

DEVELOPMENT AND APPLICATION OF A GENOMIC METHOD TO MAP  
THE POSITION, AMOUNT, AND ORIENTATION OF TRANSCRIPTIONALLY  
ENGAGED RNA POLYMERASES

A Dissertation

Presented to the Faculty of the Graduate School  
of Cornell University

In Partial Fulfillment of the Requirements for the Degree of  
Doctor of Philosophy

By

Leighton James Core

May 2009

© 2009 Leighton James Core

DEVELOPMENT AND APPLICATION OF A GENOMIC METHOD TO MAP  
THE POSITION, AMOUNT, AND ORIENTATION OF TRANSCRIPTIONALLY  
ENGAGED RNA POLYMERASES

Leighton James Core, Ph. D.

Cornell University 2009

RNA polymerases are highly-regulated molecular machines that can be modulated at the level of recruitment to a gene promoter, pre-initiation complex formation, initiation, elongation, and termination. Studies using the chromatin immunoprecipitation assay coupled to genomic DNA microarrays (ChIP-chip) or to high throughput sequencing (ChIP-seq) indicate that RNA polymerase II (Pol II) is present at disproportionately higher levels near the 5' end of many eukaryotic genes relative to downstream regions. This pattern is consistent with Pol II being either in a pre-initiation complex, or transcriptionally engaged and paused proximal to the promoter. Promoter-proximal pausing is proposed to be an important post-initiation, rate-limiting target for gene regulation, and usually occurs within the first 20-50 bases of a transcription unit. However, the ChIP assay cannot determine whether Pol II is simply promoter-bound or engaged in transcription. The goal of this dissertation project was to develop a method that would assess the generality of promoter-proximal pausing, genome-wide.

To that end, I have developed a highly-sensitive method, Global Run-On sequencing (GRO-seq), that maps the position, amount, and orientation of transcriptionally-engaged RNA polymerases across the entire genome. We

have applied GRO-seq to a primary human fibroblast cell line (IMR90). In this method, nuclear run-on reactions allow RNA polymerase to incorporate BrU affinity-tags into nascent RNA. The RNA is fragmented, purified at least 10,000 fold, and subjected to large-scale parallel sequencing. Mapping these sequences to the genome in this cell types shows that 30% of all genes have promoter-proximal paused polymerase, that transcription continues kilobases beyond the 3' cleavage for many genes, and that antisense transcription is prevalent. Surprisingly, in addition to promoter-proximal paused polymerase, most promoters also have an engaged polymerase upstream and in an orientation opposite to the annotated gene. This divergent polymerase is associated with active genes, but does not elongate effectively beyond the promoter. These results imply that the interplay between polymerases and regulators over broad promoter regions dictates the orientation and efficiency of productive transcription.

## BIOGRAPHICAL SKETCH

Leighton was born on March 14, 1977, two days after his mother's birthday, which is one day before his father's. Except for the first year of his life, he grew up in Dracut, Massachusetts, with his two older brothers, Lyle and Lee. He attended Dracut Senior High. He graduated from Boston College in 1999 with a B.S. in Biology but very little research experience. He moved to San Diego, to live the good life for a while, and ended up catching a break when he was hired to work in the lab of James Hoch and Marta Perego at the Scripps Research Institute. Literally, that's what Jim said in the interview, "I'm gonna cut you a break . . . give you a chance to work here." There Leighton got some lab experience, published a little, and decided to head back east to graduate school. He sold his surfboard, and drove across to attend Cornell University starting in 2002, because the people seemed down to earth there and he thought it looked like good fishing grounds. In the first few years he landed a few decent fish in the tributaries of Cayuga Lake, but had less luck at the bench. He was forced to declare himself a non-fisherman for the last several years so he could concentrate on the dissertation presented herein, and not concern himself with tying his flies or scouting fishing holes. It seems to have worked, but he'd like to get back to fishing someday soon.

To my family

## ACKNOWLEDGMENTS

First and foremost I'd like to thank my advisor, John Lis. His enthusiasm for science is a formidable force in the lab. He is an amazingly creative and incisive experimentalist. On a number of occasions, when I've thought I had an experiment or plan mastered, he has shown me a better way. I usually walk away - humbled, wondering, "Why didn't I think of that?" I don't think it is a stretch to surmise that John has thought through most experiments that happen in this lab long before they are made reality. He's merely waiting for a set of hands to get them done, or for technology to catch up to his brain. John's office door is always open, and he is always willing to help troubleshoot, plan, and brainstorm – I'm pretty sure he thrives on these on moments.

I'd like to thank members of my committee, Mariana Wolfner and Lee Kraus, for keeping track of my progress, and offering advice along the way.

I'd like to sincerely thank Josh Waterfall for adeptly reformulating my hypotheses into algorithms that produced answers. Nearly all the computations presented in this dissertation were performed by him, and without his contributions, I'd still be buried in frustrating heaps of unanswered questions. At the height of the flurry, we interpreted results together almost daily, and formulated new hypotheses that have never been asked at such an immense scale. So many of which, we have yet to fully explore them all. Together, we have learned a great deal about how the genome is organized and transcribed. This is only the beginning.

I'd like to thank my family for their love and support, and for trying as hard as they could not to ask too many times, "When are you finishing?" Their

enduring faith in me has sustained my will in times when I couldn't envision how I could complete this endeavor. I'd like to thank my friends, some scattered all over the place, some here in Ithaca. Finally, I'd like to thank Abbie Saunders for - when my third project was spiraling into the depths of failure - suggesting that I start doing nuclear run-ons. Most of all, I'd like to thank Abbie for making my time in Ithaca fun, exciting, and warm.

## TABLE OF CONTENTS

BIOGRAPHICAL SKETCH. . . . .	iii
ACKNOWLEDGEMENTS. . . . .	v
TABLE OF CONTENTS. . . . .	vii
LIST OF FIGURES. . . . .	xi
LIST OF TABLES. . . . .	xiii
CHAPTER1: INTRODUCTION	
1.1 Early stages of transcription . . . . .	2
1.1.1 Access of transcription factors to promoters. . . . .	2
1.1.2 Transcription initiation. . . . .	4
1.1.3 Promoter escape. . . . .	5
1.1.4 In vitro models of post-promoter escape transcription. . .	10
1.1.5 Promoter proximal pausing. . . . .	12
1.2 Regulation of Promoter proximal pausing. . . . .	14
1.2.1 Nucleosome positioning elements. . . . .	18
1.2.2 Sequence specific DNA binding factors. . . . .	19
1.2.3 The CTD, Cdk7 kinase, and the connection to RNA processing. . . . .	20
1.2.4 Promoter sequence elements. . . . .	24
1.2.5 Maintenance paused Pol II. . . . .	25
1.2.6 Escape from pausing. . . . .	28
1.3 Possible biological roles of pausing. . . . .	29
1.4 Generality of promoter-proximal pausing. . . . .	31
1.5 Concluding remarks. . . . .	32
1.6 Dissertation outline. . . . .	32

CHAPTER 2: DEVELOPMENT OF A METHOD TO MAP THE POSITION,  
AMOUNT, AND ORIENTATION OF RNA POLYMERASES IN GENOMES.

2.1 Introduction. . . . .	34
2.2 Materials and methods. . . . .	36
2.3 Development of the global nuclear run-on. . . . .	44
2.3.1 Incorporation of Br-UTP by nuclear RNA polymerases . . . . .	44
2.3.2 Control of resolution for GRO-seq. . . . .	46
2.3.3 Yield, enrichment and purity of nascent RNA after triple selection. . . . .	49
<i>Enrichment by tracking radiolabeled NRO-RNAs. . . . .</i>	49
<i>Measurement of enrichment and purity by RT-qPCR. . . . .</i>	51
2.4 The rationale for choosing sequencing vs. microarray hybridization for global analysis. . . . .	51
2.5 Overview of GRO-seq method. . . . .	52
2.6 Preliminary analyses. . . . .	57
2.6.1 Correlation between biological replicates. . . . .	57
2.6.2 Computational determination of background. . . . .	61
2.7 Transcription in nuclei: reflection of in vivo transcription status. . . . .	61
2.8 Concluding remarks. . . . .	66

CHAPTER 3: MASSIVE SEQUENCING OF NASCENT RNAS REVEALS  
DISTRIBUTIONS OF ENGAGED RNA POLYMERASE IN THE HUMAN  
GENOME

3.1 Introduction. . . . .	67
3.2 Materials and Methods. . . . .	70
3.2 Results. . . . .	74

3.2.1 GRO-seq reads relative to annotated transcript boundaries. . . . .	74
3.2.2 Comparison of GRO-seq with Pol II ChIP-chip data . . . .	74
3.2.2 Comparison of GRO-seq to microarray expression data. . . . .	79
3.2.3 Validation of GRO-seq gene activity by RT-qPCR. . . . .	82
3.2.4 Generality Promoter-proximal pausing revealed by GRO-seq. . . . .	85
3.2.5 Relationship of pausing with gene activity. . . . .	87
3.2.6 Gene Ontology of paused genes. . . . .	91
3.2.7 GRO-seq results for known paused genes. . . . .	95
3.2.8 Divergent transcription is associated with active promoters. . . . .	95
3.2.9 Transcription beyond the 3-end of genes. . . . .	108
3.2.10 Antisense transcription in gene regions. . . . .	110
3.3 Discussion and perspectives. . . . .	110
3.3.2 Pausing, termination, or both? . . . . .	110
3.3.3 Divergent transcription and histone modifications. . . . .	115
3.3.4 Possible functions for divergent transcription. . . . .	115
3.4 Concluding remarks. . . . .	116

## CHAPTER 4: FUTURE DIRECTIONS

4.1 Further Adaptations of GRO-seq. . . . .	118
4.1.1 Mapping engaged RNA polymerases with near nucleotide resolution. . . . .	118
4.1.2 Mapping Transcription start sites with GRO-seq. . . . .	121

4.2 Characterizing the genome-wide transcription response to heat shock in <i>Drosophila melanogaster</i> . . . . .	124
4.3 Cell cycle control of transcription. . . . .	125
4.4 Concluding remarks. . . . .	126
APPENDIX A: ISOLATION OF IN-VIVO FORMED PROTEIN/DNA COMPLEXES BY NUCLEOPROTEIN HYBRIDIZATIONS	
A.1 Introduction. . . . .	127
A.2 Materials and Methods. . . . .	133
A.3 Results. . . . .	138
A.4 Concluding remarks. . . . .	158
REFERENCES . . . . .	161

## LIST OF FIGURES

1.1 Transition from initiation to early elongation: promoter escape. . . . .	7
1.2 Regulation of entry and escape of Pol II at pause sites . . . . .	17
1.3 Promoter sequence elements. . . . .	27
2.1 Incorporation of Br-U by nuclear RNA polymerases. . . . .	43
2.2 Binding of Br-U RNAs to $\alpha$ BrdU beads. . . . .	45
2.3 Control of polynucleotide incorporation by polymerases. . . . .	47
2.4 Efficiency of BrU-RNA binding in response to titration of limiting nucleotide. . . . .	48
2.5 Level of enrichment $\alpha$ BrdU bead purification. . . . .	50
2.6 Overview of the GRO-seq method. . . . .	55
2.7 Denaturing PAGE analysis of fractions from GRO-seq library preparations. . . . .	58
2.8 Native PAGE analysis of an amplified NRO-library. . . . .	59
2.9 Correlation of GRO-seq biological replicates. . . . .	60
2.10 Analysis of background in low density windows. . . . .	62
2.11 Comparison of the density of sequence reads in exons vs. introns. . . . .	63
3.1 Summary of GRO-seq read distribution relative to annotated transcript boundaries. . . . .	75
3.2 Example of GRO-seq data in a large chromosome domain as viewed in the UCSC genome browser. . . . .	76
3.3 Example of a novel promoter identified by GRO-seq. . . . .	78
3.4 Comparison of gene activity by microarray vs. GRO-seq. . . . .	80
3.5 Validation of GRO-seq gene activity calls by RT-qPCR. . . . .	84
3.6 Alignment of GRO-seq reads to transcription start sites . . . . .	86

3.7 Histogram showing the distribution of pausing indexes. . . . .	88
3.8 Four classes of genes revealed by GRO-seq. . . . .	90
3.9 Box plots comparing promoter-proximal read density and pausing indexes with gene activity level. . . . .	92
3.10 Histogram of pausing indexes as a function of gene activity. . . . .	93
3.11 Gene ontology analysis of paused genes. . . . .	94
3.12 GRO-seq profiles for known paused genes. . . . .	96
3.13 GRO-seq reads aligned to transcription start sites without known bi-directional promoters. . . . .	99
3.14 Box plots comparing divergent read density with promoter-proximal peaks gene activity level. . . . .	101
3.15 Reciprocal box plots comparing divergent read density with promoter- proximal peaks. . . . .	102
3.16 Comparison of GRO-seq to ChIP profiles for Pol II and histone modifications. . . . .	104
3.17 Comparison of distance transcribed by forward vs. divergent polymerases. . . . .	106
3.18 ChIP profiles for histone modifications at genes without significant divergent transcription. . . . .	107
3.19 Aligment of transcripts to the 3' end of genes. . . . .	109
3.20 Examples of antisense transcription in the genome. . . . .	111
4.1 Schematic showing method for mapping the polymerase active site with high resolution. . . . .	120
4.2 Example to transcription start site mapping by GRO-seq. . . . .	122
A.1 Schematic of nucleoprotein hybridization experiment. . . . .	130
A.2 ChIP of HSF at Hsp70 promoter under mock denaturing conditions. . . .	139

A.3 Isolation of Hsp70 promoter with biotinylated DNA oligos. . . . .	142
A.4 Summary of HSF ChIP experiments after denaturing protocol . . . . .	143
A.5 Schematic of PNA design. . . . .	146
A.6 PNAs do not bind at concentrations feasible for large-scale experiments. . . . .	148
A.7 In vitro crosslinking HSF and HSEs with formaldehyde. . . . .	151
A.8 Denaturing of dsDNA within HSF/HSE complexes. . . . .	153
A.9 Formaldehyde crosslink stability in ionic or non-ionic detergents. . . . .	154
A.10 Thermal stability of HSF/HSE crosslinks after buffer exchange.....	156
A.11 Quenched formaldehyde maintains HSF/HSE complexes during thermal denaturing. . . . .	157

## LIST OF TABLES

2.1 Gene deserts used to calculate background of GRO-seq. . . . .	64
3.1 Comparison of the gene activity call by microarray and GRO-seq. . . . .	81
3.2 Pairwise comparison of pausing, gene activity, divergent transcription, and CpG islands. . . . .	103

## CHAPTER 1

### INTRODUCTION<sup>1</sup>

The regulation of gene expression by RNA polymerase II (Pol II) is fundamental to the growth, development and survival of an organism. While there are many cellular processes that control the final output of a gene, none are more direct and energetically efficient than at the level of transcription of the RNA itself. This RNA is often processed co-transcriptionally to produce a mature RNA that can itself be functional, but for most known genes is a message (mRNA) for translation into protein. Multiple factors regulate transcription by positively or negatively influencing the ability of RNA polymerases to access, bind and transcribe specific genes or RNAs in response to the appropriate signals. The three main stages of transcription are broadly known as initiation, elongation, and termination, and each stage is subject to regulation (Sims et al., 2004). These stages can each be further dissected into numerous discrete biochemical steps that can be targets of gene regulation. Initiation involves recruitment of Pol II to a gene promoter, formation of a preinitiation complex (PIC), and initiation of transcription. Elongation can be further divided into three distinct phases: promoter escape, promoter-proximal pausing, and productive elongation. Termination involves a

---

<sup>1</sup> Parts of this introduction appears in two previous reviews (Core and Lis, 2008;Saunders et al., 2006), and is reused or reworked here with permission from the publishers.

conformational change in the ternary complex that renders it non-processive; resulting in the release of the RNA from the ternary complex and ejection of Pol II from the template. All three main stages are also marked by changes in the modification state of the Pol II C-terminal domain (CTD) that dramatically affects its conformation and ability to associate with different factors. The phases of elongation and termination are defined by a marked difference in the stability and behavior of the ternary complex, composed of the polymerase, the template DNA, and the nascent RNA. To understand how regulation works for a particular gene, it is important to identify which of these steps is rate limiting and how signal-responsive activators and repressors act on them mechanistically.

## **1.1 Early stages of transcription**

### **1.1.1 Access of transcription factors to promoters**

The eukaryotic genome is packaged into chromatin that can occlude binding of the transcription machinery to promoter sequences. The most abundant component of chromatin is the nucleosome. The nucleosome consists of 147bp of DNA wrapped around an octameric complex of histone proteins, consisting of an H3:H4 tetramer and two H2A:H2B dimers. Nucleosomes are connected by linker DNA that can interact with a single fifth histone, H1, which helps stabilize the formation of more compact, higher order chromatin structures (Khorasanizadeh, 2004). This packaging results in a generally non-permissive environment for gene expression. Eukaryotes have developed a number of strategies that chemically and structurally transform chromatin in a way that can either prevent or allow access of the transcription machinery to gene promoters.

Histones have N-terminal tails that protrude from the core nucleosome and are substrates for modifying enzymes (Kouzarides, 2007a; Kouzarides, 2007b). Additionally, large ATP-dependent nucleosome-remodeling complexes can modify the structure and position of nucleosomes by either sliding them along, or evicting them from the template (Li et al., 2007). In the case of providing promoter access to transcription machinery these processes are intimately linked and can act synergistically with each other and the transcribing polymerase to activate gene transcription. Histone modifying and remodeling complexes are generally found to be recruited to promoters directly through interactions with DNA-binding activators, or indirectly through large coactivator complexes, but can also be effectors of certain histone modifications via specialized binding domains. In reference to the latter, acetylation of H3 and H4 tails is a strong mark for active promoters and is carried out by histone acetyltransferases (HATs) that are contained in large coactivator complexes (e.g p300, or the GCN5 subunit of SAGA), and even general transcription factors (the HAT domain TFIID subunit TAF1) (Kouzarides, 2007a). Acetylation can then both target the SWI/SNF remodeling complex via its acetyl-lysine specific bromo domain (Hassan et al., 2002), and destabilize nucleosome-DNA interactions which makes the nucleosome a more efficient substrate for the remodeling process (Brower-Toland et al., 2005; Lee et al., 1993). Tri-methylation of histone H3 at lysine 4 (H3K4me3) is also associated with active promoters, and H3K4me3 was recently shown to be important for anchoring TFIID to promoters (Vermeulen et al., 2007). Also, the SET1 methyltransferase in yeast interacts specifically with Ser5 phosphorylated Pol II (Ng et al., 2003). These observations not only point to a specific role for H3K4me3 modification during activation, but suggest

it could be part of a positive feedback loop that stimulates subsequent rounds of initiation. These examples highlight only a few of the examples how activators, histone modifiers, nucleosome remodelers, and transcribing Pol II can work together to promote efficient promoter access and binding of the general transcription machinery within the context of chromatin.

### **1.1.2 Transcription initiation**

Before transcription begins TATA Binding Protein (TBP), general transcription factors (TFIIA, B, E, F and H) and polymerase assemble on the promoter to form the preinitiation complex (PIC). PIC formation is a rate-limiting step during basal transcription, but can be circumvented in subsequent rounds of initiation during activated transcription by the maintenance of a scaffold complex, consisting of activator, Mediator, TBP, TFIIA, TFII E and TFII H, on the promoter (Reviewed in Cramer, 2004; Hahn, 2004; Orphanides et al., 1996). Several models of PIC structure have been proposed based on crosslinking, cryo-EM and crystallography studies of the Pol II with a complete set of GTFs or various subcomplexes (Asturias, 2004; Boeger et al., 2005; Chen et al., 2007; Hahn, 2004; Miller and Hahn, 2006). In all of these models the PIC clearly makes contacts over an extended region both upstream and downstream of the TATA box and initiation start site, highlighting the extensive cooperative interactions between GTFs, DNA and Pol II used to direct initiation at a specific location. The PIC is a stable complex that must be at least partially dismantled eventually for Pol II to move out from the promoter. In the start of this process, ~ 11-15 bases of the DNA around the start site are unwound in an ATP dependent event (Holstege et al., 1997; Wang et al., 1992; Yan and Gralla, 1997) to form an 'open complex' configuration that

allows the single stranded template DNA to enter the active site of Pol II (Cramer, 2004); although the precise mechanism of how the template DNA is delivered to the active site remains unknown. Transcription initiation occurs rapidly at this stage as Pol II catalyzes the first phosphodiester linkage.

### **1.1.3 Promoter escape**

Immediately following initiation, the initially transcribing complex (ITC) undergoes abortive transcript synthesis and upstream transcript slippage indicating that ITC is unstable and structurally distinct from a productive elongation complex (see Figure 1.1). Abortive initiation describes the continued synthesis and release of short RNAs, whereas transcript slippage refers to the sliding of the polymerase and nascent RNA upstream along the template, such that a sequence of bases, present only once in the template, are repeated in the RNA product. Productive elongation requires transition of the ITC through these phases and into a stable ternary complex in a process often referred to as promoter escape (sometimes called promoter clearance). Promoter escape includes a series of steps during which the polymerase breaks its contacts with promoter sequence elements and promoter-bound factors and simultaneously tightens its grip on the nascent RNA. These steps are regulated by intrinsic interactions of the polymerase, template, and nascent RNA that are dictated by core promoter structure and sequence, and are vulnerable to regulation by extrinsic factors (Dvir, 2002). An intermediate of Pol II escape from the abortive phase is evident by the formation of a metastable ternary complex after addition of the fourth nucleotide (Cai and

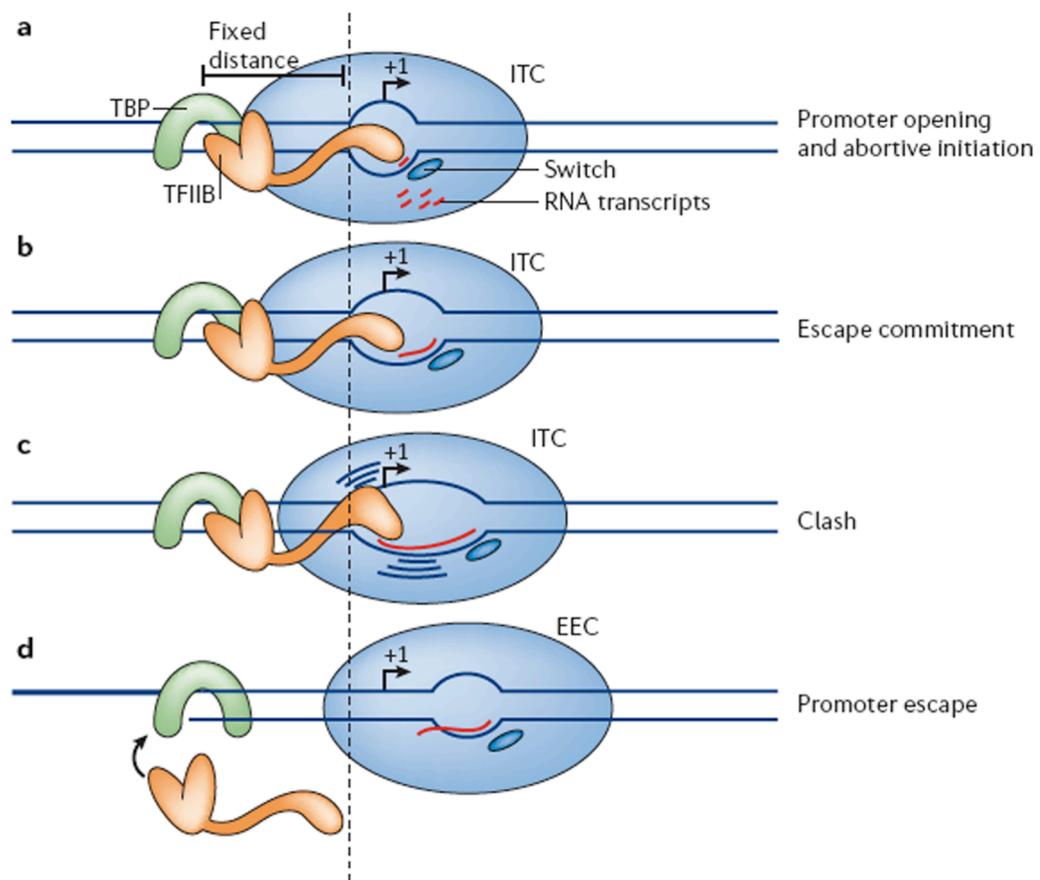
**Figure 1.1 Transition from initiation to early elongation: promoter**

**escape.** (A) The unwinding of promoter DNA to create a transcription bubble begins at a fixed position, ~20 bp downstream from the TBP binding site. The upstream bubble-edge (vertical dashed line) remains fixed until the completion of promoter escape, whereas the downstream edge expands in register with transcription. The initially transcribing complex (ITC) cycles through several rounds of abortive initiation, releasing high levels of 2-3 nucleotide-long transcripts (red).

(B) After four nucleotides are synthesized, the B-finger of TFIIB (orange) and a switch domain (dark blue) of Pol II (blue) stabilize the short RNA, reducing abortive initiation.

(C) After 5 nucleotides are added, the nascent RNA will clash with the B-finger of TFIIB, inducing stress within the ITC. This may cause increased abortive initiation, strong pausing, or transcript slippage if the 3' end of the RNA/DNA hybrid is weak, and likely contributes to the rate-limiting step of promoter escape.

(D) Stress from the growing transcription bubble and the production of an RNA that is at least 7 nucleotides long trigger bubble collapse, providing the energy to remodel the transcription complex. The B-finger is ejected from the RNA exit tunnel and TFIIB is released from the transcription complex. The RNA/DNA hybrid is at its full length of 8-9 base pairs and can make contacts with protein loops near the RNA exit tunnel. Abortive initiation ceases, as does the need for ATP hydrolysis, and transcript slippage is dramatically reduced, all indicating that the transcription complex has transitioned into an Early Elongation Complex (EEC).



