promoter proximal regulatory elements that are found within several hundred base pairs upstream of the core promoter, and distal regulatory elements that can be found up to tens of kilobase pairs upstream or downstream of the core promoter (reviewed in *(3, 29)*). The promoter proximal regulatory elements are typically bound by transcriptional activators and repressors, and the spacing between the regulatory elements and the core promoter can be important for the function of the elements. The distal regulatory elements typically contain multiple transcription factor binding sites and can act as enhancers that facilitate the activation of transcription, or as silencers that

repress transcription from the core promoter. The regulatory proteins bound to enhancers or silencers communicate with regulatory factors bound at promoter proximal elements and the general transcription machinery bound at the core promoter, often over great lengths of intervening DNA. There is also evidence that enhancers are tuned to function with the specific arrangements of core promoter elements. For example, some enhancers can activate transcription from a TATA-containing core promoter better than a core promoter containing a DPE *(30)*.

Chromatin Effects on Promoter Function

In eukaryotes, DNA is packaged with histone proteins into chromatin, which in general restricts the accessibility of the transcription machinery to promoters *(15)*. Loosely structured chromatin with sparse nucleosome occupancy promotes accessibility to a promoter region, whereas the opposite leads to transcriptional repression *(31)*. Chromatin structure in promoters is regulated by specific chromatin modifiers that can remodel nucleosomes or post-translationally modify histones *(32, 33)*. Some of these modifications can serve to recruit transcription factors to the promoter. In a broader view, it is not just the DNA that makes up eukaryotic promoters, but the chromatin. Regulation of chromatin structure is discussed later in this section.

Future Research on Promoters

Transcription is a tightly controlled process that will be an important and active area for future research. Promoter elements will continue to be discovered and characterized, especially in eukaryotes. In this quest, experimental researchers will become more dependent on computational approaches to identify DNA sequences that could function as promoter elements.

Together experimental and computational research will ultimately provide a more complete picture of promoters and their elements.

TRANSCRIPTION FACTORS AND THE TRANSCRIPTION REACTION

The transcription reaction can be divided into at least 5 steps: pre-initiation complex formation, initiation, promoter escape, elongation, and termination (Figure 1.2) *(34)*. Basal transcription begins with the recruitment of Pol II and transcription factors to form the transcription preinitiation complex (PIC). Initially the TATA-binding protein (TBP) subunit of TFIID recognizes and binds the 8-basepair TATA box for TATA-containing genes. TBP inserts phenylalanines into the minor groove and dramatically bends the promoter DNA rendering it accessible for subsequent GTF binding *(35)*. TFIID is a 750 kDa multi-subunit complex consisting of TBP and 14 TBP-associated factors (TAFs), some of which also bind the core promoter DNA *(21)*. The TAFs that recognize and bind to specific core promoter elements (e.g. TBP, TAF1, TAF2, TAF6, and TAF9) were discussed earlier in this chapter. TFIID is important for mediating interactions between many gene and cell-specific coactivators and activators and the general transcription machinery *(3)*. However, *in vitro* TATA-containing promoters need only TBP and not the entire TFIID complex to achieve basal transcription *(36, 37)*.

Next, TFIIA and TFIIB recognize the promoter-bound TBP, or TFIID, and stabilize the DNA:protein complex *(38)*. TFIIA binds upstream of the TATA box and aids in PIC formation by increasing the affinity of TBP for promoter DNA. TFIIB makes asymmetric contacts with the DNA both upstream and downstream of TBP to ensure directionality of transcription *(19)*. TFIIB also serves to recruit Pol II and TFIIF to the promoter and aids Pol II in start site selection through it's amino-terminal domain *(39, 40)*.

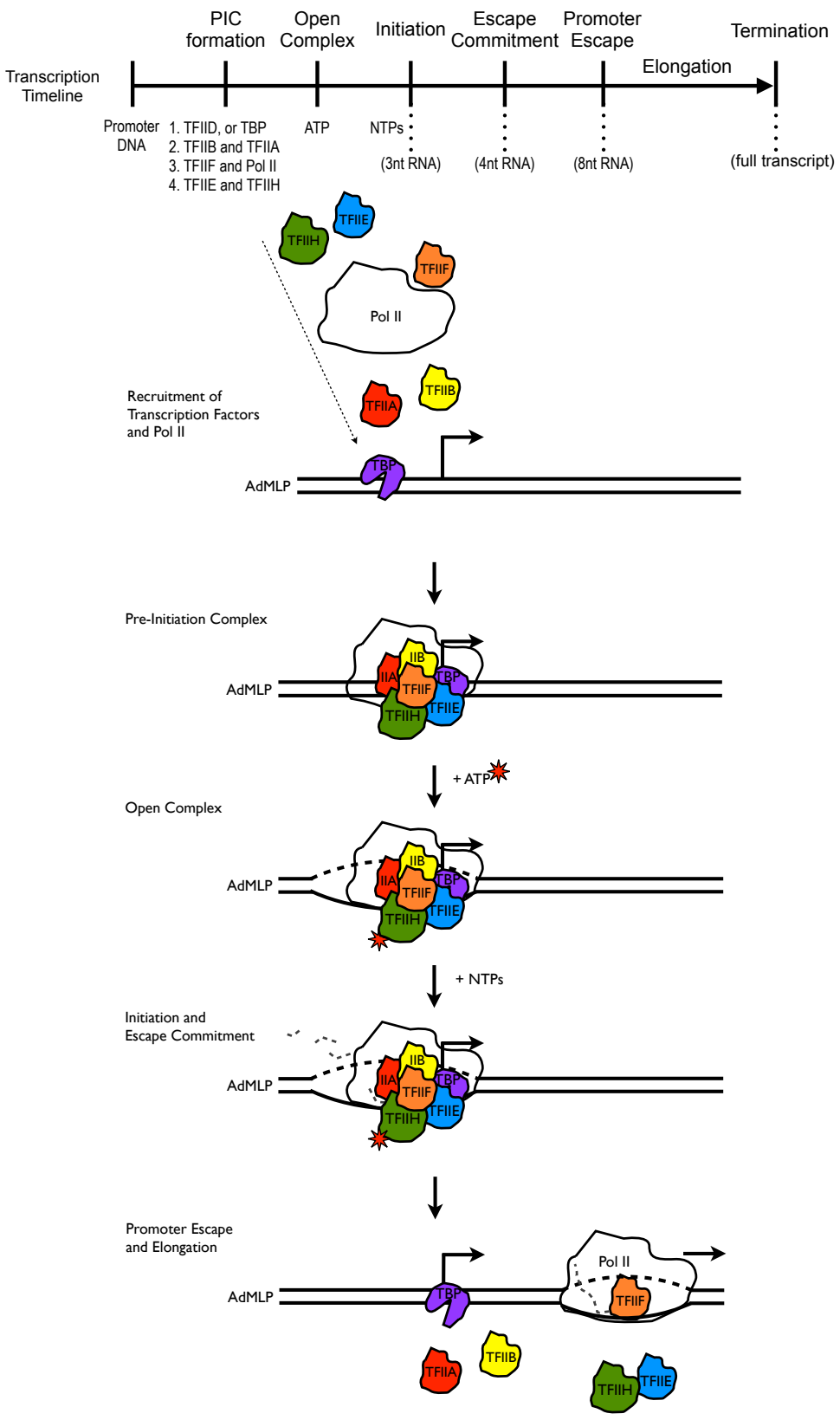


Figure 1.2. Five distinct steps of eukaryotic basal transcription.

Next, Pol II is recruited to the PIC along with TFIIF *(41)*. TFIIF consists of the RAP30 and RAP74 subunits and serves many functions throughout the transcription process including collaborating with TFIIB for start site determination, enabling promoter escape, and increasing Pol II efficiency during elongation *(42, 43)*. Recently, TFIIF has been implicated in stabilizing TFIIB to the promoter *(43)*. TFIIF also facilitates recruitment of TFIIE and TFIIH to the promoter.

Pol II is a 12 subunit complex that directs both messenger and noncoding RNA transcription. Pol II consists of four mobile lobes that, while carrying out transcription, make and break many protein-protein interactions with various GTFs and co-activators *(44-46)*. The C-terminal domain (CTD) of the largest subunit of human Pol II contains 52 heptapeptide repeats with the consensus sequence YSPTSPS and undergoes a phosphorylation cycle during the different stages of transcription where the 2nd and 5th/7th serines can be phosphorylated respectively by CDK9 of P-TEFb and CDK7 of TFIIH *(47-49)*. The CTD serves as a flexible platform that interacts with various transcription factors and other proteins including components of the splicing machinery (see reviews *(50, 51)*).

Transcription factors IIE and IIH primarily act after PIC formation. First, TFIIE enters the PIC and recruits TFIIH to facilitate promoter melting and initiation in the presence of an ATP energy source *(52, 53)*. TFIIE is a heterotetramer composed of the subunits alpha (E56) and beta (E34) *(54)*. TFIIE is thought to modulate the activity of TFIIH *(55)*. The alpha subunit is responsible for interactions with Pol II, and TFIIH, and regulates the helicase and kinase activities of TFIIH *(56, 57)*. Specifically, TFIIEα stimulates the ATPase and kinase

activities, but represses the helicase activity of TFIIH *(58-60)*. The TFIIE beta subunit interfaces with Pol II, TFIIB, RAP30 of TFIIF, TFIIH, and dsDNA *(61)*. TFIIE binds Pol II near the catalytic center and aids in preparing Pol II for initiation and facilitates promoter escape *(62)*.

Human TFIIH consists of 10 subunits. TFIIH comprises a core of XBP/ERCC3 (p89) helicase, XPD/ERCC2 (p80) helicase, p62, p52, p44, p34, and TFB5 (p8), which is bound by the cdk-activating kinase (CAK) subcomplex containing CDK7 (p40), cyclin H (p38), and the activating protein MAT1(p32) *(60, 63, 64)*. XPD can also associate with the CAK subcomplex outside of TFIIH *(65)*. TFIIH XBP helicase is responsible for promoter melting and open complex formation using the energy from ATP hydrolysis during transcription initiation *(66, 67)*. XBP is also required for nucleotide transitions leading up to promoter escape and is implicated in promoter clearance *(52, 53, 68)*. With the aid of Cyclin H, the CDK7 kinase subunit of TFIIH phosphorylates serine 5 and serine 7 within the repeats on the CTD of Pol II *(48, 69)*. This CTD phosphorylation allows the transcription complex to progress from initiation into elongation. Upon association with TFIIH, CAK exhibits substrate specificity for phosphorylating TFIIEα, TFIIFα (RAP74), and TBP *(56, 59, 65, 70)*. The CAK sub-complex can also phosphorylate other substrates to promote cell cycle progression *(71)*. In addition to functioning in transcription, the subunits of TFIIH are involved in DNA repair and E3 ubiquitin ligase activity *(72-75)*. TFIIH is a target for transcriptional activation *in vitro* to increase productive transcription, suppress promoter proximal pausing, and stimulate promoter escape *(62, 76-81)*.

In vitro transcription on a linear DNA template requires TFIIH, TFIIE, and hydrolyzable ATP to unwind the double-stranded DNA template and for Pol II to proceed through early

transcription *(53, 82-84)*. Pol II can abortively initiate transcription on a linear template without TFIIE, TFIIH, and dATP *(53)*. Without TFIIH, Pol II tends to stall in the promoter proximal region following promoter escape *(78)*. However, TFIIE and TFIIH are not necessary for transcription using a promoter contained within a negatively supercoiled template, or on DNA with a preformed bubble *(53, 82, 84-89)*. Furthermore, TFIIA is also not required for basal transcription in an *in vitro* transcription system constructed from purified factors *(90)*.

Following PIC formation, the TFIIH helicase melts the promoter DNA forming an open complex when ATP is present *(82)*. Initiation begins with the formation of the first phosphodiester bond in the RNA followed by abortive transcription of 2 and 3 nucleotide (nt) RNAs *(91-93)*. After transcription of a 4 nt RNA, the PIC, DNA, and RNA transform into a stable ternary complex; this transition is named escape commitment, and is facilitated by TFIIE and TFIIH *(34, 94-96)*. The upstream edge of the transcription bubble lies at approximately the -9 position relative to the TSS and is expanded 18 nucleotides downstream during early transcription. When the RNA becomes 7 nucleotides long the upstream region of the bubble collapses to form a ~10 nt bubble *(89)*. After this point an active TFIIH helicase is no longer required *(89, 97-99)*. During promoter escape many protein-protein interactions are broken, facilitating the advancement of the complex into elongation. For example, a recent study found TFIIB releases from the promoter during promoter clearance immediately after formation of a 13 nt RNA, and the bubble collapse transition *(100, 43)*.

After promoter escape, the transcription complex enters into elongation. TFIID, or TBP remains bound at the promoter, poised to commence another round of transcription, whereas Pol II, along with TFIIF, continues to elongate the nascent RNA *(102)*. During early

elongation the other transcription factors (TFIIB, IIE, IIH) detach *(43, 102-103)*. Once the Pol II clears the promoter, a new round of initiation can occur. Transcription termination releases the nascent RNA and Pol II, and re-initiation of transcription can occur on the promoter.

EUKARYOTIC EPIGENETIC AND CHROMATIN REGULATION

Transcription is a much more complex process than discussed above, involving activators and cofactors for specificity, as well as transcription machinery components that are cell-type and promoter-specific. Furthermore, genomic DNA is condensed into chromatin which requires activators to recruit chromatin remodeling factors and modifiers to derepress the chromatin in order to grant the GTFs and Pol II access *(31-33)*. DNA in the cell is wrapped around histone octamers to form nucleosomes consisting of 147 basepairs of DNA and two molecules each of histones H2A, H2B, H3, and H4. These can be further condensed into coiled coils to form highly condensed 30 nm chromatin fibers. Chromatin in regions of the genome exists as active euchromatin, repressed heterochromatin, or a combination of both *(32)*. Promoter regions of constitutively active genes often have nucleosome-depleted regions (NDRs) immediately upstream of their transcription start sites (TSS) and contain poly(dAdT) stretches that block nucleosome formation *(104-106)*. Furthermore, sites that bind activators frequently lie in the NDR regions.

Chromatin structure is highly dynamic and controlled by chromatin remodeling complexes and histone modifying factors in response to gene expression requirements. Chromatin remodelers are enzymes that reorganize nucleosomes by sliding, displacing, or exchanging histones to mediate transcriptional accessibility *(104)*. Nucleosome displacement occurs during transcription, along with nucleosome recycling and histone substitution during elongation *(107, 108)*. ATP-dependent chromatin remodeling complexes are surmised to aid in Pol II progression through nucleosomes in gene bodies *(107)*. Nucleosome positions vary

for different genes, and depending on the position of the nucleosome downstream of the start site it may promote promoter-proximal pausing for that gene (reviewed in *(108, 109)*).

Histone modifying complexes covalently modify the histones in nucleosomes, most often on the highly conserved, unstructured N-terminal tails. These modifications may serve to disrupt intermolecular interactions to allow structural remodeling into either active or higher-ordered repressive chromatin, and to position distinctive chromatin marks for recognition by non-histone proteins *(110)* (reviewed in *(32)*). Histone modifications include methylation (me), acetylation (ac), ubiquitination (ub), phosphorylation (p), sumoylation (su), and ADP-ribosylation (ADPr) *(31, 33, 108)*.

Genome-wide studies have found patterns of chromatin marks surrounding transcription start sites with respect to active or repressed expression (reviewed in *(31)*). Specific acetylation, methylation, phosphorylation, and ubiquitination marks are associated with active gene expression, whereas, other methylation and ubiquitination marks, as well as sumoylation are associated with repression. These modifications and their associated transcriptional responses are summarized in Figure 1.3 *(31)* (data reviewed in *(33, 108, 111)*). The following marks display the hallmarks of transcription initiation on chromatin: Pol II occupancy, the presence of nucleosomes with trimethylated lysine 4 on histone H3 (H3K4me3), and H3K9 and K14 acetylation *(110, 112)*. Activators bound at promoters additionally recruit coactivators that may also remodel and modify chromatin. Low levels of methylation (mono- and di-methylation), acetylation, and phosphorylation often mark inactive genes and require further modification in order to be transcribed. Furthermore, modifications can work in conjunction with each other to tailor GTF-activator interactions

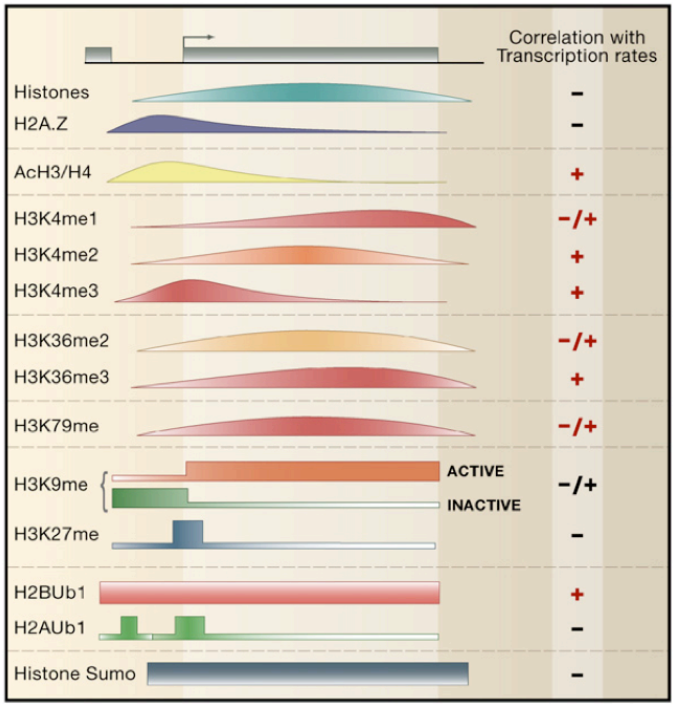


Figure 1.3. Distribution of histone modifications relative to an arbitrary gene promoter (left most) or gene body (light region). The column on the right indicates positive or negative correlation with gene expression. Curved modifications represent data from genome-wide studies and rectangle modifications from a few studies. (Figure from Li et al. 2007 *(31)*).

and expression levels. Chromatin marks also work on a broader scope by preserving inheritable epigenetic gene patterns that are vital to mammalian cellular differentiation and proliferation *(33)*. A differentiated cell is imprinted with the genetic memory from the previous generation, however these mechanisms are not fully understood.

LONG NON-CODING RNAS REGULATE THE TRANSCRIPTION REACTION

Transcriptional regulators can target each of the individual transcription steps and factors to fine-tune gene expression. Activators can recruit factors to the promoter, stabilize the PIC, and interact with the GTFs, or stimulate promoter escape to increase expression *(113)*. Furthermore, regulators can control gene expression on a global scale by regulating chromatin accessibility *(31-33, 114)*. Long noncoding RNAs (lncRNAs) have been found to act as regulators of gene expression and some lncRNAs regulate multiple regions across the genome *(114-116)*.

LncRNAs regulate gene expression

LncRNAs are RNAs longer than 200 bases that do not encode a protein. Multiple genome-wide studies have identified thousands of uncharacterized lncRNAs associated with a variety of cell types *(117-119)*. It has recently been speculated that there are over 8000 large intergenic ncRNAs (lincRNAs) and up to 40% of these are transcribed from regions containing H3K4me3/H3K36me3 marks, indicative of transcriptionally active chromatin domains *(119)*. Many lncRNAs have also been found associated with the histone modifying complex PRC2, which is responsible for the repressive trimethylation of H3K27 *(117, 118)*. These lncRNAs are believed to regulate the expression of a wide variety of genes due to the ubiquitous nature of PRC2. LncRNAs have been categorized to function as signals, decoys, guides, scaffolds, and as transcriptional effectors or enhancers through cis-acting or trans-acting mechanisms and, furthermore, some lncRNAs fit into more than one of these categories *(115, 116, 120)*. LncRNA have been identified functioning at the epigenetic level, including controlling imprinting (H19

(121), Kcnq1ot1 *(122)*), development (HOTAIR *(123)*, COLDAIR *(124)*), and also in cell cycle cycle *(125)*, pluripotency, and differentiation *(126)*.

The most widely known lncRNA in epigenetic programming is the X_i specific transcript (Xist), which is responsible, along with co-regulator lncRNAs (Tsix, Xcite, and RepA), for the inherited X chromosome inactivation in female mammalian cells *(127-129)*. Xist coats the chromosome from which it is expressed and recruits PRC2 to place repressive histone modifications for persistent transcriptional inactivation *(130)*.

Two intronic and paternally expressed lncRNAs, Kcnq1ot1 and Air, act in *cis* in a cell-type and lineage specific manner to control gene expression. The 91 kb lncRNA Kcnq1ot1 interacts with the histone methyltransferases G9a (for H3K9me3) and PRC2 (for H3K27me3) to place chromatin marks for bidirectional control to repress the Kcnq1 gene cluster *(118, 122)*. Air is a 108 kb imprinted, paternally expressed intronic lncRNA that silences 8-10 genes over a 400 kb region on the maternal allele, including the Igf2r gene cluster and the distal genes Slc22a3 and Slc22a2. Air acts through recruiting G9a methyltransferase to these loci to convert chromatin to a silenced state *(131, 132)*.

Several lincRNAs have enhancer-like functions *(133, 134)*. One group used knockdown studies to evaluate actively expressed lncRNAs greater than 1kb from known protein genes, and found seven novel lncRNAs that act as activators of neighboring genes in various cell types *(133)*. A separate large scale study found transcription of evolutionarily conserved enhancer ncRNAs (eRNAs) found in enhancer regions correlated with transcription at nearby promoters. The function of eRNAs has not been established, however the authors hypothesized that eRNAs may deliver Pol II to promoters to activate transcription of nearby genes or that possibly

transcription of these eRNAs may serve in maintaining chromatin marks *(134)*. The promoters of eRNAs are found marked with high levels of H3K4me1 and low H3K4me3 histone modifications *(134, 135)*.

Other lncRNAs regulate transcription at the promoter, such as *Evf-2 (136)*, Alu *(137)*, and the DHFR lncRNA *(138)*. *Evf-2* is an intronic ncRNA from the ultraconserved region between the Dlx-5 and Dlx-6 genes. *Evf-2* acts as a coactivator with the Dlx2 transcription factor to direct expression of the Dlx-5/6 genes *(136)*. Alu RNA is induced upon heat shock and acts to repress transcription by disrupting contacts between Pol II, the GTFs, and the promoter DNA *(137)*.

Some lncRNAs act as developmental switches to activate or inactivate families of genes during development (e.g. HOTAIR *(123)*, COLDAIR *(124)*, HOTTIP *(139)*, and HOTAIRM1 *(140)*). HOTAIR is a 2.2 kb lncRNA that is implicated in regulation of the HOXD genes during development. HOTAIR has been shown to act *in trans* as a scaffold and a guide to recruit the histone modifying complexes PRC2 and LSD1/REST/COREST, a H3K4 demethylase *(141)*. HOTAIR is temporally and spatially expressed during development in specific posterior and distal cell types. A knockdown study in primary fibroblast cells implicates HOTAIR in repressing the gene expression of a set of HOXD genes. Homeobox transcription factor (HOX) genes are dispersed across four separate chromosomal loci and are expressed collinearly and temporally along their chromosome with respect to their spacial body positioning *(142, 143)*. HOTAIR lncRNA has also been implicated in cancer progression and metastasis *(144)*.

Most of the lncRNA present in cells have yet to be investigated functionally. The scientific community is in need of a method to identify the locations of lncRNA association

with the genome. Genome-wide identification of the genomic regions directly associated with lncRNAs is a necessary next step to understanding their function as transcriptional regulators.

Transcription is a complex process requiring many factors and an intricate regulatory system and there is still much to be learned about eukaryotic gene expression. The focus of the studies discussed in these following chapters are factors that regulate eukaryotic transcription at the promoter (Chapter 2) and at the epigenetic level (Chapter 3). Chapter 2 of this dissertation describes the purification and identification of a factor that accelerates the rate-limiting step of human RNA polymerase II transcription in an *in vitro* system. The work in Chapter 3 describes the development of a chromatin oligonucleotide precipitation assay, called ChOP-seq, as a tool to identify the genomic regions with which lncRNAs associate. Moreover, I used ChOP-seq to investigate the well studied lncRNA HOTAIR.

CHAPTER 2

Identification and characterization of a factor that accelerates the rate-limiting step of human RNA polymerase II transcription

CHAPTER 2

SUMMARY

Eukaryotic transcription consists of a series of complex events that are carried out and regulated by a group of factors that include the general transcription factors, cofactors, activators, repressors, and other accessory factors. Furthermore, each step in the process of transcription can be regulated. Promoter escape has been determined to be the rate-limiting step of transcription on the adenovirus major late promoter using a minimal transcription system consisting of TBP, TFIIB, TFIIF, and RNA polymerase II (Pol II). Using an *in vitro* transcription system, I set out to identify a factor from HeLa cells that accelerates the rate of promoter escape. To do this I used a combination of column chromatography, a rate assay to test column fractions, and mass spectroscopy. TFIIE was found to be necessary for the rate-enhancing activity, but not sufficient. After extensive purification, a group of factors were identified and TFIIH was determined to be the factor responsible for this rate accelerating activity. Moreover, the amount of TFIIH needed to accelerate the rate of promoter escape is different than that required to initiate transcription on a linear template.

INTRODUCTION

In vitro transcription experiments can be highly controlled, therefore, the mechanisms of transcription regulation can be studied in detail. For example, the reaction conditions can be individually modified to create order-of-addition experiments and experiments to monitor distinct reaction steps or recruitment of factors. Furthermore, *in vitro* experiments can be used to mechanistically study how transcription is regulated.