Copyright © 2006 Nature Publishing Group  
 Nature Reviews | [Molecular Cell Biology](#)

Luse, 1987; Holstege et al., 1997; Kugel and Goodrich, 2000; Kugel and Goodrich, 2002). Recent biochemical experiments, guided by structural studies of RNA polymerases, have ascribed a positive role for TFIIB and a specific region of the polymerase near the active site in the stabilization of the early ternary complex that is functionally linked to their participation in start site selection (Bushnell et al., 2004; Chen and Hampsey, 2004; Majovski et al., 2005). The TFIIB/Pol II structure shows that the N-terminal B-finger domain of TFIIB is inserted into the polymerase active site (Bushnell et al., 2004). Modeling of RNA revealed that the B-finger is in position stabilize a short nascent RNA of 4-5 nucleotides. Indeed, in the absence of TFIIB, a ternary complex containing a five nucleotide RNA/DNA hybrid is not stable, and mutation of the B-finger results in increased abortive release of a 5 nucleotide product (Bushnell et al., 2004; Chen and Hampsey, 2004) (Figure 1.1B). Several protein loops within Pol II called 'switch' domains have been identified as lining a channel that accommodates the RNA/DNA hybrid upstream of the active site (Gnatt et al., 2001; Westover et al., 2004). It is hypothesized that interactions of the hybrid with the switch regions can transmit a conformational change within the polymerase stabilizing the ternary complex as RNA synthesis continues. In one case, a specific switch domain that contacts the template DNA immediately upstream of the active site is important for stabilizing a 5 nucleotide RNA (Majovski et al., 2005).

During the post-commitment phase of promoter escape, the ITC can undergo upstream transcript slippage, and if Pol II is artificially halted or is subjected to other challenges, the ITC can still abort the nascent RNA. These properties indicate that the ITC has not yet converted into a stable elongation complex (Keene and Luse, 1999; Pal and Luse, 2002; Pal and Luse, 2003). A

recent kinetic analysis revealed that the ITC encounters the rate-limiting step to promoter escape immediately after addition of the 8<sup>th</sup> nucleotide and when the Pol II active site is translocating to 9<sup>th</sup> position (Hieb et al., 2006).

Interestingly, several other events coincide with this transition to an Early Elongation Complex (EEC), and mark the completion of promoter escape. These events include i) a dramatic reduction in upstream transcript slippage at +8/+9 (Pal and Luse, 2003), ii) the end of abortive transcript release at +10/+11 (Holstege et al., 1997), iii) the end of the requirement for ATP and TFIIB at +8/+10 (Hieb et al., 2006; Pal et al., 2005), and iv) the sudden collapse of the transcription bubble (Figure 1.1) (Holstege et al., 1997; Pal et al., 2005). During promoter escape, the upstream edge of the transcription bubble remains fixed, while the downstream edge moves in register with transcription. Collapse of the upstream portion of the transcription bubble produces a transcription bubble that is more characteristic of a productively elongating polymerase.

Why is the final transition to the EEC a relatively slow step and how is the stabilization achieved? While the B-finger of TFIIB appears to be important for stabilization early on, the growing RNA would begin to clash somewhere beyond the 5<sup>th</sup> residue and the B-finger would need to be ejected by addition of the 10<sup>th</sup>-12<sup>th</sup> nucleotide. The slow rate of promoter escape might be the consequence of this clash (Figure 1.1C). Two parallel models could account for the stabilization of the EEC (Figure 1.1D). One model, based on crystal structures of Pol II from Kornberg and colleagues, suggests that an RNA of 9-10 nucleotides can begin to make contacts with several protein loops at the beginning of the exit tunnel that both stabilize the transcription complex and separate the RNA from the DNA so that the RNA/DNA hybrid is

maximally 8-9bp (Westover et al., 2004). These loops are also postulated to transmit a conformational change within the polymerase that converts Pol II into the EEC. It is proposed that the base of the B-finger of TFIIB might facilitate these interactions by slowing the rate of transcription at this point. Another model, presented by Luse and colleagues (Pal et al., 2005), suggests that the energy expended in unwinding the DNA is reinvested, by bubble collapse, in a remodeling event that results in the ejection of TFIIB from the exit tunnel, stabilization of the transcription complex, and completion of promoter escape (Figure 1D). In the same study, a minimal RNA length of 7 nucleotides was found to be necessary to trigger bubble collapse, and this is consistent with the position of the rate-limiting step of promoter escape (Hieb et al., 2006; Pal et al., 2005). Thus it appears that the RNA length is a critical determinant of promoter escape through its interactions within Pol II and by triggering bubble collapse. The rate of promoter escape is also sequence dependent, and a complex containing a weak RNA/DNA hybrid is slow to complete promoter escape, possibly because a weak hybrid is inefficient at competing with the B-finger for occupancy of the hybrid channel (Weaver et al., 2005). The majority of the in vitro studies described above have been performed with the Adenovirus major late promoter (AdMLP) or versions thereof. It will be useful to see how these models apply to several other promoters in the presence or absence of various core promoter elements and spacings, as well as different initially transcribed sequences.

#### **1.1.4. In vitro models of post-promoter escape transcription**

Promoter escape as described above likely constitutes the major structural changes of the transcription complex on its way to a stable

elongation complex. However, the EEC retains a measurable tendency to slip back on the template until about +23 and is susceptible to backtracking and arrest, thus indicating that formation of a fully productive elongation complex is not complete (Luse and Samkurashvili, 1998; Pal et al., 2001). Backtracked complexes at this stage are not always arrested, but are presumed to exist in dynamic equilibrium between upstream and downstream translocation since a large proportion can move out of the promoter if provided with a full complement of NTPs (Pal et al., 2001). Complexes that are arrested can be returned to competency by the transcription factor TFIIS, which stimulates the intrinsic cleavage activity of Pol II such that the 3'-OH is realigned with the active site (Fish and Kane, 2002; Izban and Luse, 1992). The relationship of the EEC distance from the transcription start site and the formation of a mature elongation complex is not understood, but it is hypothesized that the emerging RNA interacts with the polymerase and/or itself in some way that can affect the elongation potential of the polymerase (Westover et al., 2004). One study showed that Rpb7 subunit of Pol II crosslinks to the nascent RNA emerging from the exit tunnel (Ujvari and Luse, 2006). Rpb7 contains an oligonucleotide-binding domain (Orlicky et al., 2001), and is situated near the base of the CTD, and is so hypothesized to bind the nascent RNA and direct it towards the CTD where CTD dependent RNA processing enzymes await (Ujvari and Luse, 2006). Recently RNA was shown to bind the CTD of Pol II with some, but loose, sequence specificity, making the CTD a possible binding partner that plays a role in the rate of transcription at this point (Kaneko and Manley, 2005). Thus, after promoter escape, continued adjustments of the ternary complex occur as does exchange of general initiation factors for elongation and RNA processing factors, and this is often accompanied by

transcriptional pausing. Our view of the promoter clearance phase largely relies on observations made on the intrinsic properties of Pol II *in vitro*; some of the characteristics of the ternary complex during this stage display striking similarities with promoter proximal pausing as observed *in vivo*.

### **1.1.5 Promoter proximal pausing**

Although the main regulatory step in the transcription cycle has long been considered to be recruitment of Pol II to a promoter or the initiation of transcription, several studies have shown that control of an early phase of transcription elongation is also important. The first evidence for elongational control appeared more than 25 years ago in observations made by Pierre Chambon and colleagues (Gariglio et al., 1981a) that are summarized by the following quote, *“It appears therefore that RNA Polymerase B (II) molecules can be bound to DNA in the form of transcriptional complexes, not only at loci actively being transcribed in vivo, but also at loci which either have been or will be transcribed.”* This now seemingly prophetic statement was in reference to work by themselves and others that found transcriptionally-engaged polymerases on the 5'-end of genes that were supposedly not active. The concept that elongation could be rate limiting to gene expression was later extended in detail by a collection of studies that found RNA Polymerase II bound to and transcriptionally engaged near the 5'-end of several genes in *Drosophila* and human cells in the absence of gene activation (Krumm et al., 1992; Rougvie and Lis, 1988a; Rougvie and Lis, 1990). These studies suggested that control of the early elongation phase of transcription is an important regulatory step, and the phenomenon was named promoter-proximal pausing due to its similarity with a transcription regulatory mechanism

observed in prokaryotes. During promoter-proximal pausing in eukaryotes, Pol II initiates transcription but pauses at sites usually located 20-50 bases downstream of the transcription start site. Given the appropriate signals, Pol II can be released from the paused state to produce a full length RNA transcript.

The most notable examples of genes harboring a paused polymerase include heat shock-inducible genes, and the mammalian protooncogenes, *c-myc* and *c-fos* (Lis, 1998b). The first well characterized paused polymerase was that of the *Drosophila* heat shock gene, *Hsp70* (Lis, 1998b). Paused Pol II was first detected at *Hsp70* with the development of ultraviolet irradiation-crosslinking and chromatin immunoprecipitation (UV-ChIP): Pol II was found to fully occupy the promoter-proximal region of *Hsp70* (one Pol II per promoter) under conditions where gene transcription was not induced (Gilmour and Lis, 1986). Nuclear run-on analyses revealed that this polymerase was transcriptionally-engaged, and that it paused after synthesis of about 25 nucleotides of RNA (Rougvie and Lis, 1988a). Higher-resolution analysis showed that the pausing occurs at multiple sites over a region from +20 to +40 (Giardina et al., 1992; Rasmussen and Lis, 1993; Rasmussen and Lis, 1995). These corroborating assays indicated that transcriptionally-engaged Pol II exists in a paused state on at least a subset of promoters *in vivo*. Therefore, a step other than Pol II recruitment or transcription initiation is rate-limiting and a target of regulation.

Despite additional evidence that promoter-proximal pausing is a common phenomenon, this type of regulation was held in stark contrast to conclusions drawn from studies performed in the yeast, *Saccharomyces cerevisiae*. In this system transcription is regulated predominantly at the level of recruitment, that is, as long as polymerase was delivered to a gene by

classical upstream activators, activation of full-length transcript production was achieved (Barberis et al., 1995; Chatterjee and Struhl, 1995; Keaveney and Struhl, 1998; Ptashne and Gann, 1997; Xiao et al., 1995). Activation by recruitment was later reproduced in mammalian systems (Nevado et al., 1999), and given the large body of evidence from *Saccharomyces cerevisiae*, became the most widely accepted mechanism for transcriptional regulation that has appeared in textbooks over the last two decades. However, how these mechanisms relate to promoter-proximal pausing and elongation of RNA polymerase II has received less attention.

## **1.2 Regulation of promoter-proximal pausing**

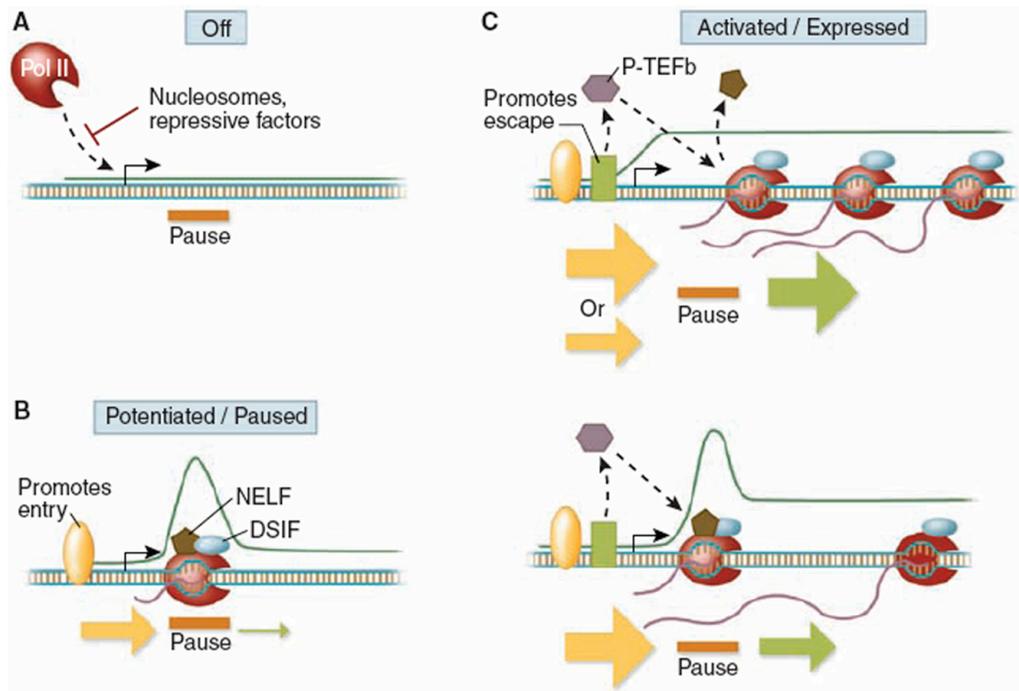
As mentioned above, several *in vitro* studies have shown that promoter-proximal pausing is a natural process that Pol II undergoes even in the absence of auxiliary factors. The DNA template and nascent RNA sequences are proposed to affect a position-dependent structural change in the transcription complex during early elongation (Pal et al., 2001; Ujvari et al., 2002). Such a conformational change may be necessary for achieving the fully processive form capable of transcribing long distances without disengaging the template or nascent RNA. The extent of intrinsic pausing *in vivo* is unclear, but the position relative to the start site is coincident with the action of known pausing factors DRB-Sensitivity Inducing Factor (DSIF) and Negative Elongation Factor (NELF) (see below) (Yamaguchi et al., 1999), which appear to further stabilize Pol II in the paused form. Also, pausing occurs at a point when several intimate contacts with the promoter are being severed. A likely *in vivo* scenario is that Pol II intrinsically pauses at specific

sites, and the extent of pausing is controlled by protein or RNA factors that compete for newly exposed surfaces or conformations of the polymerase.

Entry of Pol II to a promoter proximal-pause site requires that the transcription machinery must first gain access to the promoter and initiate transcription. Escape from the pause site occurs when Pol II moves into productive elongation, which clears Pol II from the promoter and allows sufficient space for another Pol II complex to initiate transcription. The relative rates of entry and escape combine to determine the effective level of pausing at a gene (Figure 1.2). High entry levels combined with low escape rates result in increased occupancy at promoter-proximal pause sites and is reflected by a high density of Pol II at the 5' end relative to downstream regions of genes as seen by ChIP analysis (Figure 1.2B, and C bottom). When the entry rate is less than or equal to pause site escape, a more uniform occupancy of Pol II over the gene is observed (Figure 1.1C Top). During activation of a gene containing paused Pol II the escape rate generally increases, while the intrinsic filling rate constant may or may not be altered. Several classes of transcription activators from *Drosophila* and mammals have been identified that primarily increase only initiation, only elongation, or both (Blau et al., 1996; Brown et al., 1996; Krumm et al., 1995; Yankulov et al., 1994). Activators that stimulate initiation and elongation separately have been shown to act synergistically when their binding sites are contained within the same promoter (Blau et al., 1996). Thus, activators can work independently to regulate both pause site filling and escape, leading to multiple modes of transcription activation. This theoretically imparts combinatorial control over transcription output, allowing cells to integrate more diverse signaling pathways, and synergistically upregulate genes rapidly when

**Figure 1.2. Regulation of entry and escape of Pol II at pause sites.**

The rate of pause site entry (yellow arrow—wide arrow represent fast and narrow arrow represents slow) is defined as the rate at which Pol II (red) would enter a pause site when it is freely accessible. The relative rates of entry and escape (green arrow) produce the observed patterns of Pol II density (blue line) (A) Pol II cannot access the promoter and transcription is “off”. (B) A potentiated state through the set up of a promoter-proximal paused Pol II by factors that promote entry (yellow oval). NELF (orange pentagon) and DSIF (blue oval) stabilize the paused Pol II. (C) Fully activated transcription requires factors that promote escape (green rectangle). Also, single factors can have one or both types of activation domains that in turn can be regulated by reversible modifications and associations.



needed. These next few sections describe some known examples of how protein factors and DNA elements can combine to set the level of pausing.

### **1.2.1 Nucleosome positioning elements**

As mentioned above, pausing requires efficient pause site entry. One mechanism that could influence transcription factor access to promoters is the affinity of nucleosomes for sequences that surround the core promoter and upstream control elements. Sequences have been identified that have high affinity for nucleosomes, resulting in a non-random positioning both in vitro and in vivo (Widom, 2001). Computational analyses and experimental analyses of the yeast genome have shown that many promoters (including active and inactive) have TSSs buried just inside the edge the first positioned nucleosome (+1 nucleosome relative to the TSS) (Ioshikhes et al., 2006; Segal et al., 2006). Also, highly regulated genes tend have strong positioning elements and nucleosomes that cover transcription factor binding elements, such as the TATA box (Ioshikhes et al., 2006). Interesting, the Pugh lab recently conducted a similar study with *Drosophila* embryos that had somewhat different conclusions (Mavrigh et al., 2008). They found that the majority of promoters have a positioned nucleosome on average 75bp (+135 relative to the TSS) further downstream than yeast. They hypothesize that this organization could allow access of the transcription machinery to the promoter and possible initiation followed by pausing as Pol II encounters the first nucleosome. Interestingly, genes with a paused Pol II had a +1 nucleosome that was positioned ~10bp (or 1 helical turn of DNA) even further downstream and directly contact Pol II. While this is potential evidence for direct modulation of pausing by the +1 nucleosome, it is not known whether the

nucleosome is impeding Pol II or if the affinity of Pol II for the pause site is moving the nucleosome downstream. The position of the experimentally determined nucleosomes could be recapitulated computationally using a nucleosome a positioning element similar to that used in humans (Mavrich et al., 2008). Thus, nucleosome positioning sequences encoded within the DNA may support pausing by maintaining a nucleosome free region at promoters. Consistent with the above differences between yeast and metazoans, no case of pausing in yeast has ever been documented.

### **1.2.2 Sequence-specific DNA binding factors**

The differential rate of Pol II entry and escape at a pause site is well-documented for the *Hsp70* gene of *Drosophila*. At this gene, GAGA factor (GAF) is required for efficient pause site entry under non-activating conditions, whereas binding of activated heat shock factor (HSF) is required for stimulation of escape and full activation of the gene. GAF maintains a nucleosome free promoter through by binding GA repeats, and recruiting the ATP-dependent chromatin remodeler NURF (Tsukiyama et al., 1994; Wilkins and Lis, 1997; Wilkins and Lis, 1998). GAF binding is critical for promoter-proximal pausing presumably by maintenance of this architecture, although some have suggested that GAF directly functions in pausing by interaction with the transcription apparatus (Gilmour, 2008; Lee et al., 1992; Shopland et al., 1995). Similar, although generally less defined, examples of cooperating activators that stimulate different rate limiting steps exist from mammalian and *Drosophila* systems (Bentley, 1995; Blau et al., 1996; Krumm et al., 1995; Sawado et al., 2003; Wang et al., 2005). One strong candidate that may function in a similar manner to GAF but in mammals is the transcription factor

SP1. SP1 is ubiquitously expressed, binds GC-rich sequences and, like GAF, has a glutamine-rich activation domain (Blau et al., 1996). In vitro transcription and in vivo plasmid-based assays have shown that SP1 stimulates non-productive transcription at the 5'-end of genes, but that it can act synergistically with some acidic activators to stimulate high levels of full-length transcript (Blau et al., 1996; Krumm et al., 1995). The enrichment of GC-rich elements (CpG islands) of mammalian promoters makes SP1 an attractive candidate for setting the stage for pausing in mammals (Juven-Gershon et al., 2008).

### **1.2.3 The CTD, Cdk7 kinase, and the connection to RNA processing.**

Perhaps the major difference between Pol II and the other RNA polymerases is the presence of a large, C-terminal domain that extends from the largest subunit, RPB1, that is made up repeating units of the consensus sequence  $Y_1S_2P_3T_4S_5P_6S_7$ . The number of repeats varies amongst species from 26 repeats in *Saccharomyces cerevisiae*, to 46 repeats in *Drosophila melanogaster*, and 52 in mammals (Young, 1991). Conservation of the consensus amino acids also varies with yeast showing about 73% identity among its repeats and *Drosophila* only about 4%. Genetic studies involving various alleles of RPB1 that had CTDs either deleted or truncated to different extents showed that a full length CTD is necessary for normal cell viability (Young, 1991). Early biochemical studies demonstrated that the CTD could be phosphorylated, and that the transition of Pol II from initiation to elongation was coincident with the change of the CTD from a hypophosphorylated state (Pol IIa) to a hyperphosphorylated (Pol IIo) state (Laybourn and Dahmus, 1990; O'Brien et al., 1994b; Payne et al., 1989; Weeks et al., 1993). Pol IIa

preferentially interacts at the promoter and is incorporated into the PIC, while Pol II is observed in the body of transcription units (Hengartner et al., 1998; O'Brien et al., 1994b; Weeks et al., 1993). Studies over the past two decades have demonstrated that phosphorylation is not only an indicator of the elongation competency of the polymerase, but is a controlled event that influences dynamic interactions of the CTD with a variety of factors (Prelich, 2002; Zorio and Bentley, 2004). In addition to phosphorylation, the CTD is subject to the enzymatic modification by phosphatases, prolyl isomerases, glycosylases, and ubiquitin ligases (Egloff and Murphy, 2008). Collectively, the differential modifications of the CTD create a combinatorial code, which theoretically provides an interpretable scaffold to coordinate the interaction of a diverse array of proteins that are important for co-transcriptional pre-mRNA processing events, stimulation of transcription elongation, modification of the chromatin environment, and termination of transcription.

The most highly studied modification of the CTD is phosphorylation. The CTD is phosphorylated at serines 2, 5, and 7 within the heptapeptide repeat. Ser5 phosphorylation (Ser5-P) begins early on in the transcription cycle, near the 5' end of the gene. Ser5-P normally trails off as the polymerase moves towards the 3' end of the gene, albeit to different extents depending on the gene (Boehm et al., 2003; Komarnitsky et al., 2000; Morris et al., 2005). Ser2 phosphorylation (Ser2-P) predominates in the body and towards the 3' end of a gene and occurs concomitantly with productive elongation (Boehm et al., 2003; Komarnitsky et al., 2000; Morris et al., 2005; O'Brien et al., 1994a). Phosphorylation at serine 7 (Ser7-P) is recently discovered, and has been shown to be important for proper processing snRNAs in mammals (Chapman et al., 2007; Egloff et al., 2007).

The kinases Cdk7 and Cdk8 phosphorylate the CTD at Ser5, although, Cdk8 has also been shown to phosphorylate Ser2 (Prelich, 2002). Cdk8 is a component of the mediator, and was originally thought to only be involved in repression of subsets of genes. However, it can be at least partially redundant with Kin28 in stimulating promoter escape (Liu et al., 2004). The kinase activity of Cdk7 resides within the general transcription factor TFIIH, and is only required for transcription in minimal in vitro transcription assays when its substrate, the CTD of Pol II is also required (Akoulitchev et al., 1995; Li and Kornberg, 1994; Lu et al., 1992; Makela et al., 1995; Serizawa et al., 1992 ; Serizawa et al., 1993). Li and Kornberg, observed that in crude extracts there was a block to transcription that was independent of the CTD (Li and Kornberg, 1994). Addition of Cdk7 could relieve this block but only if the polymerase had an intact CTD. Thus, it seems reasonable that Cdk7 kinase can be responsible for severing contacts with the promoter that allows Pol II to escape from the promoter and enter the pause site. Indeed, a study in yeast demonstrated that inhibition of *S. cerevisiae* Cdk7 (Kin28) prevents dissociation of the preinitiation complex (Liu et al., 2004). Consistent with the above hypothesis, the paused Pol II is phosphorylated at Ser5, and a *cdk7<sup>ts</sup>* allele reduces the occupancy of Pol II at the pause site on the *hsp70* gene in *Drosophila* (Boehm et al., 2003; Schwartz et al., 2003).

Promoter-proximal pausing occurs at a point where it may serve to coordinate transcription elongation with pre-mRNA processing (Zorio and Bentley, 2004). Indeed, pausing is coincident with mRNA capping (Rasmussen and Lis, 1993). Shortly after a nascent mRNA is extruded from the polymerase exit channel a 7 -methyl guanine 'cap' is added to the 5' end of the molecule by the successive action of three enzymes: a triphosphatase

(RT), a guanylyltransferase (GT) and a methyltransferase (MT) (Howe, 2002; Rasmussen and Lis, 1993). Capping stabilizes the mRNA by stimulating proper downstream events such as splicing, 3' processing, transport to the cytoplasm (Howe, 2002; Zorio and Bentley, 2004). Capping enzyme associates with the Ser5 phosphorylated CTD of Pol II and with Spt5 (Rodriguez et al., 2000; Wen and Shatkin, 1999), both of which also stimulate capping enzyme activity *in vitro* (Mandal et al., 2004). In fission yeast, the cap methyltransferase, Pcm1 forms a complex with fission yeast P-TEFb (Guiguen et al., 2007; Pei et al., 2006). Recruitment of the kinase activity of P-TEFb is the rate-limiting step to escape from pausing (see below). Depletion of Pcm1 blocks CTD Ser2P and abolishes recruitment of P-TEFb to chromatin, indicating an essential role for Pcm1 in P-TEFb loading (Guiguen et al., 2007). Thus, a model has emerged depicting escape from pausing as a checkpoint to ensure that the pre-mRNA is properly capped.

Another connection between early elongation and pre-mRNA processing was revealed recently by the Reinberg lab. They previously showed that the chromatin remodeling factor, Chd1, binds to H3K4me3 tails through its chromodomain (Sims et al., 2005). Interestingly, Chd1 forms a bridge between H3K4me3 and components of the spliceosome, and is functionally important for targeting the splicing machinery to chromatin and for efficient pre-mRNA processing (Sims et al., 2007). While there is no evidence that pausing is required for this efficient loading of the spliceosome or vice versa, the timing of this event parallels that of pausing.

#### 1.2.4 Promoter sequence elements.

Promoters consist of various core elements that dictate the location and strength with which transcription will initiate (Juven-Gershon et al., 2008; Sandelin et al., 2007). The core elements have been identified upstream, downstream, and overlapping the site of initiation. They adhere to a consensus sequence to varying degrees and are sometimes conserved among eukaryotes. Core promoter strength is often a function of the different combinations of elements, as they can act synergistically or antagonistically. While they are important for nucleating a functional transcription complex, the strength of promoter can be modified by short or long-range interactions with activating elements that bind sequence-specific DNA activators and repressors.

The most abundant core elements identified thus far include the TATA-box, TFIIB recognition element (BRE), the initiator (Inr), the motif ten element (MTF), and the downstream promoter element (DPE) (Juven-Gershon et al., 2008) (Figure 1.2). Mike Levine's lab recently searched for elements that correlate with a stalled polymerase, based on their genome-wide ChIP-chip study from drosophila embryos (Hendrix et al., 2008; Zeitlinger et al., 2007a). They identified a 'Pause Button' motif that is most often found from +25 to +35 relative to the TSS and has the consensus sequence KCGRWCG. This overlaps the position of the DPE, which also correlates with stalling on its own. In addition, the presence of a button or DPE in combination with GAF binding sites and an Inr were strong predictors of stalling. The presence of GAGA factor binding sites is expected to be a predictor of stalling, since it is necessary for pause site entry (Lee et al., 2008; Lee et al., 1992; Shopland et al., 1995). Both the Inr and DPE interact with subunits of the general

transcription factor TFIID (Purnell et al., 1994; Sypes and Gilmour, 1994), which points a function for TFIID in pausing. Perhaps simultaneous interaction of TFIID with the promoter and Pol II impedes the progress of transcription. The pause button could fulfill a similar purpose, but it is also possible that it directly modulates the conformation of Pol II.