The transcription reaction can be examined using a purified *in vitro* transcription system. We have used a minimal transcription system consisting of recombinant TBP, TFIIB, TFIIF, and purified native Pol II along with a negatively supercoiled DNA template. In this system the negatively supercoiled DNA negates the need for TFIIE and TFIIH *(53, 85, 87)*. Furthermore, TBP can be used in place of holo-TFIID on a TATA containing promoter for examination of basal level transcription *(36, 37, 85)*. Using this *in vitro* transcription system and the adenovirus major late promoter (AdMLP), our lab has kinetically determined that the Pol II transcription reaction minimally consists of the following steps: preinitiation complex formation, initiation, escape commitment, promoter escape, transcript elongation (Figure 2.1) *(34)*. Activators and repressors can potentially regulate each of these steps *(145-152)*.

Initially, Pol II and the general transcription factors must be recruited to assemble on the promoter DNA to form preinitiation complexes. In an ordered assembly at TATA-containing genes, TBP binds in the DNA minor groove and bends the DNA *(14, 35)*. TFIIA and TFIIB can then bind their respective DNA response elements and enter the complex *(38)*. TFIIF and Pol II then bind and recruit TFIIE and TFIIH *(58)*. Once assembled, the TFIIH helicase melts the promoter DNA to form an open DNA complex and the PIC is then ready to engage in

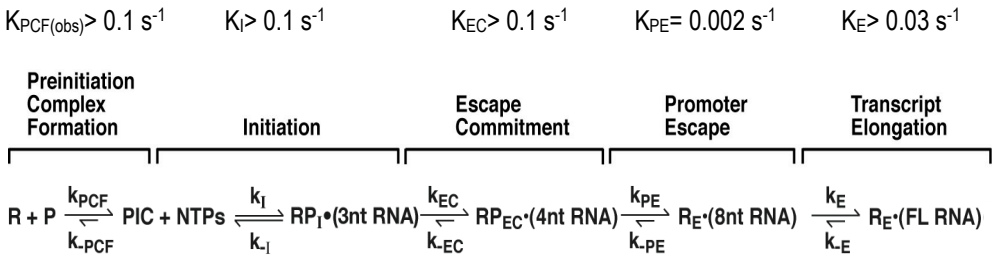


Figure 2.1. Rate constants of eukaryotic transcription on negatively supercoiled AdMLP templates. Experimentally determined rate constants using the AdMLP are listed above each step, where R= general transcription factors (TBP, TFIIB, TFIIF, and Pol II), P= promoter DNA (AdMLP), PIC= preinitiation complex, RP_I= initiated complex containing a 3 nt RNA, RP_{EC}= escape-committed complex containing a 4 nt RNA, R_E= elongation complex containing a 8 nt RNA or 390 nucleotide RNA *(34, 97)*.

initiation *(67)*. During initiation, the polymerase produces 2 to 3 nucleotide abortive transcripts *(93)*. As the 4th nucleotide is added, stable ternary complexes form between the RNA, DNA, and proteins *(95, 153)*. This is termed escape commitment *(34)*. The transcripts then elongate to 8 nucleotides where the complexes transition from initiation complexes into elongation complexes, completing promoter escape *(97)*. During early transcription, multiple protein-protein and protein-nucleic acid interactions are broken and new ones are created as the polymerase moves away from the transcription start site *(154)*.

Our kinetic studies determined promoter escape to be the rate-limiting step on the AdMLP in the minimal transcription system *(34, 91)*. More specifically, we have determined that the rate-limiting step during promoter escape on the AdMLP promoter occurs after the synthesis of an 8 nt RNA and during the translocation of the Pol II active site to the 9th register *(97)*. The rate of promoter escape varies on several tested promoters, however, it remains the rate-limiting step *in vitro (155)*. Other events that occur during promoter escape include changes in the transcription bubble, release of TFIIB, transcript slippage, pausing of the ternary complex, and stable RNA:DNA hybrid formation *(89, 98, 102, 156, 157)*. After promoter escape, the transcript is elongated and ultimately transcription is terminated.

Being the rate-limiting step, promoter escape is a likely target for regulation *(91)*. How the rate of this step can be enhanced was not understood and therefore was the goal of the studies described in this chapter. Using chromatography techniques and transcription rate assays, I purified a factor that accelerates the rate of promoter escape. The purified factor containing this rate-enhancing activity was identified to be the previously characterized and well-studied general

transcription factor, TFIIH. These findings illustrate a new activity for TFIIH in regulating the early steps of transcription.

RESULTS

HeLa and Jurkat nuclear extracts contain an activity that accelerates the rate of promoter escape on an AdMLP template

The transcription assays performed in these studies utilized the adenovirus major late core promoter (AdMLP) from -53 to +10, fused to a 380-base pair G-less cassette in a negatively supercoiled plasmid. This promoter contains A nucleotide bases at +1 and +16 (non-template strand), enabling us to make a 16 nt RNA *in vitro* by initiating transcription with a limited nucleotide mixture containing ApC, UTP, [α-^{32}P]-CTP, and the chain terminator 3'-O-methyl ATP (Figure 2.2A). Figure 2.2B depicts the transcription assay timeline. Preinitiation complexes containing purified TBP, TFIIB, TFIIF, and Pol II were allowed to assemble on the promoter DNA for 10 minutes at 30^{o}C then the nuclear extract or sample being tested for rate-enhancing activity was added and allowed to incubate for 5 minutes after which limited nucleotides were added. Transcription was quenched at multiple time points to obtain the rate of transcript formation. Transcript quantity was determined using gel electrophoresis and phosphorimagery, and plotted versus time to determine the single-exponential rate constant for the reaction. The rate constants were calculated using the equation $y = y_{max}(1-e^{-kt})$, where y is the quantity of transcript. Negative controls for rate-enhancement used buffer instead of nuclear extract or chromatography fraction, and positive controls included the input to the column.

Addition of HeLa nuclear extract (HNE) to the transcription assay outlined in Figure 2.2B caused an 8-fold increase in the rate of transcription, taking it from 2.1 x 10^{-3} s^{-1} to 17.6 x 10^{-3} s^{-1} as seen in Figure 2.3. Jurkat nuclear extract also increased the rate (data not shown). The rate-limiting step of transcription on the AdMLP was previously found to be promoter escape, or the step between escape commitment and transcript elongation, therefore we proposed the rate

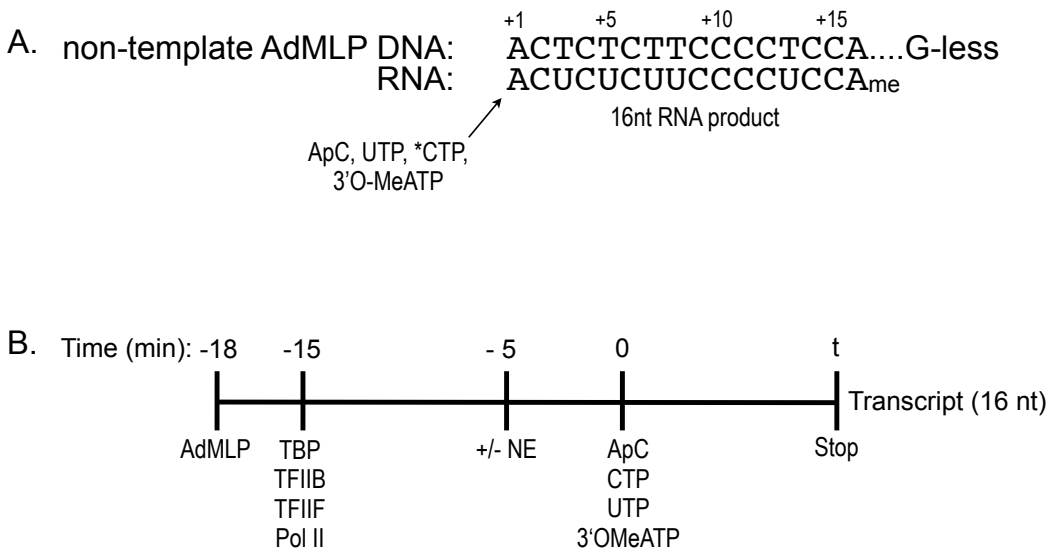


Figure 2.2. A) A method for separating early steps of transcription from transcript elongation. The rate of promoter escape is determined using ^{32}P-labeled CTP and the dinucleotide ApC to initiate transcription and the chain terminator 3'O-Methyl ATP to produce a 16-nucleotide radiolabeled RNA easily resolved by gel electrophoresis and detected by phosphorimagry. The AdMLP non-template strand sequence is shown above the transcript sequence. B) Experimental time course shows optimized order of addition and incubation times. NE= nuclear extract or fraction from chromatography.

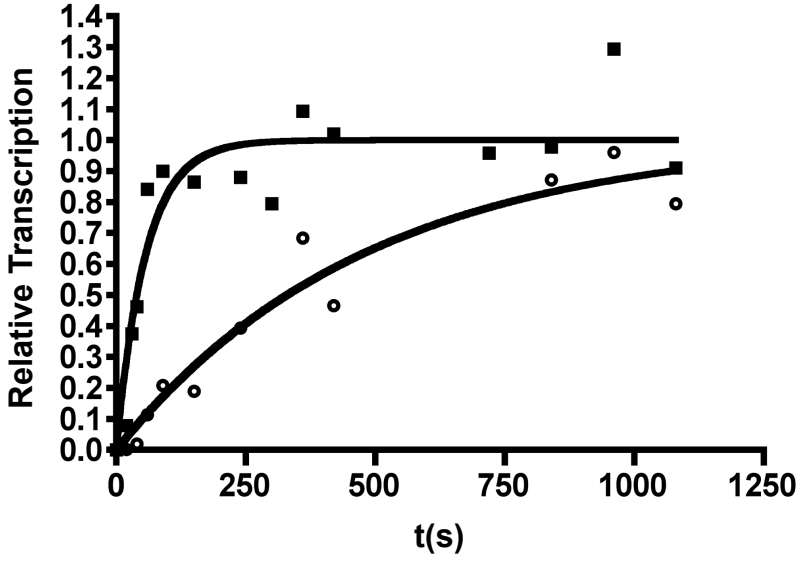


Figure 2.3. HeLa nuclear extract accelerates the rate of promoter escape. Rate of full length transcript synthesis with (squares) and without (circles) HeLa nuclear extract. $k_{+extract}$= 17.6 x 10^{-3} s^{-1}, $k_{-extract}$= 2.1 x 10^{-3} s^{-1}. Data were fit to the equation $y= y_{max}(1-e^{-kt})$.

increase was due to the effect of increasing the rate constant of promoter escape. Control reactions showed that the small amount of nuclear extract added to reactions did not produce transcription without the addition of purified transcription factors (data not shown). It was previously determined in our lab that the transcription factors TFIID, TFIIA, TFIIE, TFIIH, mediator, or the elongation factor TFIIS at levels used for typical *in vitro* transcription assays were not sufficient to produce the accelerated rate (data not shown).

A two-timepoint activity assay can be used to detect the rate enhancement.

I can quantify the rate of transcription by performing a transcription assay in which the reaction is terminated at numerous time points and the radiolabeled transcript for each time point is measured. However, obtaining a comprehensive rate curve for fractions eluting from a chromatography column would be slow and tedious. In order to easily screen chromatography fractions, I developed the two-timepoint activity assay shown in Figure 2.4. Comparing the data from the full rate curve described above to that from our two-timepoint assay, I found that quantifying or visualizing the ratio of transcript produced at 15 and 600 seconds gave an easily discernible difference between rate-accelerated and basal level transcription. I used this two-timepoint assay throughout purification of the activity. This assay also included TFIIE, TFIIH, and dATP in addition to the factors in the minimal system previously described. Including TFIIE, TFIIH, and dATP throughout the purification enabled transcription to occur even if factors in the extract nicked the supercoiled template.

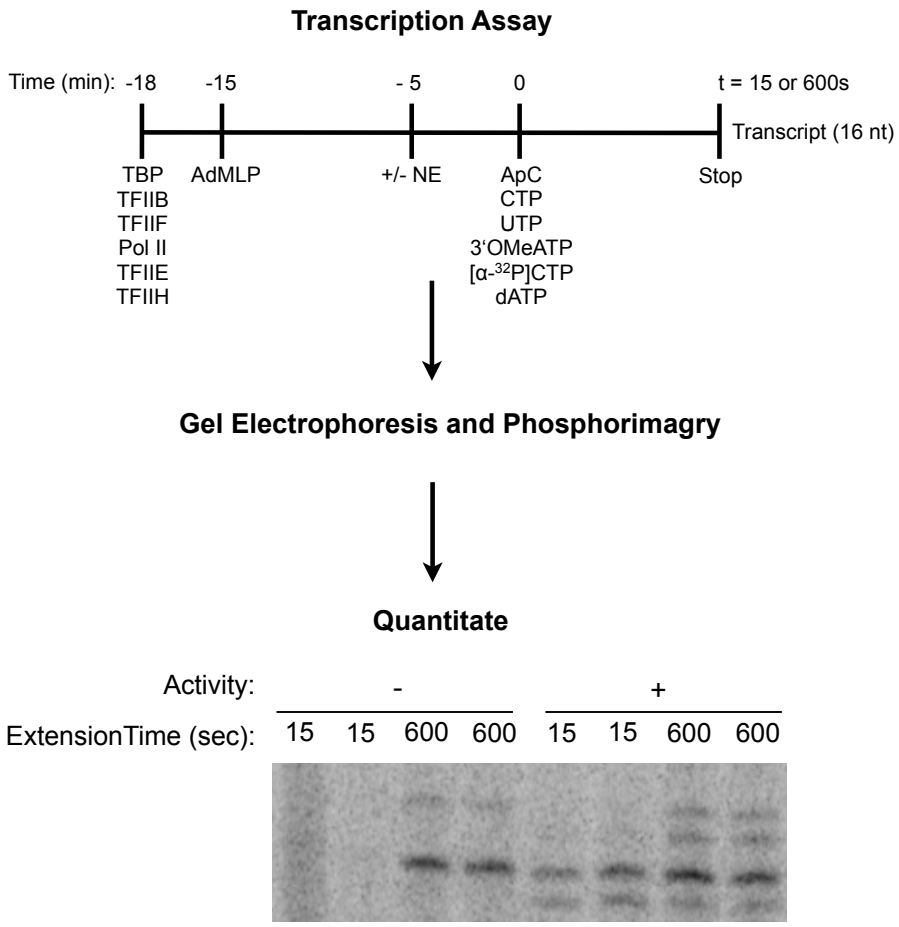


Figure 2.4. Activity assay used to identify the factor(s) in HeLa nuclear extracts (HNE) responsible for accelerating the rate of promoter escape. The *in vitro* transcription assay uses the AdMLP promoter and radiolabeled CTP to monitor 16-nucleotide product formation at 15 and 600 second timepoints. By calculating or visualizing the ratio of product at 15 s to 600 s I can identify the chromatography fractions that contain the activity.

Purification of the activity from HeLa nuclear extracts

After developing the two-timepoint screening assay that would be used to identify the chromatography fractions that contained the activity during purification, I began pilot studies using small batches of HeLa nuclear extract to test how the activity fractionated after step-wise salt elution over various chromatography columns. Fractions were assayed for total protein concentration using the Bradford assay and activity was detected using the two-timepoint assay. The samples were pooled according to activity. Titration of the eluates into the transcription assay determined the volume required to observe the accelerated rate of transcription. A variety of chromatography columns were tested and the final purification scheme that was used in two subsequent large scale purifications from 50 L of HeLa cells is shown in Figure 2.5. Total protein and activity units (U), defined by the volume used to elicit an accelerated rate of transcription, were used to determine percent recovery and fold purification (Table 2.1). The minimum volume required for rate-enhancement, and protein concentration, was not determined for the later purification steps due to dilution of activity, or in some instances to preserve material, however total activity, percent recovery, and fold purification for the latter chromatography steps were expected to be much greater.

DEAE chromatography

I utilized DE52, di-ethyl-amino-ethyl (DEAE), gravity flow anion exchange resin with a step-wise salt elution to begin purifying the activity. The 50 mL column was equilibrated and the nuclear extract loaded with Buffer D containing 50 mM KCl (D.05) and then the column was washed with 5 column volumes of Buffer D.05, D.3 (0.3 M KCl), and D.5 (0.5 M KCl) for

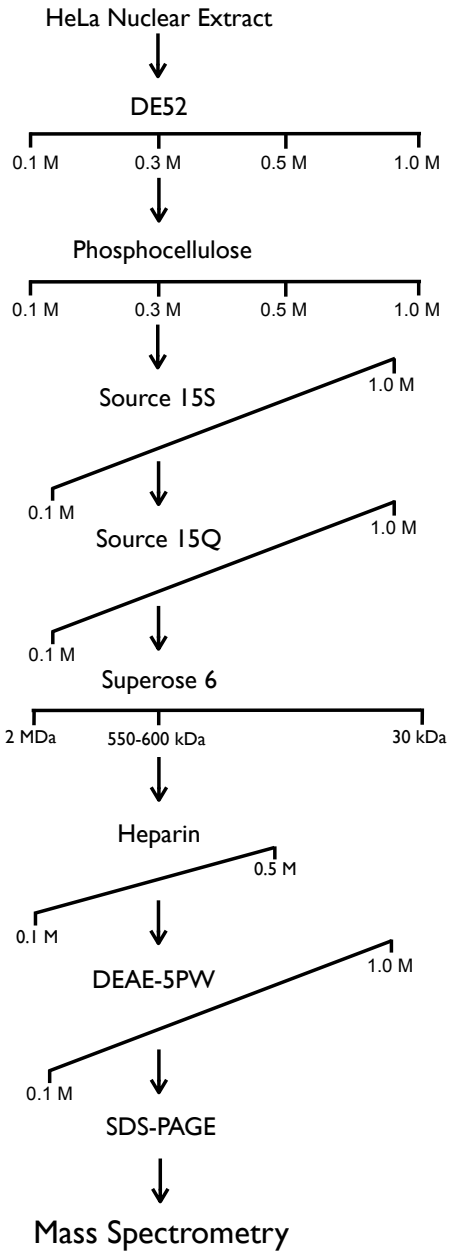


Figure 2.5. Purification scheme of activity that accelerates the rate of promoter escape from HeLa nuclear extracts. Arrows indicate the approximate KCl concentrations (angled lines indicate gradient salt elution) where activity eluted for each column. Eluate from the DEAE-5PW was prepared and submitted for mass spectrometry.

Chromatography Column	Protein (mg)	Volume (ml)	Concentration (mg/ml)	Activity (U)	Total Activity (U*mL)	Spec. Act. (U/mg)	% Recovery	Fold Purification
HeLa Nuclear Extract	238	56	4.2	1.0	56000	236	100	1
DEAE 52	9.1	90	0.10	1.0	90000	9934	161	42
Phosphocellulose	nd	25	nd	1.0	25000	nd	45	nd
Source 15S Sepharose	1.9	4.5	0.42	1.0	4500	2400	8	10
Source 15Q Sepharose	1.3	1.4	0.99	0.5	2700	2022	5	9
Superose 6	nd	1.5	nd	2.0	750	nd	1.3	nd
Heparin	nd	0.3	nd	1.0	260	nd	0.5	nd
DEAE-5PW	nd	1.5	nd	2.0	750	nd	1.3	nd

Table 2.1. Purification table of rate-enhancing activity. Protein concentration was measured by Bradford Assay. Activity unit (U) was measured as the volume in µL used to elicit accelerated transcription. Note: the minimum volume required for rate-enhancement was not determined, therefore total activity, specific activity, and fold purification numbers are lower limits. nd= not determined.

stepwise elution. Eluates were assayed for protein concentration using a Bradford protein assay and pooled accordingly for each salt step. The pooled fractions for each salt elution were tested in the two-timepoint rate assay. Figure 2.6A shows that the activity eluted in the 0.3 M KCl fraction, as seen by similar transcript levels in the 15 and 600 s timepoints. An S Sepharose (Spool) from an earlier test fractionation shows a low level of activity as a positive control. Some of the activity was lost in the flow through (DE52.ft), however this was barely detectable in the experiment shown. Lack of activity in the input was likely due to an insufficient quantity added to the reaction. The assay also confirms that the activity survives dialysis into buffer D.1, which was needed for the next purification step (sample D.3d in Figure 2.6A, where d= dialysis into Buffer D.1).

Purification of the activity using phosphocellulose chromatography

I next ran the dialyzed 0.3 M KCl DE52 pool (D.3d) over a phosphocellulose (P11) resin which contains an orthophosphate functional group that acts to resolve proteins through phosphate affinity and weak cation exchange. Phosphocellulose (Pcell) columns are often used as initial columns for purifying transcription factors. The D.3d fraction was loaded onto a 20 mL phosphocellulose column followed by step elutions of 0.1 M KCl, 0.3 M KCl, 0.5 M KCl, and 1.0 M KCl in Buffer D. Each pooled salt elution was tested and the activity was found to be in the 0.5 M KCl fraction (D.3P.5) (Figure 2.6A). This fraction was then dialyzed to 0.1 M KCl prior to running the next chromatography column.

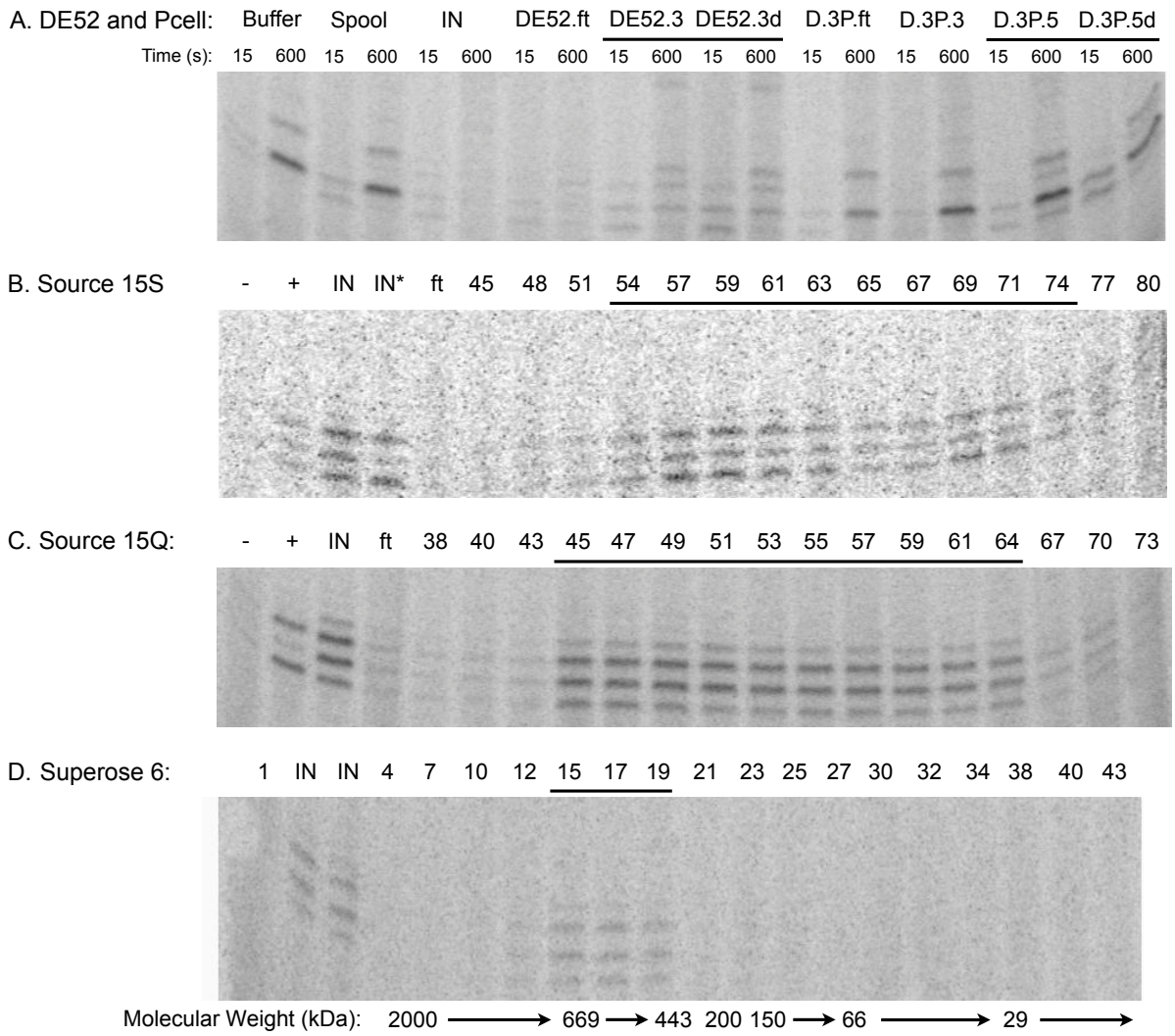


Figure 2.6. Purification of activity over first five chromatography columns. A) Fractions from DE52 and Phosphocellulose columns tested at 15 and 600 seconds. DE52.3d and D.3P.5d fractions were dialyzed to 0.1 M KCl. B) Fractions from Source 15S column. Negative control samples substituted buffer for sample tested and positive control samples contained unpurified HNE. *IN= input + unpurified HNE. C) Fractions from Source 15Q column. D) Fractions from Superose 6 column. For B), C), and D) the activity was detected by comparing levels of transcript production after 15 seconds of transcription. Protein concentration was also determined (not shown). Numbers above gels correspond to chromatography fractions and the lines below the numbers indicate fractions that contain accelerated rate activity. IN= input to column. ft = column flow through.

Purification of the activity using Source 15S chromatography

I next tried purification of the D.3P.5 fraction using a 1.6 mL Source 15S resin and HPLC chromatography. S sepharose is a strong cation exchanger containing methyl sulfonate functional groups. I used a mobile phase of 10 column volumes of Buffer D with an increasing salt gradient from 0.1 M KCl to 0.7 M KCl. Preliminary S sepharose pilot studies using step-elution showed that the activity eluted in the 0.3 M salt step, therefore I tested a variety of fractions spanning from ~0.15 M to 0.3 M KCl (Figure 2.6B) looking at transcript production after 15 seconds with NTPs. Fraction 45 was determined to contain 0.1 M KCl and fraction 94 contained 0.3 M KCl. Fractions 53-75 were pooled. The corresponding chromatograph showed one large A_{280} peak spanning these fractions (data not shown). The flow through did not elicit any rate-enhancing activity. The S-pool was then diluted ~2.22-fold to 0.1 M KCl with Buffer D (0 M KCl) for fractionation over a Source 15Q column.

Purification of the activity using Source 15Q chromatography

I next used a Source 15Q anion exchange resin on our HPLC. This resin contains an immobilized quarternary amine functional group. I used a 10 column volume linear salt elution from 0.1 M to 0.7 M KCl, similar to the Source 15S column. Again, previous small scale purifications indicated that elution of the rate-enhancing activity would be between 0.2 and 0.3 M KCl, and the fractions were tested in the transcription assay accordingly (Figure 2.6C). Fractions 45-64 showed robust activity and fractions 44-65 were pooled. The corresponding A_{280} chromatograph contained 3 peaks, peak 1 at ~0.28 M, peak 2 at 0.4 M, and a third at 0.7 M KCl.

Activity correlated with peak 1, however peaks 1 and 2 overlapped and therefore did not resolve fully. Samples were stored at -80^{o}C prior to the next fractionation.

Purification of the activity using Superose 6 size exclusion chromatography, and small scale batch test of phenyl sepharose and hydroxyapatite CHT

Next I subjected the D.3P.5SQ activity to Superose 6 (S6) gel filtration chromatography using Buffer D.1 as the mobile phase. Superose 6 is a size exclusion column of highly cross-linked agarose beads. The Q-pool was concentrated ~2.5 times for loading onto a 24 mL Superose 6 column. Figure 2.6D shows that the activity eluted in the fractions corresponding to a 550-600 kDa protein, as determined by a set of standards fractionated on a separate, but identical chromatography run. Silver stain analysis of the activity indicated the fraction contained a mixture of multiple proteins (not shown). Based on the activity assay and similar silver stain patterns, fractions 16-18 were pooled for the next purification step.

Next, small scale pilot purifications were tried with the P.5D.3 pool. I tested phenyl sepharose and hydroxyapatite CHT resins in batches using 50 µL or 100 µL of resin. The resins were nutated with P.5D.3 for 30 min at 4^{o}C, prior to step elutions. Hydroxyapatite resin contains positively charged calcium and negatively charged phosphate binding sites and dually poses as a cation exchange and a calcium affinity column. Following the acidic protein purification protocol from BioRad, the proteins were eluted with increased phosphate buffer concentration (10 mM, 250 mM, then 500 mM phosphate buffer, pH 7.6) and two stepwise salt elutions (0.25-0.5 M KCl). The hydroxyapatite CHT resin did not result in greater purification as almost all the fractions contained activity. Hydroxyapatite also posed a further complication as it required a

buffer change into phosphate buffer. However, silver stain of the flow-through and eluate did show the majority of proteins resided in the high phosphate fraction. Activity did not bind the phenyl sepharose resin (50 mM buffer, pH 7.6, 1.5 mM KCl, then 50 mM buffer with 0 M salt) and was found in the flow through. Data from these two batch purifications is not shown.

Purification of the activity using Heparin chromatography

I next fractionated the D.3P.5SQ6 fraction over a 100 µL heparin affi-gel HPLC affinity column using our SMART system. I eluted with a linear salt gradient from 0.1 M to 0.5 M KCl. Figure 2.7A shows a preliminary heparin chromatography run using D.3P.5SQ, but prior to the Superose 6 column. The A_{280} chromatograph for this run showed a sharp peak corresponding to the flow through and a broad peak centered at fraction 26 and spanning fractions ~16 to ~35. There was some activity in the flow through, most likely due to surpassing the column's capacity. For later runs, a larger column was packed and used. Fractions 20-31 were pooled. For large scale purification after the Superose 6 column, activity was seen in fractions pertaining to 0.32 M - 0.49 M KCl. The heparin column also served to concentrate the activity 5-fold.

Purification of the activity using DEAE-5PW chromatography

The D.3P.5SQ6Heparin pool was diluted 3-fold to 0.1 M KCl with Buffer D.0 (0 M KCl) and loaded onto a 100 µL DEAE-5PW weak anion exchange column. A 10 column volume salt gradient from D.1 to D1.0 was used. Fractions were subjected to the transcription rate assay and analyzed by SDS-PAGE and silver staining (Figure 2.7B and C). The A_{280} chromatograph showed two partially overlapping peaks. The activity was present throughout most of the

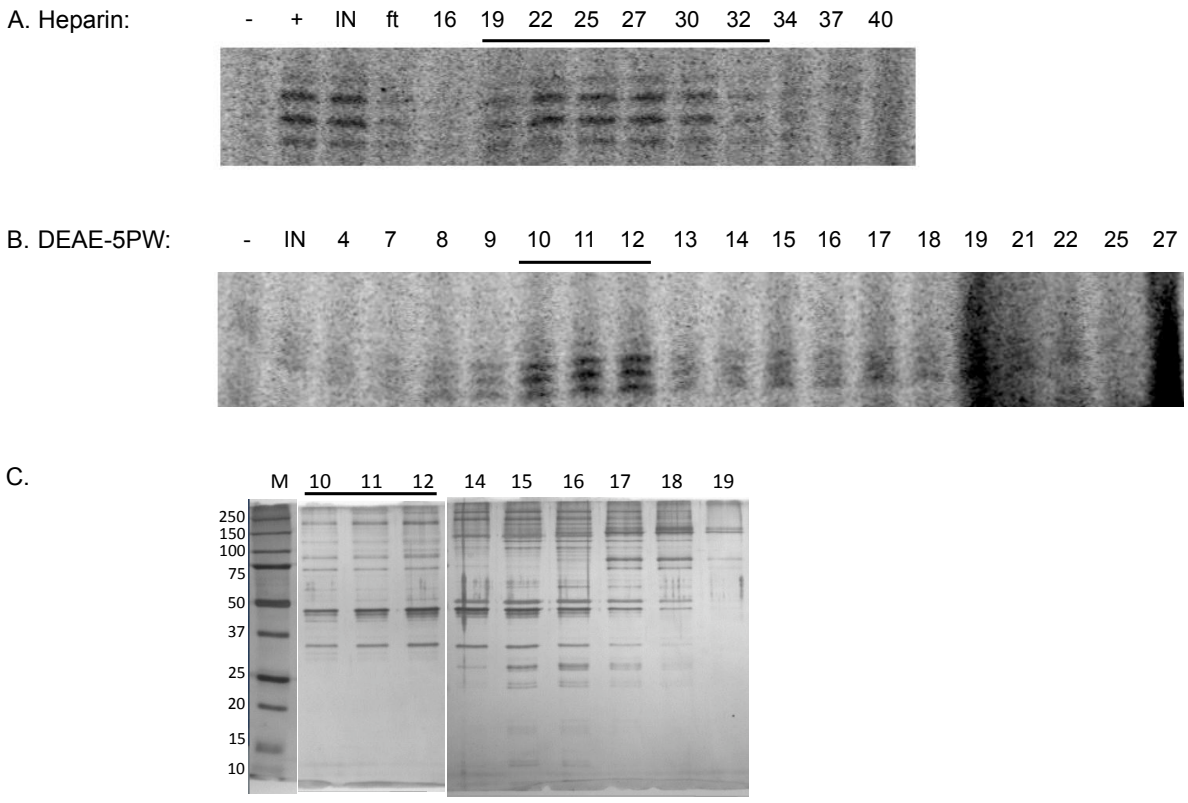


Figure 2.7. A) Transcription assay testing fractions from heparin column. Figure shows transcript production after 15 seconds of transcription. + is a positive control from another fractionation. B) Fractions from the DEAE-5PW column. C) Silver stain of DEAE-5PW fractions. DEAE-5PW fractions 10, 11, and 12 showed robust transcript production after 15 seconds. M= molecular weight markers, and corresponding molecular weights are listed on the left of the gel.

fractions, but was concentrated in the first peak (fractions 8-12). The saddle between the two peaks was fraction 13, and was about 5/6 the height of the first peak and about 1/3 height of the second peak. A longer column and salt gradient may have yielded a higher resolution purification. Fractions 10-12 were pooled for mass spectrometry analysis.

Mass spectrometry and protein identification

The DEAE-5PW pool (D.3P.5SQ6H-5PW) containing the activity was TCA precipitated and Figure 2.8 shows the sypro ruby stained SDS gel of the sample. Twenty-two gel sections were excised, in-gel trypsin digested, and eluted before submission to the University of Colorado Mass Spectrometry facility for analysis by LC-MS/MS with electrospray ionization and an ion trap. MASCOT analysis of the peptides is shown in Table 2.2. Among the proteins, 9 subunits of TFIIH were identified with abundant peptides and high MASCOT scores. Western blot analysis verified the presence of TFIIH in the final purified fractions (data not shown).

Activity on a negatively supercoiled versus linear template

Given that I identified TFIIH, I wanted to know whether the rate-enhancing activity was specific to the supercoiled template or if the activity could also evoke rate acceleration on linear DNA. Transcription from a linear template requires TFIIE, TFIIH, and ATP (or dATP) energy to form an open complex and advance through promoter escape, whereas a supercoiled template does not have this requirement. I tested the partially purified activity on a linear AdMLP template, quenching transcription at 15 and 600 seconds. Upon addition of the partially purified activity I saw accelerated rate activity (Figure 2.9). Therefore, the activity was not specific to a

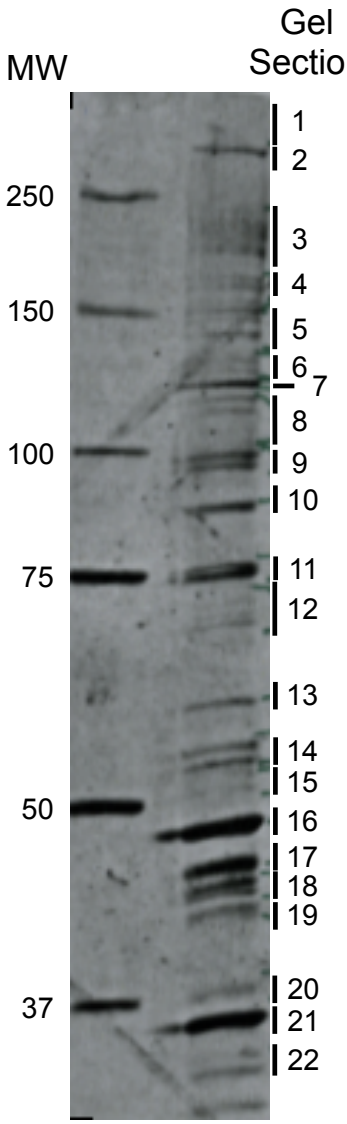


Figure 2.8. Sypro ruby stain of mass spectrometry gel. Numbers to the right indicate gel sections that were submitted as separate samples for mass spectrometry identification. TFIIH peptides were identified in gel sections: 10, 11, 13, 15, 19, 20, and 22.

Gel Section	Mass Spectrometry Identification	Alias	Mass (kDa)	emPAI	MASCOT Score	# Peptides matched
2	Isoform 1 of DNA (cytosine-5)-methyltransferase 1	DNMT1	190	0.02	60	6
	Isoform 1 of DNA-dependent protein kinase catalytic subunit	PRKDC	469	0.01	51	15
3	Isoform 3 of DNA (cytosine-5)-methyltransferase 1	DNMT1	144	0.02	72	4
	Isoform 1 of DNA (cytosine-5)-methyltransferase 1	DNMT1	189	0.05	68	6
4	TOP2A 183 kDa protein	TOP2A	183	0.07	91	9
	TOP2B Isoform Beta-2 of DNA topoisomerase 2-beta	TOP2B	183	0.04	76	8
	1-phosphatidylinositol-4,5-bisphosphate phosphodiesterase beta 3	PLCB3	139	0.02	49	5
5	SCC-112 protein	SCC-112	151	0.02	43	7
6	Putative pre-mRNA-splicing factor ATP-dependent RNA helicase	DHX16	119	0.06	71	7
7	Putative pre-mRNA-splicing factor ATP-dependent RNA helicase	DHX16	119	0.5	336	26
8	Superkiller viralicidic activity 2-like	SKIV2L2	118	0.09	63	11
9	Kinesin-like protein	KIF2A	78	0.45	154	23
10	**TFIIH basal transcription factor complex helicase XPB subunit**	ERCC3	89	0.2	143	14
11	Isoform 1 of WD repeat protein 48	WDR48	76	0.29	193	14
	TFIIH basal transcription factor complex helicase subunit	ERCC2	87	0.16	97	8
	KIF2A Kinesin-like protein	KIF2A	78	0.04	73	3
	HSPA5 protein	HSPA5	72	0.05	44	3
12	Lysyl-tRNA synthetase	KARS	68	0.05	75	3
	DNA-binding protein	RFX5	65	0.1	55	5
13	**TFIIH basal transcription factor complex p62 subunit**	GTF2H1	62	0.44	134	12
14	Eukaryotic translation initiation factor 2 subunit 3	EIF2S3	51	0.37	88	9
	Histone-binding protein	RBBP7	48	0.07	53	4
	Histone-binding protein	RBBP4	48	0.07	53	4
15	EIF2S2 protein	EIF2S2	38	0.95	282	16
	TFIIH basal transcription factor complex p52 subunit	GTF2H4	52	0.2	88	9
16	Eukaryotic translation initiation factor 2 subunit 3	EIF2S3	51	1.4	464	65
	Hypothetical protein	n/a	51	0.77	221	31
17	Eukaryotic translation initiation factor 2 subunit 2	EIF2S2	38	3.43	477	54
	Eukaryotic translation initiation factor 2	EIF2S3	51	0.13	92	3
18	Eukaryotic translation initiation factor 2 subunit 2	EIF2S2	38	0.64	155	8
	Pre-mRNA-processing factor 19	PRPF19	55	0.19	81	7
	hypothetical protein LOC55006	FLJ20628	53	0.06	64	2
19	**TFIIH basal transcription factor complex p44 subunit**	GTF2H2	44	0.33	77	12
	Isoform 2 of Basic leucine zipper and W2 domain-containing protein 1	BZW1	41	0.08	44	4
20	**Cell division protein kinase 7 (TFIIH subunit)**	CDK7	39	0.3	142	9
	Eukaryotic translation initiation factor 2 subunit 1	EIF2S1	36	0.42	190	9
	DNA-binding protein RFX5	RFX5	65	0.05	54	3
21	Eukaryotic translation initiation factor 2 subunit 1	EIF2S1	36	2.13	688	77
	Regulatory factor X-associated protein	RFXAP	28	0.4	79	10
	Hypothetical protein DKFZp313O1018 (Fragment)	HECTD2	17	0.19	51	7
22	Eukaryotic translation initiation factor 2 subunit 1	EIF2S1	36	0.55	158	10
	CDK-activating kinase assembly factor MAT1 (TFIIH subunit)	MNAT1	36	0.86	128	17
	Cyclin-H (TFIIH subunit)	CCNH	38	0.8	128	17
	TFIIH basal transcription factor complex p34 subunit	GTF2H3	34	0.58	122	12

Table 2.2. Mascot identification of peptides. LC/MS/MS was performed on gel slices of purified activity. Mascot search engine identified the above proteins using the IPI human database. MASCOT scores of at least 40 and proteins with 3 or more peptides matched were included in this table. TFIIH subunits are highlighted in bold. emPAI = exponentially modified protein abundance index.

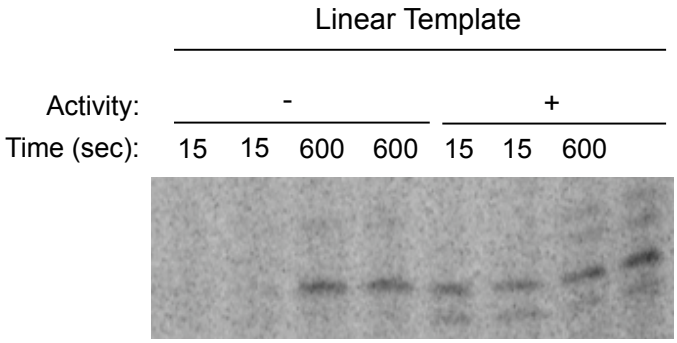


Figure 2.9. Comparison of 15 and 600 seconds of transcription with partially purified activity on a linear AdMLP template. All lanes contained TFIIE, TFIIH, and dATP.

supercoiled template.

The rate-enhancing activity requires TFIIE.

I added recombinant TFIIE, native purified TFIIH, and dATP to the activity assay to determine if either, or both, were sufficient for the activity or necessary to observe rate-enhancement by the purified activity. Analysis determined TFIIE was necessary to produce the rate-enhancing activity (Figure 2.10A). The activity-containing fraction used in Figure 2.10A was from the final DEAE-5PW column. dATP was not required. Furthermore, TFIIE, TFIIH, and dATP without the activity-containing fraction were not sufficient under standard *in vitro* transcription concentrations to produce the accelerated rate. Further tests showed that the factor requires both subunits of TFIIE for activity (Figure 2.10B). TFIIH alone does not accelerate the rate of promoter escape when used at concentrations sufficient to produce transcript from a linear template (Figure 2.10C). From this I concluded that the activity required TFIIE, but not TFIIH or dATP.

Immunodepletion studies

Using the information from the MASCOT identifications, I performed immunodepletion experiments on a few of the top scoring protein matches with documented roles in transcription, including topoisomerase II A (TOP2A) and B (TOP2B), Kif2A, and TFIIH. Antibodies immobilized on protein A/G beads were used to pull down the respective proteins after which the flow-through and beads were assayed for activity using the transcription assay, silver stained,

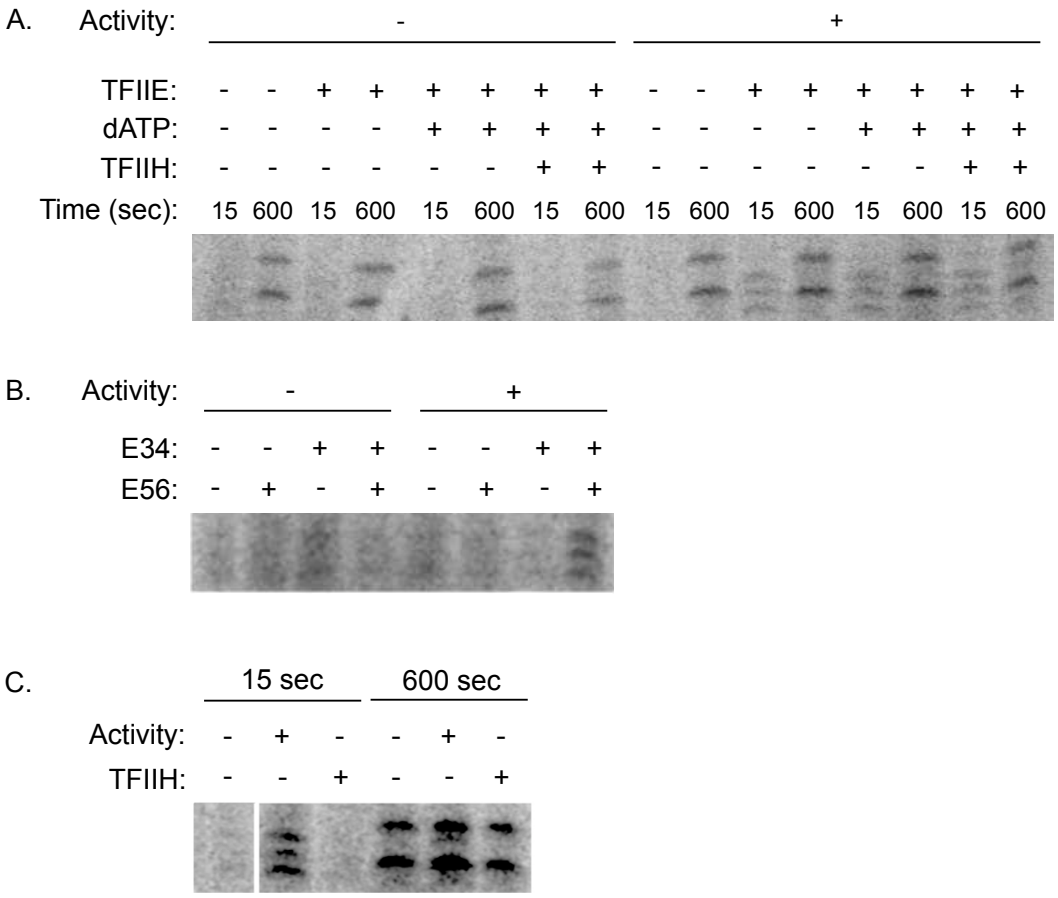


Figure 2.10. A) Partially purified activity requires TFIIE, but not TFIIH or dATP to accelerate the rate of promoter escape. 16 nucleotide product is shown. B) The partially purified activity requires both subunits of TFIIE. Transcription proceeded for 15 seconds. C) Transcription at 15 seconds and 600 seconds comparing fractionated activity and purified TFIIH reveals that TFIIH does not accelerate the rate of promoter escape when used at transcription-dependent concentrations. TFIIE was present in all of these reactions.

and western blotted for the protein of interest. Activity remained in the flow through for TOP-A, TOP-B, and Kif2A immunodepletions (data not shown). The immunoprecipitated TFIIH was eluted with a TFIIH epitope peptide and the activity was found to elute with TFIIH (Figure 2.11).

TFIIH can provide rate enhancement in a concentration dependent manner

Because the activity immunodepleted with TFIIH and required TFIIE, I decided to further investigate the possibility of TFIIH involvement. Figure 2.12A shows 15 seconds of transcription with a titration of two different TFIIH preparations (JAG and Taatjes) alongside the partially purified activity. The JAG and Taatjes TFIIH stocks were prepared from HeLa nuclear extracts following two different purification schemes *(91, 158)*. At the higher concentrations, all three show similar levels of transcript, however, at the lower titration points, the JAG TFIIH does not show product. The volume of JAG TFIIH added at all these points is sufficient to produce transcript from linear templates given 10-20 minutes of transcription. Figure 2.12B compares transcription in reactions containing a variety of rate-enhancing fractions and TFIIH at 15 and 600 second timepoints. None of the preparations required dATP to exhibit activity. In this experiment the JAG TFIIH exhibits rate-enhancing activity like the others, however at a volume greater than 10 fold higher than we typically add to our assays. In a transcription assay on a linear template, transcript levels plateaued with 0.04 μL of the JAG TFIIH (data not shown). Therefore, approximately 10 times more TFIIH is required to see rate-enhancement with the JAG TFIIH compared to the amount needed to observe maximal levels of transcription from a linear template. I conclude that TFIIH is the rate-enhancing factor. Our assays contained a low amount of TFIIH during the purification; the chromatography fractions containing the activity

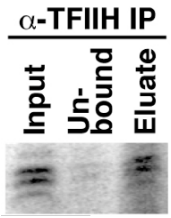


Figure 2.11. Transcription assay shows that the activity immunodepletes with α-TFIIH and can be eluted with a TFIIH peptide. All lanes contained TFIIE.

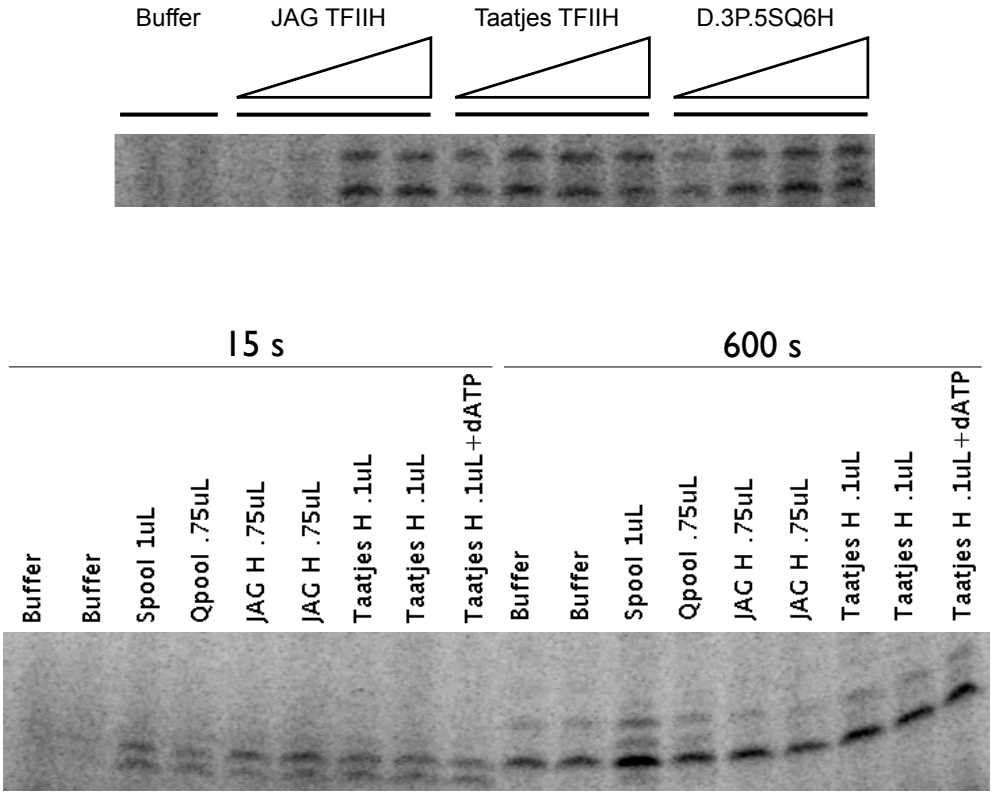


Figure 2.12. A) Transcription assay titrating (0.05, 0.1, 0.5, and 1.0 uL) different TFIIH preparations shows the activity is consistent with a high concentration of TFIIH. Transcription was terminated after 15 seconds. JAG TFIIH and Taatjes TFIIH are TFIIH preps previously purified by the Goodrich and Taatjes labs, respectively. D.3P.5SQ6H is activity purified through the Heparin column. B) Various TFIIH preparations give similar accelerated transcription rates on an AdMLP template. JAG TFIIH sample shows rate-enhancing activity when added at greater than 10 times higher concentration (0.75 uL) than required for TFIIH-dependent transcription on a linear template. Samples were assayed for 15 and 600 seconds of transcription. S- and Q-pools are partially purified nuclear extracts, D.3P.5S and D.3P.5SQ respectively. JAG H and Taatjes H are JAG TFIIH and Taatjes TFIIH, respectively.

substantially increased the concentrations of TFIIH in the transcription reactions and hence enhanced the rate of promoter escape.