### **1.2.5 Maintenance of paused Pol II**

Candidate pausing factors were discovered during attempts to reconstitute *in vitro* transcription that displays the same sensitivity for a kinase inhibitor, commonly known as DRB, seen *in vivo* (Peterlin and Price, 2006; Price, 2000). DRB inhibits elongation but not initiation. DRB sensitivity-inducing factor (DSIF) (Wada et al., 1998a) and negative elongation factor (NELF) (Yamaguchi et al., 1999) were both required for the inhibition of transcription by DRB *in vitro*, and it was later found that that DSIF and NELF cooperative to repress transcription elongation (Yamaguchi et al., 1999), and that their negative effects are overcome by the action of P-TEFb, which is inhibited by DRB (Wada et al., 1998b) (Figure 1.3). DSIF consists of the elongation factors Spt4 and Spt5, which are conserved from yeast to humans (Yamaguchi et al., 2001). NELF comprises four subunits, NELF-A, B C/D and E, and is conserved between mammals and *D. melanogaster*, but is not present in *Caenorhabditis elegans*, *Saccharomyces cerevisiae* or *Arabidopsis thaliana* (Price, 2000). *In vivo*, DSIF and NELF are present at uninduced *D. melanogaster* heat shock genes (Andrulis et al., 2000; Saunders et al., 2003; Wu et al., 2005), and DSIF and NELF are important for *Hsp70* promoter-proximal pausing *in vitro* (Wu et al., 2005). The position of paused polymerases correlates with where DSIF and NELF begin to exert their

**Figure 1.3 Core promoter elements and the Pause Button.** A Schematic diagram of some common core promoter elements and their positions relative to the transcription start site. Shown are the TATA-box (-28 to -32), TFIIB recognition element (BRE<sup>u</sup>: immediately upstream; and BRE<sup>d</sup>: immediately downstream of the TATA box), the initiator (Inr: flanking the TSS), the motif ten element (MTF: +18 to +27), the downstream promoter element (DPE: +28 to +32), and the newly identified Pause Button (Hendrix et al., 2008). General transcription factors that are known to interact with these elements are shown. The pause button overlaps the DPE, but it is not known whether TFIID interacts with it as efficiently as with the DPE. The figure is adapted from (Juven-Gershon et al., 2008), with permission from RightsLink.



negative effects and also with when the nascent RNA is long enough to protrude from Pol II. One mechanism, therefore, by which DSIF and NELF could specifically mediate pausing in the promoter proximal region is through interaction with the nascent RNA, and consistent with this, NELF-E binds to RNA (Yamaguchi et al., 2002). The observation that NELF is not present in all metazoans indicates that NELF might be a less general elongation factor than Spt5 and that there might be other factors involved in promoter-proximal pausing.

### **1.2.6 Escape from pausing**

The signals that lead to escape of Pol II from pause sites are generally initiated by additional sequence-specific DNA binding proteins that directly modulate the transcription complex, and/or possibly by manipulating the chromatin environment such that transcription through nucleosomes is possible. The primary executor of escape from pausing is the kinase activity of Positive Elongation Factor b (P-TEFb) (Marshall and Price, 1995; Peterlin and Price, 2006). This factor phosphorylates multiple targets within the transcription complex including Ser2 of the Pol II CTD, NELF, and DSIF, and is crucial for relief of the NELF and DSIF-dependent block to transcription elongation (Peterlin and Price, 2006). At this transition, NELF dissociates from the transcription complex, but the modified DSIF remains associated and enhances elongation. Not surprisingly, cells have developed a number of ways to bring P-TEFb to genes. Several gene-specific regulators have been shown to interact with P-TEFb (Mancebo et al., 1997; Oven et al., 2007; Zhou et al., 2006; Zhu et al., 1997), but activators do not ubiquitously recruit P-TEFb (Lis et al., 2000). It has been shown that, in human cells, the bromodomain

protein, Brd4, is likely responsible for recruitment of P-TEFb to most genes, through its interaction with acetylated histones (Jang et al., 2005; Yang et al., 2005). As mentioned previously, the *S. Pombe* cap methyltransferase can also bring in P-TEFb to the transcription complex (Guiguen et al., 2007).

Additionally, the activity of TFIIS is important for the efficient escape of Pol II from the pause (Adelman et al., 2005b). A fraction of the paused polymerases are susceptible to backtracking, whereby Pol II moves upstream on the template and the 3' OH of the RNA transcript becomes misaligned with respect to the Pol II active site (Fish and Kane, 2002). TFIIS stimulates the intrinsic RNA cleavage activity of Pol II to create a new RNA 3' OH in the active site and once again enable transcription elongation.

### **1.3 Possible biological roles of pausing**

A paused polymerase represents a transitional state in expression of a gene: it is past the steps of PIC formation, initiation, and promoter escape, but is not productively elongating the transcript. One can envision several advantages that this could have for response to external cues, and in maintenance of cellular homeostasis. As mentioned above, pausing has been connected with control over pre-mRNA processing. Considerable evidence also suggests maintenance of pausing at a promoter is key for full activation of a gene. Studies on the *Drosophila Hsp70* and human FOS and MYC genes, have shown that removal of the sequences that cause pausing result in decreased transcription factor accessibility and defective activation (Fivaz et al., 2000; Lee et al., 1992; Shopland et al., 1995). Also, RNAi knockdown of NELF decreases Pol II occupancy and increases histone occupancy at a subset of genes (Gilchrist et al., 2008). Thus, pausing is apparently important

for maintaining a permissive chromatin environment for transcription. How a paused Pol II grants accessibility of promoter DNA to regulatory factors is unclear, but it is possible that Pol II exerts this effect by either directly preventing nucleosomes from obstructing DNA binding sites, or by recruiting other factors that modify the chromatin architecture around the promoter.

Past and recent studies have suggested that pausing might also be crucial for the timing of gene expression in response to environmental or developmental stimuli. The heat shock genes are robustly transcribed within seconds of thermal upshift. The importance of this response for survival of the organism and the degree with which it responds invoked a 'potentiated promoter' hypothesis early on (Lis, 1998a). That is, since the transcription complex is already past the earliest stages of transcription, at least the first wave of transcription can happen immediately. Genes involved in other response pathways also seem to fit into this category. Of the genes identified as likely having a paused polymerase in *Drosophila* by the genomic studies (see below), genes that respond rapidly to developmental and cell signaling were overrepresented (Muse et al., 2007; Zeitlinger et al., 2007b). Also, recent evidence from the Kraus lab also suggests that pausing could be important for the timing of the physiological response to estrogen in human breast cancer cells (Kininis et al.). Genes that are preloaded with Pol II prior to estrogen stimulation respond with faster kinetics than genes that have Pol II recruited after estrogen treatment. Finally, additional evidence suggests that pausing can serve to limit the expression of genes involved estrogen signaling (Aiyar et al., 2004) and in the immediate early response to growth factors (Aida et al., 2006). Together, these observations raise the likelihood that potentiation through pausing prior to activation, or fine-tuning of activated

transcription are a fundamental steps for rapidly controlling physiological and developmental programs.

#### **1.4 Generality of promoter-proximal pausing**

Under the hypothesis that recruitment of Pol II is sufficient for gene activation, Pol II levels at a gene promoter should correlate to some degree with mRNA levels. While this holds true for the vast majority of genes in *S. cerevisiae* (Robert et al., 2004), several genome-wide or more focused analyses have recently revealed that this is not always the case in mammalian and *Drosophila* cells (Guenther et al., 2007; Kininis et al., 2007; Muse et al., 2007; Zeitlinger et al., 2007a). These studies used the chromatin immunoprecipitation assay coupled with genomic microarray technologies (ChIP-chip) to examine Pol II density along genes. These studies found that approximately 20-30 percent of genes have enriched Pol II density at the 5'-end relative to the body of the genes. This class included genes with either detectable or undetectable expression. Identification of this latter subclass, which has Pol II bound without full-length transcript production, suggests that a post-recruitment step of the transcription cycle is rate limiting at these genes. Whereas the ChIP assay can detect the density of Pol II across a gene, it cannot necessarily determine whether or not Pol II is transcriptionally engaged, that is, the 5'-skewed distribution of Pol II could represent Pol II in either the pre-initiation form or initiated, but paused form. Four of the genome-wide studies presented additional assays of permanganate footprinting, which maps the transcription bubble in the wake of a transcriptionally-engaged Pol II, or analysis of short RNA products as evidence that Pol II had progressed beyond initiation at multiple candidate genes (Guenther et al., 2007; Lee et al.,

2008; Muse et al., 2007; Zeitlinger et al., 2007a). Although the validated genes in these studies were mainly in the class of low or non-detectable expression levels, it is important to emphasize that highly expressed genes were also identified as candidates for pausing. Thus, regulation at the level of pausing appears to occur broadly and over a large dynamic range of transcript production.

### **1.5 Concluding remarks**

Transcription regulation is a multi-step process that is controlled at the level of recruitment, initiation, pausing, and elongation of RNA polymerase II. A number of genome-scale studies have identified large classes of genes that are likely to be regulated by promoter-proximal pausing, and thus have provided us with a large set of model genes with which to study distinct aspects of this mode of regulation. Future investigation, directed at determining how promoter-specific binding proteins affect the initiation/pausing and pausing /productive elongation transitions, will provide important insights into the role of cell signaling events in the mechanics of transcription regulation.

### **1.6 Dissertation outline**

This dissertation presents a highly-sensitive method (GRO-seq) that maps the position, amount, and orientation of transcriptionally-engaged RNA polymerases across the entire genome. I have applied it to a primary human lung fibroblast cell line, IMR90. This method not only permits the analysis of the generality of promoter-proximal pausing, but also detects transcription beyond the 3'-end of genes, antisense transcription, and has revealed that

most human promoters initiate in both directions. The data presented here can also be used to identify the direct transcriptional outcome of different transcription factor and sequence element combinations.

Chapter 2 presents the development of method. Results and considerations of the control of resolution, the efficiency of the protocol, and the choice of genomic platform are presented and discussed.

Chapter 3 presents the data analysis, results, and conclusions from the first complete run-through of the method.

Chapter 4 presents future directions for this project, including important biological questions that can be asked with this new technology, and future adaptations of GRO-seq that will provide transcription start site information as well as near nucleotide mapping of the polymerase active site.

Appendix A presents my attempt to use nucleoprotein hybridization as a method to isolate cross-linked chromatin in a sequence-dependent manner for identification of proteins via mass spectrometry.

## CHAPTER 2<sup>2</sup>

### DEVELOPMENT AND APPLICATION OF A GENOMIC METHOD TO MAP THE POSITION, AMOUNT, AND ORIENTATION OF RNA POLYMERASES

#### 2.1 Introduction

The goal of this dissertation is to map transcriptionally engaged polymerases (specifically Pol II) across any genome, thus a brief review of methodologies that are currently used is warranted. Polymerases can be mapped *in vivo* or *in vitro* by a number of techniques. Chromatin Immunoprecipitation Assays (ChIP) can map the location of polymerases to within 100-300 bases. However, one cannot determine the direction the polymerase is oriented, or whether or not it is transcriptionally engaged. Potassium permanganate footprinting maps the unwound portion of DNA, known as the transcription bubble, that is associated with a transcriptionally engaged polymerase. This has an obvious benefit over ChIP, since KMnO<sub>4</sub> sensitivity indicates that the polymerase is engaged in transcription. However, like ChIP it cannot specify the orientation of transcription. In addition, KMnO<sub>4</sub> footprinting has high background, and other DNA binding factors that cause torsional stress on the DNA can have footprints with this assay. I therefore chose the nuclear run-on assay as the method of choice to adapt for genome-wide studies and have named the assay Global run-on sequencing (GRO-seq). Conventional run-ons and GRO-seq are reviewed below.

Nuclear Run-On (NRO) assays have been used to measure the density of transcribing polymerases over specific targeted regions of the genome, and

---

<sup>2</sup> Information in this chapter is largely from ((Core et al., 2008)).

variations of the assay are capable of mapping the position of polymerases with high precision (Gariglio et al., 1974; Gariglio et al., 1981b; Rasmussen and Lis, 1993; Rougvie and Lis, 1988a). Traditionally, nuclei are isolated, endogenous nucleotides are removed by washing, and radionucleotides are added back allowing transcriptionally engaged polymerases to resume elongation. The incorporated radiolabel is restricted to sequences immediately downstream of the original position of the transcriptionally-engaged polymerase by keeping run-on reaction times short. The anionic detergent sarkosyl, which does not interfere with elongating polymerases, is often added to the nuclear run-on reaction to ensure that new transcription initiation events do not occur, and to remove physical impediments that can block elongation (Hawley and Roeder, 1985; Rougvie and Lis, 1988a). Thus all new transcription is produced by polymerases that are engaged at the time of nuclear isolation. The RNA is then isolated and hybridized to filters containing genes or gene regions of interest. These measurements have been shown to represent the level of transcriptionally-engaged polymerase on genes at the time of nuclei isolation, and have also been used to identify Pol II that is paused at the 5' ends of genes as well as the distance Pol II travels beyond the 3'-ends of genes prior to termination (Faro-Trindade and Cook, 2006; Gromak et al., 2006; Lis, 1998b; Proudfoot, 1989).

Previous attempts at scale-up have hybridized radiolabeled NRO RNAs to cDNA probes spotted on macroarrays to analyze how steady state transcription of genes relates to mRNA accumulation (Garcia-Martinez et al., 2004; Schuhmacher et al., 2001). These methods can give reasonable approximations for steady state transcription levels for some genes, however, they suffer from low sensitivity, lack of whole genome coverage, and no

resolution within gene regions. Whole genome coverage is important for detection of novel transcription units as well as transcripts that are not present in cDNA libraries. The lack of resolution of cDNA arrays is of concern since genes that have a promoter-proximal paused Pol II, and do not produce full-length transcripts will produce detectable signal that does not reflect actual levels of full-length transcription of those genes (Schilling and Farnham, 1994). In addition, the distribution of transcribing polymerases within genes provides information on how a particular gene is regulated, and when combined with our knowledge of promoter DNA sequences, transcription factor binding sites, and nucleosomes and their modifications, can further our knowledge of how these elements cooperate to specify distinct transcriptional outcomes.

## **2.2 Materials and methods**

### ***Isolation of nuclei.***

Isolation of nuclei was carried out as described in (Strobl and Eick, 1992), (39), with several modifications. 15cm<sup>2</sup> plates of IMR90 cells (~6X10<sup>6</sup> cells at 80% confluency) were washed directly on the plate 3X with ice cold PBS. 10ml of ice cold swelling buffer (10mM Tris-cl pH7.5, 2mM MgCl<sub>2</sub>, 3mM CaCl<sub>2</sub>) was added and allowed to swell on ice for 5 min. Cells were removed from the plate with a plastic cell scraper, transferred to a 15 ml conical tube, and pelleted for 10 min at 4°C at setting 3 on an IEC clinical centrifuge. Cells were resuspended in 1ml of lysis buffer (swelling buffer + 0.5% Igepal, + 10% glycerol + 2units/ml SUPERase In (ambion)), and gently pipetted up and down 20 times using a p1000 tip with the end cut off to reduce shearing. The volume was brought to 10 ml and nuclei pelleted at setting 4 on an IEC clinical centrifuge. The nuclei were washed and pelleted once in Lysis buffer,

resuspended in 1ml Freezing buffer (50mM Tris-CL pH 8.3, 40% glycerol, 5mM MgCl<sub>2</sub>, 0.1 mM EDTA), and transferred to a 1ml tube. Nuclei were pelleted at 1000Xg, and resuspended in 100ul of Storage Buffer / 5X10<sup>6</sup> nuclei.

### ***NRO-RNA library construction.***

Construction of a NRO-library for sequencing involves the run-on reaction, base hydrolysis, immuno-purification, end repair, 5'- and 3'- adapter ligation, amplification, and PAGE purification.

### ***NRO reaction.***

5X10<sup>6</sup> IMR90 nuclei (100ul) were mixed with an equal volume of reaction buffer (10mM Tris-Cl pH 8.0, 5mM MgCl<sub>2</sub>, 1mM DTT, 300mM KCL, 20 units of SUPERase In, 1% sarkosyl, 500uM ATP, GTP, and Br-UTP, 2μM CTP and 0.33μM α-32P-CTP (3000Ci/mmol)). The reaction was allowed to proceed for 5 min at 30°C, followed by the addition of 23ul of 10X DNaseI buffer, and 10ul RNase free DNase I (Promega). Proteins were digested by addition of an equal volume of Buffer S (20mM Tris-Cl pH 7.4, 2% SDS, 10mM EDTA, 200ug/ml Proteinase K (invitrogen)), followed by incubation at 55°C for 1 hour. RNA was extracted twice with acid Phenol:chloroform, and once with chloroform, and precipitated at a final concentration of 300mM NaCl, with 3 volumes of -20°C ethanol. The pellet was washed in 75% ethanol before resuspending in 20ul of DEPC-treated water.

### ***Base hydrolysis of RNA.***

Base hydrolysis was performed on ice by addition of 5ul 1M NaOH and incubated on ice for 30min. The reaction was neutralized by addition of 25 ul 1M Tris-Cl pH 6.8. The reaction was then run twice through a p-30 RNase-free spin column (BioRad), according to the manufacturer's instructions. Before moving on to the immuno-purification, DNA was further removed by another digestion with RNase-free DNaseI for 10 min at 37°C, and the reaction stopped by addition of 10mM EDTA.

### ***Immuno-purification of Br-U RNA.***

Anti-deoxyBrU beads (Santa Cruz Biotech) were blocked in 0.5X SSPE, 1mM EDTA, 0.05% tween, 0.1% PVP, and 1mg/ml ultrapure BSA (Ambion). NRO-RNAs were heated to 65°C, added to 100ul beads in 500ul of binding buffer (0.5XSSPE, 1mM EDTA, 0.05% tween), and allowed to bind 1hour while rotating (Labquake rotator, 8rpm). The beads were washed once in low salt buffer (0.2X SSPE, 1mM EDTA, 0.05% Tween), twice in high salt buffer, 0.5% SSPE, 1mM EDTA, 0.05% Tween, 150mM NaCl), and twice in TET buffer (TE + 0.05% Tween). The Br-U RNA is then eluted 4X 125ul of Buffer E (20mM DTT, 300mM NaCl, 5mM Tris-cl pH 7.5, 1mM EDTA, and 0.1% SDS). The RNAs are then extracted and precipitated as above.

### ***End Repair.***

Enriched RNAs were resuspended in 20ul DEPC-treated water, and incubated with 2.5ul Tobacco acid pyrophosphatase (TAP, Epicentre Biotechnologies), 1X TAP buffer, and 1ul SUPERase Inhibitor in a final volume of 30ul at 37°C for 1hour. 1ul of Polynucleotide Kinase (PNK, NEB), and 0.5ul

of 5mM MgCl<sub>2</sub> is then added and the reaction continued for 30min. 20 ul PNK buffer, 2ul 100mM ATP, and 145ul water, and 1ul PNK is then added and the reaction continued for another 30 min. 90ul water and 10ul 500mM EDTA, is then added, followed by extraction and ethanol precipitation of the RNA.

### ***Adapter ligations.***

For adapter ligations the RNA was resuspended in 8.5ul, and incubated with 2.5ul of either the 5'- or 3'- adapter oligo (Small RNA Isolation Kit, Illumina), 1ul SUPERase In, 2ul RNA ligase-1 buffer, 5ul 50% PEG 8000, and 1.5ul of T4 RNA ligase-1 (NEB). The reactions were incubated on the lab bench for 4 hours. After both the first and second adapter ligations the RNAs were enriched over anti-deoxy-BrU beads as described above.

### ***Reverse transcription, amplification and PAGE purification of NRO-RNA libraries.***

The RNAs were reverse transcribed (otherwise according to the manufacturer's specifications) in two separate 10ul reactions, with 0.5ul 100uM RT-Primer (Illumina Small RNA Isolation Kit), and 1ul SIII reverse transcriptase (Invitrogen), at 44°C for 15min, followed by 52°C for 45 min. The RNAs were degraded by addition of RNase cocktail (Ambion), and RNase H (Ambion), and amplified 15 cycles, with Phusion high fidelity DNA polymerase (Finnzymes) using the PCR primers specified by Illumina. The NRO-cDNA libraries were then run on a non-denaturing 1XTBE, 8% acrylamide gel, and cDNAs greater than 90 nucleotides were excised from the gel and eluted by incubating in TE + 300mM NaCl overnight while rotating. The library was then

extracted, precipitated, and then sent to Illumina for sequencing on the 1G Genome Analyzer.

### ***Data analysis***

#### ***Alignment of GRO-seq reads to the human genome.***

Two independent biological replicates were submitted for sequencing at Illumina. Library 1 was sequenced on three channels and yielded 13,818,931 total reads while library 2 was sequenced on two channels and yielded 9,389,058 reads. All reads were 33 bases long. Alignments to the hg18 assembly of the human genome were performed with the Eland alignment tool from Illumina. 5,316,960 full length reads from library 1 aligned uniquely to the human genome and 4,459,581 full length reads from library 2 aligned uniquely to the human genome. Alignments allowed up to two mismatches per sequence to account for sequencing errors and SNPs between the IMR90 cell line and the sequenced genome. To increase the coverage of our libraries, we then iteratively trimmed one base from the 3' end of reads that did not align uniquely and checked if they now aligned uniquely at the reduced length. Trimming was done from the 3' end, because the quality score for reads was highest at the 5' end and lowest in the 3' end, and because it is possible that some of our amplified library was shorter than the 33 bases sequenced. Analysis of the correlation between the two libraries as a function of trimming extent showed that 29 bases was the optimal minimum length to be included. Alignments were done to the full (non-repeat masked) human genome. While unique alignments can be achieved in repeat masked sequences, we analyzed the number of reads mapping to such repeat masked sequences to be sure they were trust worthy. With the exception of rRNA repeats, the

density of alignments to repeat regions mirrored the average overall density of surrounding regions, suggesting that they were indeed accurate. The highly transcribed and repetitive rRNAs, as expected, had an average density roughly five orders of magnitude above the average genome wide level. Since rRNA is the most abundant mature RNA in the cell, we reasoned that this would be the major non-NRO RNA contaminant in our purifications, and thus we removed all alignments to rRNA repeats in the genome. These steps increased the total number of reads aligned to the genome to 5,800,577 for library 1 and 4,950,956 for library 2, for a total of 10,751,533 unique alignments. Since sequencing was performed from the 5' end of the BrU purified NRO RNA, the 5' coordinate of each read was used as the position of engaged polymerase for all future analyses.

### ***Identifying mappable bases in the genome.***

To assess the fraction of the genome where reads could be expected to align, all unique 32 base sequences from both strands of the hg18 assembly were identified. This is a total of 2,414,845,175 32-mers per strand from a total possible 3,080,436,051 per strand. A mappable or unmappable base refers to the 5' base of a given mappable or unmappable 32-mer. All calculations of read densities in future analyses were relative to these mappable bases.

### ***Background calculation from low-density windows.***

To assess the background GRO-seq density, the genome was divided into 500 kbp windows and the density of reads in each window was calculated. The distribution of low-density windows is described very well by placing 3% of the total GRO-seq reads randomly on the mappable portion of the genome (Figure 2.10). The blue theoretical curve is described by

$$p(x) = \frac{\lambda^{x * l} e^{-\lambda}}{(x * l)!}$$

where  $x$  is the density of reads on both strands per base pair,  $l$  is the window size (500 kb), and  $\lambda$  is the background density of reads (in units of reads/bp).

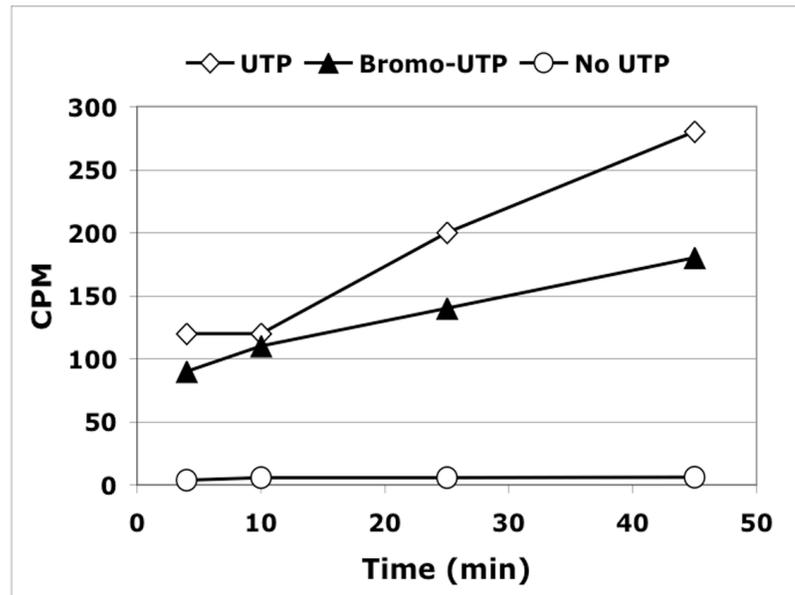
$$\lambda = \frac{f * N_{reads}}{L_{mappable}}$$

$f$  is the fraction of all reads that are from background (0.03 in Fig. S27),  $N_{reads}$  is the total number of reads aligning to the genome (10,751,533) and  $L_{mappable}$  is the total number of mappable 32-mers in the genome summed over both strands (4,829,690,350).

### ***Background calculation from gene deserts.***

Sixteen separate 'gene deserts' were identified where most GRO-seq alignments should represent background. These regions ranged in size from roughly 500 kb to nearly 7 Mb. The details of the coordinates of these gene deserts and the number of GRO-seq reads are in Table S2.