DISCUSSION

In this chapter I described purification of a factor that accelerates the rate of promoter escape on negatively supercoiled and linear AdMLP templates *in vitro*. A two-timepoint transcription activity assay was developed in order to monitor the fractionation of the factor throughout purification. A seven column chromatography scheme was developed and described. The purified fraction was subjected to mass spectroscopy analysis and was found to contain 9 subunits of TFIIH. The activity immunodepleted with a TFIIH antibody. Moreover, I found that titration of two independent TFIIH preparations produced similar rate-accelerating activity, however required greater than 10 times more TFIIH than necessary for maximal transcription on a linear template.

The activity consistently eluted between 0.2 M and 0.3 M KCl during all the ion exchange columns used, consistent with some previous purifications of TFIIH *(158)*, but not others *(58)*. Furthermore, the sepharose 6 gel exclusion column indicated that the activity eluted as a complex approximately 550 to 600 kDa in size. Holo-TFIIH consists of 10 subunits and is greater than 500 kD in size *(159)*. Upon characterizing the activity, I found that TFIIE was required, but not sufficient to produce the accelerated rate. This is consistent with the activity being TFIIH because TFIIE is required to recruit TFIIH into preinitiation complexes *(160)*.

Two different TFIIH preparations showed rate-enhancing activity upon titration into the transcription assay. Since all of the tested TFIIH preparations were purified from human cells, it would be interesting to compare the purity of the two preparations with my fraction. This would provide insight into whether any other factors may contribute to the activity. Furthermore, each TFIIH lot was prepared differently and therefore contained different concentrations of active

TFIIH. It appears in these studies that a higher concentration of TFIIH leads to an acceleration in the rate of transcription. Western blots for TFIIH confirmed that differences in TFIIH levels correlate with differences in rate-enhancing activity for JAG TFIIH, Taatjes TFIIH, and my purified activity (data not shown). It is also possible that the TFIIH in each preparation is a heterogenous mixture of TFIIH complexes, for example with multiple phosphorylation states, and the complex responsible for the activity is more or less abundant in the different purifications *(161)*.

TFIIH is required for transcription beyond 2-3 nucleotides on a linear template *(53, 84, 88, 162)*. Our studies indicate a different requirement for TFIIH in enhancing the rate of promoter escape. Using our *in vitro* system, I needed greater than 10 times more TFIIH to maximize the rate of promoter escape compared to the amount needed to proceed past initiation on a linear template. Perhaps the TFIIH releases after initiation and a new TFIIH binds to facilitate promoter escape (Figure 2.13). This model explains the higher concentration dependence of the rate-enhancing activity of TFIIH. One could test this through release and rebinding assays on immobilized templates, or with the addition of competitor DNA to capture the released TFIIH.

It is interesting that TFIIH produces the rate enhancing activity in our *in vitro* system since the negatively supercoiled DNA negates the need for TFIIH helicase activity *(82, 85-87)*. TFIIH may create a more stable ternary complex via interactions with Pol II and other transcription factors that facilitates structural transformations during the rate-limiting step of promoter escape *(163)*. During promoter escape, the upstream region of the transcription bubble

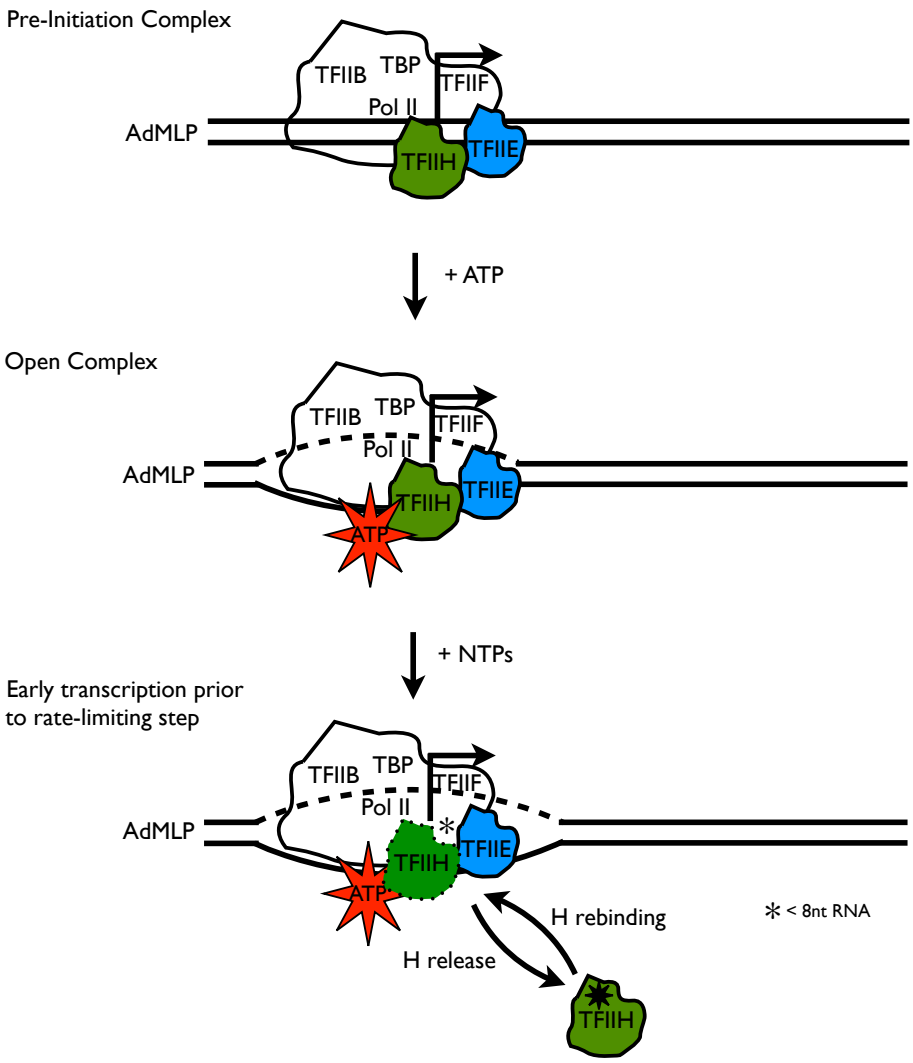


Figure 2.13. Transcription complexes show different requirements for TFIIH during the early steps of transcription.

begins to close, which may drive the transcription machinery forward, promoting promoter escape *(100)*. TFIIH may participate in this closure which could enhance the rate of promoter escape. Interestingly, a follow-up study indicated that TFIIH cannot enhance the rate of transcription on a pre-melted mismatched bubble template that cannot close (data not shown). This suggests a link between bubble collapse and TFIIH-mediated rate enhancement.

TFIIH is known to be a target for transcriptional regulation by activators. The TFIIH helicase activity is a target for transcription stimulation by the FUSE binding protein (FPB) transcription factor, which acts at the c-myc promoter *(147, 164)*. TFIIH has also been implicated in TBP-mediated activation by PC4 and BRCA1 in a minimal transcription system *(76, 77)*. p53 binds the p62 and helicase subunits of TFIIH, and these interactions may be important for transcriptional activation *(55, 56)*. In addition, activator influence on TFIIH may stabilize PICs to facilitate transcription *(101)*. It is understandable that TFIIH is a prime target for activation, as it is one of the few general transcription factors with enzymatic activity *(102)*.

MATERIALS AND EXPERIMENTAL METHODS

DNA constructs

The negatively supercoiled plasmid used as the DNA template in these studies contained the AdMLP promoter from -53 to +10 fused to a 380 basepair G-less cassette *(53)*. The linear template used the above plasmid cut with Hind III and EcoRI .

Preparation of Pol II and transcription factors

Recombinant human TBP, TFIIB, TFIIE, and TFIIF were expressed in *E. coli* and purified as previously described *(91)*. Native human RNA polymerase II and JAG human TFIIH were purified from HeLa cells as previously described *(91, 165)*. Taatjes TFIIH, courtesy of D. Taatjes, was purified from HeLa cells using the flow through from a Q column followed by peptide elution off an anti-ERCC3 monoclonal antibody column as previously described *(158)*.

Transcription reactions

Transcription reactions were performed in 20 µL of Buffer A: 10% glycerol, 10 mM Tris (pH 7.9), 50 mM KCl, 10 mM HEPES (pH 7.9), 4 mM $MgCl_2$, 1 mM DTT, 20 µg/ml BSA, 160 U RNaseOUT™ (Invitrogen), and 0.02% NP-40. Transcription factors were added at the following final concentrations: 3.5 nM TBP, 10 nM TFIIB, 2 nM TFIIF, 9 nM TFIIE-34, 5 nM TFIIE-56, 0.7 µL TFIIH, and 1-3 nM Pol II. Proteins (in 10 µL) were incubated at 30°C for 3 minutes and then 2 nM template DNA (in 10 µL) was added and preinitiation complexes were allowed to form for 10 minutes at 30°C. Transcription was initiated by adding a limited nucleotide mix, giving final concentrations of 20 µM CTP, 1 µM [α-^{32}P]CTP (5 µCi/reaction),

650 μM UTP, 500 μM ApC dinucleotide (Sigma-Aldrich), 100 μM dATP (Invitrogen), and 350 μM 3'O-methyl-ATP (3'O-Methyl-ATP) chain terminator (TriLink BioTechnologies).

When indicated, nuclear extract or chromatography fractions were added 5 minutes prior to NTP addition. At defined time points, transcription was quenched with a stop solution containing 3.1 M ammonium acetate, 20 mM EDTA, 10 μg carrier yeast RNA, and 15 μg proteinase K. Transcripts were ethanol precipitated using 100% ethanol on dry ice for 25 minutes then centrifuged 13,000xg for 50 minutes followed by a 70% ethanol wash. The 15 nucleotide product was denatured and resolved with 14% PAGE after which the gel was dried under heat and vacuum.

Determination of rate constants

Gels were scanned using a Molecular Dynamics Typhoon™ Phosphorimager and quantitated using ImageQuant™ software. Phosphorimage units (PI) were plotted versus time, and the data was fit with the equation $PI=PI_{max}(1-e^{-kt})$. Prism™ software was used to solve for k and PI_{max}. Experiments were normalized using the PI_{max} calculated from the data sets.

Nuclear extraction and chromatography buffers

HeLa nuclear extracts were prepared and column chromatography performed using the following buffers: Buffer A (10 mM HEPES, pH 7.9, 1.5 mM $MgCl_2$, 10mM KCl), Buffer C (20 mM HEPES (pH 7.9), 25% glycerol, 417 mM NaCl, 1.5 mM $MgCl_2$, 0.2 mM EDTA), Buffer D (40 mM HEPES, (pH 7.9), 0.05-1 M KCl, 20% glycerol, 1.5 mM $MgCl_2$, 0.2 mM EDTA, 1 mM DTT, 0.2 mM PMSF), PBSM (1X PBS, 1.2 mM $MgCl_2$). All buffers were filtered using a

0.2 micron filter and used at 4°C. For CHT hydroxyapitate column chromatography (BioRad) the following buffers were used: 10 mM, 250 mM, or 500 mM Na-Phosphate buffer (pH 7.6) with 20% glycerol, and 0.2 mM EDTA. Na-Phosphate buffer contained a 1:1 ratio of Na_2HPO_4 monobasic and dibasic. Phenyl sepharose column buffer contained 0.05 M Na-Phosphate buffer (pH 7.6) and 1.5 M KCl.

Cell culture and isolation of nuclei

HeLa cells, courtesy of D. Taatjes, were grown to a density of 0.7 million cells per mL in DMEM media and split 1:1 every day, until 100 L of cells were obtained. Cells were harvested in 8-12 L batches and nuclei were isolated immediately. The cells were resuspended in 5 pellet column volumes (PCV) of buffer and incubated on ice for 20 minutes and then centrifuged for 10 minutes at 3000xg. The pellet was next resuspended with 2 PCV of buffer, dounce homogenized 7 times using a 20 mL dounce, and centrifuged for 15 minutes at 15,000xg. The isolated nuclei were stored as pellets at -80°C. Protease inhibitors and reducing agent were added in all buffers throughout the cell harvest and processing (1X Roche complete protease inhibitor cocktail, 1mM $Na_2S_2O_4$, 1 mM DTT, 0.2 mM PMSF). After isolating the nuclei from 100 L of cell culture, the nuclei were dounce homogenized to isolate nuclear extract as previously described and dialyzed into buffer D.1 *(166)*.

Chromatography

All chromatography was performed at 4°C in the presence of protease inhibitors and reducing agents. All columns were equilibrated with 5 column volumes of buffer prior to sample

loading and during each KCl elution step, where applicable. Phosphocellulose P11 (Whatman) and DEAE DE52 cellulose (Whatman) columns were prepared and pre-treated according to manufacturers specifications and chromatography was executed using gravity flow or peristaltic pump at no greater than 2 mL/min.

All other chromatography columns were run on either a BioRad HPLC (Source 15S column) or PharmaciaBiotech SMART micro-purification system (Source 15Q, Superose 6, Heparin, DEAE-5PW columns) and monitored for absorbance at 260 nm and 280 nm. The superose 6 PC 3.2/30 24 mL gel filtration column (GE Healthcare) was equilibrated and run using Buffer D.1 as the mobile phase at a constant flow for both the sample and standard run on the SMART system. A mix of 29,000-700,000 molecular weight standards were used prior to sample purification to calibrate MW elution time. The input to the column was filtered and spin concentrated to ~500 µL using 5K NMWL membrane micro-concentrator (Millipore) prior to loading.

LC-MS/MS

Samples were resolved by 8.5% SDS-PAGE, visualized with Sypro Ruby stain. Bands were in-gel reduced with DTT, alkylated with iodoacetamide, and then trypsin digested for identification by the University of Colorado at Boulder Mass Spectroscopy Facility. The samples were run on a LC-MS fitted with an electrospray ionizer and an ion-trap detector and peptides were analyzed using the MASCOT search engine for the human IPI database.

Immunodepletions

Immunodepletion of TFIIH, TOPIIA, TOPIIB, or Kif2A was done at 4°C using HGEN wash buffer (15% glycerol, 20 mM HEPES pH 7.9, 0.5mM EDTA, 1 mM DTT, 0.2 mM PMSF). A 40 µL portion of Protein A/G agarose beads (Santa Cruz) was incubated with the manufacturer's recommended amount of antibody for 1 hour while nutating, followed by an additional 1-3 hours of incubation with the partially purified activity. Beads were washed twice with HGEN containing 0.1 M KCl followed by three times with HGEN containing 0.3 M KCl, and the flow throughs from each wash, along with the beads, were analyzed via western blot and silver stain. TFIIH was eluted with a TFIIH peptide as published *(158)*.

Western blot and silver stain detection

Western blots were performed using standard procedures using TBS wash buffer (137 mM NaCl, 20 mM Tris pH 6.5) first with 0.05% Tween and then without Tween. The following antibodies were used according to the manufacturer's specifications: αTOPIIA (Santa Cruz), polyclonal αTOPIIB (Santa Cruz), polyclonal αKif2A (Novus Biologicals), and αTFIIH p89 antibodies (Santa Cruz, or an in house preparation) were used. Secondary antibodies with a horseradish peroxidase conjugate, were used at a 1:10,000 dilution, and then detected using ECL Plus western blot detection reagents (GE Healthcare). Silver stained samples were SDS-PAGE resolved and silver stained following a standard silver stain process.

CHAPTER 3

Chromatin oligonucleotide precipitation-sequencing (ChOP-seq): Developing a new technique to monitor lncRNA occupancy on chromatin

CHAPTER 3

SUMMARY

Recently thousands of long noncoding RNAs (lncRNAs) have been discovered in mammalian cells *(117-119)*. Researchers have used genome-wide methods including ChIP-sequencing and RNA tiling experiments to discover these lncRNAs and the regions from which they are transcribed, however, at the time this work was begun none had used a genome-wide method to determine whether lncRNAs interact with chromatin, and if so, where they are located. Here I describe the development of a ChIP-like method we have termed Chromatin Oligonucleotide Precipitation, or ChOP, that uses antisense oligonucleotides to precipitate the lncRNA of interest from cross-linked cells thereby pulling down any associated chromatin and proteins. The oligonucleotide precipitated material is then sequenced using Illumina technology. In this study I targeted the well characterized lncRNA HOTAIR. An initial sequencing experiment using one candidate oligonucleotide probe to pull down HOTAIR identified a variety of regions in the genome as potential HOTAIR binding regions. These results indicate that the ChOP assay selectively pulls down some chromatin regions in an RNA-dependent manner. This chapter describes a new method to identify the genome-wide association of ncRNAs on chromatin, which provides insight into the mechanism and function of uncharacterized ncRNAs.

INTRODUCTION

Long non-coding RNAs (lncRNAs) are greater than 200 nucleotides in length and can be up to tens of kilobases long *(167)*. They are typically transcribed by RNA polymerase II, 5'capped, polyadenylated, often spliced, and are evolutionary conserved in mammals *(168, 169)*. However, they are not translated into proteins like mature mRNAs. Some lncRNAs have been shown to be cell-type and lineage specific (maternal or paternal) *(118, 122, 169)*.

LncRNAs are emerging as key regulators of gene expression *(116, 170)*. Hundreds of lncRNAs co-immunoprecipitate with chromatin remodeling complexes, such as PRC2, CoREST, and JARID1C/SMCX *(118)*. Several lncRNAs have been implicated in transcriptional regulation (Evf-2 *(136)*, Alu *(137)*, HOTAIR *(123)*, HOTTIP *(139)*, HOTAIRM1 *(140)*, SRA *(171)*, ANRIL *(172)*), including gene imprinting (XIST/RepA *(128, 173)*, Air *(131, 132)*, Kcnq1ot1 *(122)*), often of developmentally regulated genes or gene family loci. It has been confirmed that a few of these lncRNAs are targeted to chromatin (Alu *(137)*, Kcnq1ot1 *(122)*, Air *(132)*, and Xist *(173)*) in a loci specific manner. In order to determine where lncRNAs might localize on chromatin in a global manner, we needed a genome-wide method. I set out to create an assay to identify genomic regions associated with a lncRNA in order to determine the potential scope of lncRNAs in regulating gene expression. I used HOTAIR lncRNA to develop this assay.

HOTAIR is a 2158-nucleotide spliced and polyadenylated lncRNA transcribed from the HOXC locus. HOTAIR was first identified in microarray experiments designed to screen the transcriptional activity across human HOX loci *(123)*. Further study showed that siRNA knock-down of HOTAIR upregulated gene expression in the HOXD locus and decreased H3K27me3 chromatin marks (Figure 3.1) *(123)*. Later studies proposed a model in which HOTAIR serves as

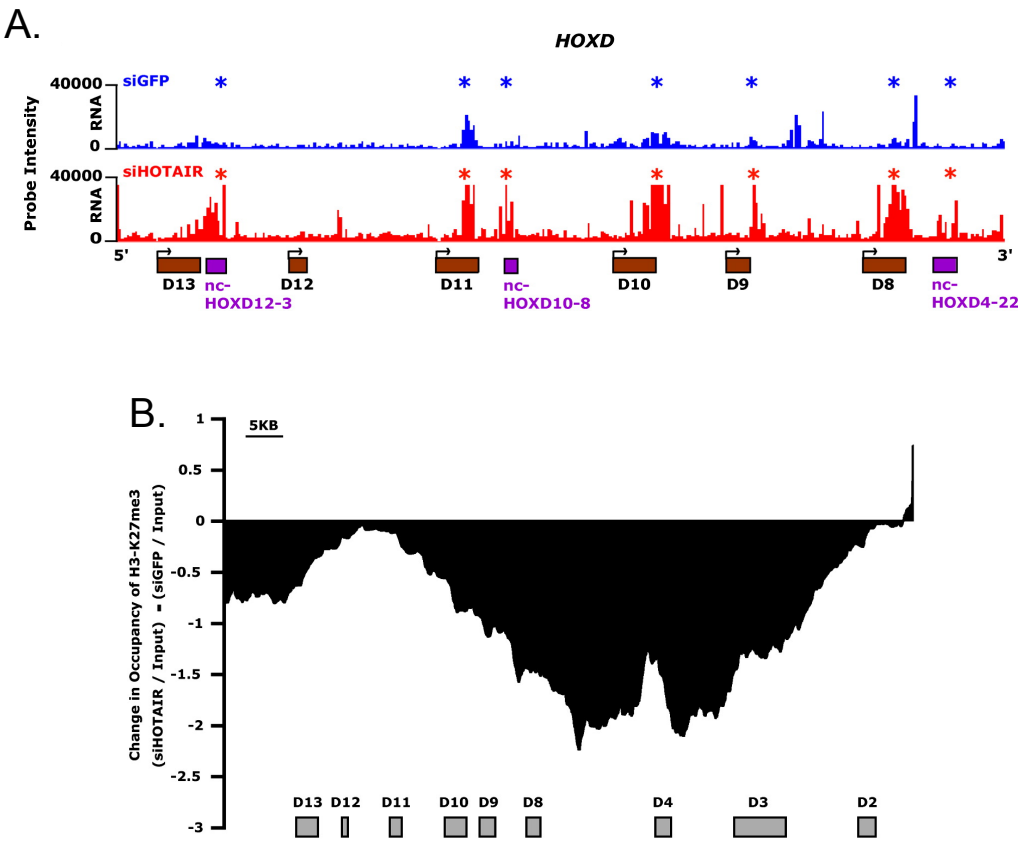


Figure 3.1. HOTAIR knockdown in primary human fibroblasts A) upregulates transcription in HOX D genes and B) decreases H3K27me3 in ChIP-chip experiments. * indicates significant increase in transcription over siGFP transfection. (Figure from Rinn et al. 2007 *(123))*.

a tether or a scaffold on which to recruit different chromatin modifying complexes to genomic loci to suppress gene expression *(141)*. The first 300 nucleotides at the 5'-end of HOTAIR were found to bind PRC2 and the 700 nucleotides at the 3'-end were found to bind the LSD1/ CoREST/REST complex, which contains a H3K4me2 demethylase (Figure 3.2) *(141*). This model strongly suggests HOTAIR will function on the genome, however, its presence on chromatin had not been investigated. HOTAIR lncRNA has also been implicated in epigenetic and transcriptional disregulation in breast and colorectal cancer during metastasis *(144, 175)*. It is therefore important to understand the mechanisms by which HOTAIR regulates transcription and epigenetics to more fully understand cancer metastasis and potentially unearth new therapeutic targets.

To better understand the breadth of transcriptional regulation by HOTAIR, I set out to develop and conduct a chromatin oligonucleotide precipitation assay targeting HOTAIR lncRNA to purify and identify the genomic regions that HOTAIR is associated with. Oligonucleotide probes were designed against HOTAIR and used to pull down the lncRNA and associated chromatin from formaldehyde cross-linked cells. The chromatin was tested for recovery of HOTAIR-regulated HOXD genes. Eluted chromatin was then deep-sequenced using Illumina technology and candidate genes were evaluated for HOTAIR occupancy.

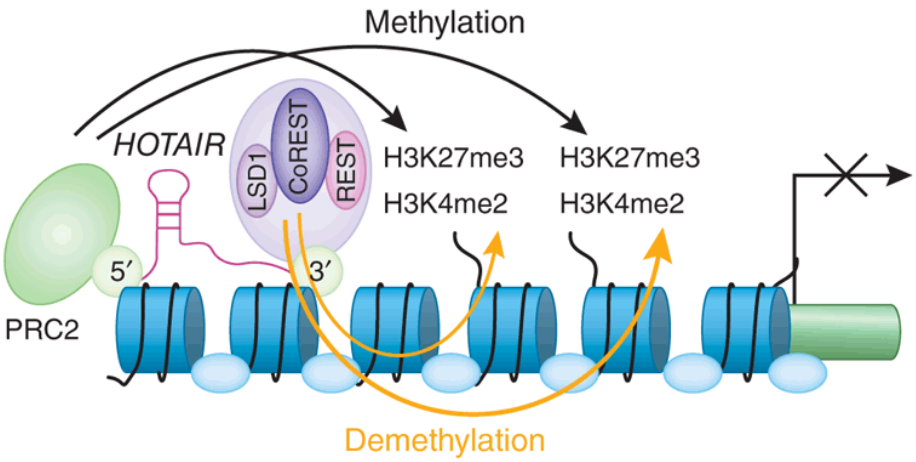


Figure 3.2. HOTAIR acts as a scaffold in HeLa and primary foreskin fibroblasts for H3K27 methylation (PRC2) and H3K4 demethylation (LSD1) factors to silence genes *(141)*. HOTAIR interacts with the LSD1-CoREST complex at its 3'-end and with PRC2 at the 5'-end. (Figure modified from Croce 2010 *(174))*.