Incorporation of nucleotide analogs during run-on time course



**Figure 2.1. Incorporation of Br-UTP in a nuclear run-on.**

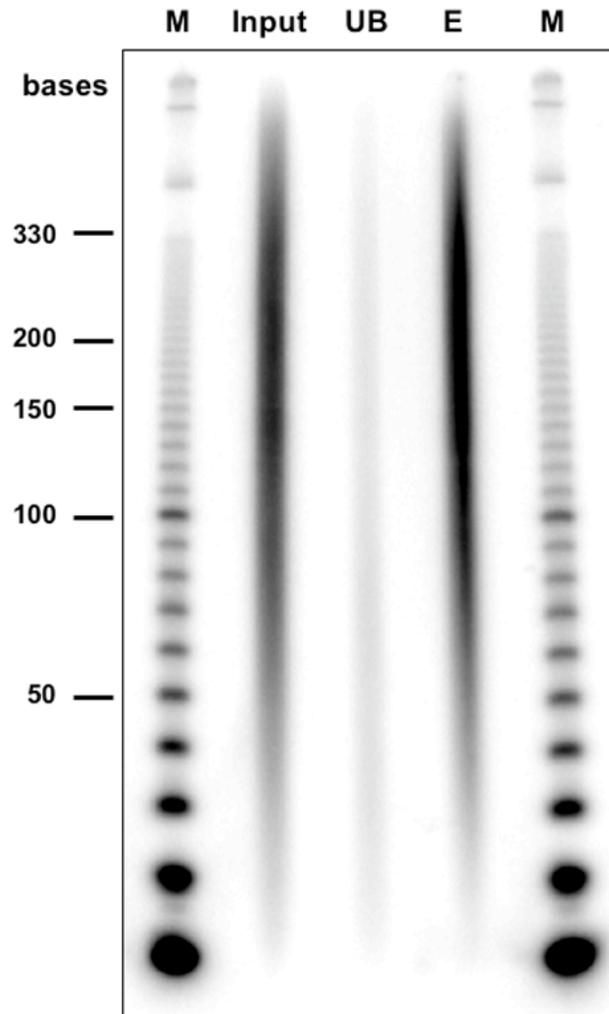
Polymerases were run-on in nuclei supplemented with Sarkosyl, ATP, GTP,  $\alpha$ -<sup>32</sup>P-CTP and UTP (open diamonds), Br-UTP (closed triangles), or no UTP (open circles). Separate reactions were setup for each timepoint and the reactions were stopped at 5, 10, 25 or 45 min. The RNAs were isolated, and the radioactivity incorporated was assayed by scintillation counting.

## **2.3 Development of the global nuclear run-on**

### **2.3.1 Incorporation of Br-UTP by nuclear RNA polymerases.**

Given that the NRO-RNA represents a small fraction of the total RNA in nuclei (see below), analysis of NRO-RNA with conventional genomic platforms requires specific isolation of this RNA. To adapt nuclear run-ons for a global analysis, we reasoned that a nucleotide with an affinity purifiable tag could be added to the run-on reaction, and sought to test the incorporation and purification efficiencies as outlined below.

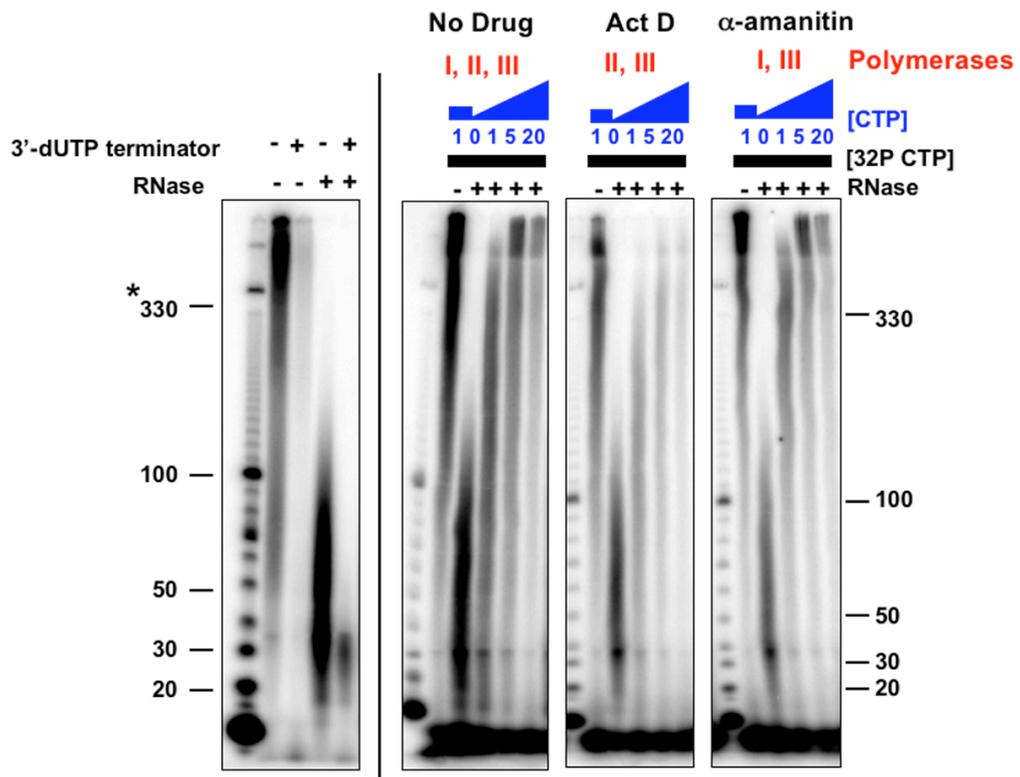
I first tested whether 5-Bromo-UTP (BrUTP) could be efficiently incorporated into RNA by nuclear RNA polymerases by also incorporating a radioactive nucleotide ( $\alpha^{32}\text{P}$ -CTP) in a run-on time course experiment. Consistent with previous results (Iborra et al., 1996), addition of Br-UTP allowed incorporation of  $\alpha^{32}\text{P}$ -CTP at ~80% efficiency compared with UTP, and approximately 10 fold over the control that lacked both UTP and Br-UTP (Figure 2.1). These radiolabeled RNAs made in the presence of Br-UTP bind very well to anti-Br-deoxy-U beads, which cross-reacts well with BrU (Figure 2.2). Although BrU is sometimes used as a mutagen, sequenced clones from GRO-seq libraries indicated the misincorporation rate by nuclear RNA polymerases is low. I also tested the propensity of BrU to cause misincorporation during reverse transcription by comparing sequencing results of cDNA clones that were generated from RT reactions that contain a BrU or U RNA template of known sequence. The results showed that there is no appreciable level of misincorporation by reverse transcriptase when BrU is incorporated into the RNA template. We thus chose BrU as our affinity tagged nucleotide for further development of the assay.



**Figure 2.2. Binding and elution of base-hydrolyzed BrU-RNA to  $\alpha$ -BrdU beads.** Isolated RNA from a nuclear run-on containing Br-UTP and  $\alpha$ - $^{32}\text{P}$ -CTP was base hydrolyzed to an average size of 100 bases, and then bound to agarose beads that are conjugated with an antibody specific for  $\alpha$ -BrdU. The beads were washed several times and then eluted. Equivalent amount of each fraction were run on an 8% denaturing PAGE gel to assess the efficiency of bead binding. Lane demarcations: M, 10bp ladder; UB, unbound fraction; E, Elution fraction.

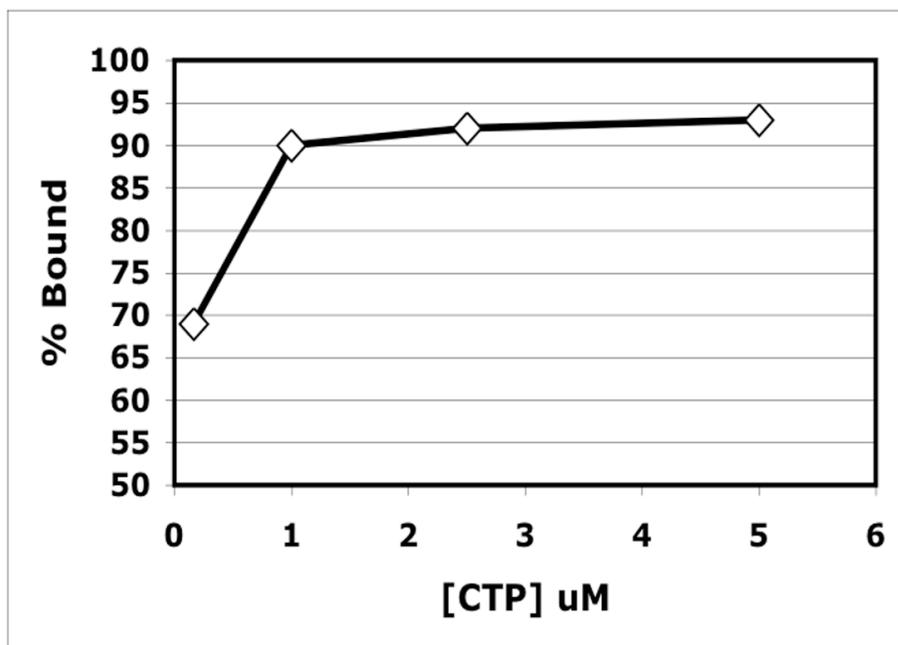
### 2.3.2 Control of resolution for GRO-seq.

The goal of the GRO-seq method is to isolate and obtain a high resolution and unbiased map of all RNAs as they are being transcribed. High resolution requires that run-on distances are kept short, whereas unbiased mapping requires efficient incorporation of the affinity-tagged nucleotide analog into all RNAs. I titrated nucleotide concentrations during the run-on step and defined the minimum distance for library preparation as the lowest concentration that allows maximum binding of the run-on RNAs to beads. To determine the length of the run-on transcription, nuclei were first pre-treated with RNase A and T1 in order to trim the nascent RNAs (Jackson et al., 1998). RNA polymerases can protect the nascent RNA from 15-20 bases upstream from the active site (Gu et al., 1996; Kireeva et al., 2005). The RNase activity was then removed through extensive washing and treatment with RNase inhibitor. The distance polymerases run-on was then controlled by titrating limiting concentrations of CTP. Since I primarily wanted to identify locations of RNA polymerase II (Pol II), I also examined the distance transcribed by polymerases in the presence of  $\alpha$ -amanitin and actinomycin-D.  $\alpha$ -amanitin is an efficient inhibitor of Pol II, but works much less effectively on Pol III, and is completely innocuous for Pol I transcription (Jackson et al., 1998). Actinomycin-D, when added to cells prior to nuclei isolation, primarily inhibits Pol I. By comparing the length of nascent transcripts produced from RNase treated nuclei and in the presence of inhibitors, I was able to deduce the distance Pol II transcribes under various limiting nucleotide concentrations, (Figure 2.3). Analysis of the efficiency of bead binding under similar conditions shows that with nuclei from IMR90 cells, 1  $\mu$ M CTP is sufficient to allow near maximum bead binding (Figure 2.4). This corresponds to a run-on



**Figure 2.3. Control of polynucleotide incorporation by RNA**

**Polymerases.** Nuclei were pre-treated with RNase to reduce the nascent RNA to ~20 nucleotides, washed, and then allowed to run-on in separate reactions containing a  $\alpha$ - $^{32}$ P-CTP and cold CTP for a total of 0.65 $\mu$ M (Lane 2), 1 $\mu$ M (lane 3), 5 $\mu$ M (lane 4) or 25 $\mu$ M (lane 5). Non-RNase treated nuclei supplemented with 1  $\mu$ M total CTP were used as a control (Lane 1). Cells were treated with Act-D and nuclei were treated with  $\alpha$ -amanatin.



**Figure 2.4. Efficiency of BrU-RNA binding in response to titration of limiting nucleotide.** Nuclear run-on were performed as described in figure 2.3, but without pre-treatment with RNase. Run-on RNAs from each sample were base hydrolyzed and bound to equivalent amounts of beads. The bound and unbound fractions were monitored for radioactivity by scintillation counting. The percent bound (y-axis) was calculated relative to input fractions and is displayed relative to the concentration of CTP in the reaction (x-axis).

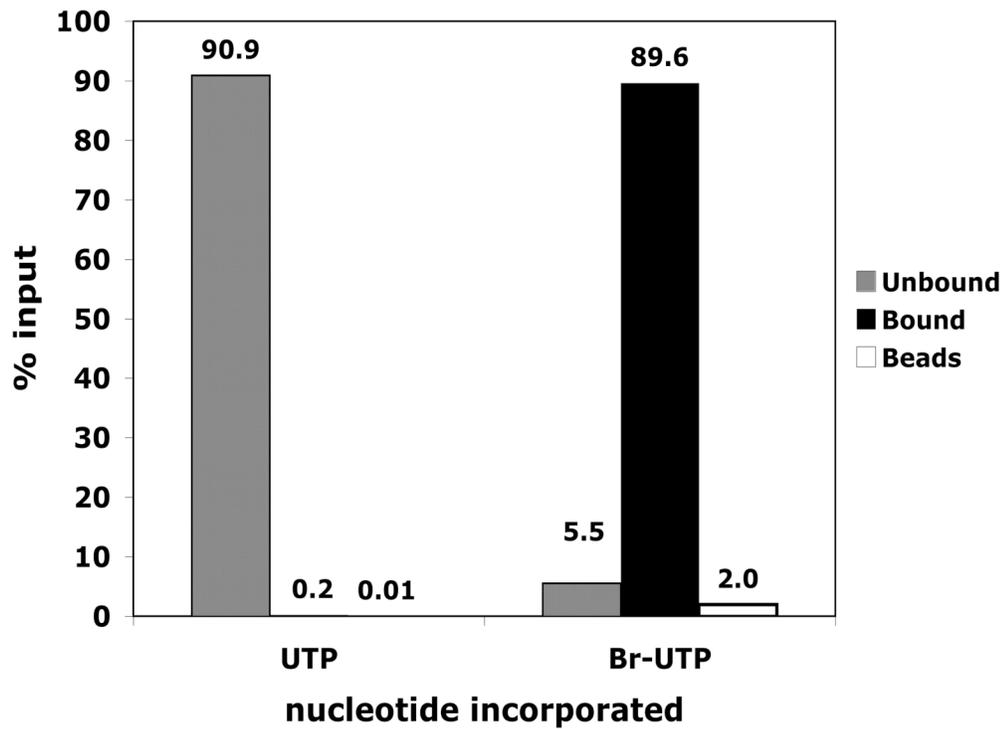
extension of ~80-100 nucleotides (Figure 2.3), which is the average length of the RNAs (~100 -120 nucleotides) subtracted by the length of RNAs protected by the polymerase (~20 nucleotides). I therefore considered 1uM CTP as the optimum concentration for these nuclei. In non-RNase treated nuclei (which are used for creating the NRO-library), base hydrolysis of the nascent RNAs to an average size that is equal to the length of the run-on transcripts will then result in a final mapping resolution of approximately half this distance. Base hydrolysis of the RNA improves the resolution of this assay by severing the extended portions of the nascent RNA transcript that contain the nucleotide analog from distal regions that were transcribed prior to the run-on reaction. For preparing GRO-seq libraries, I allowed Pol II to run-on approximately 80-100 bases, thus we estimate our resolution to be 40-50bp from the location of the polymerase active site at the time of the assay.

### **2.3.3 Yield, enrichment and purity of nascent RNA after triple selection**

High sensitivity and specificity is desired in any genomic assay in order to decrease both false negative and false positive results. These parameters require that both the yield and enrichment of run-on RNAs be high relative to contaminant RNAs.

#### ***Enrichment by tracking radiolabeled NRO-RNAs.***

To assess the specificity and efficiency of the purification, I first measured the enrichment of the nascent RNAs by incorporating a radiolabeled nucleotide ( $\alpha$ -<sup>32</sup>P-CTP) in run-on reactions performed in the presence of either UTP or Br-UTP. Quantification of the bound and unbound fractions from each reaction by scintillation counting showed that the enrichment by this method is



**Figure 2.5. Level of enrichment from  $\alpha$ -BrdU bead purification.**

Run-ons were performed in the presence of either UTP or Br-UTP, and handled as described previously. RNAs from each fraction were quantified by scintillation counting.

~450 fold for a single round of bead binding (Figure 2.5). Successive enrichment could not be examined because the amount of radioactivity in the UTP-RNA was below the limit of detection in the bound fraction after binding to a new set of beads. In order to assess whether contaminant RNA was able to cross-hybridize with BrU-RNA, I also performed a bead binding with  $\alpha^{32}\text{P}$ -CTP radiolabeled, UTP-containing RNA in the presence of non-radioactive, BrU RNA. Under these conditions the level of radioactivity in the bound fraction was the same as CTP-labeled samples containing only UTP suggesting that cross-hybridization is negligible.

#### ***Measurement of enrichment and purity by RT-qPCR.***

Since the amount of radiolabeled NRO-RNA measured in the above experiments is a minor fraction of the total RNA isolated from nuclei, it is possible that a significant amount of contaminant RNA still exists after triple selection. The total mass of RNA in the bound fraction after triple selection was near the limit of detection, and beyond the limit of detection for Br-U and U-RNA, respectively, thus I could not reliably measure the enrichment by UV spectrometry alone. I could determine that there was 50 $\mu\text{g}$  in the starting pool and 300ng in the elution from the third round of bead binding for the Br-UTP samples. We therefore added spiking controls, consisting of multiple small (~100base) RNAs that were in vitro transcribed in the presence of either UTP or Br-UTP. The cDNAs used for in vitro transcription were reverse transcribed and amplified from *Arabidopsis thaliana* total RNA. U-RNAs were added in 10-fold dilutions from  $1 \times 10^{10}$ - $1 \times 10^7$  copies and a BrU-RNA was added at  $1 \times 10^7$  copies. After triple selection, reverse transcription followed by quantitative PCR (RT-qPCR) was carried out on the final elution for each RNA. The Br-U

RNA was present at 50% relative to input, and all U-RNAs were at or below background for the assay. The lowest amount of the input that we could detect was 1:10,000, therefore non-BrU RNAs are present at >1:10,000<sup>th</sup> relative to the starting amount. This corresponds to 5ng since the procedure starts with 50µg of nuclear RNA. Since the final elution contains 300ng of RNA, U-RNA represents 1.6% of the final mass, corresponding to >98% purity for BrU-RNA.

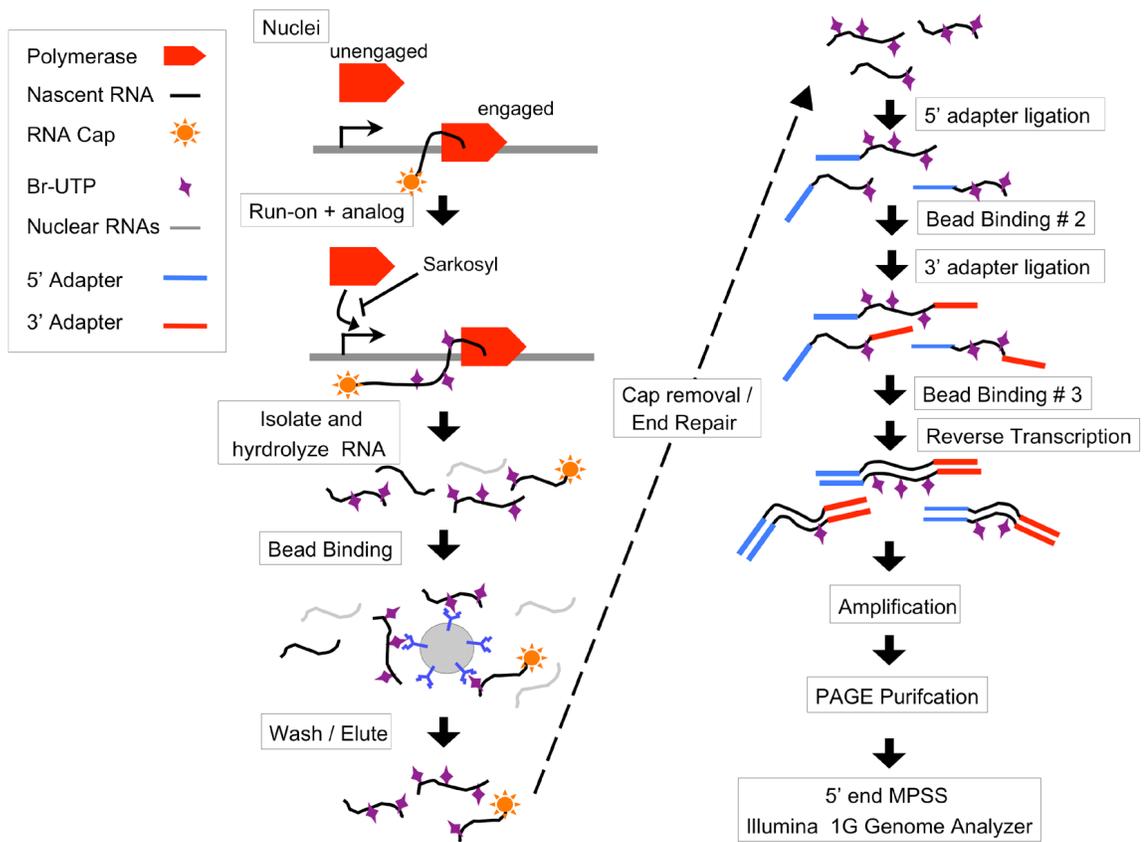
## **2.4 The rationale for choosing sequencing vs. microarray hybridization for global analysis.**

The global analysis of nascent transcripts by our GRO-seq method is compatible with microarray hybridization platforms as well as high throughput sequencing methods. We have chosen to adapt our global survey of nascent transcripts for analysis by high throughput next generation sequencing over conventional microarray hybridization platforms for a number of reasons. Conceptually, sequencing not only maps the location of transcribing polymerases, but also provides the direction of transcription since the direction of sequencing is controlled by to which end the sequencing primer template is added. Whole genome coverage is then obtained by simply mapping the sequence to a reference genome, thus one can apply any genome wide method to any sequenced genome. Even if you were interested in studying the transcriptome of - say - the grey fox, for which there is no reference genome, there is significant enough homology with the canine genome to get an appreciable amount of alignment (LJ, Core, unpublished results). In contrast, whole genome analysis on a microarray platform requires synthesizing on the slide representative segments from both strands of the

entire genome, and preparation of the sample requires an extra step of in vitro transcription of the cDNA to ensure that only the transcribed strand is hybridized to the array. From a technical standpoint, whole-genome microarray hybridizations require pooling of samples from multiple plates of cells (150 cm<sup>2</sup> plates) to obtain large amounts of material, on the order of 50ug (for hybridization to 38 separate arrays (Li et al., 2003), whereas we have carried out our initial analysis of GRO-seq with 5X10<sup>6</sup> cells (one plate), and less than 1 ug of a purified library. Finally, analysis of microarray data requires extensive normalization and statistical testing to obtain a relative level of signal, but sequencing allows the counting of signals for easy quantification and provides a signal that is more linearly responsive to the amount of an RNA species.

A number of high-throughput next generation sequencing technologies are currently available, and it is expected that more will come to market in this rapidly developing field (Shendure and Ji, 2008). The currently available platforms vary in the number and lengths of sequence reads per run, reagent costs, and library preparation protocols. 454-sequencing, available from Roche, offers the longest reads (~250 bases), but has the disadvantage of a lower number of reads (~3X10<sup>5</sup>/run) and high reagent costs compared to the platforms offered by Illumina and Applied Biosystems. These two systems offer shorter reads (33-35 bases) but obtain larger numbers of reads per run (4X10<sup>7</sup> and 8X10<sup>7</sup> for Illumina and ABI, respectively). The greater depth of coverage afforded by high numbers of reads is critical for efficient quantification of nascent transcripts, and the shorter read lengths are sufficient for accurate mapping of the transcripts to genomes. This is important for coverage. Peter Cook's Lab has estimated that there are ~90,000 active RNA

**Figure 2.6 Overview of the GRO-seq method.** Polymerases are allowed to run-on ~100 bases in isolated nuclei in the presence of sarkosyl and Br-UTP. The RNA is then base hydrolyzed to ~100 bases and bound to agarose beads that are coated with an  $\alpha$ -BrdUTP antibody. 5'-7meG caps are then removed, and the ends of the RNA are prepared for adapter ligations. Illumina small RNA adapters are added to the 5' end followed by the 3' end, with an additional round of immuno-enrichment after each adapter ligation. The RNAs are then reverse transcribed, amplified, and PAGE purified prior to sequencing from the 5' end on the Illumina 1G genome analyzer. See text for a more detailed description and methods for protocol.



polymerases in HeLa cells: 15,000 Pol I, 65,000 Pol II, and 10,000 Pol III (Faro-Trindade and Cook, 2006). Ensuring that genes containing low levels of transcriptionally-engaged polymerases are detected, requires the sequencing of millions of run-on RNAs. Our core facility has purchased an Illumina 1G genome analyzer, and we will conduct our initial experiments on this platform, however GRO-seq is easily compatible with any sequencing technology.

## **2.5 Overview of GRO-seq method**

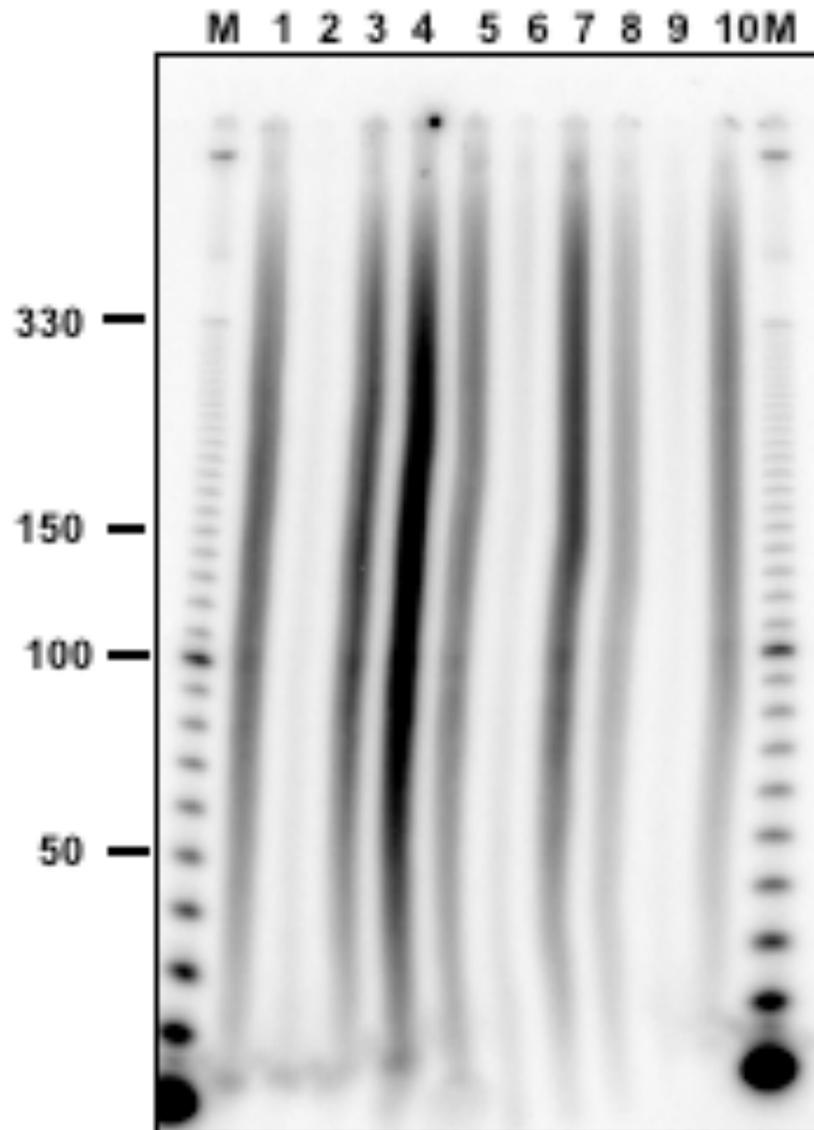
This is intended to be a walk-through of the GRO-seq method and results (Figure 2.6), for a detailed description of the steps involved in preparing NRO-libraries, please see the methods in section 2.2. Nuclei isolation and run-on reactions are performed using standard protocols with the exception that 0.5% sarkosyl is added, 5-Bromo-UTP is used in place UTP, and the concentration of CTP is adjusted to  $1\mu\text{M}$  to keep the run-on distance to  $\sim 100$  nucleotides (see above).  $\alpha\text{-}^{32}\text{P}\text{-CTP}$  is also used as a tracer in order to follow the purification steps, and analyze the products on denaturing PAGE. RNA is isolated and base hydrolyzed to the desired size. RNA fragments are then isolated by binding to anti-deoxy-BrU beads to select against accumulated nuclear RNAs, washed several times, and eluted from the beads. Because base hydrolysis of RNA leaves a molecule with a 5'-hydroxyl and a 3'-phosphate, neither of which are substrates for ligation of adapter oligos, the RNA ends must be repaired. First, the RNAs are treated at low pH with tobacco acid pyrophosphatase to remove 5-methyl guanosine caps (Rasmussen, and Lis, 1993), and then are treated at low pH with T4 polynucleotide kinase (PNK) to remove the 3'-phosphate (Cameron, and Uhlenbeck, 1977). The pH is then raised and the RNA is treated again with

PNK, except now in the presence of ATP, to add a 5'-phosphate. An adapter is then added to the 5'-end with T4-RNA ligase and the RNA is bound to anti-deoxy-BrU beads to remove excess linkers and further enrich the RNA. This process is then repeated for the addition of a 3'-adapter. The affinity-enriched RNAs are then reverse transcribed, amplified, and PAGE purified. Analysis of a fraction of each step by denaturing polyacrylamide gel electrophoresis (Figure 2.7) shows that the RNA remains largely intact throughout the procedure. After amplification and PAGE purification (Figure 2.8), the library appears to be, on average, 100 bases in length (~190 base – 90 base adapters). A known amount of the library is re-amplified to determine the primer efficiency from which the original complexity of the cDNA library can be extrapolated. In the two libraries I constructed for this study, I obtained complexities of  $1 \times 10^9$  molecules prior to amplification. I also cloned and sequenced by conventional methods 50 molecules to verify the size and ensure the quality of the library before massively parallel sequencing on the Illumina 1G genome analyzer.

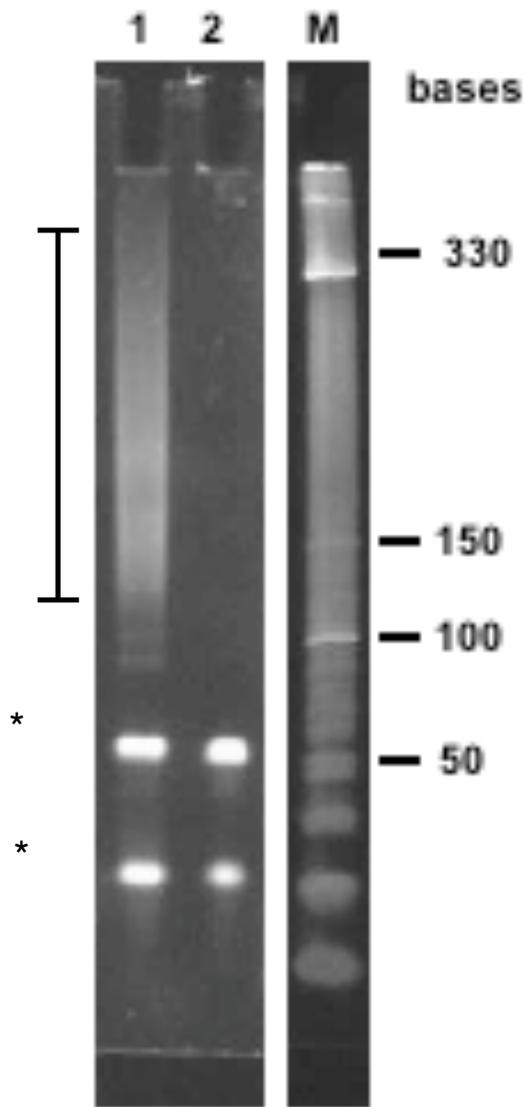
## **2.6 Preliminary analyses**

### **2.6.1 Correlation between replicates.**

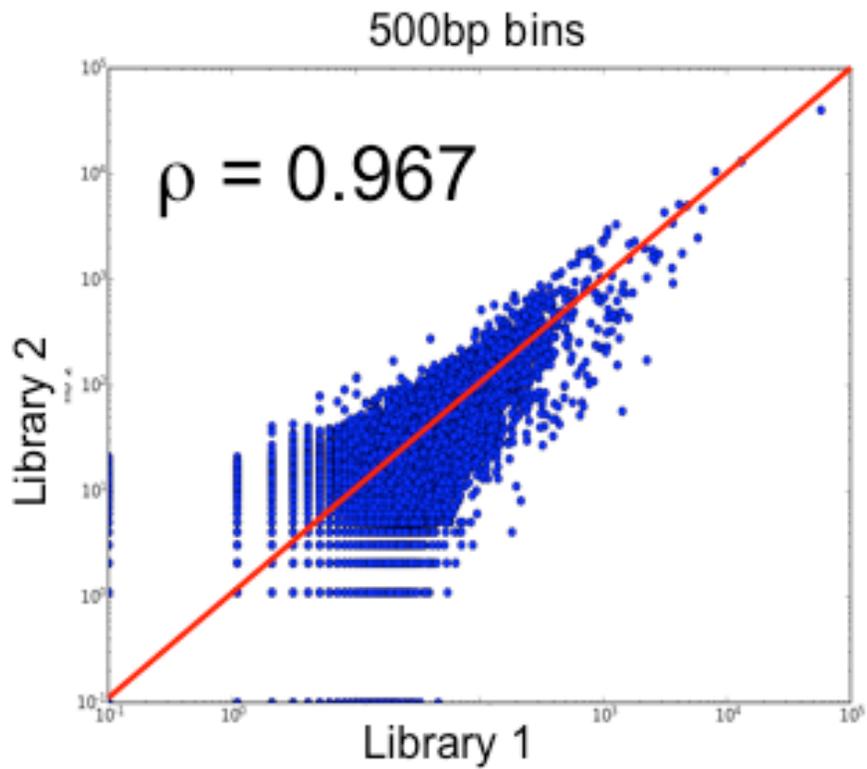
In total,  $\sim 2.5 \times 10^7$  33bp reads were obtained from two independent replicates prepared from sarkosyl-treated human IMR90 nuclei, of which  $\sim 1.1 \times 10^7$  (40%) of reads mapped uniquely to the human genome. Only unique reads were further analyzed. Correlation of the read densities between the two replicates produced in this study show that replicates agree remarkably well (Spearman correlation = 0.96, Figure 2.9).



**Figure 2.7. 8% Denaturing PAGE analysis of fractions from GRO-seq library preparation.** Lanes: 1) Input, 2) Unbound-1, 3) Elution-1, 4) After TAP-PNK treatment, 5) 5' adapter ligation, 6) Unbound 2, 7) Elution 2, 8) 3' adapter ligation, 9) Unbound 3, 10) Elution 3.



**Figure 2.8. Example of amplified NRO-library cDNA.** After the third elution the library was reverse transcribed amplified by 15 cycles of PCR, and then run on an 8% PAGE gel for purification away from the primers (\*) Lane 1 cDNA library, Lane 2) No template control. Bracket indicates region cut from gel.



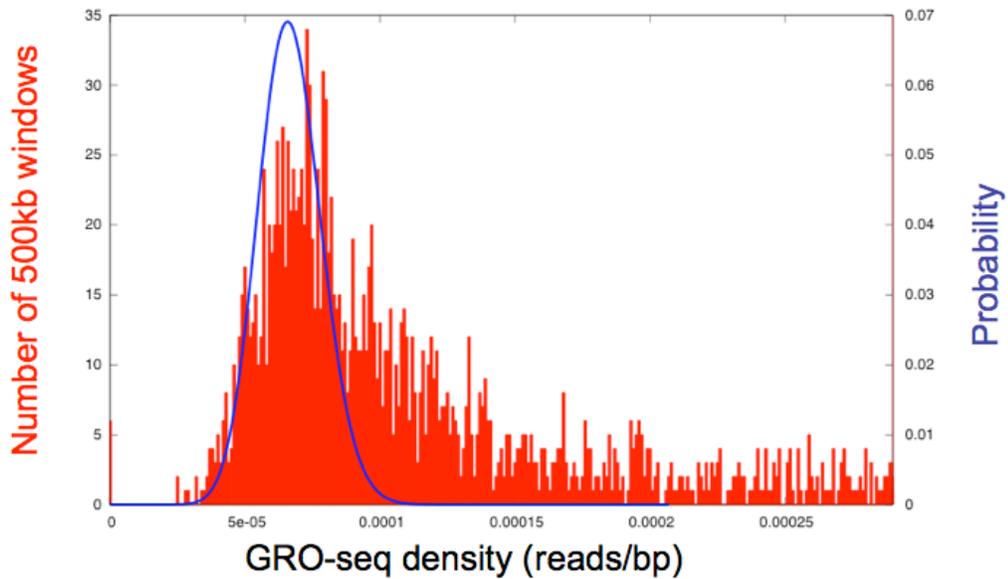
**Figure 2.9. Correlation of GRO-seq biological replicates.** GRO-seq transcript reads were mapped to the genome and unique reads were binned in 500bp windows. Of the 6,160,849 windows, 3,458,076 windows had no reads in each replicate. The replicates show a correlation coefficient of 0.967 (Spearman correlation).

### **2.6.2 Computational analysis of background.**

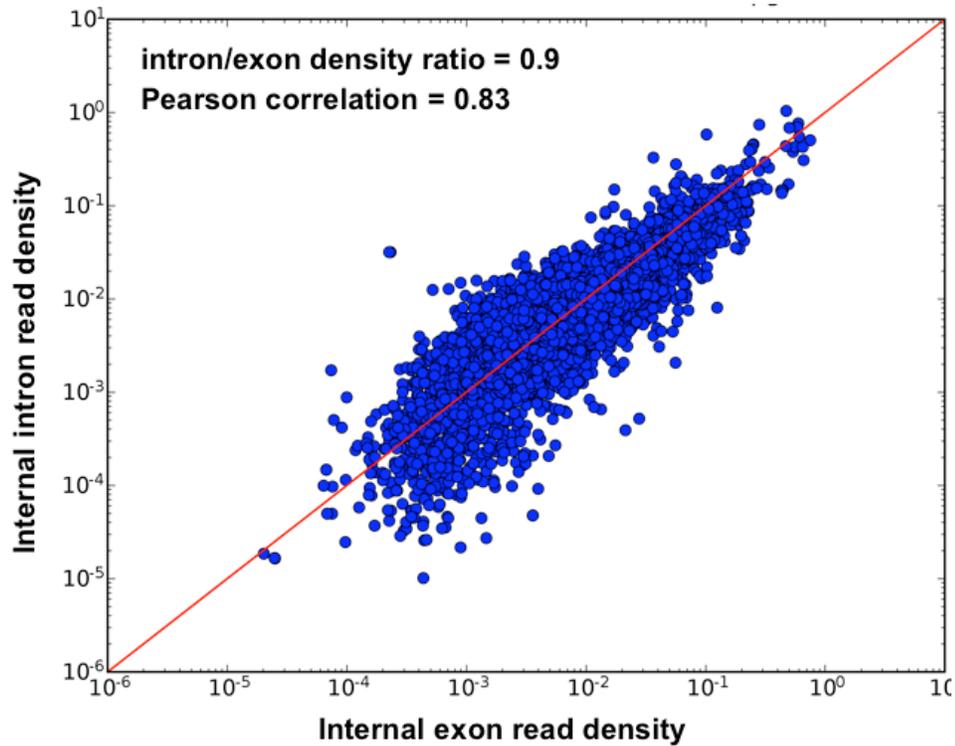
In addition to the experimental results presented in section 2.3.3 , several computational analyses suggest that our NRO-RNA libraries were highly enriched for NRO-RNA relative to accumulated RNAs. First, an estimation of background was determined by binning reads in 500kb windows genome wide. The distribution of windows with the lowest densities fits a Poisson distribution corresponding to spreading 2-3% of the aligned reads randomly over the mappable portion of the genome, agreeing well with the above experimental results and suggesting that background for the assay approaches 0.04 reads on a single strand per 1kb (Figure 2.10). Second, transcription is detected in regions of transcription units that are not present in fully processed mRNAs, including introns and regions beyond the site of nascent RNA cleavage and polyadenylation. The ratios of read density within introns vs. exons is 0.9 (pearson correlation = 0.83), and is not significantly different from 1 ( $P = 0.71$ , Figure 2.11). Third, known gene deserts ranging from 0.6 Mb to 3 Mb, have an average density of reads on both strands together of 0.07 reads/1kb, which also agrees well with our experimental and computational analyses of background (Table 2.1).

### **2.7 Transcription in nuclei: reflection of in vivo transcription status.**

Given that the nuclear run-on assay is performed in vitro, it is conceivable that some polymerases might bind and initiate transcription and/or elongate during isolation of the nuclei - prior to the addition of sarkosyl. However, several considerations suggest that very little if any transcription initiation or elongation occurs during nuclei isolation. Immediately before



**Figure 2.10. Background calculation by low-density windows.** After aligning reads to genome, the density of GRO-seq reads was assessed in 500kb windows. Shown in red is a histogram of the lowest density windows and in blue is a Poisson distribution with a mean given by placing 3% of all GRO-seq reads at random throughout the mappable portion of the genome.



**Figure 2.11. Comparison of GRO-seq read density in Exons vs. introns.** Scatter plot showing the density of GRO-seq reads within introns (y-axis) vs exons (x-axis) for each RefSeq gene. Axes are in  $\log_{10}$  scale. Only internal exons and introns were used in the analysis to avoid inflation of signal due to promoter-proximal pausing or build up of polymerases that can occur near the 3'-end of genes.

**Table 2.1: Background calculation in gene deserts.** The indicated large intergenic spaces were analyzed for the number of GRO-seq reads on either strand and the number of mappable bases.

Chromosome	Start	Stop	Read count	Mappable Length	Read density (reads/bp)
Chr4	27900001	3500000	1667	6900326	$2.42 \cdot 10^{-4}$
Chr2	144700001	148400000	927	3425237	$2.71 \cdot 10^{-4}$
Chr1	79311112	81964443	170	2407985	$7.06 \cdot 10^{-5}$
Chr1	185879619	188333419	149	2234374	$6.67 \cdot 10^{-5}$
Chr2	139254282	140705466	56	1328410	$4.22 \cdot 10^{-5}$
Chr2	56466815	57988288	67	1344339	$4.98 \cdot 10^{-5}$
Chr2	33700268	36420000	125	2384667	$5.24 \cdot 10^{-5}$
Chr2	139254283	140705464	56	1328410	$4.21 \cdot 10^{-5}$
Chr2	155421262	156585290	42	1057143	$3.97 \cdot 10^{-5}$
Chr2	192775891	196025184	147	2902767	$5.06 \cdot 10^{-5}$
Chr2	222155254	222762851	21	550326	$3.82 \cdot 10^{-5}$
Chr4	44473369	45702544	43	1099820	$3.91 \cdot 10^{-5}$
Chr4	104870422	105599015	21	640337	$3.28 \cdot 10^{-5}$
Chr4	116264481	118214158	71	1772986	$4.00 \cdot 10^{-5}$
Chr4	135352353	137135534	61	1605865	$3.80 \cdot 10^{-5}$
Chr5	104744250	106704250	65	1785211	$3.64 \cdot 10^{-5}$

preparing nuclei, the cells are brought to  $\sim 4^{\circ}\text{C}$  within seconds of removing the media. Under these conditions, even if a polymerase comes into contact with a promoter, it is unlikely to form a proper preinitiation complex (PIC) within the timeframe of the procedure (30min), and due to the high energy requirements of promoter DNA unwinding, even less likely to initiate transcription. Nucleotides are removed by washing within in the first 15 min of the procedure, thus initiation becomes impossible after this point. Also, experiments from Peter Cook's lab that utilized a combination of in vivo labeling of nascent transcripts with BrU followed by in vitro labeling with biotin-CTP have shown that no new initiation occurs in nuclei, since all biotin-CTP sites also labeled with BrU (Jackson et al., 1998). These experiments by the Cook lab were carried out in the absence of sarkosyl, thus we think the event of observing initiation in isolated nuclei in the presence of sarkosyl, is very unlikely. Finally, high-resolution mapping experiments of pausing at the *Drosophila Hsp70* gene have shown that Pol II does not elongate during the nuclei isolation step (Rasmussen and Lis, 1993; Rougvie and Lis, 1988b).

Further support that the nuclear-run-on reflects the in vivo state of transcription can be obtained by comparing the GRO-seq results with other assays that start with whole cell preparations. In a parallel study, Seila et al. (Seila and et al., in press), show that small transcription start site RNAs (TSS-RNAs) are produced by promoters in both the forward and divergent direction (see chapter 3). This is evidence that the transcription we detect at promoters with GRO-seq occurs in vivo. Also, ChIP data that show that promoter regions are bound by Pol II is generated by cross-linking whole cells, thus Pol II-DNA interactions are occurring in vivo at the time of cross-linking (Kim et al., 2005b). These peaks of polymerase binding show nearly complete overlap

with promoters called active by GRO-seq (See chapter 3). GRO-seq identifies additional active promoters because of the increased sensitivity afforded by sequencing. In addition, recent Pol II ChIP-seq data from Sultan et al. (Sultan et al., 2008) shows that Pol II is present in a peak that is resolvable from the peak at the transcription start site. Sultan et al. hypothesize that the upstream peak could be an upstream pre-initiation complex, or some sort of storage site for Pol II. We show that this peak represents transcriptionally engaged Pol II complexes that are oriented in the opposite direction of gene transcription. This ChIP-seq data is further evidence that divergent polymerases can be detected from whole-cell preparations, and are not a consequence of polymerase binding during preparation of nuclei.

## **2.8 Concluding remarks**

Imagine our excitement and anticipation when we were first uploaded the data onto the UCSC genome browser. Of course, the first gene we went to was the HSP70 gene (HSPA1A in humans), and sure enough, we saw a beautifully sharp and prominent promoter-proximal peak. Our excitement has since been tempered (only slightly) by the daunting task of analyzing such a large dataset. Not only is this first genome-wide data set for our lab, it is a unique dataset forcing us to adapt or develop the proper analysis tools. For this we procured a faster-than-average computer, and enlisted the computational and computer programming powers of Josh Waterfall.

## CHAPTER 3<sup>3</sup>

### MASSIVELY-PARALLEL SEQUENCING OF NASCENT RNAS REVEALS DISTRIBUTIONS OF ENGAGED RNA POLYMERASE IN THE HUMAN GENOME

#### 3.1 Introduction

Transcription of coding and non-coding RNAs by RNA polymerases requires the collaboration of hundreds of transcription factors with specific DNA sequence elements over short and long distances to direct and control polymerase recruitment, initiation, elongation and termination. The advent of whole-genome microarrays and ultra high-throughput sequencing technologies provide remarkable advances in the efficiency of mapping the distribution of transcription factors, nucleosomes and their modifications, as well as RNA transcripts throughout genomes (ENCODE Project Consortium et al., 2007; Wold and Myers, 2008). Several studies using the chromatin immunoprecipitation assay coupled to genomic DNA microarrays (ChIP-chip) have shown that RNA polymerase II (Pol II) is present at disproportionately higher levels near the 5' end of many genes relative to downstream portions (Guenther et al., 2007; Kim et al., 2005b; Muse et al., 2007; Schones et al., 2008; Zeitlinger et al., 2007a). This technique locates Pol II complexes but cannot determine whether they are engaged in transcription or not. Small-scale analyses using independent methods have shown that this distribution likely represents a transcriptionally engaged but paused or arrested Pol II (Lee

---

<sup>3</sup> Information in this chapter is largely from ((Core et al., 2008)).

et al., 2008; Muse et al., 2007; Zeitlinger et al., 2007a). This promoter-proximal pausing is a potential mechanism through which transcription of genes can be regulated at the stage of elongation rather than recruitment of Pol II (Saunders et al., 2006). However, no assay exists to confirm this hypothesis on the genomic scale.

Whole genome mapping of accumulated RNA transcripts by microarray hybridization (Bertone et al., 2004; Kapranov et al., 2007a; Kapranov et al., 2007b) and their transcription start sites (TSSs) by selection of full-length transcripts (Carninci et al., 2005; Carninci et al., 2006; Katayama et al., 2005; Kimura et al., 2006) are beginning to show that the genome is highly transcribed compared to previous estimates, with some notable features being novel transcription units and unannotated sense/antisense transcript pairs. These recent discoveries indicate that the origin and function of transcribed RNAs is still being defined, thus independent methods that can comprehensively document sites of transcription are of utmost importance for understanding genome function. An alternative approach to analyzing accumulated RNAs is to examine transcriptionally-engaged polymerase density throughout the genome by tracking the associated nascent RNA. This would allow both the detection of paused polymerases and the detection of rare or unstable transcripts that are not easily detected in accumulated RNA pools. In addition, tracking of steady-state production of nascent RNA is valuable data that can be compared to accumulated mRNA levels in order to examine the extent with which particular genes are regulated by mRNA turnover. We therefore sought to develop a nuclear run-on assay (NRO) that would document the above phenomena in a more comprehensive manner.

Here, I present a Global Run-On-Sequencing (GRO-seq) assay to map and quantify transcriptionally-engaged polymerase density genome-wide. These measurements provide a snapshot of genome-wide transcription and directly evaluate promoter-proximal pausing on all genes. I used nuclear run-on assays (NRO) to extend nascent RNAs that are associated with transcriptionally-engaged polymerases under conditions where new initiation is prohibited. To specifically isolate NRO-RNA, we added a ribonucleotide analog (BrUTP) to BrU tag nascent RNA during the 'run-on' step. The length of the incorporated polynucleotide was kept short and the NRO-RNA was chemically hydrolyzed into short fragments (~100 bases) to facilitate high-resolution mapping of the polymerase origin at the time of assay. BrU-containing NRO-RNA was triple selected through immuno-purification with an antibody that is specific for this nucleotide analog, resulting in a >10,000-fold enrichment of NRO-RNA pool that was determined to be >98% pure. A NRO-cDNA library was then prepared for sequencing from what represents the 5'-end of the fragmented RNA molecule using the Illumina 1G high-throughput sequencing platform. The origin and orientation of the RNAs, and therefore the associated transcriptionally-engaged polymerases was documented genome-wide by mapping the reads to the reference human genome.

I will describe genes that have transcriptionally engaged Pol II accumulated at the 5'-end as 'paused' since this pattern mirrors that of several human and drosophila genes that have been identified as paused (see below). Pausing refers to a polymerase that is engaged in transcription, is either not moving forward or moving slowly, but nonetheless retains its elongation potential. Since the NRO assay that I have used here requires the polymerase to be transcriptionally competent, it is fitting to describe the polymerases that

are accumulated at promoters as paused. The term ‘stalled’ is sometimes used to describe a polymerase that is found at higher levels at the 5’ ends of a gene (Muse et al., 2007; Nechaev and Adelman, 2008; Zeitlinger et al., 2007a). Stalling refers to an engaged polymerase complex, but makes no assumption about whether that polymerase is competent to resume elongation (Fish and Kane, 2002). That is, a stalled polymerase could be paused, backtracked and arrested, or could exist in some form of dynamic equilibrium between the two states. The potassium permanganate footprinting assay can be used to map the location of a paused or stalled polymerase (Giardina and Lis, 1993). This technique maps the unwound portion of the template DNA that is associated with an engaged polymerase. In the absence of further experimentation that examines transcriptional competence, genes that have excessive permanganate reactivity at the 5’ end corresponding to the position of a paused polymerase are generally described as experiencing stalling (Lee et al., 2008; Muse et al., 2007; Zeitlinger et al., 2007a).

### 3.2 Materials and Methods

#### *Calculation of gene activity*

Gene activity was defined as  $N/L$  where  $N$  is the number of coding strand GRO-seq reads from +1kb (relative to the TSS) to the end of each gene, and  $L$  is the number of mappable bases in this region. The significance of a given gene’s activity level was determined by the probability of observing at least  $N$  reads in an interval of length  $L$  from a Poisson distribution of mean  $\lambda = 0.04$  reads/kb (the background density of our libraries).

$$P = \sum_{n=N}^{\infty} \frac{(\lambda * L)^n e^{-\lambda * L}}{n!}$$

If the probability was less than 0.01, the gene was called active. The first kilobase of each gene was omitted to better gauge the density of polymerase that actively elongates through the gene and to avoid over-counting from the increased density of paused polymerase in the 5' end of the gene. All analyses were done with the complete RefSeq gene list for the hg18 assembly of the human genome reduced to include only genes at least 3kb in length so that the measurement of GRO-seq density in the body of the gene would be robust.

### ***Correlation of GRO-seq densities with microarray expression data***

The previous expression microarray work (Kim et al., 2005a) was performed on the Affymetrix U133Plus2 array. To correlate the GRO-seq data with this expression array data, the original array data was downloaded from the supplementary material of that paper, and the knownToRefSeq and knownToU133Plus2 tracks from the UCSC genome browser were used to map RefSeq genes to probe IDs. The analysis of the array data was performed as in the original paper (Kim et al., 2005b). That is, a probe had to be present or absent in both replicates to be called present or absent. If all probes mapping to a particular gene are absent then the gene is absent and if any probes mapping to a particular gene are present then the gene is present. All other genes are considered ambiguous and removed from future analyses.