RESULTS

Overview of chromatin oligonucleotide precipitation method

I developed a ChIP-like assay (Figure 3.3) that we termed chromatin oligonucleotide precipitation, or ChOP, to directly pulldown the lncRNA HOTAIR and any associated chromatin, with the goal of identifying genomic regions associated with, and presumably regulated by, HOTAIR. In brief, I formaldehyde cross-linked Jurkat, HeLa, or primary foreskin fibroblast cells, isolated the insoluble chromatin, solubilized it using sonication and fragmentation with DNase 1, and then used the chromatin for ChOP experiments. The ChOP input chromatin was first pre-cleared with unblocked beads, incubated with biotin-conjugated oligonucleotide probes, and then precipitated with blocked NeutrAvidin agarose beads. The bead-bound chromatin was then stringently washed and the lncRNAs, and associated chromatin and proteins, were eluted and cross-links were reversed. The recovered DNA was then purified and PCR analyzed for a subset of HOXD genes previously shown to be regulated by HOTAIR *(123)*. For Illumina sequencing, a library of the eluted DNA was prepared using a commercially available ChIP-seq kit.

Optimization of chromatin oligonucleotide precipitation assay

The original ChOP assay was previously published by our lab, and had been used with mouse cells for the 182 nt B2 RNA and with human cells for the 280 nt Alu RNA *(137)*. Besides the obvious difference in length between B2 or Alu RNA and HOTAIR, all ncRNAs invariably differ in cellular localization, intracellular interactions, structure, stability, sensitivity to buffer conditions, and cell-type specificity. I initially set out to optimize the ChOP assay for the

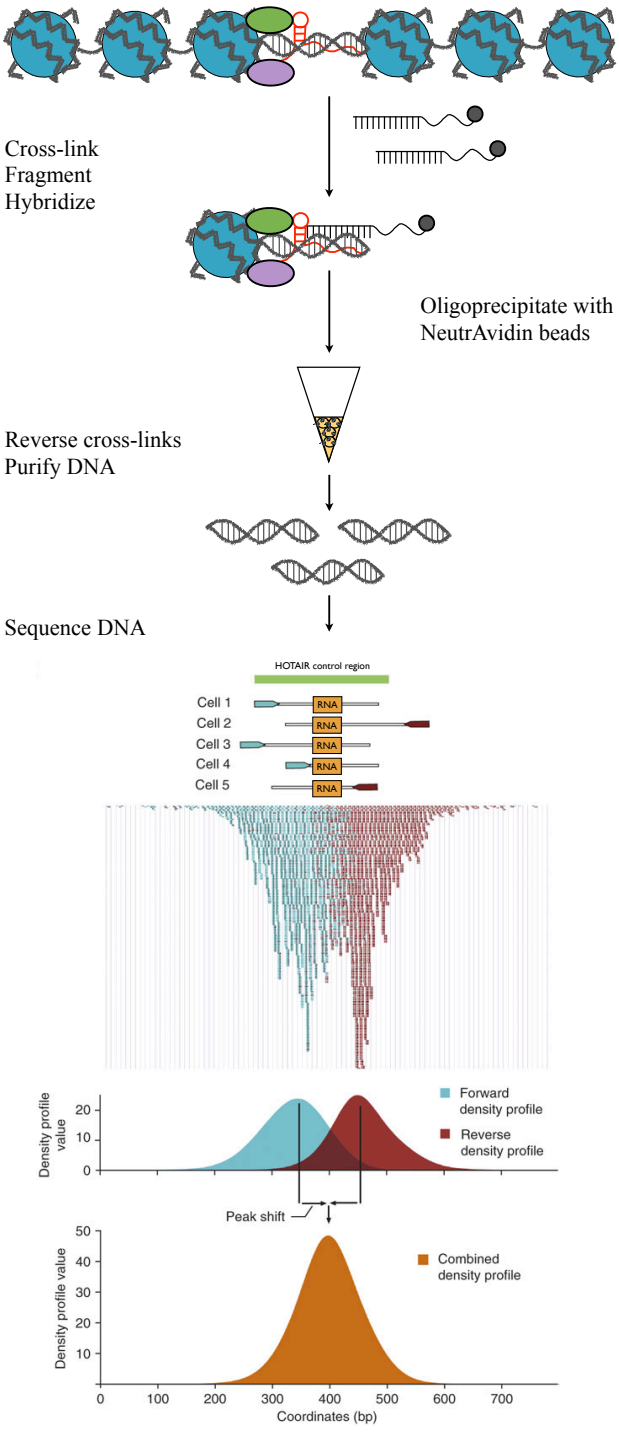


Figure 3.3. Schematic of the ChOP-seq assay. Chromatin is cross-linked, fragmented, and hybridized to oligonucleotide probes. Chromatin-bound oligo-probes are then precipitated with blocked beads. The bound genomic DNA is then purified and sequenced using an Illumina Sequencer. (Figure adapted from Valouev et al. 2008 *(176)*).

lncRNA HOTAIR. I first needed to find a cell type easily cultured in large quantities that contained detectable amounts of HOTAIR. Using RT-PCR, I found that both HeLa and Jurkat suspension cells express HOTAIR as shown in Figure 3.4. Both HeLa and Jurkat cells were used in the initial optimization experiments.

To monitor the stability of HOTAIR during the isolation of chromatin, I performed RT-PCR on cross-linked and non-cross-linked Jurkat cells at various steps in the isolation procedure. Figure 3.5 shows HOTAIR levels during cellular fractionation into cytoplasm (cyto), nuclei (NP), and then fractionating the nuclei into soluble (NE) and insoluble parts (NE*). Ultimately, the input chromatin would be taken from the NE* fraction. I found that cross-linking, and the subsequent reversal of cross-links during which the samples are subjected to 65^{o}C for 4-5 hours, decreased or eliminated detection of HOTAIR by RT-PCR. This prevented us from monitoring HOTAIR during the ChOP assay, which requires cross-linking.

Technical modifications of original ChOP

Our initial course of action was to optimize the ChOP assay for pulling down the lncRNA HOTAIR. To limit RNA degradation, I performed all ChOP processing steps and incubations at 4^{o}C and subjected samples to the shortest, most efficient incubation periods sufficient to see results. The input chromatin and ChOP-purified DNA were stored at -80^{o}C when not being used. I determined that two hour incubations, including the oligonucleotide precipitation step, showed comparable amounts of precipitated chromatin detected by PCR to that of traditional overnight precipitations.

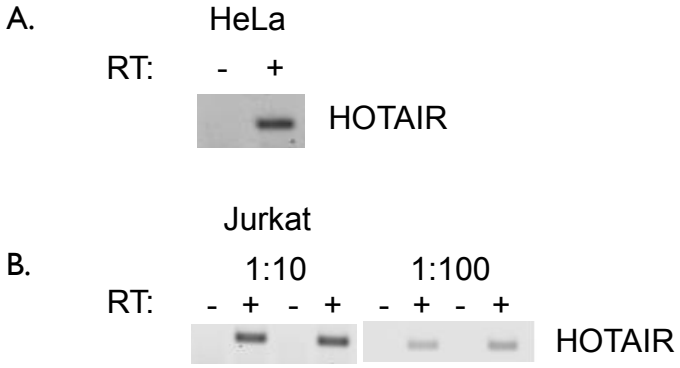


Figure 3.4. RT-PCR of RNA from A) HeLa and B) Jurkat cells shows HOTAIR is present in both. In the case of Jurkat cells, two amounts of RT sample was tested.

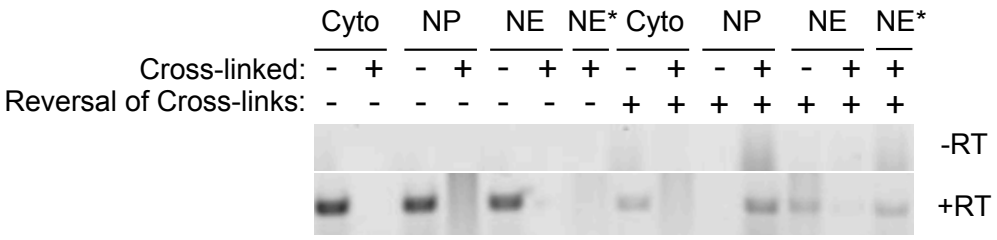


Figure 3.5. Detection of HOTAIR RNA by RT-PCR decreases after reversing cross-links for 5 hours at 65^{o}C, making it difficult to assess HOTAIR levels after ChOP as cross-linking is necessary to precipitate and purify the genomic DNA. Cyto= cytoplasm, NP= Nuclear Pellet, NE= nuclear extract, NE*= pellet from nuclear extract.

LncRNAs are anticipated to have secondary and higher order structures. To relax RNA structure and make the RNA sequence accessible to the oligonucleotide probes, the probes and cross-linked chromatin were incubated at 65^{o}C for 10 minutes and then slowly cooled to 4^{o}C. The incubation was also tried at 4^{o}C and at 90^{o}C for 3 minutes, but the 65^{o}C with a slow cool resulted in less background in the PCR analysis.

ChOP wash conditions were extended to 5 minutes with nutation at 4^{o}C with 1mL of buffer and 15 µL bead volume, whereas the original ChOP assay utilized 300 µL washes without nutation. This volume and incubation time reduced background. ChOP washes included standard low salt, high salt, LiCl, and two TE low washes which served to decrease salt concentration for negligible ionic interference during the elution step *(137)*.

I tested various elution conditions, including the traditional $NaHCO_3$ (pH 8.0) buffered elution, an RNase A/T1 and RNase H-assisted elution. The best results came from a published ChIP-seq method that eluted during the reverse cross-linking step with 1% SDS *(177)*. This elution streamlined the ChOP procedure and showed less background during PCR analysis.

Fragmentation of chromatin for ChOP assays

I was initially concerned that the extensive sonication of chromatin that is necessary to produce the 200-500 basepair fragments required for ChIP assays would also fragment the HOTAIR RNA. RT-PCR of sonicated chromatin, shown in Figure 3.6A, appeared to support this assumption, as HOTAIR could not be detected after sonication. So I decided to use DNase I as a

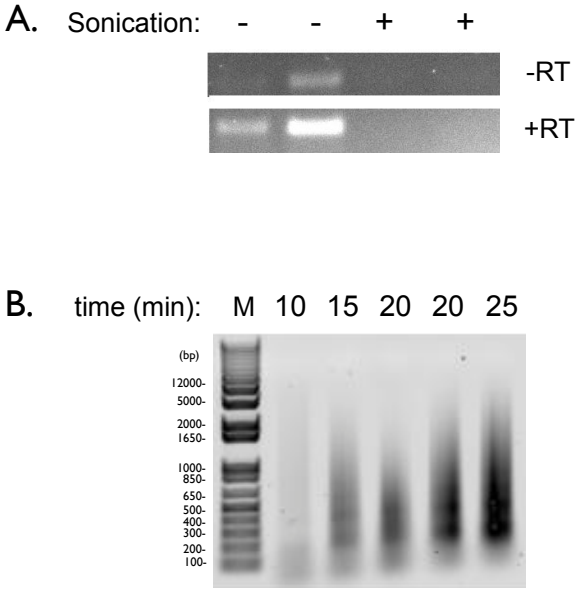


Figure 3.6. A) RT-PCR of chromatin from HeLa cells with and without sonication reveals no detection of HOTAIR RNA after cup sonication. B) Optimization of Jurkat chromatin fragmentation using DNase 1.

more gentle approach to fragmenting the chromatin. I had initial difficulty obtaining appropriately sized fragments of Jurkat chromatin, and found that limited cup sonication (1 to 2 times of 1 minute sustained sonication) followed by limited DNase 1 treatment produced satisfactory fragmentation of DNA between 200 and 600 basepairs. Figure 3.6B shows DNase 1 treatment for 10 to 25 minutes after limited sonication. Our concern about fragmentation of HOTAIR upon sonication was later validated by Chu et al. *(178)*.

Oligonucleotide probe design

I created two pools of DNA oligonucleotide probes that were antisense to HOTAIR with the goal of pulling down the RNA in two separate reactions for comparison and validation of precipitated chromatin. This comparative analysis was aimed at reducing false positive results because identifying overlapping regions of occupancy from the two pulldowns would indicate bonafide HOTAIR-associated genes or regions. The probes were designed with high T_m values to ensure that the DNA probe:HOTAIR RNA hybrid would survive the ChOP's stringent washing conditions, whereas nonspecific probe interactions would not. The 12 probes, listed in Table 3.1 and shown in Figure 3.7, were subjected to strict hairpin, and homo- and heterodimer analysis so that no hairpins would form during the incubation at 65^{o}C and minimal hetero- and homo-dimer base pairing would occur between the oligonucleotide probes. The oligo-probes were also analyzed using NCBI Blastn human genomic + transcript pairwise alignment to verify that no annealing would occur with endogenous RNAs other than HOTAIR.

	OLIGO	SEQUENCE: 5'- /5BioTEG/...
HOTAIR Pool A	HOT E1	CTTTCGGATCAAGCTCCAGAG
	HOT E3	CTTCTAAATCCGTTCCATTCCACTG
	HOT E5	GTGTAATTGCTGGTTTAGGTTGCAG
	HOT E6.2	ACAAGCCTCATCATAAAGATGGAGA
	HOT E6.4	CACTGCATAATCACTCCTGTATGGA
HOTAIR Pool B	HOT E2	GGGATATTAGGGACCTGAGGGTCTA
	HOT E4	GTCCTCCATTTCAGCCTTTTCTCT
	HOT E6.1	CTCTCTGTACTCCCGTTCCCTAGAT
	HOT E6.3	ATTCTTAAATTGGGCTGGGTCTACA
	HOT E6.5	AAGTGCATACCTACCCAATGTATGG
Other	E6.662	GAACCCTCTGACATTTGCCT
	E6.947	GACCTTTGCTTCTATGTTCCTCT
	Random	GAGTCGTGATACGGTTAGTGTGAGTG
	E4 RevComp	AGAGAAAAGGCTGAAATGGAGGAC

Table 3.1. Oligonucleotide probes used in ChOP assays. Probes were 5' conjugated with biotin and a triethyleneglycol linker (/5BioTEG/). E=exon followed by exon number (e.g. E4= exon 4).

Figure 3.7. DNA sequence shown is the template strand from which HOTAIR RNA is transcribed (reverse-complement of the RNA). Exon sequences are displayed (from top to bottom E6, E5, E4, E3, E2, and E1, respectively), and introns are shown as lines for simplicity. Oligonucleotide probes are underlined and in bold.

Evaluation of chromatin oligonucleotide precipitation

To test the effectiveness of our assay during optimization, I PCR amplified the ChOP-purified DNA for genes in the HOXD cluster (HOXD8, HOXD10, HOXD12, HOXD13) that were previously found to be regulated by HOTAIR *(123)*. As a negative control I also PCR amplified regions that were presumed not to be controlled by HOTAIR and, therefore, would not purify with a HOTAIR probe (IL-2, GAPDH (not shown), and Actin) *(123)*. Since we did not know the mechanism of how HOTAIR regulated HOXD genes, I designed three main sets of primer pairs to amplify HOXD genes either 200-500 basepairs upstream of the transcription start site (TSS) (_p), within 200 bases downstream of the TSS (_a), or in the last exon of each gene (_d). These are the primers used in my subsequent optimization experiments.

Optimization of cross-linking for ChOP assay

I used formaldehyde to create networks of cross-links between RNAs and proteins associated with chromatin. As HOTAIR associates with chromatin modifying factors, we predicted the HOTAIR RNA would be interacting with chromatin, either directly or indirectly *(123)*. A concern was that by under- or over-cross-linking, I would either miss weak interactions or yield a high level of background *(179)*. To optimize the cross-links for ChOP, I decided to titrate formaldehyde into *in vivo* cross-linking reactions for 10 or 20 minutes and test the capture of HOXD chromatin by the ChOP assay. I tested 0.1, 0.5, and 1% formaldehyde using a pool of HOTAIR oligo-probes and compared recovery of various HOXD loci to the SFPQ lncRNA locus as a negative control. Figure 3.8 shows that 0.1% formaldehyde cross-linking of Jurkat cells for

Figure 3.8. A ChOP was performed with Jurkat cells testing pool B of HOTAIR oligo-probes for HOXD gene recovery after formaldehyde cross-linking. Cells were cross-linked for 10 or 20 minutes with varying percentages of formaldehyde by volume followed by isolation of insoluble chromatin that was fragmented with DNase 1 for 25 minutes at 37^{o}C. Promoter regions (_p) or downstream (_d) regions of the genes were targeted in PCR of the eluted chromatin. R= probe with random sequence. All other lanes had a pool of HOTAIR oligo-probes.

20 minutes produced the greatest recovery of the various HOXD gene regions. The random oligo-probe did not pull down any of the HOXD gene regions. The HOTAIR probes did not pull down the SFPQ locus. I decided to continue the lncRNA ChOP optimization using the 20 minute, 0.1% formaldehyde cross-linking condition.

NeutrAvidin beads

Our previously published ChOP assay captured the biotinylated oligo-probes using an antibody against biotin immobilized on agarose beads (Ag) *(137)*. I compared this technology to using NeutrAvidin beads (NAg). NeutrAvidin is a chemically modified form of avidin (a cationic glycoprotein (pI 10.0) that tightly binds to biotin) in which the carbohydrate moiety is removed or deglycosylated. NeutrAvidin has no Arg-Tyr-Asp sequence and has a neutral isoelectric point of pH 6.3, thereby supporting very low non-specific binding (ThermoScientific). NeutrAvidin has a higher affinity for biotin than an α-biotin antibody, and therefore a reduced binding time.

I compared the NeutrAvidin beads with α-biotin-bound protein A/G beads using our ChOP assay (Figure 3.9). The ChOP compared a HOTAIR oligo-probe pool to a pool targeting another lncRNA (SFPQ) and to the random probe. Upstream region of HOXD12 shows reduced background for the random probe in the NAg bead samples but similar levels of HOXD gene pulldown with the lncRNA probes. It is not known whether SFPQ lncRNA regulates HOXD, but the data show SFPQ localizing to HOXD12. Another control experiment using beads-only versus lncRNA specific probes or a random probe also showed reduced nonspecific binding for the NAg compared to the antibody-bound Ag beads (not shown). Based on this increased signal to

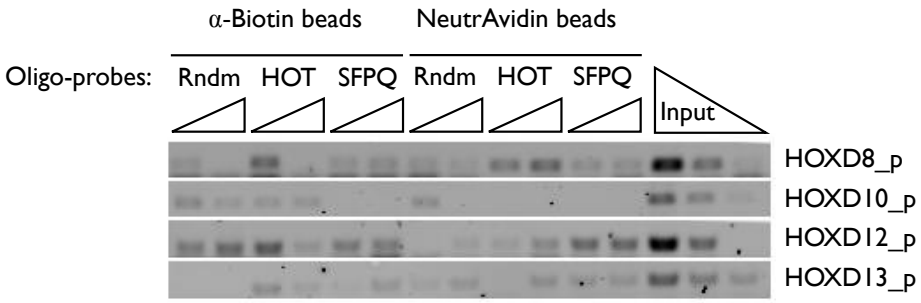


Figure 3.9. ChOP assay titrating oligo-probes for random, HOTAIR pool B, or SFPQ (pool of 5) compares α-Biotin and NeutrAvidin beads. Assay shows NeutrAvidin agarose beads exhibit reduced nonspecific binding compared to αbiotin-bound agarose beads. PCR of HOXD loci is shown.

background ratio, I used the NeutrAvidin beads in subsequent ChOPs.

Analysis of oligonucleotide probes

I wanted to compare the efficiency with which the two different oligonucleotide pools pulled down the HOXD locus in the ChOPs. Splitting the HOTAIR probes into two sets containing five probes each, I compared the two pools (A and B) and found that the B pool had significantly higher genomic recovery at the HOXD12 promoter, and comparable levels at the HOXD10 promoter to pool A (Figure 3.10A). Next I compared the oligo-probes within pool B to determine if a subset of the probes was more effective for pulling-down HOTAIR, or if any probe combinations appeared to decrease precipitation. Figure 3.10B shows a pool of probes containing E4, E6.1, and E6.3 precipitated more HOXD loci, and the addition of probes E2 and E6.5 were not inhibitory. Testing each probe individually revealed the probe targeting HOTAIR exon 4 (HOT E4 probe) was primarily responsible for the higher amount of HOXD pulldown in pool B (Figure 3.10C). Oligo-probe E6.3 also pulled down a significant amount of HOXD-containing chromatin. The E4 and E6.3 probes, along with the random probe, were used in subsequent studies.

Illumina sequencing

Jurkat cell ChOPs, followed by PCR, repeatedly suggested that the E4 oligo-probe was pulling down the HOXD locus in a specific manner. I decided to sequence an E4 oligo ChOP sample. The sample was prepared using the Illumina ChIP-sequencing amplification kit. Prior to submitting the sample for Illumina sequencing, we PCR analyzed the ChOP sample before (E4) and after (Seq) amplification with the kit (Figure 3.10D). The E4 and Seq samples showed

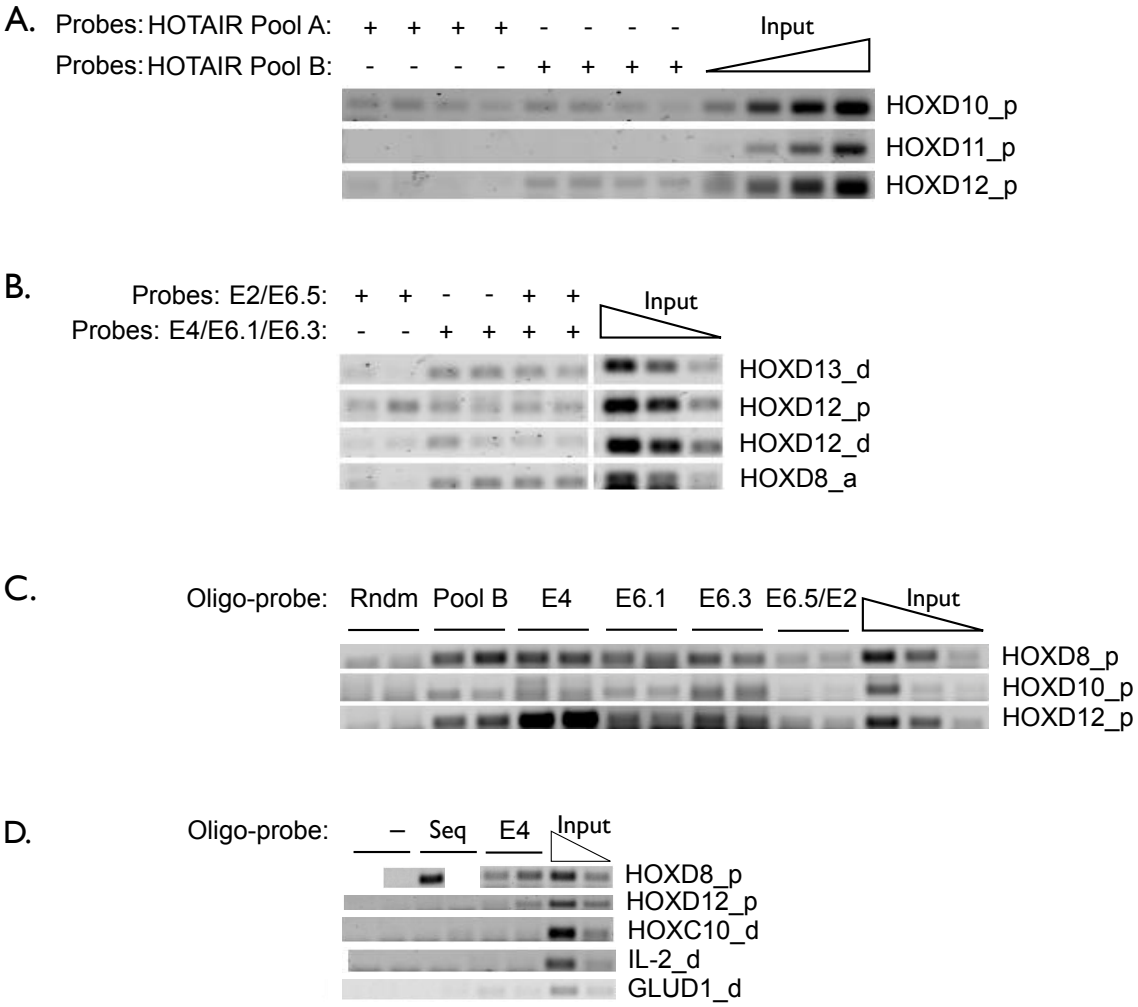


Figure 3.10. Investigation of probe specificity for pulling down chromatin in the ChOP assay. Equal amounts of probes were pooled and tested in ChOP assays followed by PCR for recovery of HOXD genes. A) HOTAIR pool B recovered greater quantities of DNA than pool A. B) Pool B was further divided into two pools and compared. C) Analysis of individual oligo-probes revealed that probe E4 precipitates the greatest amount of HOXD12 and HOXD8 promoter regions. D) E4 ChOP tested before (E4) and after (Seq) amplification in Illumina sequencing kit. Promoter regions (_p), upstream (_a) or downstream (_d) regions of the genes were targeted for PCR.

specificity at the HOXD8 loci, and no signal at the control regions tested (HOXC10, IL-2, and GLUD1). The returned Illumina sequences (1377 Mbases) were mapped with Bowtie *(180)* and the mapped sequences were run through the QuEST peak calling software *(176)*, which found peaks indicative of HOTAIR binding sites. I opted to use QuEST because I can specify ChOP-seq peak-calling parameters that represent different peak distribution morphologies: punctate (similar to sequence-specific transcription factors), somewhat broad distribution (Pol II-like), or very broad distribution (similar to histone marks).

I used the web-based data analysis program Galaxy to analyze the peaks *(181)*. Peak calls were joined with hg19 UCSC Main RefSeq genes which found that approximately 85% of the 3732 peaks called by QuEST were within RefSeq annotated genes. 1844 different genes contained peaks and approximately 53% of the peaks were within introns, whereas 52% of the total peaks were within 1000 basepairs of a TSS. Figure 3.11A shows these peak distributions along with those in other genomic regions. Figure 3.11B shows the binned tag density of peaks within 1000 bases of annotated gene transcriptional start sites. The peaks of occupancy concentrated approximately 200 bases upstream and downstream of the TSS. This signature is similar to the occupancy patterns of several histone modifications found in ChIP-seq experiments *(182, 183)*.

I performed MEME (Multiple EM for Motif Elicitation) motif analysis of two sets of peaks: 1) those from -1000 to +1000 with respect to annotated TSSs (promoter proximal) and 2) those not in the 1st category (promoter distal) (Figure 3.12) *(184)*. I ran multiple MEME

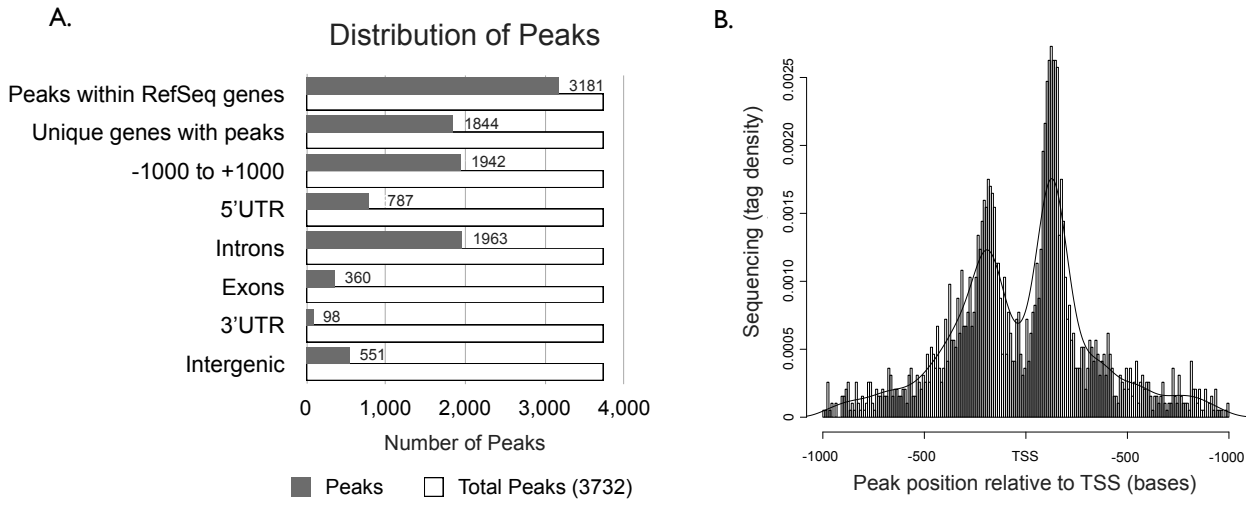
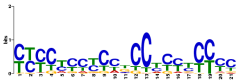
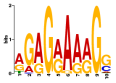
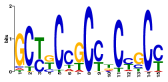


Figure 3.11. A) Analysis of sequencing peak distribution relative to human RefSeq genes reveals varied occupancy. B) Analysis of HOTAIR distribution surrounding the TSS of genes containing peaks between -1000 and +1000. The y-axis shows sequence tag density for HOTAIR occupancy.

analyses, with the maximum 300 sequences at a time, using 200 bp of sequence centered around each peak. Identified motifs in the promoter proximal data set included purine-rich motifs (motifs 1-3), a motif complimentary to the E4 probe (motif 3), and a sequence with 8 bases complimentary to the E4 oligo-probe (motif 5). Promoter distal peaks exhibited a GGAAT repeat sequence (motifs 9 and 10) and purine-rich sequences (motifs 6-7). MAST (Motif Alignment and Search Tool) revealed that both the promoter proximal and promoter distal peaks exhibited multiple repeats of the various motifs throughout the 200 bp sequences analyzed *(184)*. The HOTAIR-E4 probe had a high CT-content which may have resulted in the precipitation of purine-rich motifs from direct interaction with genomic DNA.

Evaluation of the sequencing data uncovered a few complicating issues. First, two of the motifs found in our MEME analysis matched the E4 oligo, which alludes to either direct oligo-DNA binding or the possibility of the oligo triplexing to the dsDNA. Second, no peaks were called within the HOXD loci. New ChOP assays were performed to validate the peaks identified in the sequencing data. These ChOPs showed occupancy at several of the Illumina-identified peaks, but also at control regions on various chromosomes not identified in the sequencing data. Figure 3.13A-C shows three different ChOP experiments testing six HOTAIR oligo-probes or beads only for recovery of Illumina peaks (NUBP1, ELP3, and PRMT2) versus IL-2 and CIITA negative control regions. ELP3 and NUBP1 were consistently recovered with all the probes, however, the control IL-2 region was also repeatedly precipitated.

Promoter-proximal peaks

Motif	Score	sites per 300 sequences	
1	5.3e-273	218	
2	4.9e-212	268	
3	1.2e-36	122	
4	1.3e-10	76	
5	2.3e-03	16	

Promoter-distal peaks

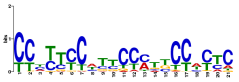
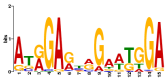
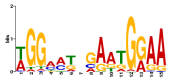
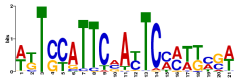
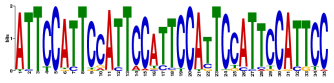
Motif	Score	sites per 300 sequences	
6	4.5e-351	295	
7	1.6e-373	267	
8	2.3e-264	264	
9	8.6e-508	178	
10	9.2e-863	67	

Figure 3.12. MEME analysis comparing promoter-proximal and promoter-distal (>1000 basepairs from a TSS) sequence motifs. 200 bp regions centered around the Illumina-identified peaks were analyzed 300 at a time. Many of the motifs were found in repeats within the 200 basepair regions. Motif 3 is complementary to a section of the E4 oligo-probe and Motif 5 contains a partial E4 oligo sequence (GGCTGAAA).

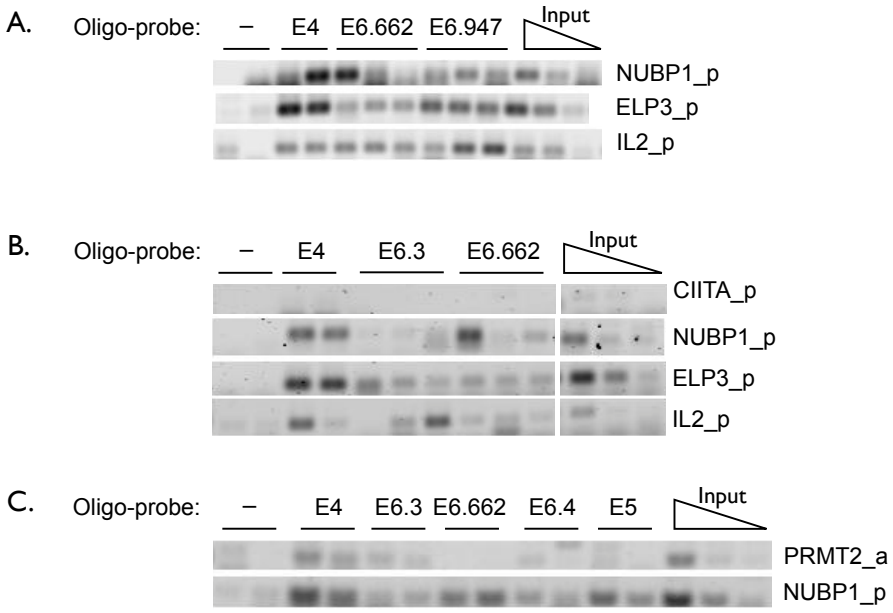


Figure 3.13. ChOP-PCR of gene regions identified from Illumina sequencing.

HOTAIR occupies genes outside of the HOXD locus

Concurrently to sequencing the Jurkat ChOP enriched DNA, a publication from the Chang lab *(141)* identified a set of genes de-repressed by siRNA knockdown of HOTAIR in primary foreskin fibroblasts. From this list I designed primers for promoter regions of the following genes: JUD1, SCN2A, BDNF, SIRT2, and GATA1. I then purchased the same foreskin fibroblast cell line they used and amplified those genes after ChOP assays. Figure 3.14A shows RT-PCR verifying HOTAIR expression in the foreskin fibroblast cells. An initial ChOP in the foreskin fibroblasts showed patterns of HOTAIR occupancy at the HOXD genes similar to Jurkat cells (Figure 3.14B). This experiment included a control oligo-probe that was antisense to the E4 oligo. I did not expect precipitation with this antisense probe as no RNA is produced that is antisense to HOTAIR in this genomic region *(123)*, however the anti-E4 probe pulled down select HOXD regions, perhaps by directly binding genomic DNA.

Figure 3.14C shows percent precipitation from a ChOP using foreskin fibroblast chromatin pre-treated with RNase A/T1 or untreated. I tested for recovery of chromatin in the peak regions of genes identified to be regulated by HOTAIR *(141)*. The RNase samples were included to determine that the ChOP enrichment depended on RNA and was not due to the oligo-probe interacting with genomic DNA. Comparing the RNase-treated to untreated samples, the E4 oligo-probe appeared to successfully pull down the SCN2A promoter in an RNA-dependent manner and the E6.3 oligo-probe appeared to pull down all but the SCN2A promoter. The results indicate that together the E4 and E6.3 probes pull down all the regions tested in an RNA-specific manner.

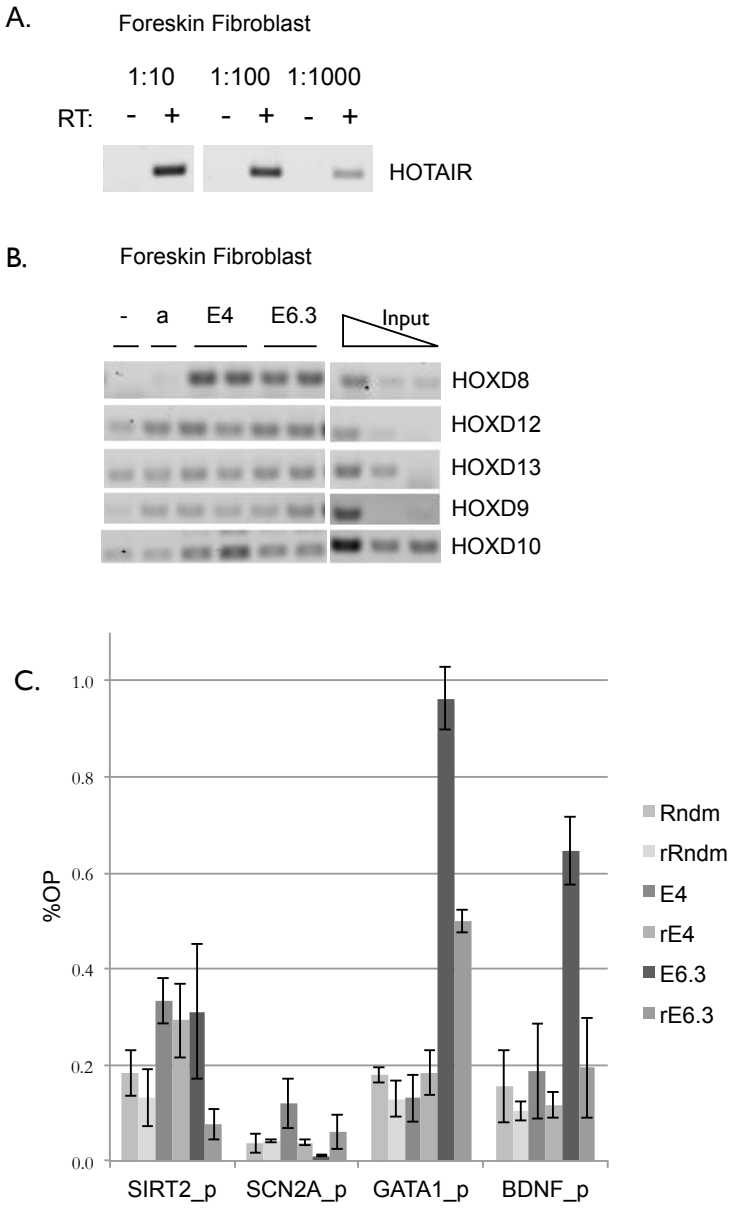


Figure 3.14. A) RT-PCR shows HOTAIR in human foreskin fibroblasts. B) ChOP-PCR for promoter regions in foreskin fibroblasts shows HOTAIR occupancy at several loci. Negative ChOP contained no probe (beads only) and a=oligo-probe antisense to the E4 oligo. C) Foreskin fibroblast ChOP-PCR showing %OP for genes identified by Tsai et al. *(141)*. A ChOP assay using probes E4, E6.3, or a random sequence was performed on input chromatin pre-treated (rRndm, rE4, rE6.3) with RNase A/T1 or untreated chromatin (Rndm, E4, E6.3). PCR products of duplicate ChOPs were ran on an agarose gel and quantitated using ImageJ.

Lastly, a paper published by Chu et al. in October of 2011, described a ChIRP assay, or Chromatin Isolation by RNA Purification, that was nearly identical to our ChOP assay *(178)*. They too targeted HOTAIR and sequenced their precipitated DNA. They identified 832 genomic sites occupied by HOTAIR in HOTAIR-overexpressed MDA-MB-231 breast cancer cells including NFKBIA, SERINC5, ABCA2, and an intergenic region between HOXD3 and HOXD4. I tested these four regions for HOTAIR occupancy using the ChOP assay and foreskin fibroblasts. Again, comparing ChOPs using RNase pre-treated chromatin alongside untreated chromatin, I tested the E4, E6.3, and random probes and found HOTAIR occupancy at all four gene loci (Figure 3.15A). The E4 oligo-probe pulled down HOXD3-4 and SERINC5, whereas E6.3 oligo pulled down all the regions. Figure 3.15B shows that both the E4 and E6.3 oligo-probes pulled down HOXD8, but probe E6.3 also pulled down the ACTB negative control region above the random oligo. Oligo-probe E4 did not pull-down ACTB or IL-2 above the random oligo. From this result, and others, I concluded that both probe E4 and E6.3 pulled-down verified HOTAIR-associated genomic regions, however the background precipitation of negative control regions remained problematic.

Lastly, I also compared the 832 ChIRP-seq identified sites to our 3732 ChOP-seq sites and found that only 14 overlapped. Of these 14 overlapping regions, 6 were in centromeres, 5 were mRNA genes and the other 3 were intergenic regions (data not shown). Because the ChIRP method was very similar to our ChOP assay, and the Chang lab already used ChIRP to map HOTAIR occupancy to the genome, we decided not to pursue further development of the ChOP-seq assay for investigation of the lncRNA HOTAIR.

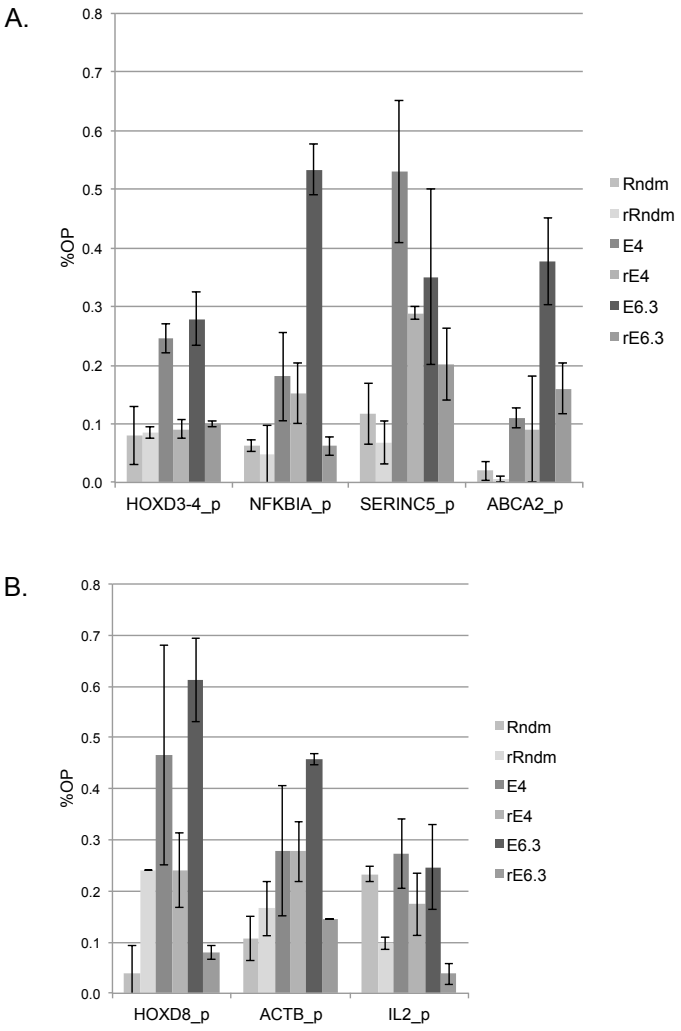


Figure 3.15. A) Foreskin fibroblast ChOP-PCR of genes identified by ChIRP. A ChOP assay using probes E4, E6.3 or the random sequence was performed on input chromatin pre-treated with RNase A/T1 alongside untreated chromatin. B) ChOP-PCR of HOXD8, along with the ACTB and IL2 negative control regions.

DISCUSSION

Here I describe the development of the ChOP assay and the genomic regions identified by ChOP-seq that are occupied by HOTAIR lncRNA. I optimized many facets of the technique, including cross-linking conditions, chromatin fragmentation, bead-type, and other parameters. The Illumina sequencing trial of our ChOP from Jurkat cells using the E4 oligo identified 3732 regions. However, it proved difficult to definitively confirm that the peaks were sites at which HOTAIR was localized as opposed to background. Our plan was originally to utilize a second oligonucleotide probe for a new ChOP-seq and to identify overlapping peaks, or to look at HOTAIR occupancy by ChOP-seq after HOTAIR siRNA knockdown. After Tsai et al. published their studies using foreskin fibroblasts, I purchased and began using this cell line *(141)*. In these cells, it appeared the ChOP assay was selectively pulling down chromatin in an RNA-dependent manner. I had planned to sequence material precipitated from these cells for comparison with the Jurkat data, however the recent publication of the ChIRP-seq paper *(178)* led us to make the decision to not further pursue ChOP-seq on HOTAIR.

To test the effectiveness of our lncRNA ChOP assay, we operated under the assumption that regions of the HOXD locus would be precipitated, and therefore detectable with PCR. Based on previous knockdown studies it was clear that HOTAIR regulated gene expression across the HOXD locus, however, any specific binding regions for HOTAIR were unknown *(123)*. One question we wanted to answer was how HOTAIR was distributed across the HOXD locus. It was unknown whether HOTAIR is required at promoters or throughout genes to function in transcriptional silencing, or if it uses a central global control region to more broadly control expression *(185)*. A single HOTAIR RNA may be capable of associating with PRC2 and relaying

an information cascade across multiple nucleosome to cover large genomic regions. The recent ChIRP-seq paper *(178)* identified an intergenic HOTAIR binding event in the region between HOXD3 and HOXD4 that may provide evidence that HOTAIR is indeed acting in this manner as they did not precipitate any other HOXD regions.

The ChOP assay could be further improved through use of other cross-linking reagents, including the cleavable amine-reactive NHS-esters EGS (Ethylene glycol-bis[succinimidyl succinate], 16.1 angstrom spacer) and DSP (N-hydroxysuccinimide (NHS) ester, 12 angstrom spacer), or dual cross-linking using formaldehyde and glutaraldehyde, which would facilitate expanded networks of cross-links (DNA to histones to chromatin modifying factors to lncRNA) *(186)*. The Chang group found glutaraldehyde gave superior results and low background in their ChIRP assay *(178)*.

Our method of chromatin fragmentation via DNase 1 treatment provides a nice alternative to obtain small fragments. DNase 1 treatment raises the obvious concerns that DNase hypersensitive regions of open chromatin, including enhancer and promoter regions, could bias the assay *(187)*. However, previous publications have shown DNase 1 elicits no sequence specific cleavage bias using naked DNA *(188)*. Moreover, sequencing sonicated chromatin has also revealed a bias toward open regions including promoters *(189)*.

A concern during oligonucleotide probe selection is that the oligonucleotide probes have the potential to nonspecifically bind to mRNAs, other ncRNAs, or DNA in open genomic regions. In addition, HOTAIR RNA exhibits secondary structure and is bound by PRC2 and the LSD1/REST/CoREST complex *(141)*, therefore, regions of the lncRNA may be blocked from binding some probes. The use of two different oligo-probes or two different pools of oligo-

probes in separate ChOP assays would circumvent some of these concerns. Our MEME analysis in Figure 3.12 identified motifs (Motif 3 and 5) found in a fraction of the regions with similar sequence to our HOTAIR E4 probe, which indicates the probe may have pulled down DNA directly by partial annealing with the chromatin. The two probe validation method would have excluded these regions during data analysis, as they would have only been pulled down by one probe set. Another recently published ChOP-like assay, termed CHART *(190)*, also saw direct binding of oligos to endogenous DNA regions in *Drosophila* chromatin.

The technical differences between ChOP, ChIRP, and CHART methods are outlined in Table 3.2. The ChIRP method *(178)* uses a series of oligonucleotide probes tiled across the entire HOTAIR RNA to circumvent the fragmentation of the RNA observed during sonication. They also used two pools of oligo-probes to compare recovered DNA sequences and the overlap was less than half the peaks in each individual pool. Our method of using DNase to fragment chromatin leaves the RNA intact and therefore it is unnecessary to target the entire RNA with a large quantity of costly modified probes. Furthermore, for the ChIRP assay the Chang lab used cells in which HOTAIR was over-expressed. It is possible that using HOTAIR over-expression may have provided results that do not reflect actual HOTAIR binding and activity. The vast majority of their regions did not not coincide with our ChOP-seq data. Differences may lie in the cell type used; my sequencing data was from Jurkat cells and theirs was from overexpressed breast cancer cells. Lastly, the genomic regions identified in their ChIRP-sequencing study do not appear to correlate with genes found to be regulated by HOTAIR in their previous fibroblast ChIP-chip and microarray experiments *(141)*. In addition, the ChIRP-seq findings were not compared to their previous studies that found genes to be regulated by HOTAIR in cells

	ChOP (see Ch3 Methods for specifics*)	**ChIRP** (see Chu et al. 2011)	**CHART** (see Simon et al. 2011)
lncRNA target	HOTAIR	HOTAIR	rox2
Cells	Jurkat, HeLa, or Foreskin Fibroblasts	MDA-MB-231 Breast cancer cells over-expressing HOTAIR	Drosophilia S2 expressing MSL3
Cross-linking	0.1% formaldehyde, 18-20 min room temperature (RT) in PBS	1% glutaraldehyde, 10 min RT in PBS	1% formaldehyde, 10 min RT, followed by nuclei in 3% formaldehyde, 30 min, RT in PBS
Chromatin enrichment	2 dounce steps	dounce followed by nuclei disruption in Bioruptor	dounce followed by nuclei disruption in sonicator
Fragmentation	Misonix cup sonicator 2×60s followed by limited DNase 1 treatment to 200-500 bp average size fragments	Bioruptor to 100-500 bp average size fragments	Covaris S2 instrument 2-3K bp average size fragments
Probes	5'-Biotin-TEG 24- or 25-mer probes 1 probe per ChOP compare 2 ChOPs	3'-Biotin 20-mer probes with 18 C spacer (against full-length HOTAIR sequence) 2 probe sets with 24 probes each per ChIRP	3'-Biotin-TEG 24- or 25-mer probes with RNase H elution 3 probes per CHART
Beads	NeutrAvidin	MyOne Streptavidin C1 magnetic	MyOne Streptavidin C1
Hybridization, binding and capture	hybridization 10 min 65°C, slow cool to RT* followed by 2 hour, 4°C capture 2 hour, 4°C	(750 mM NaCl, 1% SDS, 50 mM Tris 7.0, 1 mM EDTA, 15% formamide, add DTT, PMSF, PI, and Superase-in) hybridization 4 hours, 37°C capture 30 min 37°C	(20 mM Hepes pH 7.5, 817 mM NaCl, 1.9 M urea, 0.4% SDS, 5.7 mM EDTA, 0.3 mM EGTA, 0.03% sodium deoxycholate, 5×Denhardt's solution) capture overnight at RT
Wash buffers	low KCl, high KCl, LiCl, TE×2*	5 times with wash buffer (2x SSC, 0.5% SDS, add DTT and PMSF)	5 times with WB250 buffer (250 mM NaCl, 10 mM Hepes pH 7.5, 2 mM EDTA, 1 mM EGTA, 0.2% SDS, 0.1% N-lauroylsarcosine)
Elution	65°C reverse cross-linking followed by RNase A/T1 and Proteinase K	100 µg/ml RNase A and 0.1 Units/µl RNase H. 2× 37°C (100 mM NaCl, 10 mM Tris 7.0, 1 mM EDTA, 0.5% SDS)	10 Units RNase H/100 µl buffer 10 min RT (50 mM Hepes pH 7.5, 75 mM NaCl, 3 mM MgCl2, 0.125% N-lauroylsarcosine, 0.025% sodium deoxycholate, 20 Units/mL SUPERasIN, 5 mM DTT)

Table 3.2. Summary of technical differences between ChOP, ChIRP, and CHART.

containing over expressed HOTAIR *(144)*.

Overall, using different ChOP-like methods, our results, along with those of the Chang group further support broad epigenetic control by HOTAIR lncRNA. It will be interesting to determine where on the genome other PRC2-associated lncRNAs localize and how this contributes to the global control of gene expression. Another ChOP-like assay, termed CHART, was also recently published *(190)*. These methods are slightly different. It is likely that different methods may work better for different ncRNAs, which will depend on the function, length and other properties of the ncRNA of interest. We are likely to see many more studies of this kind determining genomic occupancies of a variety of ncRNAs.