### ***Identification of promoter proximal peaks***

The exact position of many TSSs are not precisely annotated and many promoters in fact do not have a single well defined TSS (Carninci et al., 2006). Therefore, in order to identify the peak of promoter proximal coding strand

GRO-seq reads, we tiled around each annotated TSS 1kb upstream and downstream in 50 bp windows, shifting by 5 bp. In each window we counted the number of coding strand reads and the number of mappable bases. We could then calculate the significance of the density in each window by comparing to the background density of 0.04 reads/kb in a manner similar to how gene activity significance was calculated (see above). The most significant window was chosen as the promoter proximal peak, and if multiple windows had the same significance, then the most 5' of these windows was chosen. If the promoter proximal peak had a p value less than 0.001, the gene was identified as having a significant promoter proximal activity. To identify the divergent peak, a similar approach was used but tiling was done +/- 1 kb from the identified promoter proximal peak and only reads on the noncoding strand were counted. The same p value cutoff of 0.001 was used to classify genes as having a significant peak of divergent transcription.

### ***Identification of paused genes***

Significantly paused genes were identified by using the Fisher exact test to compare the density of reads in the sense strand promoter proximal peak to the density of reads in the body of the gene as compared to a uniform distribution of all these reads based on the number of mappable bases. A p value cutoff of 0.01 was used to call significantly paused genes.

### ***Extending peaks to transcribed regions:***

To measure how far the significant promoter proximal peaks could be extended into transcribed regions we began by identifying the 3' most read within the peak (in a strand specific manner), and calculated  $d(n)$ , the distance

from the current read to the  $n^{\text{th}}$  downstream read on the same strand. If this distance was less than the cutoff distance, the 3' boundary of the peak was extended to this  $n^{\text{th}}$  read and the process was repeated by shifting one read downstream. This process continued until the peak could no longer be extended. The value of  $n$  used in this analysis was 5 and the length cutoff was 2.5 kb.

***Correlation of GRO-seq and ChIP-chip data:***

The ChIP-chip from the Ren lab data was reported for positions relative to the hg16 assembly of the human genome (Kim et al., 2005b). The UCSC liftOver tool was used to convert these coordinates to the hg18 assembly. To assess GRO-seq levels around the TAF1 peaks identified in the previous work, we looked either at the GRO-seq density of the associated gene for the transcript-matched promoters, or 1kb upstream and downstream for the novel promoters. For the transcript-matched promoters, gene activity values and significance were calculated as described above. For the novel promoters, we counted the total number of reads on both strands and the number of mappable bases. To identify significant transcription, we used a p value cutoff of 0.01 when comparing to the probability of obtaining that number of reads or more from a Poisson distribution with a rate of  $\sim 0.08$  reads/kb because both strands are being counted.

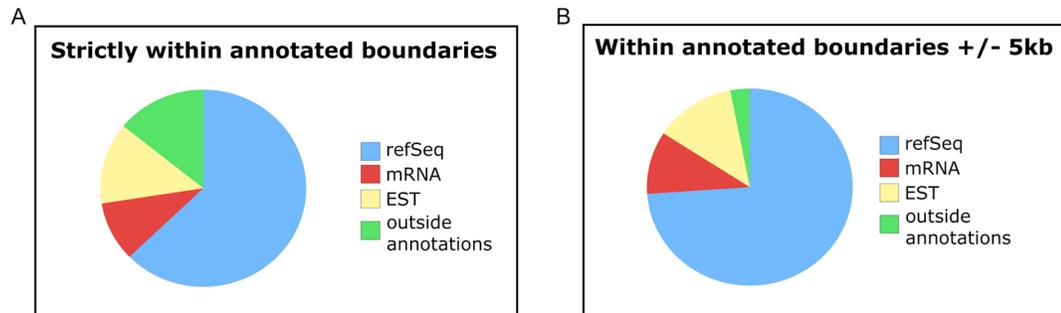
## **3.2 Results**

### **3.2.1 GRO-seq reads distribution relative to annotated transcript boundaries**

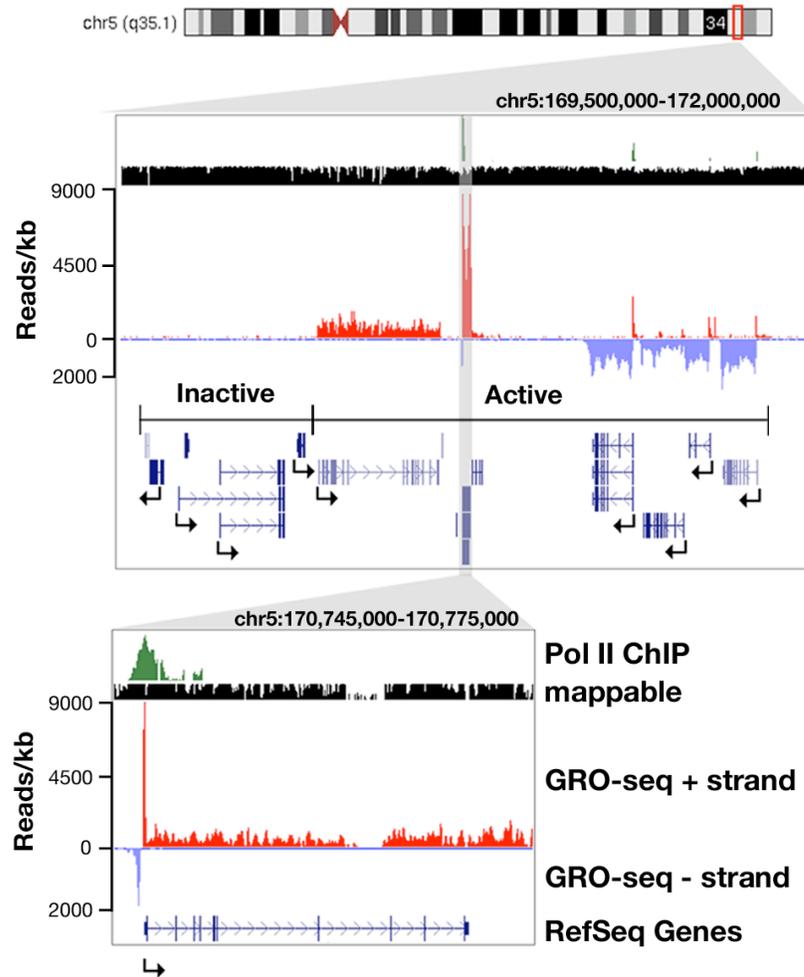
Most reads align within boundaries of known transcription units. 62.8% of reads align on the coding strand within Refseq genes. An additional 9.6% of reads align to the coding strand within the boundaries of Human mRNA, and a further 13.4% within EST coding regions (Figure 3.1). These values increase to 74.0%, 10.2%, and 12.8%, respectively, for a total of 97%, if the boundaries are expanded by 5kb from both the 5' and 3' ends of the annotated features. Manual inspection of several large regions shows that GRO-seq can differentiate between transcriptionally active and inactive regions in large chromosomal domains (Figure 3.2).

### **3.2.2 Comparison of GRO-seq with Pol II ChIP-chip data from the Ren lab**

To assess the relationship between promoters identified by transcription factor binding (i.e. ChIP) assays and the presence of engaged polymerase, we compared our GRO-seq densities with the list of over 10,000 active promoters identified in a previous study performed in the same cell line (Kim et al., 2005b). Active promoters in that study were identified genome-wide by binding of TAF1, a component of the general transcription factor TFIID that is critical for specifying most sites of initiation by Pol II (Kim et al., 2005b). That study identified 9,324 TFIID binding sites within 2.5kb of annotated transcripts (referred to as transcript-matched) and 1,239 novel promoters that were greater than 2.5kb from known 5'-ends of genes. Of the promoters associated with annotated transcripts, 9,217 (98.9%) have coding-strand GRO-seq densities within the body of the associated gene significantly above



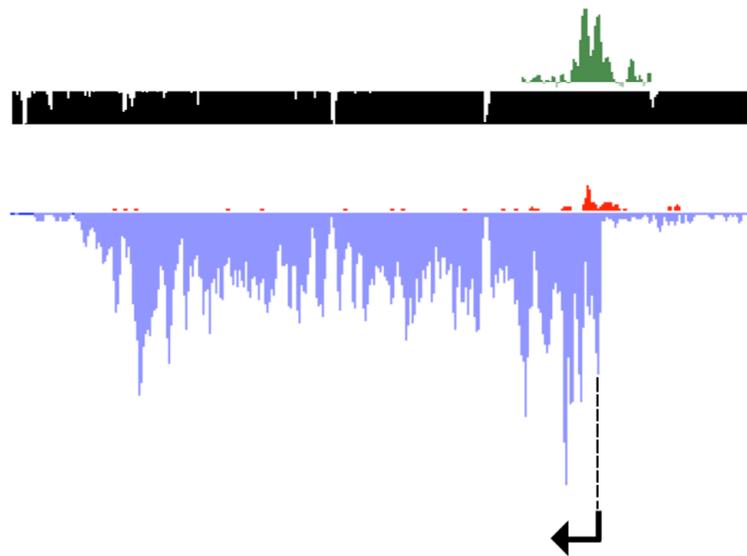
**Figure 3.1. Summary of GRO-seq data relative to annotated transcript boundaries.** The fraction of reads aligning to the coding strand and strictly within the annotated boundaries (A) or within the annotated boundaries expanded by 5 kb (B). Reads were first mapped to RefSeq genes (blue), then unmapped reads were mapped to Human mRNA (red), then reads that were still unmapped were mapped either to Human ESTs (yellow) or outside annotations (green).



**Figure 3.2. Sample of GRO-seq data viewed on the UCSC genome browser.** A 2.5 Mb region on chromosome 5 showing GRO-seq reads aligned to the genome at 1bp resolution, followed by an up-close view around the NPM1 gene. Pol II ChIP results are shown in green, mappable regions (black), GRO-seq hits on the plus strand (left to right; red), GRO-seq hits on the minus strand (light blue), RefSeq gene annotations (dark blue).

background. Because the novel promoters have no associated orientation by CHIP, we assayed the neighboring +/- 1 kb region and found that 1,185 (95.6%) had GRO-seq densities significantly above background. Details of the statistical methods are described in the Methods section below. GRO-seq not only confirms these sites as active promoters, but also provides the direction and extent of transcription from these novel promoters (Figure 3.3). When we used GRO-seq densities alone to identify the number of active promoters within +/- 1 kb of RefSeq annotated 5'-ends, we find 16,882 active promoters (see below). The increase in active promoters found here could be a consequence of different sensitivities, but may also reveal a class of promoters that are independent of TFIID (Huisinga and Pugh, 2004). The Kim et al. study also reported that Pol II was bound to 97% of confirmed TFIID binding sites by performing ChIP-chip with an antibody that recognizes Pol II (antibody: 8WG16). This represented the most comprehensive Pol II ChIP data set at the time we began GRO-seq development, which is why we chose the IMR90 cell line. The 8WG16 antibody preferentially recognizes the hypophosphorylated form of the largest subunit of Pol II that is found at the 5'ends of genes. It has been demonstrated at many genes that as Pol II progresses further into a gene, the CTD of RPB1 becomes hyperphosphorylated, and thus a less suitable substrate for the antibody. Thus, in some cases the antibody will show a reduction in the signal the downstream portions of a gene, that actually reflects a reduced affinity for Pol II in these regions. Therefore, we cannot directly compare GRO-seq density and ChIP density in the downstream region of most genes, since GRO-seq detects transcriptionally engaged Pol II regardless of the phosphorylation state. In addition, the array used to analyze the Pol II ChIP data was essentially a promoter array, so there

ChrX: 45,475,000-45,530,000bp

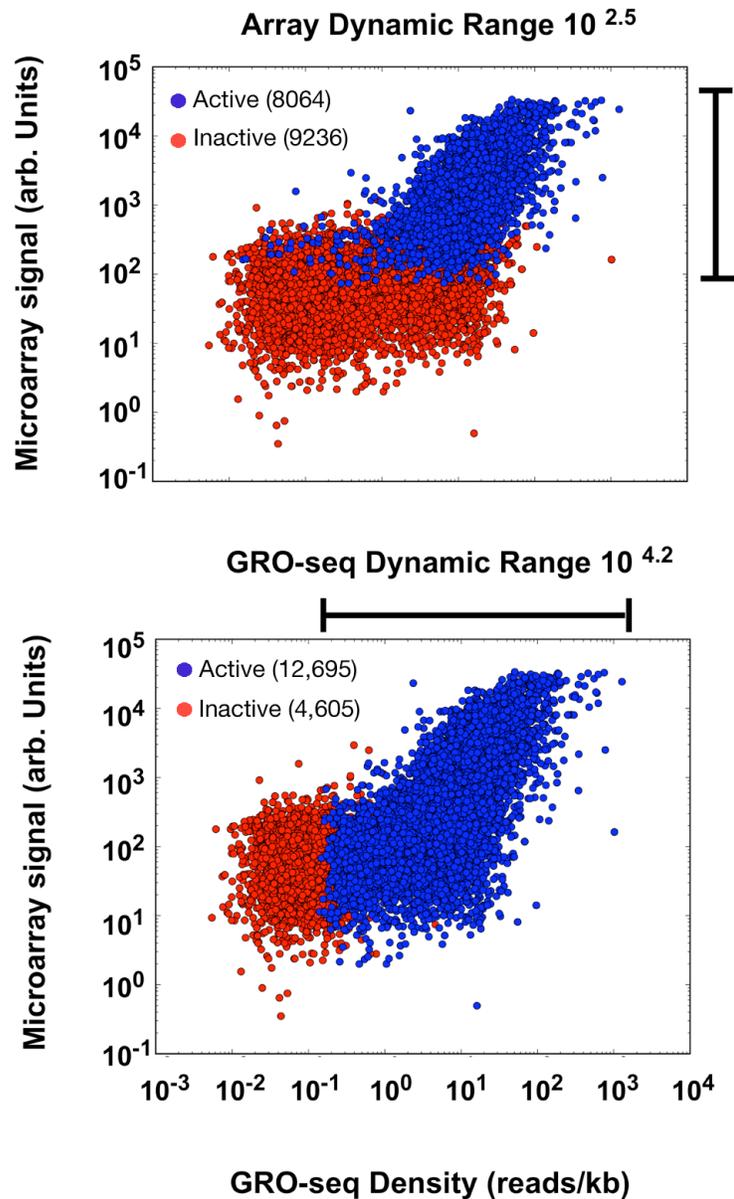


**Figure 3.3. Identification of a novel promoter by ChIP and GRO-seq.** A novel transcription unit on chrX: 45,475,000- 45,530,000bp is shown that is not annotated by any of the major databases or gene prediction tools. The promoter was identified as putative by Pol II ChIP shown in green.

is no data in the downstream portion of longer genes. The above reasons explain why, in some of the figures presented herein, Pol II ChIP signal appears concentrated only at the promoter regions, when this in fact is a result of the antibody used and the extent of the array design.

### **3.2.2 Comparison of GRO-seq to microarray expression data**

We additionally determined how GRO-seq transcript densities in the sense orientation within gene regions compared to the microarray expression data available for this cell line (Kim et al., 2005b). First, microarray expression values plotted against GRO-seq densities reveal that accumulated, fully processed mRNA levels generally correlate with steady state transcription of genes obtained by GRO-seq (Figure 3.4). However, GRO-seq densities have a wider dynamic range that extends below the limit of detection by microarray. To gauge the increase in sensitivity, we compared genes called absent or present by microarray to genes that could be called active or inactive by GRO-seq. For a gene to be called active by GRO-seq, we required the density within the downstream portions of genes to be significantly above background ( $P < 0.01$ ). The first 1 kb was excluded from the analysis to avoid signals produced by promoter-proximal paused polymerases. When considering all RefSeq genes, 16,882 genes (68%) were classified as active by GRO-seq. When considering the genes covered by probes on the microarray, 16,858 genes were called active by GRO-seq, while only 8,438 were called active by microarray hybridization (Figure 3.4, Table 3.1). Active gene calls for GRO-seq span more than four orders of magnitude, whereas microarray experiments are restricted to approximately 2.5 orders of magnitude. The increased number of active genes in our GRO-seq analysis can be attributed



**Figure 3.4. GRO-seq activity versus expression microarray.** Scatter plots of gene expression levels by microarray versus GRO-seq. Inactive genes are colored in red and active genes are colored in blue. Only the 17,300 genes that were unambiguous by both methods are shown in the plots. The range for which genes can be called significantly active is shown by the brackets. The number of active genes and inactive genes called by (A) microarray, and (B) GRO-seq are inset in each panel.

**Table 3.1. Gene activity calls by GRO-seq vs. Microarray.** GRO-seq gene categories (left column),

compared with Microarray categories (Top row). Totals for each GRO-seq category are shown in the right most column; Microarray category totals in bottom row. Genes that are less than 3kb cannot be analyzed by GRO-seq. Genes that are ambiguous do not fall in the same category between two array experiments.

	Expression Microarray				Total
	Present	Absent	Ambiguous	Not on array	
Active	7983	4712	4163	24	16882
Inactive	81	4524	1101	6	5712
Less than 3kb	374	1159	683	3	2219
Total	8438	10395	5947	33	24813

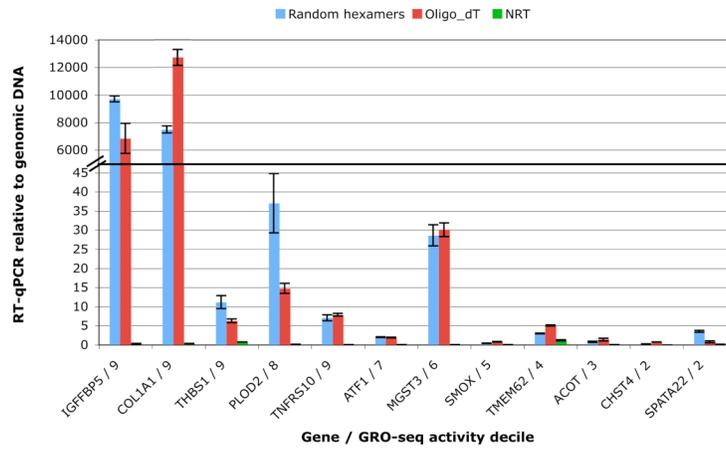
to the increased sensitivity of sequencing technologies versus hybridization methodologies (Wilhelm et al., 2008; Wold and Myers, 2008), and possibly due to the fact that nascent RNA libraries may be enriched for rare or unstable transcripts relative to highly accumulated RNAs.

### **3.2.3 Validation of GRO-seq gene activity by RT-qPCR**

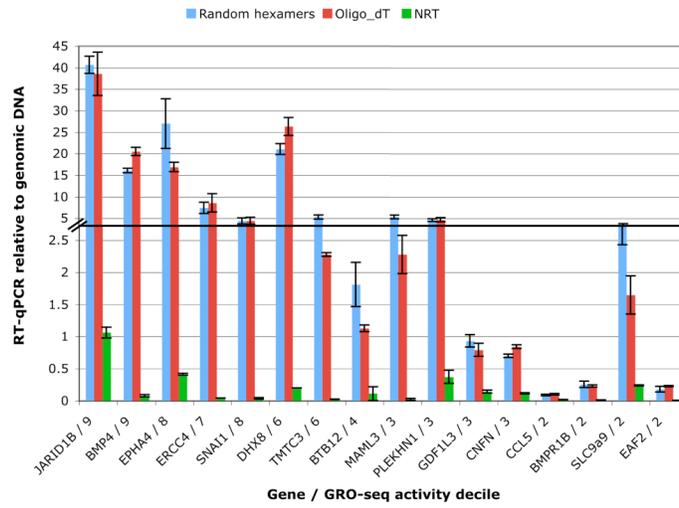
Transcripts that are regulated by post-transcriptional mRNA turnover can be identified by comparing mRNA levels to GRO-seq densities. A highly stable transcript would be expected to have a high level of mRNA expression compared to the GRO-seq density within the corresponding gene, while unstable transcripts would be expected to have higher GRO-seq densities relative to mRNA expression level. By comparing GRO-seq with expression microarray data, I identified candidates as stable or unstable transcripts by searching for genes that were microarray active : GRO-seq inactive or microarray inactive : GRO-seq active, respectively. I then compared several of these genes to genes that were found to be active in both assays by performing RT-qPCR. I first ranked the genes from each class into deciles of gene activity as determined from the GRO-seq density within gene bodies. Genes were then chosen from a range of activity deciles to validate. The results show that all genes tested that are called active by GRO-seq can be detected by RT-qPCR after priming the reverse transcription with either random hexamers or oligo-dT to extents that generally mirror their level of GRO-seq transcription (Figure 3.5). In addition, genes that were not detected by the microarray had similar RT-qPCR levels as those that were not detected by the arrays. These results verify GRO-seq as a general and sensitive method for detecting active genes, and suggest that many genes are not detected by the microarray due to insufficient sensitivity or incorrect probe design. Two genes (COL1A1, IGFBP5) may be highly stabilized transcripts

**Figure 3.5. RT-qPCR validation of GRO-seq levels.** Genes that were active by microarray and GROseq (A), inactive by microarray – active by GRO-seq (B), and active by microarray but not by GRO-seq (C) were analyzed by RT-qPCR. Reverse transcription was performed with random primers (blue), or oligo-dT (red), and compared to a known amount of genomic DNA. No reverse transcription reactions (NRT) (green). Error bars represent standard error of the mean, n=3. Note the breaks in the y axis in (A) and (B).

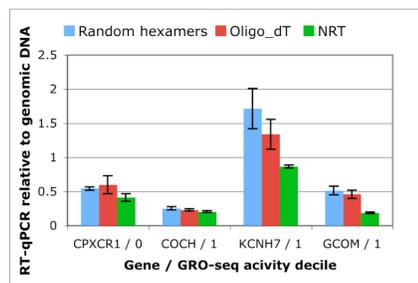
A



B



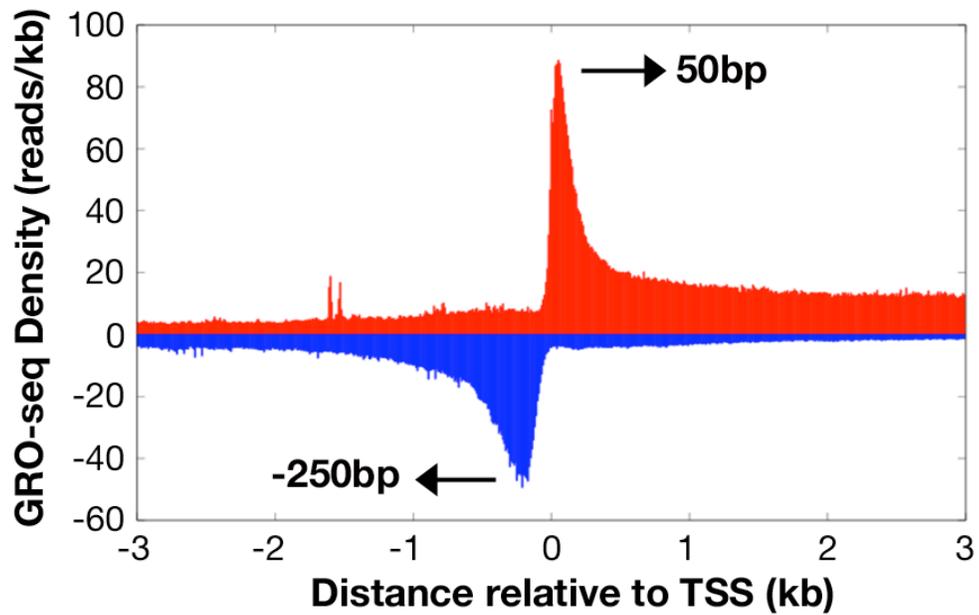
C



because they are called active by both microarray and GRO-seq, but were detected by microarray at much higher levels than other genes that are inactive by microarray but have similar GRO-seq densities. Accumulated mRNA levels and GRO-seq density on the body of genes, generally showed a strong concordance in IMR90 cells (Figures 3.4, 3.5). The relatively limited dynamic range and sensitivity of the microarray data may have caused some less stable RNAs to be missed. Also, classes of genes that are regulated by mRNA stability might be more readily detectable in response to changing environments (Garcia-Martinez et al., 2004; Schuhmacher et al., 2001). Comparison of GRO-seq to RNA-seq data should also improve the efficiency of identifying mRNAs that are regulated by mRNA turnover rates (Mortazavi et al., 2008; Nagalakshmi et al., 2008; Wilhelm et al., 2008; Wold and Myers, 2008).

#### **3.2.4 Generality Promoter-proximal pausing revealed by GRO-seq.**

We next aligned the GRO-seq hits relative to RefSeq TSSs and found that the highest density of hits clustered around TSSs in both the sense and antisense directions (Figure 3.6). The sense strand distribution peaks at ~50 bp downstream of TSSs, which mirrors that of recent global Pol II ChIP assays. To identify all genes that show a peak of engaged Pol II that is characteristic of promoter-proximal pausing, we assessed whether each gene showed significant enrichment of read density in the promoter-proximal region relative to the density in the body of each gene. The ratio of these densities is called the pausing index (Muse et al., 2007; Zeitlinger et al., 2007a) and significant pausing indices range from 2 to  $10^3$  (Figure 3.7). 7,522 genes were identified as having a significant enrichment of GRO-seq transcript hits within

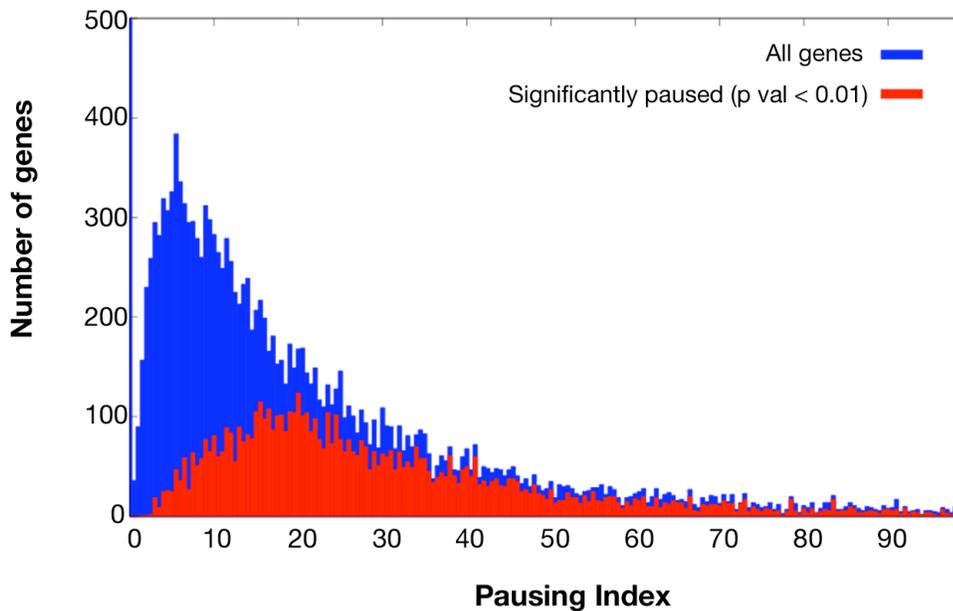


**Figure 3.6. Alignment of GRO-seq hits to TSSs.** GRO-seq reads aligned to Ref-seq TSSs in 10bp windows in both the sense (red) and antisense (blue) directions relative to the direction of gene transcription.

the defined promoter region relative to the body of genes ( $p$ -value  $< 0.01$ ), representing 28.3% of total genes (41.7% of active genes). Comparison of paused gene calls by GRO-seq profiles to microarray expression data revealed four main classes of genes: I) not paused and expressed, II) paused and expressed, III) paused and not expressed, and IV) inactive (not paused and not expressed) (Figure 3.8). Class III was severely depleted when we used GRO-seq to classify gene activity, likely due to a matter of sensitivity, since the few genes left within this class have very low signal at their promoters. Therefore, the overwhelming majority of genes with a paused polymerase also produce significant transcription throughout the gene, albeit often to levels not detectable by expression microarrays. A recent comparison of Pol II ChIP-seq data to RNA-seq also supports the view that virtually all genes that are bound by Pol II produce full length transcripts (Sultan et al., 2008).