MATERIALS AND EXPERIMENTAL METHODS

Cell culture

Suspension HeLa (courtesy of D. Taatjes) and Jurkat cells were grown to near confluency in JMEM (Sigma) or RPMI (Invitrogen) media, respectively, supplemented with 5% NCS or 10% FBS, respectively, and 100 U/mL penicillin (Invitrogen) and 100 µg/mL streptomycin (Invitrogen). Foreskin fibroblasts (ATCC CRL-2091) were cultured in EMEM (Gibco) media supplemented with 10% FBS and 100 U/mL penicillin (Invitrogen) and 100 µg/mL streptomycin (Invitrogen) according to ATCC guidelines, and split and harvested during the exponential growth phase. Cells were treated with 0.1% formaldehyde for 20 minutes at room temperature. Foreskin fibroblasts were first rinsed with 1X PBS to remove any cells in the suspension and then adherent cells were resuspended in PBS for the cross-linking procedure. Glycine was added to a final concentration of 0.125 M to stop the cross-linking reaction. The cells were washed twice with cold PBSM (1X PBS, 1.2 mM $MgCl_2$, 1X Roche complete protease inhibitor cocktail, 1mM $Na_2S_2O_4$, 1 mM DTT, 0.2 mM PMSF) and frozen with liquid nitrogen and stored at -80^{o}C.

Primer pairs used in RT-PCR and PCR analysis

All primers are listed 5’ to 3’. RT-PCR was done using the HOTAIR primer pairs forward AAGTGAAACCAGCCCTAGC and reverse CCCATGTGTCTCAAGATGC. PCR analysis of the ChOP reactions included the following primer sets. (_p) promoter, (_a) upstream, or (_d) downstream indicate region of gene targeted for PCR. HOXD8_p forward GGTGCGTCAAGGGTAAATGT, reverse CCAGATCCTGGCGTTATCAA; HOXD8_a forward GCTGTGGTGCGAAAATGC, reverse TACAAACCCCGGCTCTGG; HOXD9_p forward

TCAACCTCACCTCTGTAGGG, reverse GCGGTTCGCACTTAAAGG; HOXD10_p forward ACAACGCAGGGACCAACC, reverse GGGTCCAAAGGAAAGATCC; HOXD10_d forward CATGGCATTTTGAATACATCC, reverse ATCTCACCAAGGACAACTGC; HOXD11_p forward GGTGAACAAGAAGCAACAAGC, reverse ACTAGGTGTGAGGGTGTGAGG; HOXD12_p forward AAAATTGGGTAGTTTGTGATGC, reverse CCACAGAGAGGGTTTCTCC; HOXD12_d forward GTCAACGAATTCATCAACAGG, reverse GCGCTTCTTCTTCATACGC; HOXD13_p forward CCTCTTACCTGTGTGAAATGC, reverse TTTTTATCCTGTTGCTGAATCC; HOXD13_d forward GGTGGGAACTTACATACAGAACC, reverse TATTTTTCTTCTTGGACCTTGG; BDNF_p forward TTTAATGAGACACCCACCGCTGCT, reverse AGTCACATCGTGGTTCCGATTCTG; BDNF_d forward AAACTCTCAACCACCTTGGC, reverse GGGCGTTTGCGTAAATCTAT; SCN2A_a forward TGCAGTCTTCTTGGTGCCAGCTTA, reverse AAAGTGGTCCCTCTGCAGA; SCN2A_d forward CGTGTTTCAAGGCTACAGCA, reverse CTCTAGCCTCCCAACCTTCC; SIRT2_p forward ACTCTAGGGCTCAAATCGGGAACT, reverse AACCTGACTCCAGTCCTGAGACTT; GATA2_p forward TAGGTAACTGCGCTCGGACTGA, reverse AGCAGTAACTAACCACCAACTGCC; GATA2_d forward CCCTCTGAGCCCTTTGTTTA, reverse TTACAAAGGGAGGGCAAACT; NUBP_p forward AAAGGCGACGGAATGGAGGAGGT, reverse AGTTAGAGCAGCTCCCGACACTTT; ELP1_p forward CTCTATCCAGGAACTCTGTCCCAT, AACGTCTGAACGGAGTGGCATCTT reverse ; IL2_p forward CATACAGAAGGCGTTAATTGCATGA, reverse

CCCAAAGACTGACTGAATGGATGT; CIITA_p forward TTTGGTGTGGGAGTAGGCATGGTA, reverse CAGATGCCTCAAGACAAGCTGAGA; HOXD3-4_p forward ATAAGTGCGCAGGCAGAAGT, reverse CCGCGATTCAATACACACAA; NFKBIA_p forward ACAAGTTCAGCCAGGTGGTC, reverse GCCTAGCTGCACTGAGCAAT; SERINC5_p forward TTGACTGTTCTGCCCTGATG, reverse AGGGACGGAAAATGAGTCCT; ABCA2_p forward CTCTGCCTGTCAAACATGGA, reverse CAACCCCAAAATCCTGTACG. SFPQ_d forward TCGTACTGTTAGGCCCTTGG, reverse ACCCTTGCATGAAGAGCACC; GLUD1_d (see Mariner et al. *(137))*

Design of DNA oligonucleotide probes for oligonucleotide precipitation

DNA oligonucleotides were designed using Integrated DNA Technologies OligoAnalyzer 3.1, NCBI Blastn, and Oligo Perfect designer software (Invitrogen). Software recommended oligonucleotides were analyzed for the following: optimal melting temperatures at various salt concentrations, percent GC content, heterodimers, homodimers, secondary structure formation, 5'- and 3'-end GC content, GGG and nucleotide runs, and genome-wide basic local alignment using NCBI Blastn. Probes were ordered with a 5'-end biotin tag with a TEG linker (5'Biotin-TEG) from IDT. HOTAIR probes: E1, CTTTCGGATCAAGCTCCAGAG; E2, GGGATATTAGGGACCTGAGGGTCTA;
E3, CTTCTAAATCCGTTCCATTCCACTG; E4, GTCCTCCATTTCAGCCTTTTCTCT;
E5, GTGTAATTGCTGGTTTAGGTTGCAG; E6.1, CTCTCTGTACTCCCGTTCCCTAGAT;
E6.2, ACAAGCCTCATCATAAAGATGGAGA; E6.3, ATTCTTAAATTGGGCTGGGTCTACA;

E6.4, CACTGCATAATCACTCCTGTATGGA; E6.5, AAGTGCATACCTACCCAATGTATGG; E6.662, GAACCCTCTGACATTTGCCT; E6.947, GACCTTTGCTTCTATGTTCCTCT; and reverse complement to E4, AGAGAAAAGGCTGAAATGGAGGAC. SFPQ probes: GATAAAGATTGGGGGTAAGTTACAG, GGGTATATGAAGTAAGAGTTCCCTG, CTCTTTATTGGGGAAAAGTGAGAC, CCCATCAGGGTTCTATGTAATTTAG, TACCTCATTCTTTCCCATCTAAGTC, AAATCATTAAGTGAAAGGCAGTAAA, TCTAACACCCAATCACTAAACCAC, AAGTTGTGCTTTTGAAGACTTAGGT. Random probe: GAGTCGTGATACGGTTAGTGTGAGTG.

Preparation of chromatin

Cells were resuspended in 5 packed cell volumes (PCV) of Buffer A (10 mM HEPES, pH 7.9, 1.5 mM $MgCl_2$, 10 mM KCl, 1X Protease Inhibitor Cocktail, 1 mM $Na_2S_2O_4$, 1 mM DTT, 0.2 mM PMSF) and incubated for 20 minutes on ice, followed by centrifugation for 10 minutes at 4^{o}C at 3000 rpm. Pellets were resuspended in 2.5 PCVs of Buffer A and dounce homogenized 7 times on ice. Nuclear pellets were recovered by spinning 10 min at 3000 rpm. Pellets were resuspended in 2 PCV Buffer C (20 mM HEPES, pH 7.9, 25% glycerol, 417 mM NaCl, 1.5 mM $MgCl_2$, 0.2 mM EDTA, 1X Protease Inhibitor Cocktail, 1 mM $Na_2S_2O_4$, 1 mM DTT, 0.2 mM PMSF) and dounce homogenized 20 times on ice, then incubated with nutation for 30 minutes at 4^{o}C. Nuclear extract and the nuclear insoluble chromatin pellet were separated by centrifugation at 15,000 rpm for 30 min at 4^{o}C then frozen with liquid nitrogen and stored at -80^{o}C for up to 1.5 months.

Fragmentation of chromatin

Chromatin was fragmented using a Misonix XL2020 sonicator cup horn at power 4.5 two times for 1.5 minutes with 1 minute rest in between sonications. Insoluble chromatin pellets were resuspended to a concentration of 60 million cells per 1.5 mL with DNase1 buffer (10 mM Tris-HCl, pH 7.9, 2.5 mM $MgCl_2$, 0.5 mM $CaCl_2$, 10 mM NaCl, 0.05% NP40, 25 U/mL SUPERase-IN (Invitrogen)) in 2 mL sonication tubes. 10 units of RQ1 RNase-Free DNase1 (Promega) was added and allowed to incubate for 20 minutes at 37^{o}C. The fragmentation was quenched by adding 10 mM EDTA at 4^{o}C followed by centrifugation for 15 minutes to recover the fragmented supernatant, which was then stored at -80^{o}C. Fragmentation was confirmed by running reverse cross-linked, purified DNA on a 1% agarose gel at 130 V for 25 minutes. Chromatin was reverse cross-linked in 200 mM NaCl for 4 hours at 65^{o}C, RNase A/T1 (Ambion) treated for 30 minutes at 37^{o}C, Proteinase K treated for 70 minutes at 55^{o}C, phenol chloroform extracted, and ethanol precipitated to obtain reverse cross-linked, purified DNA *(177)*.

Chromatin oligonucleotide precipitation assays

ChOPs were done using chromatin prepared from 1 x 10^7 cells for Jurkat and HeLa or 2.5 x 10^6 cells for foreskin fibroblasts and resuspended in 1 mL of DB40 buffer (16.7 mM Tris-HCl, pH 7.9, 167 mM NaCl, 10 mM EDTA, 0.5% NP40, 0.2 mM PMSF, 25 U/mL SUPERase-IN (Invitrogen)). NeutrAvidin beads were washed 3 times with DB40 buffer and 50 µL 50% bead slurry was used to pre-clear chromatin for 2 hours at 4^{o}C. 100 µL of input chromatin (10%) was placed at 4^{o}C for later analysis. NAg beads were centrifuged at 2,000 rpm unless specified.

30 µL/sample washed NeutrAvidin beads were blocked for 2 hours at 4°C with 800 µg/mL yRNA and 1600 µg/mL BSA.

Pre-cleared chromatin was centrifuged at 3,500 rpm for 10 minutes and the chromatin supernatant was transferred to new tubes and centrifuged again. 75 pmoles oligonucleotide probe were incubated for 10 minutes with chromatin at 65°C, then slowly cooled to 4°C by placing a 65°C sample heatblock at room temperature for 10 minutes. The samples were then nutated at 4°C for the remainder of the 2 hour incubation. After washing the blocked beads 3 times with DB40, 30 µL of the 50% NAg bead slurry was added to the pre-incubated oligonucleotide probe and chromatin tubes and nutated at 4°C for 2 hours for oligonucleotide precipitation.

The oligonucleotide precipitation was then subjected to a series of washes to reduce background. The samples were washed for 5 minutes while nutating at 4°C using, sequentially, 1 mL of low salt (20 mM Tris-HCl, pH 7.9, 150 mM NaCl, 10 mM EDTA, 1X Triton, 0.1% SDS), high salt (20 mM Tris-HCl, pH 7.9, 500 mM NaCl, 10 mM EDTA, 1X Triton, 0.1% SDS), LiCl wash buffer (10 mM Tris-HCl, pH 7.9, 250 mM LiCl, 10 mM EDTA, 1% Deoxycholate, 1% NP40), followed by two quick washes with TE (10 mM Tris-HCl, pH 7.9, 1 mM EDTA) to wash away remaining salt in the samples. The washes were centrifuged for 3 minutes at 2,000 rpm and any remaining supernatant removed with a 10 µL pipette. For the second TE wash the samples were transferred to new 0.7 mL eppendorf tubes to reduce background.

100 µL input samples were diluted with 300 µL TE and treated the same as ChOP samples for the remainder of the protocol. Purification of DNA from these samples followed the method outlined by Schmidt et al. *(177)*. Samples were eluted and cross-links reversed for 12-15 hours at 65°C with elution buffer (50 mM Tris-HCl, pH 7.9, 10 mM EDTA, 1% SDS) with gentle

resuspension every 5 minutes for the first 15 minutes. Samples were diluted 2-fold with TE and RNase treated with 8 µL of an RNase A/T1 cocktail (Ambion) for 30 minutes at 37°C. Proteins in the sample were degraded using 80 µg Proteinase K in 5mM $CaCl_2$ for 65 minutes at 55°C. After a 5 minute centrifugation at 8,000 rpm the eluate was transferred to new 1.5 mL eppendorf tubes for phenol:chloroform extraction followed by ethanol precipitation. The samples were resuspended in 50 µL milliQ water and stored at -80°C for PCR analysis.

Preparation of sample for Illumina sequencing

After purification of the oligonucleotide-precipitated genomic DNA, the DNA was subjected to a series of modifications (blunt end the DNA, add A bases, and ligate sequencing adapters) using the Illumina ChIP-Sequencing Kit (Illumina) as per the Illumina instructions, except the DNA was subjected to PCR amplification before gel purification and not after. The modified DNA was then quantitated using a calf-thymus DNA standard curve, Quant-iT Picogreen reagent (Invitrogen), and a Tecan Safire II spectrophotometer. Samples were re-tested via PCR using controls to determine retention of sample quality post sequencing kit amplification. Samples were analyzed on an Illumina-based GAII system at the Tufts University School of Medicine genomics facility, yielding 40 nucleotide reads.

Analysis of sequencing data

The Illumina sequencing produced 1377 Mbases of sequence reads. Illumina sequences were aligned with the hg18 and hg19 human assemblies using Bowtie software yielding ~80% aligned reads *(180)*. Peak calls were generated using QuEST v2.4 under 1) transcription factor-

like, 2) Pol II-like, or 3) histone-like peak parameters and a 4 basepair peak call window *(176)*. Peak calls were then analyzed using the web-based platform Galaxy *(181)*.

BIBLIOGRAPHY

1. G. Orphanides, T. Lagrange, D. Reinberg, The general transcription factors of RNA polymerase II, *Genes Dev.* **10**, 2657 (1996).

2. G. J. Narlikar, H.-Y. Fan, R. E. Kingston, Cooperation between complexes that regulate chromatin structure and transcription, *Cell* **108**, 475–487 (2002).

3. A. M. Näär, B. D. Lemon, R. Tjian, Transcriptional coactivator complexes, *Annu. Rev. Biochem.* **70**, 475–501 (2001).

4. S. T. Smale, Core promoters: active contributors to combinatorial gene regulation, *Genes Dev.* **15**, 2503–2508 (2001).

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

6. E. Lehmann, F. Brueckner, P. Cramer, Molecular basis of RNA-dependent RNA polymerase II activity, *Nature* **450**, 445–449 (2007).

7. J. A. Goodrich, J. F. Kugel, Non-coding-RNA regulators of RNA polymerase II transcription, *Nat. Rev. Mol. Cell Biol.* **7**, 612–616 (2006).

8. T. Mercer, M. Dinger, J. Mattick, Long non-coding RNAs: insights into functions, *Nature Rev. Genet.* **10**, 155–159 (2009).

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

10. R. Breathnach, P. Chambon, Organization and expression of eucaryotic split genes coding for proteins, *Annu. Rev. Biochem.* **50**, 349–383 (1981).

11. M. Goldberg, Sequence analysis of Drosophila histone genes, Stanford University, Stanford (1979).

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

13. D. B. Starr, D. K. Hawley, TFIID binds in the minor groove of the TATA box, *Cell* **67**, 1231–1240 (1991).

14. M. Horikoshi, C. Bertuccioli, R. Takada, J. Wang, T. Yamamoto, R. G. Roeder, Transcription factor TFIID induces DNA bending upon binding to the TATA element, *Proc. Natl. Acad. Sci. USA* **89**, 1060-1064 (1992).

15. T. I. Lee, R. A. Young, Transcription of eukarytoic protein-coding genes, *Annu. Rev. Genet.* **34**, 77–137 (2000).

16. Y. Tokusumi, Y. Ma, X. Song, R. H. Jacobson, S. Takada, The new core promoter element XCPE1 (X Core Promoter Element 1) directs activator-, mediator-, and TATA-binding protein-dependent but TFIID-independent RNA polymerase II transcription from TATA-less promoters, *Mol. Cell. Biol.* **27**, 1844–1858 (2007).

17. S. T. Smale, D. Baltimore, The "initiator" as a transcription control element, *Cell* **57**, 103–113 (1989).

18. B. A. Purnell, P. A. Emanuel, D. S. Gilmour, TFIID sequence recognition of the initiator and sequences farther downstream in Drosophila class II genes, *Genes Dev.* **8**, 830–842 (1994).

19. T. Lagrange, A. N. Kapanidis, H. Tang, D. Reinberg, R. H. Ebright, New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB, *Genes Dev.* **12**, 34–44 (1998).

20. W. Deng, S. G. E. Roberts, A core promoter element downstream of the TATA box that is recognized by TFIIB, *Genes Dev.* **19**, 2418–2423 (2005).

21. T. Juven-Gershon, J. Hsu, J. T. Kadonaga, Perspectives on the RNA polymerase II core promoter, *Biochem. Soc. Trans.* **34**, 1047–1050 (2006).

22. T. W. Burke, J. T. Kadonaga, Drosophila TFIID binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters, *Genes Dev.* **10**, 711–724 (1996).

23. J. Kadonaga, The DPE, a core promoter element for transcription by RNA polymerase II, *Exp. Mol. Med.* **34**, 259–264 (2002).

24. U. Ohler, G. Liao, H. Niemann, G. M. Rubin, Computational analysis of core promoters in the Drosophila genome, *Genome Biol.* **3**, research0087.1-0087.12 (2002).

25. C. Y. Lim, The MTE, a new core promoter element for transcription by RNA polymerase II, *Genes Dev.* **18**, 1606–1617 (2004).

26. D.-H. Lee, N. Gershenzon, M. Gupta, I. P. Ioshikhes, D. Reinberg, B. A. Lewis, Functional characterization of core promoter elements: the downstream core element is recognized by TAF1, *Mol. Cell. Biol.* **25**, 9674–9686 (2005).

27. T. Zhou, C. M. Chiang, The intronless and TATA-less human TAF(II)55 gene contains a functional initiator and a downstream promoter element, *J. Biol. Chem.* **276**, 25503–25511 (2001).

28. T. Juven-Gershon, J. Hsu, J. Theisen, J. Kadonaga, The RNA polymerase II core promoter-the gateway to transcription, *Curr. Opin. Cell Biol.* **20**, 253–259 (2008).

29. M. Carey, The enhanceosome and transcriptional synergy, *Cell* **92**, 5–8 (1998).

30. J. E. Butler, J. T. Kadonaga, Enhancer-promoter specificity mediated by DPE or TATA core promoter motifs, *Genes Dev.* **15**, 2515–2519 (2001).

31. B. Li, M. Carey, J. L. Workman, The role of chromatin during transcription, *Cell* **128**, 707–719 (2007).

32. B. R. Cairns, The logic of chromatin architecture and remodelling at promoters, *Nature* **461**, 193–198 (2009).

33. T. Kouzarides, Chromatin modifications and their function, *Cell* **128**, 693–705 (2007).

34. J. F. Kugel, J. A. Goodrich, A kinetic model for the early steps of RNA synthesis by human RNA polymerase II, *J. Biol. Chem.* **275**, 40483–40491 (2000).

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

36. M. G. Peterson, N. Tanese, B. F. Pugh, R. Tjian, Functional domains and upstream activation properties of cloned human TATA binding protein, *Science* **248**, 1625–1630 (1990).

37. C. C. Kao, P. M. Lieberman, M. C. Schmidt, Q. Zhou, R. Pei, A. J. Berk, Cloning of a transcriptionally active human TATA binding factor, *Science* **248**, 1646–1650 (1990).

38. D. Reinberg, M. Horikoshi, R. G. Roeder, Factors involved in specific transcription in mammalian RNA polymerase II. Functional analysis of initiation factors IIA and IID and identification of a new factor operating at sequences downstream of the initiation site, *J. Biol. Chem.* **262**, 3322–3330 (1987).

39. E. J. Cho, S. Buratowski, Evidence that transcription factor IIB is required for a post-assembly step in transcription initiation, *J. Biol. Chem.* **274**, 25807–25813 (1999).

40. T. S. Pardee, C. S. Bangur, A. S. Ponticelli, The N-terminal region of yeast TFIIB contains two adjacent functional domains involved in stable RNA polymerase II binding and transcription start site selection, *J. Biol. Chem.* **273**, 17859-17864 (1998).

41. O. Flores, E. Maldonado, D. Reinberg, Factors involved in specific transcription by mammalian RNA polymerase II: Factors IIE and IIF independently interact with RNA polymerase II, *J. Biol. Chem.* **264**, 8913-8921 (1989).

42. M. C. Thomas, C.-M. Chiang, The general transcription machinery and general cofactors, *Crit. Rev. Biochem. Mol. Biol.* **41**, 105–178 (2006).

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

44. J. Fu, A. L. Gnatt, D. A. Bushnell, G. J. Jensen, N. E. Thompson, R. R. Burgess, P. R. David, R. D. Kornberg, Yeast RNA polymerase II at 5 Å resolution, *Cell* **98**, 799–810 (1999).

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

46. J. A. Davis, Y. Takagi, R. D. Kornberg, F. J. Asturias, Structure of the yeast RNA polymerase II holoenzyme: mediator conformation and polymerase interaction, *Mol. Cell* **10**, 409–415 (2002).

47. Y. Liu, C. Kung, J. Fishburn, A. Z. Ansari, K. M. Shokat, S. Hahn, Two cyclin-dependent kinases promote RNA polymerase II transcription and formation of the scaffold complex, *Mol. Cell. Biol.* **24**, 1721-1735 (2004).

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

49. D. Bentley, The mRNA assembly line: transcription and processing machines in the same factory, *Curr. Opin. Cell Biol.* **14**, 336–342 (2002).

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

51. M. J. Muñoz, M. de la Mata, A. R. Kornblihtt, The carboxy terminal domain of RNA polymerase II and alternative splicing, *Trends Biochem. Sci.* **35**, 497–504 (2010).

52. M. Sawadogo, R. G. Roeder, Energy requirement for specific transcription initiation by the human RNA polymerase II system, *J. Biol. Chem.* **259**, 5321–5326 (1984).

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

54. Y. Ohkuma, H. Sumimoto, M. Horikoshi, R. G. Roeder, Factors involved in specific transcription by mammalian RNA polymerase II: purification and characterization of general transcription factor TFIIE, *Proc. Natl. Acad. Sci. USA* **87**, 9163–9167 (1990).

55. P. Di Lello, L. M. Miller Jenkins, C. Mas, C. Langlois, E. Malitskaya, A. Fradet-Turcotte, J. Archambault, P. Legault, J. G. Omichinski, p53 and TFIIEalpha share a common binding site on the Tfb1/p62 subunit of TFIIH, *Proc. Natl. Acad. Sci. USA* **105**, 106–111 (2008).

56. Y. Ohkuma, S. Hashimoto, C. K. Wang, M. Horikoshi, R. G. Roeder, Analysis of the role of TFIIE in basal transcription and TFIIH-mediated carboxy-terminal domain phosphorylation through structure-function studies of TFIIE-alpha, *Mol. Cell. Biol.* **15**, 4856-4866 (1995).

57. D. Forget, B. Coulombe, Site-specific protein-DNA photocross-linking of purified complexes: topology of the RNA polymerase II transcription initiation complex, *Methods Enzymol.* **370**, 701–712 (2003).

58. O. Flores, H. Lu, D. Reinberg, Factors involved in specific transcription by mammalian RNA polymerase II: Identification and characterization of factor IIH, *J. Biol. Chem.* **267**, 2786–2793 (1992).

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

60. R. Drapkin, J. Reardon, A. Ansari, J. Huang, L. Zawel, K. Ahn, A. Sancar, D. Reinberg, Dual role of TFIIH in DNA excision repair and in transcription by RNA polymerase II, *Nature* **368**, 769–772 (1994).

61. T. Okamoto, S. Yamamoto, Y. Watanabe, T. Ohta, F. Hanaoka, R. G. Roeder, Y. Ohkuma, Analysis of the role of TFIIE in transcriptional regulation through structure-function studies of the TFIIEβ subunit, *J. Biol. Chem.* **273**, 19866–19876 (1998).

62. Y. Ohkuma, Multiple functions of general transcription factors TFIIE and TFIIH in transcription: possible points of regulation by trans-acting factors, *J. Biochem.* **122**, 481–489 (1997).

63. W. Feaver, J. Svejstrup, N. Henry, R. Kornberg, Relationship of CDK-activating kinase and RNA polymerase II, *Cell* **79**, 1103–1109 (1994).

64. P. Schultz, S. Fribourg, A. Poterszman, V. Mallouh, D. Moras, J. M. Egly, Molecular structure of human TFIIH, *Cell* **102**, 599–607 (2000).

65. M. Rossignol, I. Kolb-Cheynel, J. M. Egly, Substrate specificity of the cdk-activating kinase (CAK) is altered upon association with TFIIH, *EMBO J.* **16**, 1628–1637 (1997).

66. H. Qiu, E. Park, L. Prakash, S. Prakash, The Saccharomyces cerevisiae DNA repair gene RAD25 is required for transcription by RNA polymerase II, *Genes Dev.* **7**, 2161-2171 (1993).