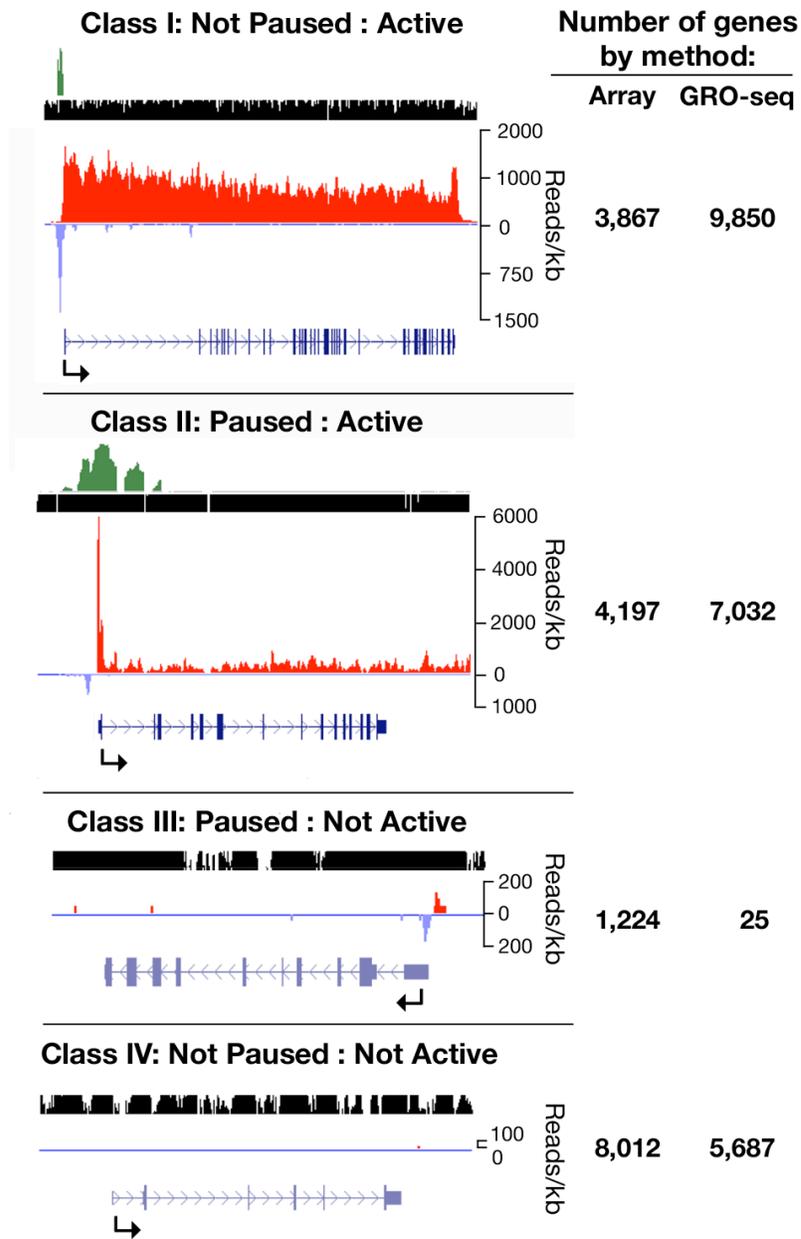
### **3.2.5 Relationship of pausing with gene activity**

To further investigate the relationship between pausing and gene activity, we compared the number of hits contained within the promoter-proximal region of all genes with the lowest, middle and highest deciles of gene activity. The result of this analysis shows that the density of polymerases within the promoter-proximal region generally correlates with the level of gene activity when all genes (Figure 3.9 A), or only genes with a paused polymerase are considered (Figure 3.9 C). However, pausing indices have an inverse correlation with gene activity (Figure 3.9 B, D). This relationship could reflect that highly expressed genes either do not experience pausing, or they transition through pausing faster, allowing more polymerase



**Figure 3.7. Histogram of pausing indices.** Pausing indices for all genes (blue) or significantly paused genes (red) were binned in windows of width 0.5 from 0 to 100. There are 3,907 genes with a pausing index less than 0.5. The smallest pausing index amongst the significantly paused genes is 1.65 and the largest pausing index is 8661.2 in both distributions.

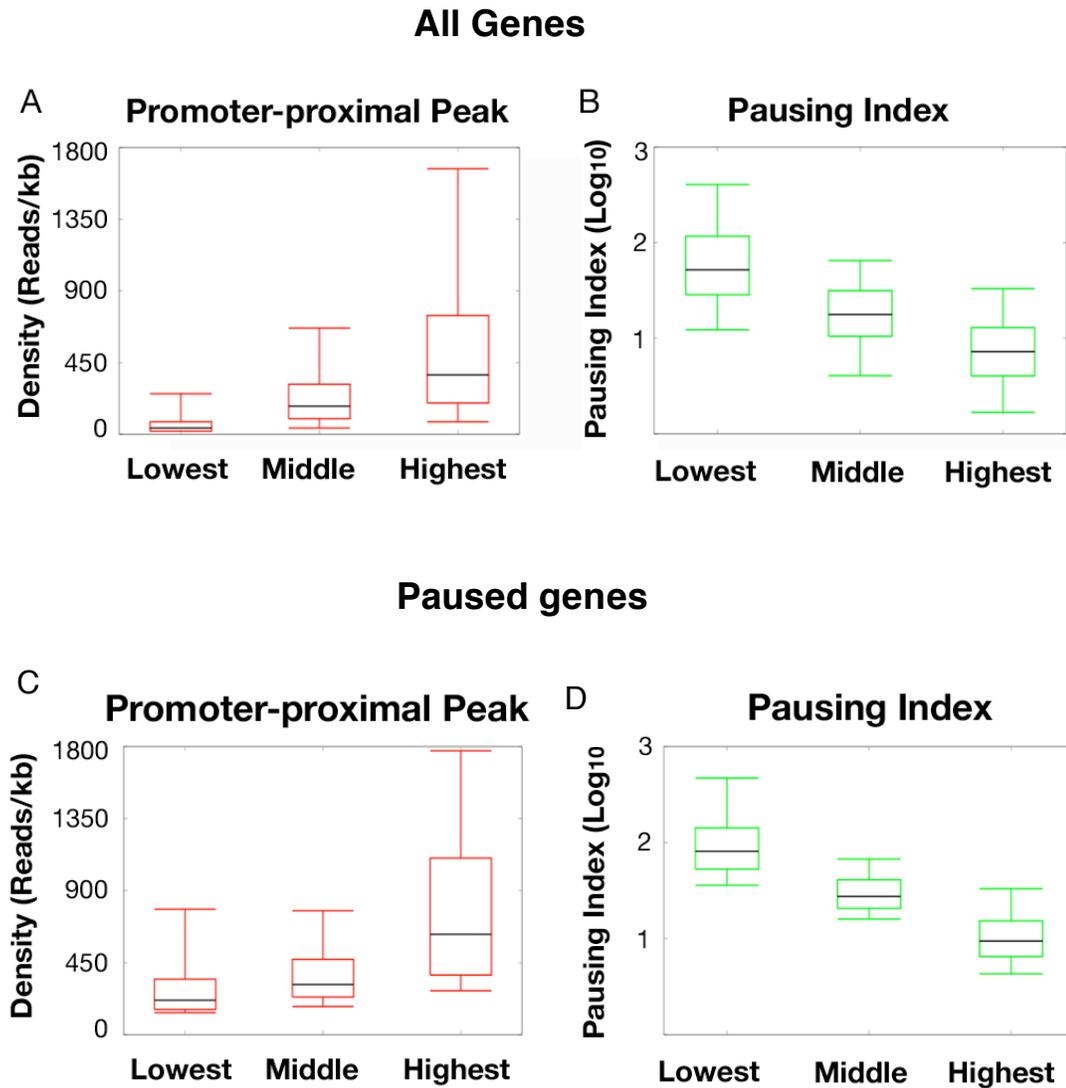
**Figure 3.8. Comparison of pausing with gene activity.** Four classes of genes are found when comparing genes with a paused polymerase and transcription activity either by microarray or GRO-seq density in the downstream portions of genes. An example of each class is shown, with tracks shown in the UCSC genome browser as in figure 1. The gene name, pausing index, and P value, from top to bottom respectively are as follows: TRIO, 1.1, 0.62; FUS, 41,  $2.8 \times 10^{-43}$ ; IZUMO1, 410,  $7.6 \times 10^{-3}$ ; and GALP, -1, 1. The number of genes represented in each class is shown to the right.



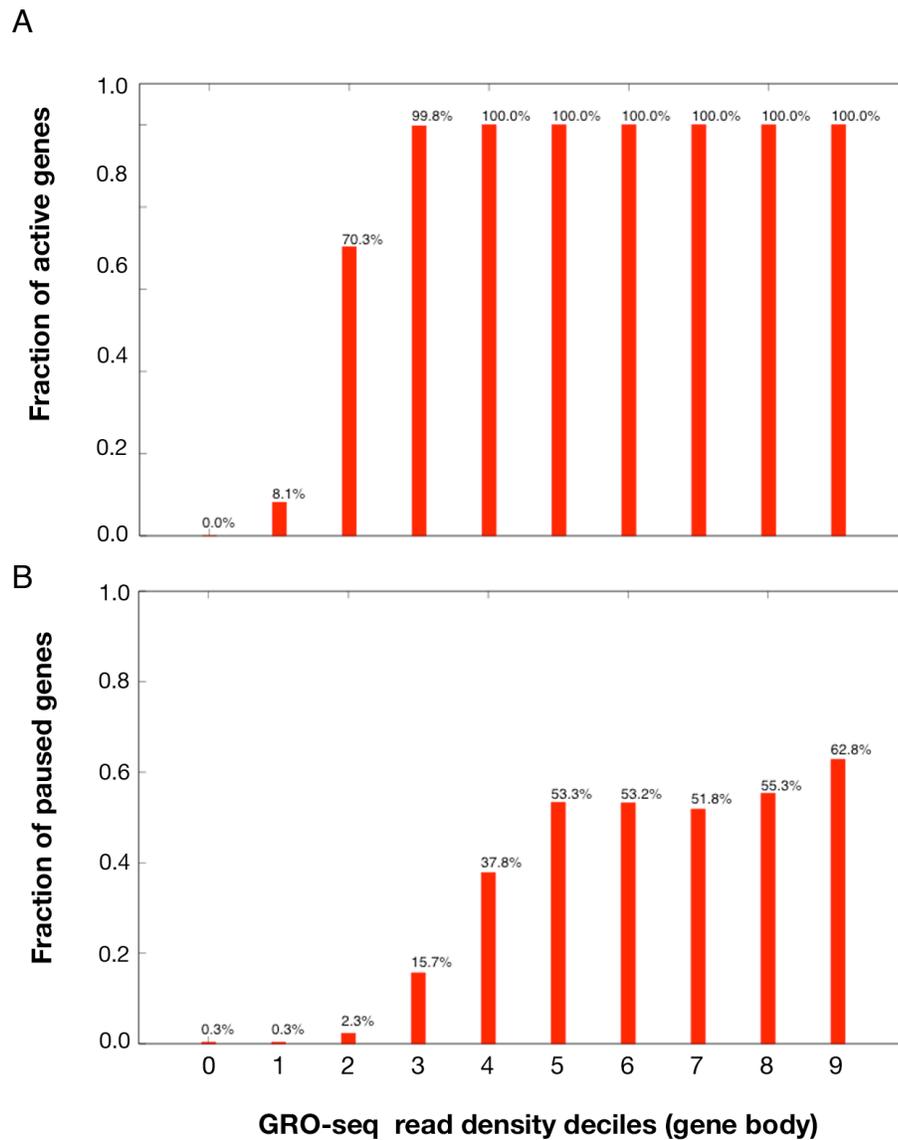
to enter into productive elongation. When we examine the fraction of paused genes according to gene activity deciles (Figure 3.10), we find that the fraction of paused genes increases with increasing gene activity and represent 63% of the highest decile of gene transcription. This result, in combination with the inverse correlation between gene body density and pausing indexes, indicates that highly active genes, relative to genes with lower activity, not only recruit more polymerase and stimulate faster pause site entry rates, but they must also increase pause site escape to a greater extent in order to account for these profiles.

### **3.2.6 Gene Ontology of paused genes**

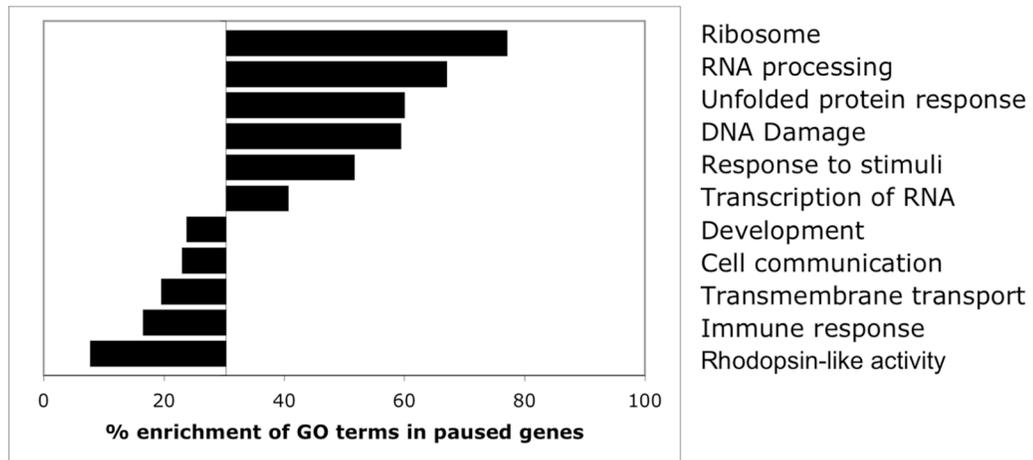
Gene ontology (GO) analysis of significantly paused genes reveals enrichment of biological processes such as cell cycle regulation, and stress response, and molecular function categories such as zinc-finger DNA binding proteins, and ribosomal proteins (Figure 3.11). Although previous studies identified developmentally regulated genes as enriched in the paused class (Guenther et al., 2007; Muse et al., 2007; Zeitlinger et al., 2007a), these studies used either embryonic stem cells, an embryonic-derived cell line, or developmentally staged *Drosophila* embryos. The fact that we do not see an enrichment of developmentally regulated genes in the paused class may reflect the more differentiated state of the primary fibroblasts used in this study.



**Figure 3.9. Correlation of promoter-proximal transcription patterns with gene activity.** (A-D) Box plots (each showing the 5<sup>th</sup>, 25<sup>th</sup>, 50<sup>th</sup>, 75<sup>th</sup>, and 95<sup>th</sup> percentiles) that show the relationship of Promoter-proximal (PP) sense peaks (red) and Pausing indices (green) with the top, middle, and bottom deciles of gene activity. The plots in A and B represent all genes; C and D represent paused genes. All deciles are significantly different from each other ( $P < 10^{-9}$ )



**Figure 3.10. Fraction of paused genes and active genes by gene activity decile.** The percentage of significantly active (A) and significantly paused (B) genes in each decile of gene activity. See methods for calculation of gene activity levels and the criteria for significant pausing and significant gene activity.



**Figure 3.11. Gene ontology of paused genes.** Bar plot show the summary of enriched and de-enriched gene ontology (GO) terms of significantly paused genes. The Y-axis is set to 28.3%. GO terms that are enriched in paused genes are to the right of the axis, and GO that are de-enriched are to the left. All terms are significant ( $P < 10^{-10}$ ). GO analysis was performed with GStat.

### 3.2.7 GRO-seq results for known paused genes

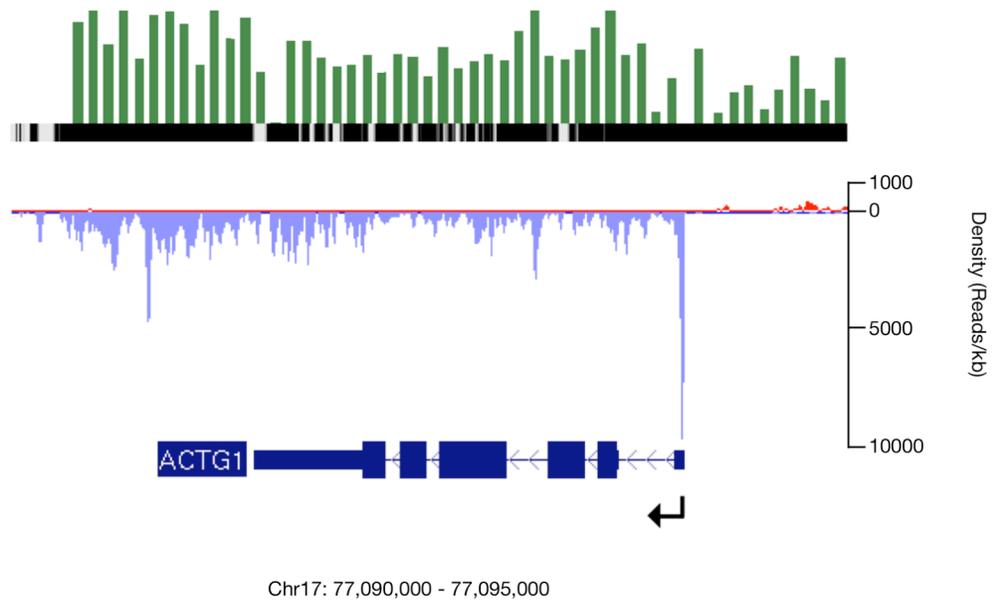
Several human genes have been shown to have a high level of transcriptionally engaged Pol II at the 5'-end relative to the downstream portions either by traditional NRO-hybridization assays, or by potassium permanganate footprinting. The genes include ACTG1 ( $\gamma$ -Actin) (Cheng and Sharp, 2003), FOS (Fivaz et al., 2000), DHFR (Cheng and Sharp, 2003), MYC (Krumm et al., 1992; Strobl and Eick, 1992), and HSPA1A (HSP70) (Brown et al., 1996). The first four genes do exhibit a pattern consistent with pausing (Figure 3.12), and are called significantly paused by our analysis. The MYC gene displays a broad peak over the first exon, consistent with there being multiple TSSs within this region (Krumm et al., 1992). Interestingly, there is another sharp peak ~2.5kb upstream of the annotated gene that may represent another promoter that has not been annotated. The human genome has two nearly identical copies of the HSP70 gene, and was analyzed because reads mapping to multiple locations were removed before any analysis performed. If the sequence reads mapping to the two copies are averaged, the HSP70 gene does have a paused Pol II (PI: 94 P:  $2 \times 10^{33}$ ).

### 3.2.8 Divergent transcription is associated with active promoters

A prominent and surprising feature of the GRO-seq profiles around transcription start sites is the robust signal from an upstream, divergent, engaged polymerase. RNAs generated by these divergent polymerases can be identified at low levels when small RNAs are isolated from whole cells (Seila and et al., in press). These divergent polymerases cannot be accounted for by the 10% of known bidirectional promoters that are less than 1kb apart (Trinklein et al., 2004) (Figure 3.13). 13,633 genes (55% of all genes, 77% of

**Figure 3.12. GRO-seq profiles for known paused genes.** Snapshots from the UCSC genome browser showing the regions around genes previously characterized as paused. Gene names, pausing indexes, and associated P values are as follows: (A) ACTG1, 6.3,  $8 \times 10^{-30}$ ; (B) FOS, 43,  $1.7 \times 10^{-4}$ ; (C) DHFR, 25,  $7.8 \times 10^{-4}$ ; and (D) MYC, 5.7,  $3.2 \times 10^{-3}$ . Pol II ChIP results are shown in green and the start site and direction of transcription of the gene is shown by the arrow (black). Y-axis (Reads/kb) is shown to delineate the scale between the images.

A



B

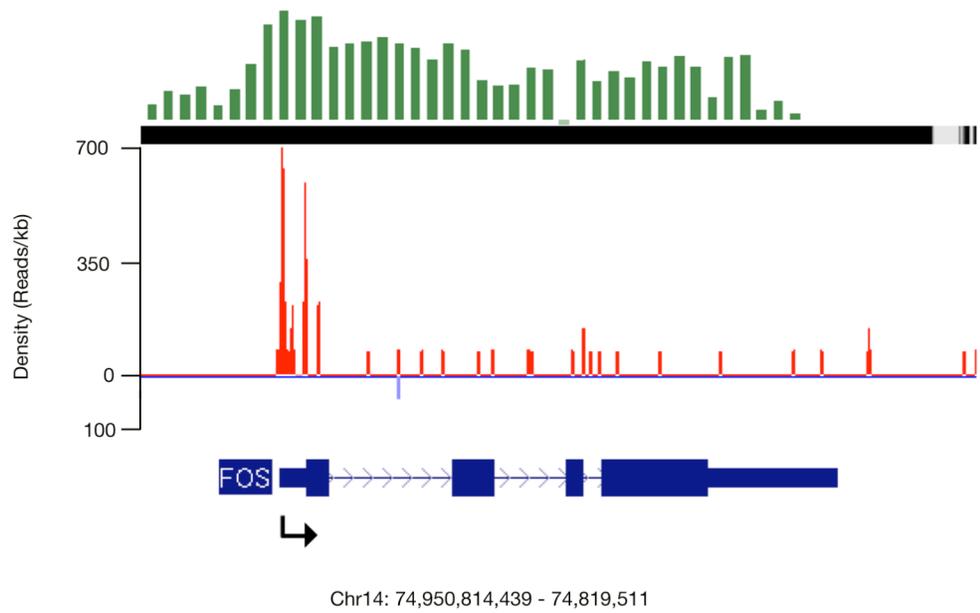
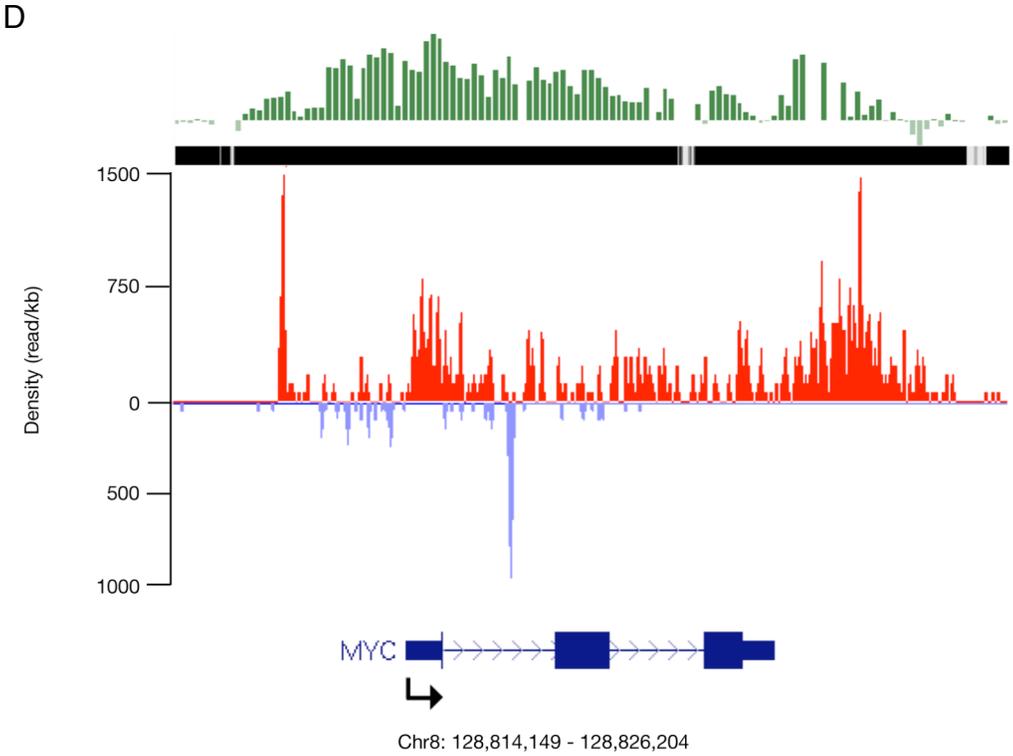
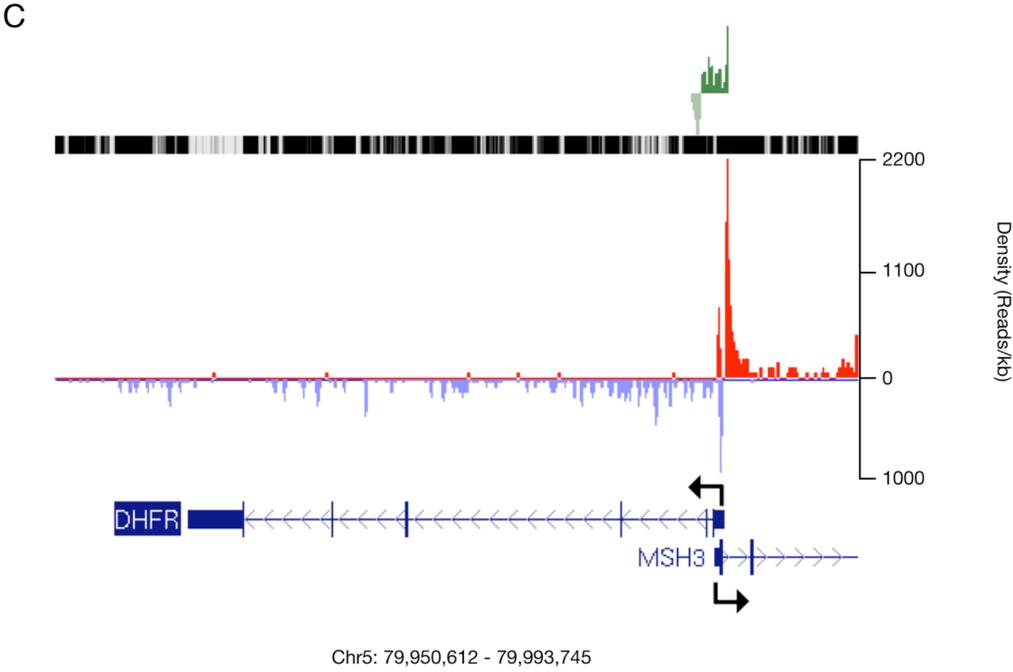
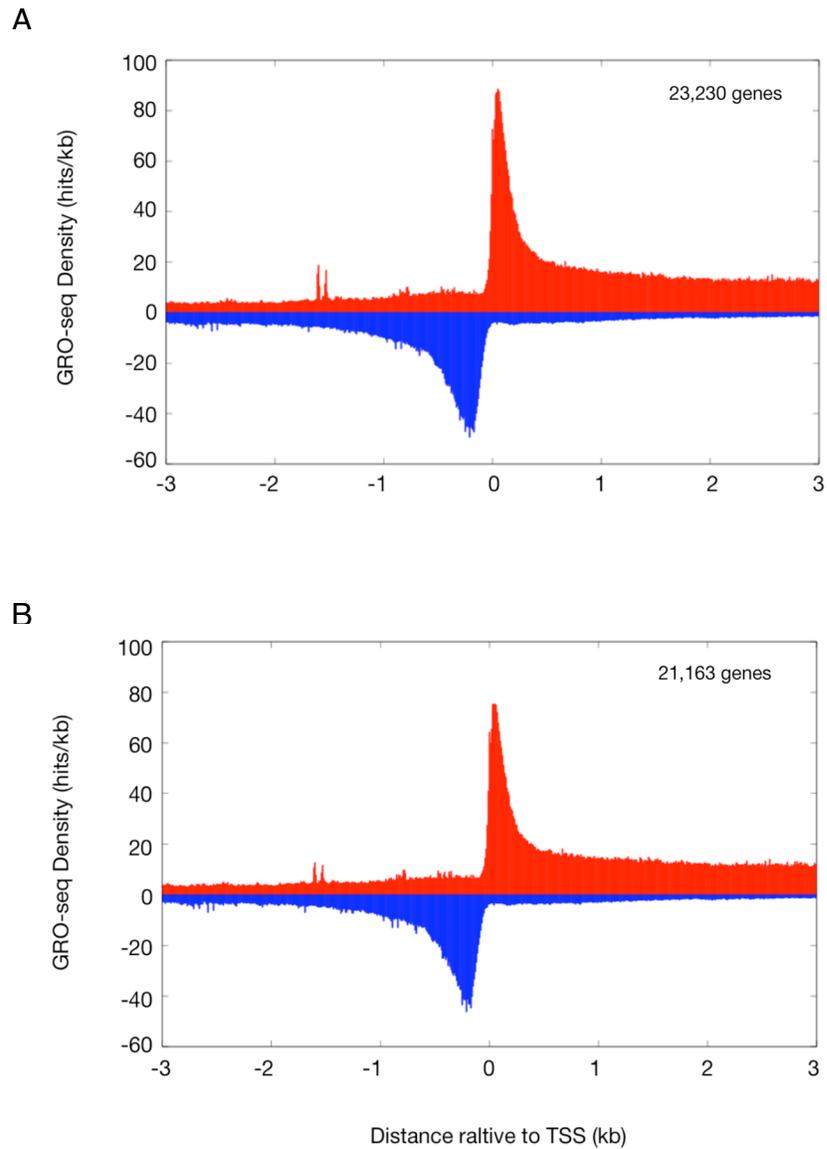