67. T. K. Kim, R. H. Ebright, D. Reinberg, Mechanism of ATP-dependent promoter melting by transcription factor IIH, *Science* **288**, 1418-1421 (2000).

68. F. Tirode, D. Busso, F. Coin, J. M. Egly, Reconstitution of the Transcription Factor TFIIH: Assignment of Functions for the Three Enzymatic Subunits, XPB, XPD, and cdk7, *Mol. Cell* **3**, 87–95 (1999).

69. T. Makela, J. Parvin, K. Jaisang, P. Sharp, R. Weinberg, A Kinase-deficient transcription factor TFIIH is functional in basal and activated transcription, *Proc. Natl. Acad. Sci. USA* **92**, 5174–5178 (1995).

70. K. Y. Yankulov, D. L. Bentley, Regulation of CDK7 substrate specificity by MAT1 and TFIIH, *EMBO J.* **16**, 1638–1646 (1997).

71. D. O. Morgan, Principles of CDK regulation, *Nature* **374**, 131–134 (1995).

72. W. L. De Laat, N. G. J. Jaspers, J. H. J. Hoeijmakers, Molecular mechanism of nucleotide excision repair, *Genes Dev.* **13**, 768-785 (1999).

73. A. Devault, A. Martinez, D. Fesquet, J. Labbe, N. Morin, J. Tassan, E. Nigg, J. Cavadore, M. Doree, MAT1 ("menage à trois") a new RING finger protein subunit stabilizing cyclin H-cdk7 complexes in starfish and Xenopus CAK, *EMBO J.* **14**, 5027-5036 (1995).

74. S. Larochelle, J. Chen, R. Knights, J. Pandur, P. Morcillo, H. Erdjument-Bromage, P. Tempst, B. Suter, R. P. Fisher, T-loop phosphorylation stabilizes the CDK7-cyclin H-MAT1 complex in vivo and regulates its CTD kinase activity, *EMBO J.* **20**, 3749–3759 (2001).

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

76. D. T. Haile, J. D. Parvin, Activation of transcription in vitro by the BRCA1 carboxyl-terminal domain, *J. Biol. Chem.* **274**, 2113–2117 (1999).

77. S. Y. Wu, E. Kershnar, C. M. Chiang, TAFII-independent activation mediated by human TBP in the presence of the positive cofactor PC4, *EMBO J.* **17**, 4478–4490 (1998).

78. K. P. Kumar, S. Akoulitchev, D. Reinberg, Promoter-proximal stalling results from the inability to recruit transcription factor IIH to the transcription complex and is a regulated event, *Proc. Natl. Acad. Sci. USA* **95**, 9767–9772 (1998).

79. A. Fukuda, Y. Nogi, K. Hisatake, The regulatory role for the ERCC3 helicase of general transcription factor TFIIH during promoter escape in transcriptional activation, *Proc. Natl. Acad. Sci. USA* **99**, 1206-1211 (2002).

80. H. Xiao, A. Pearson, B. Coulombe, R. Truant, S. Zhang, J. L. Regier, S. J. Triezenberg, D. Reinberg, O. Flores, C. J. Ingles, Binding of basal transcription factor TFIIH to the acidic activation domains of VP16 and p53, *Mol. Cell. Biol.* **14**, 7013–7024 (1994).

81. J. Blau, H. Xiao, S. McCracken, P. O'Hare, J. Greenblatt, D. Bentley, Three functional classes of transcriptional activation domains, *Mol. Cell. Biol.* **16**, 2044–2055 (1996).

82. H. T. Timmers, Transcription initiation by RNA polymerase II does not require hydrolysis of the beta-gamma phosphoanhydride bond of ATP, *EMBO J.* **13**, 391-399 (1994).

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

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

85. C. Tyree, C. George, L. Lira-DeVito, Identification of a minimal set of proteins that is sufficient for accurate initiation of transcription by RNA polymerase II, *Genes Dev. 7, 1254-1265* (1993).

86. J. D. Parvin, H. T. M. Timmers, P. A. Sharp, Promoter specificity of basal transcription factors, *Cell* **68**, 1135–1144 (1992).

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

88. G. Pan, J. Greenblatt, Initiation of transcription by RNA polymerase II is limited by melting of the promoter DNA in the region immediately upstream of the initiation site, *J. Biol. Chem.* **269**, 30101–30104 (1994).

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

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

91. J. F. Kugel, J. A. Goodrich, Promoter escape limits the rate of RNA polymerase II transcription and is enhanced by TFIIE, TFIIH, and ATP on negatively supercoiled DNA, *Proc. Natl. Acad. Sci. USA* **95**, 9232–9237 (1998).

92. A. Dvir, R. C. Conaway, J. Weliky Conaway, Promoter Escape by RNA Polymerase II: A role for an ATP cofactor in supperssion of arrest by polymerase at promoter-proxial sites, *J. Biol. Chem.* **271**, 23352-23356 (1996).

93. D. S. Luse, G. A. Jacob, Abortive initiation by RNA polymerase II in vitro at the adenovirus 2 major late promoter, *J. Biol. Chem.* **262**, 14990–14997 (1987).

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

95. D. S. Luse, T. Kochel, E. D. Kuempel, J. A. Coppola, H. Cai, Transcription initiation by RNA polymerase II in vitro: At least two nucleotides must be added to form a stable ternary complex, *J. Biol. Chem.* **262**, 289–297 (1987).

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

97. A. R. Hieb, S. Baran, J. A. Goodrich, J. F. Kugel, An 8 nt RNA triggers a rate-limiting shift of RNA polymerase II complexes into elongation, *EMBO J.* **25**, 3100–3109 (2006).

98. F. C. Holstege, U. Fiedler, H. T. Timmers, Three transitions in the RNA polymerase II transcription complex during initiation, *EMBO J.* **16**, 7468–7480 (1997).

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

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

101. P. Čabart, D. S. Luse, Inactivated RNA polymerase II open complexes can be reactivated with TFIIE, *J. Biol. Chem.* (2011), **287**, 961-967.

102. L. Zawel, K. P. Kumar, D. Reinberg, Recycling of the general transcription factors during RNA polymerase II transcription, *Genes Dev.* **9**, 1479–1490 (1995).

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

104. D. E. Schones, K. Cui, S. Cuddapah, T.-Y. Roh, A. Barski, Z. Wang, G. Wei, K. Zhao, Dynamic regulation of nucleosome positioning in the human genome, *Cell* **132**, 887–898 (2008).

105. E. Segal, J. Widom, Poly(dA:dT) tracts: major determinants of nucleosome organization, *Curr. Opin. Struct. Biol.* **19**, 65–71 (2009).

106. K. Struhl, Naturally occurring poly (dA-dT) sequences are upstream promoter elements for constitutive transcription in yeast, *Proc. Natl. Acad. Sci. USA* **82**, 8419-8423 (1985).

107. J. L. Workman, R. E. Kingston, Alteration of nucleosome structure as a mechanism of transcriptional regulation, *Annu. Rev. Biochem.* **67**, 545–579 (1998).

108. J. L. Workman, Nucleosome displacement in transcription, *Genes Dev.* **20**, 2009–2017 (2006).

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

110. M. Vermeulen, K. W. Mulder, S. Denissov, W. W. M. P. Pijnappel, F. M. A. van Schaik, R. A. Varier, M. P. A. Baltissen, H. G. Stunnenberg, M. Mann, H. T. M. Timmers, Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4, *Cell* **131**, 58–69 (2007).

111. A. Shilatifard, Chromatin modifications by methylation and ubiquitination: implications in the regulation of gene expression, *Annu. Rev. Biochem.* **75**, 243–269 (2006).

112. R. Margueron, D. Reinberg, Chromatin structure and the inheritance of epigenetic information, *Nature Rev. Genet.* **11**, 285–296 (2010).

113. Y.-J. Kim, S. Björklund, Y. Li, M. H. Sayre, R. D. Kornberg, A multiprotein mediator of transcriptional activation and its interaction with the C-terminal repeat domain of RNA polymerase II, *Cell* **77**, 599–608 (1994).

114. C. Kanduri, Long noncoding RNA and epigenomics, *Adv. Exp. Med. Biol.* **722**, 174–195 (2011).

115. T. Nagano, P. Fraser, No-nonsense functions for long noncoding RNAs, *Cell* **145**, 178–181 (2011).

116. K. C. Wang, H. Y. Chang, Molecular mechanisms of long noncoding RNAs, *Mol. Cell* **43**, 904–914 (2011).

117. M. Guttman, I. Amit, M. Garber, C. French, M. F. Lin, D. Feldser, M. Huarte, O. Zuk, B. W. Carey, J. P. Cassady, M. N. Cabili, R. Jaenisch, T. S. Mikkelsen, T. Jacks, N. Hacohen, B. E. Bernstein, M. Kellis, A. Regev, J. L. Rinn, E. S. Lander, Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals, *Nature* **458**, 223–227 (2009).

118. A. M. Khalil, M. Guttman, M. Huarte, M. Garber, A. Raj, D. Rivea Morales, K. Thomas, A. Presser, B. E. Bernstein, A. van Oudenaarden, A. Regev, E. S. Lander, J. L. Rinn, Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression, *Proc. Natl. Acad. Sci. USA* **106**, 11667–11672 (2009).

119. M. N. Cabili, C. Trapnell, L. Goff, M. Koziol, B. Tazon-Vega, A. Regev, J. L. Rinn, Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses, *Genes Dev.* **25**, 1915–1927 (2011).

120. M. Baker, Long noncoding RNAs: the search for function, *Nature Methods* **8**, 379–383 (2011).

121. P. Leighton, R. Ingram, J. Eggenschwiler, A. Efstratiadis, S. Tilghman, Disruption of imprinting caused by deletion of the H19 gene region in mice, *Nature* **375**, 34–39 (1995).

122. R. R. Pandey, T. Mondal, F. Mohammad, S. Enroth, L. Redrup, J. Komorowski, T. Nagano, D. Mancini-Dinardo, C. Kanduri, Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation, *Mol. Cell* **32**, 232–246 (2008).

123. J. L. Rinn, M. Kertesz, J. K. Wang, S. L. Squazzo, X. Xu, S. A. Brugmann, L. H. Goodnough, J. A. Helms, P. J. Farnham, E. Segal, H. Y. Chang, Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs, *Cell* **129**, 1311–1323 (2007).

124. J. B. Heo, S. Sung, Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA, *Science* **331**, 76–79 (2011).

125. T. Hung, Y. Wang, M. F. Lin, A. K. Koegel, Y. Kotake, G. D. Grant, H. M. Horlings, N. Shah, C. Umbricht, P. Wang, Y. Wang, B. Kong, A. Langerød, A.-L. Børresen-Dale, S. K. Kim, M. van de Vijver, S. Sukumar, M. L. Whitfield, M. Kellis, Y. Xiong, D. J. Wong, H. Y. Chang, Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters, *Nature Genet.* **43**, 621–629 (2011).

126. M. Guttman, J. Donaghey, B. W. Carey, M. Garber, J. K. Grenier, G. Munson, G. Young, A. B. Lucas, R. Ach, L. Bruhn, X. Yang, I. Amit, A. Meissner, A. Regev, J. L. Rinn, D. E. Root, E. S. Lander, lincRNAs act in the circuitry controlling pluripotency and differentiation, *Nature* **477**, 295–300 (2011).

127. C. J. Brown, A. Ballabio, J. L. Rupert, R. G. Lafreniere, M. Grompe, R. Tonlorenzi, H. F. Willard, A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome, *Nature* **349**, 38–44 (1991).

128. J. Zhao, B. K. Sun, J. A. Erwin, J. J. Song, J. T. Lee, Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome, *Science* **322**, 750–756 (2008).

129. G. D. Penny, G. F. Kay, S. A. Sheardown, S. Rastan, N. Brockdorff, Requirement for Xist in X chromosome inactivation, *Nature* **379**, 131–137 (1996).

130. J. T. Lee, The X as model for RNA's niche in epigenomic regulation, *Cold Spring Harb. Perspect. Biol.* **2**, a003749 (2010).

131. F. Sleutels, R. Zwart, D. Barlow, The non-coding Air RNA is required for silencing autosomal imprinted genes, *Nature* **415**, 810–813 (2002).

132. T. Nagano, J. A. Mitchell, L. A. Sanz, F. M. Pauler, A. C. Ferguson-Smith, R. Feil, P. Fraser, The Air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin, *Science* **322**, 1717–1720 (2008).

133. U. A. Ørom, T. Derrien, M. Beringer, K. Gumireddy, A. Gardini, G. Bussotti, F. Lai, M. Zytnicki, C. Notredame, Q. Huang, R. Guigo, R. Shiekhattar, Long noncoding RNAs with enhancer-like function in human cells, *Cell* **143**, 46–58 (2010).

134. T.-K. Kim, M. Hemberg, J. M. Gray, A. M. Costa, D. M. Bear, J. Wu, D. A. Harmin, M. Laptewicz, K. Barbara-Haley, S. Kuersten, E. Markenscoff-Papadimitriou, D. Kuhl, H. Bito, P. F. Worley, G. Kreiman, M. E. Greenberg, Widespread transcription at neuronal activity-regulated enhancers, *Nature* **465**, 182–187 (2010).

135. F. De Santa, I. Barozzi, F. Mietton, S. Ghisletti, S. Polletti, B. K. Tusi, H. Muller, J. Ragoussis, C.-L. Wei, G. Natoli, A large fraction of extragenic RNA pol II transcription sites overlap enhancers, *PLoS Biol.* **8**, e1000384 (2010).

136. J. Feng, C. Bi, B. S. Clark, R. Mady, P. Shah, J. D. Kohtz, The Evf-2 noncoding RNA is transcribed from the Dlx-5/6 ultraconserved region and functions as a Dlx-2 transcriptional coactivator, *Genes Dev.* **20**, 1470–1484 (2006).

137. P. D. Mariner, R. D. Walters, C. A. Espinoza, L. F. Drullinger, S. D. Wagner, J. F. Kugel, J. A. Goodrich, Human Alu RNA is a modular transacting repressor of mRNA transcription during heat shock, *Mol. Cell* **29**, 499–509 (2008).

138. I. Martianov, A. Ramadass, A. Serra Barros, N. Chow, A. Akoulitchev, Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript, *Nature* **445**, 666–670 (2007).

139. K. C. Wang, Y. W. Yang, B. Liu, A. Sanyal, R. Corces-Zimmerman, Y. Chen, B. R. Lajoie, A. Protacio, R. A. Flynn, R. A. Gupta, J. Wysocka, M. Lei, J. Dekker, J. A. Helms, H. Y. Chang, A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression, *Nature* **472**, 120–124 (2011).

140. X. Zhang, Z. Lian, C. Padden, M. B. Gerstein, J. Rozowsky, M. Snyder, T. R. Gingeras, P. Kapranov, S. M. Weissman, P. E. Newburger, A myelopoiesis-associated regulatory

intergenic noncoding RNA transcript within the human HOXA cluster, *Blood* **113**, 2526–2534 (2009).

141. M.-C. Tsai, O. Manor, Y. Wan, N. Mosammaparast, J. K. Wang, F. Lan, Y. Shi, E. Segal, H. Y. Chang, Long noncoding RNA as modular scaffold of histone modification complexes, *Science* **329**, 689–693 (2010).

142. B. Tarchini, D. Duboule, Control of Hoxd genes' collinearity during early limb development, *Dev. Cell* **10**, 93–103 (2006).

143. P. Tschopp, D. Duboule, A regulatory “landscape effect” over the HoxD cluster, *Dev. Biol.* **351**, 288–296 (2011).

144. R. A. Gupta, N. Shah, K. C. Wang, J. Kim, H. M. Horlings, D. J. Wong, M.-C. Tsai, T. Hung, P. Argani, J. L. Rinn, Y. Wang, P. Brzoska, B. Kong, R. Li, R. B. West, M. J. van de Vijver, S. Sukumar, H. Y. Chang, Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis, *Nature* **464**, 1071–1076 (2010).

145. Y. Jiang, M. Yan, J. D. Gralla, A three-step pathway of transcription initiation leading to promoter clearance at an activation RNA polymerase II promoter, *Mol. Cell. Biol.* **16**, 1614–1621 (1996).

146. P. M. Lieberman, A. J. Berk, A mechanism for TAFs in transcriptional activation: activation domain enhancement of TFIID-TFIIA-promoter DNA complex formation, *Genes Dev.* **8**, 995–1006 (1994).

147. J. Liu, S. Akoulitchev, A. Weber, H. Ge, S. Chuikov, D. Libutti, X. W. Wang, J. W. Conaway, C. C. Harris, R. C. Conaway, D. Reinberg, D. Levens, Defective interplay of activators and repressors with TFIH in xeroderma pigmentosum, *Cell* **104**, 353–363 (2001).

148. F. Sauer, S. K. Hansen, R. Tjian, Multiple TAFIIs directing synergistic activation of transcription, *Science* **270**, 1783-1788 (1995).

149. S. Y. Wu, C. M. Chiang, TATA-binding protein-associated factors enhance the recruitment of RNA polymerase II by transcriptional activators, *J. Biol. Chem.* **276**, 34235–34243 (2001).

150. S. Narayan, S. G. Widen, W. A. Beard, S. H. Wilson, RNA polymerase II transcription. Rate of promoter clearance is enhanced by a purified activating transcription factor/cAMP response element-binding protein, *J. Biol. Chem.* **269**, 12755-12763 (1994).

151. Y. S. Lin, M. R. Green, Mechanism of action of an acidic transcriptional activator in vitro, *Cell* **64**, 971–981 (1991).

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

153. S. Borukhov, V. Sagitov, C. Josaitis, R. Gourse, A. Goldfarb, Two modes of transcription initiation in vitro at the rrnB P1 promoter of Escherichia coli, *J. Biol. Chem.* **268**, 23477-23482 (1993).

154. A. Dvir, Promoter escape by RNA polymerase II, *Biochim. Biophys. Acta* **1577**, 208–223 (2002).

155. J. R. Weaver, J. F. Kugel, J. A. Goodrich, The sequence at specific positions in the early transcribed region sets the rate of transcript synthesis by RNA polymerase II in vitro, *J. Biol. Chem.* **280**, 39860–39869 (2005).

156. M. Pal, D. S. Luse, Strong natural pausing by RNA polymerase II within 10 bases of transcription start may result in repeated slippage and reextension of the nascent RNA, *Mol. Cell. Biol.* **22**, 30-40 (2002).

157. M. Pal, D. S. Luse, The initiation–elongation transition: Lateral mobility of RNA in RNA polymerase II complexes is greatly reduced at+ 8/+ 9 and absent by+ 23, *Proc. Natl. Acad. Sci. USA* **100**, 5700-5705 (2003).

158. M. T. Knuesel, K. D. Meyer, C. Bernecky, D. J. Taatjes, The human CDK8 subcomplex is a molecular switch that controls Mediator coactivator function, *Genes Dev.* **23**, 439–451 (2009).

159. R. Drapkin, D. Reinberg, The multifunctional TFIIH complex and transcriptional control, *Trends Biochem. Sci.* **19**, 504–508 (1994).

160. S. Buratowski, S. Hahn, L. Guarente, P. A. Sharp, Five intermediate complexes in transcription initiation by RNA polymerase II, *Cell* **56**, 549–561 (1989).

161. S. Akoulitchev, D. Reinberg, The molecular mechanism of mitotic inhibition of TFIIH is mediated by phosphorylation of CDK7, *Genes Dev.* **12**, 3541-3550 (1998).

162. Y. Jiang, J. D. Gralla, Nucleotide requirements for activated RNA polymerase II open complex formation in vitro, *J. Biol. Chem.* **270**, 1277–1281 (1995).

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

164. J. Liu, L. He, I. Collins, H. Ge, D. Libutti, J. Li, J. M. Egly, D. Levens, The FBP interacting repressor targets TFIIH to inhibit activated transcription, *Mol. Cell* **5**, 331–341 (2000).

165. G. LeRoy, R. Drapkin, L. Weis, D. Reinberg, Immunoaffinity purification of the human multisubunit transcription factor IIH, *J. Biol. Chem.* **273**, 7134–7140 (1998).

166. J. Dgnam, P. Martin, B. Shastry, R. Roeder, Eukaryotic gene transcription with purified components, *Methods Enzymol.* **101**, 582–598 (1983).

167. M. Furuno, K. C. Pang, N. Ninomiya, S. Fukuda, M. C. Frith, C. Bult, C. Kai, J. Kawai, P. Carninci, Y. Hayashizaki, J. S. Mattick, H. Suzuki, Clusters of internally primed transcripts reveal novel long noncoding RNAs, *PLoS Genet.* **2**, e37 (2006).

168. T. Ravasi, H. Suzuki, K. C. Pang, S. Katayama, M. Furuno, R. Okunishi, S. Fukuda, K. Ru, M. C. Frith, M. M. Gongora, S. M. Grimmond, D. A. Hume, Y. Hayashizaki, J. S. Mattick, Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome, *Genome Res.* **16**, 11–19 (2006).

169. J. Ponjavic, C. P. Ponting, G. Lunter, Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs, *Genome Res.* **17**, 556–565 (2007).

170. J. Wilusz, H. Sunwoo, Long noncoding RNAs: functional surprises from the RNA world, *Genes Dev.* **23**, 1494-1504(2009).

171. R. Lanz, N. McKenna, S. Onate, U. Albrecht, J. Wong, S. Tsai, M. Tsai, B. O'Malley, A steroid receptor coactivator, SRA, functions as an RNA and is present in an SRC-1 complex, *Cell* **97**, 17–27 (1999).

172. K. L. Yap, S. Li, A. M. Muñoz-Cabello, S. Raguz, L. Zeng, S. Mujtaba, J. Gil, M. J. Walsh, M.-M. Zhou, Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a, *Mol. Cell* **38**, 662–674 (2010).

173. N. Brockdorff, A. Ashworth, G. F. Kay, V. M. McCabe, D. P. Norris, P. J. Cooper, S. Swift, S. Rastan, The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus, *Cell* **71**, 515–526 (1992).

174. C. M. Croce, LINCing chromatin remodeling to metastasis, *Nature Biotech.* **28**, 931–932 (2010).

175. R. Kogo, T. Shimamura, K. Mimori, K. Kawahara, S. Imoto, T. Sudo, F. Tanaka, K. Shibata, A. Suzuki, S. Komune, S. Miyano, M. Mori, Long noncoding RNA HOTAIR regulates polycomb-dependent chromatin modification and is associated with poor prognosis in colorectal cancers, *Cancer Res.* **71**, 6320–6326 (2011).

176. A. Valouev, D. S. Johnson, A. Sundquist, C. Medina, E. Anton, S. Batzoglou, R. M. Myers, A. Sidow, Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data, *Nature Methods* **5**, 829–834 (2008).

177. D. Schmidt, M. D. Wilson, C. Spyrou, G. D. Brown, J. Hadfield, D. T. Odom, ChIP-seq: Using high-throughput sequencing to discover protein–DNA interactions, *Methods* **48**, 240–248 (2009).

178. C. Chu, K. Qu, F. L. Zhong, S. E. Artandi, H. Y. Chang, Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions, *Mol. Cell* **44**, 667–678 (2011).

179. C. Klockenbusch, J. Kast, Optimization of Formaldehyde Cross-Linking for Protein Interaction Analysis of Non-Tagged Integrin β1, *Journal of Biomedicine and Biotechnology* **2010**, 1–14 (2010).

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

181. J. Goecks, A. Nekrutenko, J. Taylor, Galaxy Team, Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences, *Genome Biol.* **11**, R86 (2010).

182. A. Barski, S. Cuddapah, K. Cui, T.-Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev, K. Zhao, High-resolution profiling of histone methylations in the human genome, *Cell* **129**, 823–837 (2007).

183. Z. Wang, C. Zang, J. A. Rosenfeld, D. E. Schones, A. Barski, S. Cuddapah, K. Cui, T.-Y. Roh, W. Peng, M. Q. Zhang, K. Zhao, Combinatorial patterns of histone acetylations and methylations in the human genome, *Nature Genet.* **40**, 897–903 (2008).

184. T. L. Bailey, M. Bodén, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li, W. S. Noble, MEME SUITE: tools for motif discovery and searching, *Nucleic Acids Research*, **37**, W202-W208 (2009).

185. F. Spitz, F. Gonzalez, D. Duboule, A global control region defines a chromosomal regulatory landscape containing the HoxD cluster, *Cell* **113**, 405–417 (2003).

186. P. Zeng, C. R. Vakoc, Z. Chen, G. A. Blobel, S. L. Berger, In vivo dual cross-linking for identification of indirect DNA-associated proteins by chromatin immunoprecipitation, *Biotech.* **41**, 694-698 (2006).

187. G. Crawford, I. Holt, J. Mullikin, D. Tai, E. Green, T. Wolfsberg, F. Collins, Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites, *Proc. Natl. Acad. Sci. USA* **101**, 992-997 (2004).

188. G. E. Crawford, S. Davis, P. C. Scacheri, G. Renaud, M. J. Halawi, M. R. Erdos, R. Green, P. S. Meltzer, T. G. Wolfsberg, F. S. Collins, DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays, *Nature Methods* **3**, 503–509 (2006).

189. R. K. Auerbach, G. Euskirchen, J. Rozowsky, N. Lamarre-Vincent, Z. Moqtaderi, P. Lefrançois, K. Struhl, M. Gerstein, M. Snyder, Mapping accessible chromatin regions using Sono-Seq, *Proc. Natl. Acad. Sci. USA* **106**, 14926–14931 (2009).

190. M. D. Simon, C. I. Wang, P. V. Kharchenko, J. A. West, B. A. Chapman, A. A. Alekseyenko, M. L. Borowsky, M. I. Kuroda, R. E. Kingston, The genomic binding sites of a noncoding RNA, *Proc. Natl. Acad. Sci. USA* (2011), **108**, 20497-20502.