Figure 3.12 (Continued)





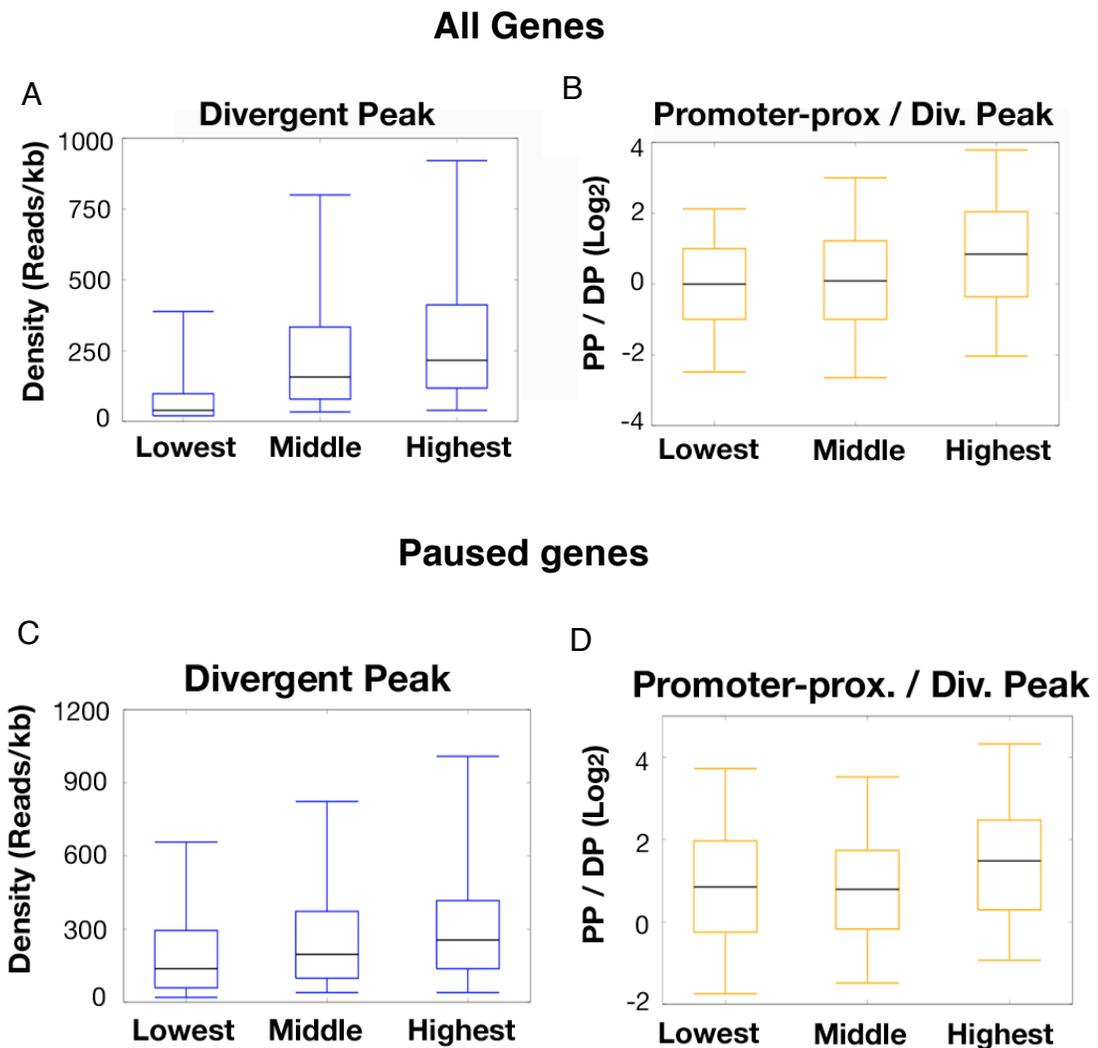
**Figure 3.13. GRO-seq aligned to TSS without bidirectional promoters.**

(A) Comparison of composite GRO-seq profiles aligned to TSSs of all RefSeq genes, or (B) RefSeq genes minus the annotated bidirectional promoters (genes arranged head-to-head and within 1kb).

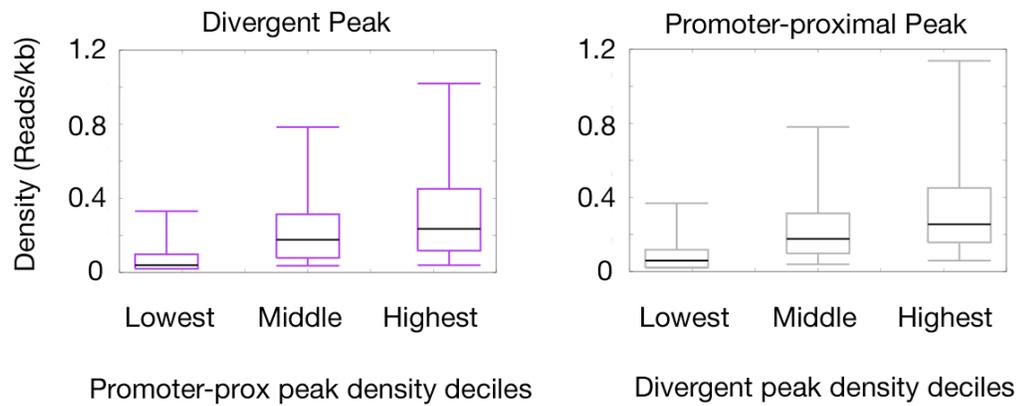
active genes) display significant divergent transcription within 1 kb upstream of sense-oriented promoter-proximal peaks ( $P < 0.001$ ), indicating that the number of bidirectional promoters exceeds even the highest estimates (Kapranov et al., 2007a; Rada-Iglesias et al., 2008). However, since it appears that the majority of these promoters produce mRNAs in only one direction (see below), we refer to this new class of promoters as divergent. Although the top 10% of active genes have, on average, a slightly larger promoter-proximal than divergent peak (Figure 3.14B,D), levels of divergent transcription generally correlate with both the promoter-proximal signal (Figure 3.15) and the transcription level of the associated gene (Figure 3.14A,C). Thus, divergent transcription is a mark of most active promoters.

Gene activity, pausing, and divergent transcription correlate with each other and with promoters containing a CpG island. These four characteristics co-occur significantly more often than would be expected by chance ( $P < 10^{-52}$ ) (Table 3.2). Previous mapping of capped mRNA transcripts has shown that at CpG island promoters initiation occurs broadly over hundreds of base pairs (Carninci et al., 2006), and GRO-seq now shows that polymerases initiate and accumulate on this large class of promoters in both orientations.

Given the ubiquity and prominence of this divergent polymerase peak in our GRO-seq data, we asked whether there was any sign of it in the ChIP-chip data (Kim et al., 2005b). Manual inspection of a number of genes and comparison with composite profiles aligned to TSSs shows that the Pol II ChIP peak at promoters is clearly accounted for by the two divergent peaks uncovered by GRO-seq (Figure 2.16A). Higher resolution ChIP-seq data in different cell lines has identified Pol II upstream of promoters that were



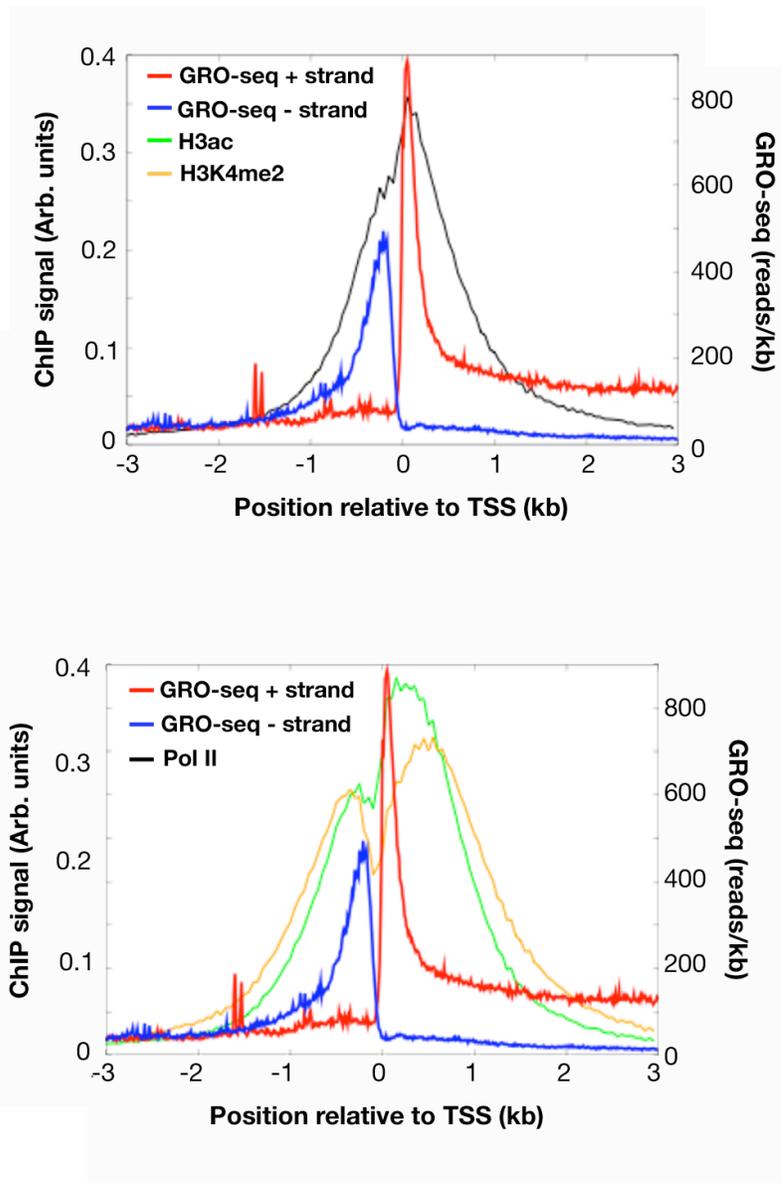
**Figure 3.14. Correlation of promoter-proximal transcription patterns with gene activity.** Box plots displayed as in Figure 3.9 that show the relationship of divergent peaks (DP) (blue) and PP/DP ratios (orange) to the bottom, middle and highest deciles of gene activity. The plots in A and B represent all genes; C and D represent paused genes. All deciles are significantly different from each other:  $P < 10^{-9}$  for all comparisons except between the lowest and middle deciles in D ( $P < 10^{-3}$ ).



**Figure 3.15. Correlation between divergent read density and promoter-proximal read density.** Reciprocal box plots that show the relationship of (A) divergent peak height to the lowest, middle, and highest deciles of Promoter-proximal (PP) sense peak heights (violet), and (B) Promoter-proximal (PP) sense peak heights with the lowest, middle and highest divergent peaks (DP) heights (gray). All deciles are significantly different from each other:  $p < 10^{-9}$  for all comparisons.

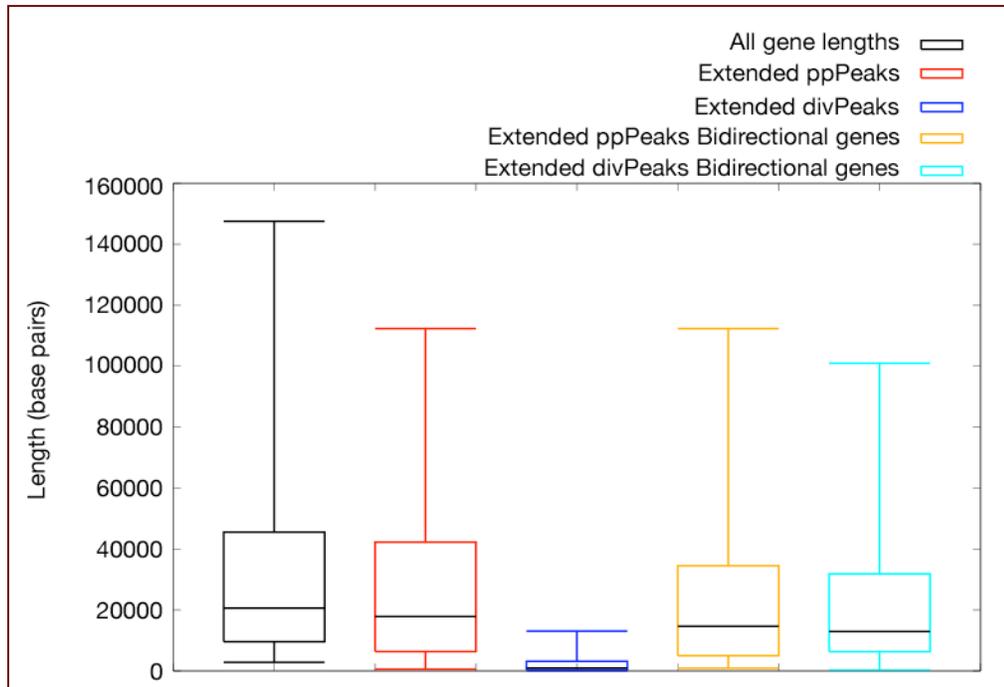
**Table 3.2. Pairwise correlations between Gene Activity, Pausing, Divergent transcription, and CpG island promoters.** Four qualities of individual genes were found to significantly cooccur by pairwise tests. The four qualities were significant levels of gene activity, significant levels of pausing, a significant peak of divergent transcription, and having a CpG island-type promoter. The criteria for gene activity, pausing, and divergent transcription are described in the methods. To define whether a given promoter had a CpG island the CpG Islands track was downloaded from the UCSC Genome Browser. If there was an annotated CpG island within 1 kb of a given TSS, the gene was classified as having a CpG island-type promoter. The percentages listed in the Table are the fraction of genes from the category on the left that are also in the category on the top.

	Active	Paused	Divergent	CpG island
Active	16,882 (100%)	7,032 (99.6%)	13,087 (96.0%)	13,773 (85.2%)
Paused	7,032 (41.7%)	7,057 (100%)	6,614 (48.5%)	6,304 (39.1%)
Divergent	13,087 (77.5%)	6,614 (93.7%)	13,633 (100%)	12,053 (74.8%)
CpG island	13,773 (81.6%)	6,304 (89.3%)	12,053 (88.4%)	16,118 (100%)

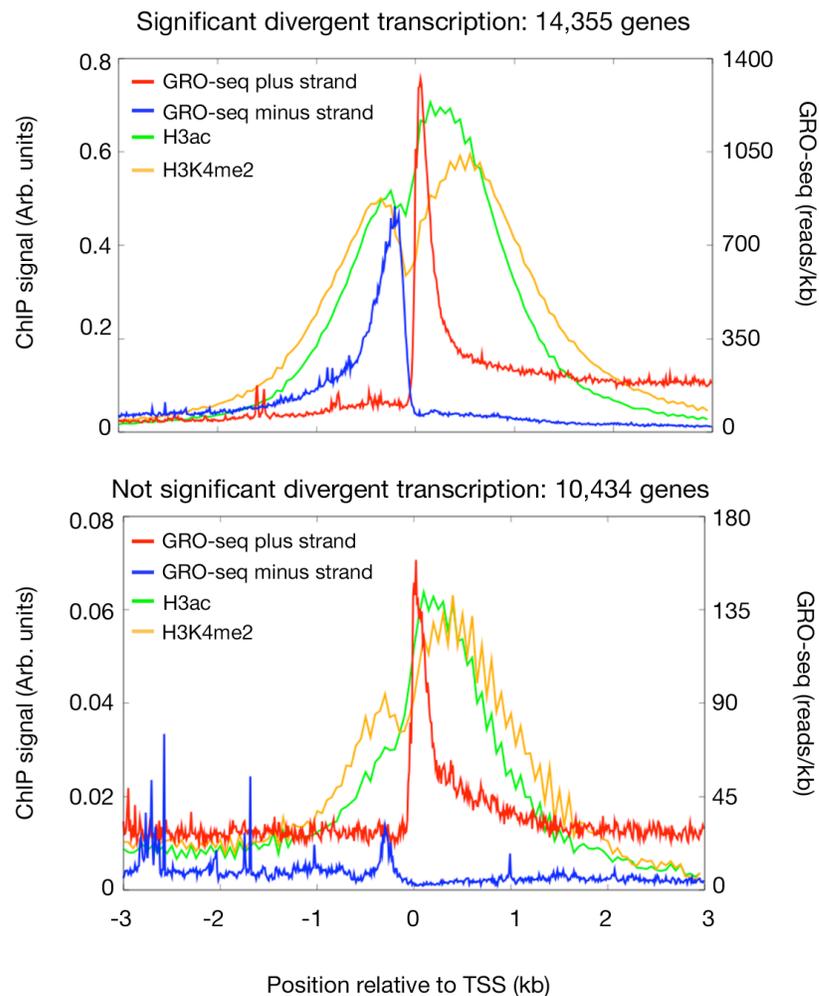


**Figure 3.16. Comparison of GRO-seq profiles with ChIP-chip data in IMR90 cells.** Plotted are GRO-seq reads on both DNA strands with (E) ChIP profiles of Pol II and (F) ChIP profiles of H3ac and H3K4me2 and GRO-seq. Data are aligned with respect to the transcription start site for all RefSeq genes.

proposed to be in the same orientation of the annotated gene, but are likely representative of the divergent promoters identified by GRO-seq (Sultan et al., 2008). Additionally, active promoters are typically marked by histone modifications such as di- and tri-methylation of H3-Lysine 4 (H3K4me2, H3K4me3) as well as acetylation of histone H3 and H4 (H3ac, H4ac). These modifications show a bimodal distribution around TSSs, with the trough representing a nucleosome free region encompassing the TSS (Barski et al., 2007; Guenther et al., 2007; Kim et al., 2005b). Comparison of available H3ac and H3K4me2 data in this cell line (Kim et al., 2005b) with GRO-seq suggests that both the upstream and downstream peaks of these histone modifications are associated with active transcription, with each peak of histone modifications being adjacent and downstream of an engaged polymerase (Figure 3.16B). When promoter-proximal sense strand peaks and divergent peaks are extended to include contiguously transcribed regions, the sense strand peaks can be extended roughly the length of the full gene while the regions extended from the divergent peaks are on average 10 times shorter (Figure 3.17). Consistent with this, other studies have shown that histone modifications associated with transcription elongation (e.g. H3K36me3 and H3K79me3) do not associate in a bimodal fashion around TSSs (Barski et al., 2007; Guenther et al., 2007). To further examine the relationship of divergent promoters and histone modifications, we replotted the histone modification data, but this time removed genes that do not have significant levels of divergent transcription ( $P > 0.001$ ). As shown in Figure 3.18, the peak of H3K4me2 and H3ac at these genes is less defined in the upstream region compared to the region downstream of the TSS. The low levels of these modifications in the upstream region can likely be accounted for by the small



**Figure 3.17. Comparison of distance transcribed by forward vs. divergent polymerases.** In black is the distribution of all RefSeq gene lengths to provide a scale for the other distributions. In red and orange are the transcribed regions extended from promoter proximal peaks on the sense strand of genes while dark blue and cyan are transcribed regions extended from the divergent peaks. The genes used for the red and dark blue data sets do not include pairs of annotated genes oriented head-to-head with less than 1kb between TSSs. The orange and cyan distributions are from just those annotated gene pairs alone.

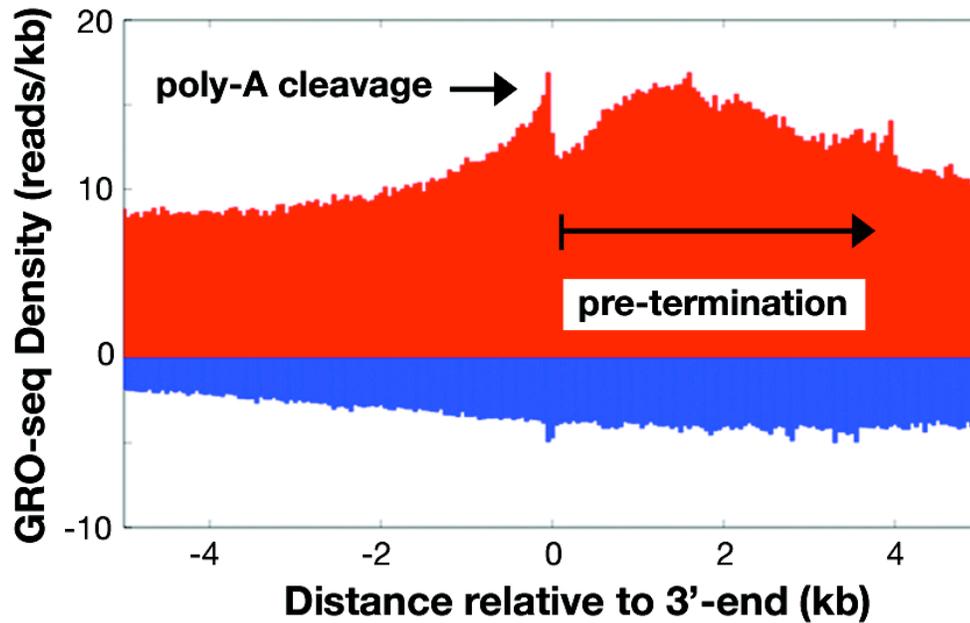


**Figure 3.18. Histone modifications at promoters with or without significant divergent transcription.** Genes were separated based on whether they had significant divergent transcription ( $P < 0.001$ ) (A) or not ( $P > 0.001$ ) (B). The profiles for histone modifications H3ac (green) and H3K4me2 (orange) were then plotted in arbitrary units against GRO-seq read density (reads/kb) for the plus strand (red) and minus strand (blue) reads. X-axis represents the distance (kb) relative to the TSS, which is set to zero.

but identifiable peak of anti-sense GRO-seq reads at  $\sim -250$ . These observations indicate that the majority of divergent promoters experience initiation in the upstream direction, but that these polymerases do not productively elongate transcripts. Also, the mechanism of placing these histone modifications at promoters appears to be tightly associated with the mechanism of forming divergent elongation complexes.

### **3.2.9 Transcription beyond the 3-end of genes**

Another interesting feature of transcription that is revealed by GRO-seq is the distance polymerase continues to elongate after the 3'-end, cleavage/polyadenylation site (Proudfoot, 1989). Alignment of GRO-seq reads to annotated 3'-ends of genes revealed a broad peak that is maximal at approximately +1.5kb and can extend greater than 10kb downstream of poly-A sites (Figure 3.19). This peak distance is consistent with previous and recent estimates (Lian et al., 2008; Proudfoot, 1989). Interestingly, a small peak followed by a sharp drop off is observed at the site of polyadenylation. This most likely represents the occurrence of cleavage prior to polyadenylation of the RNA, which defines the 3'-end of the pre-mRNA and is important for the eventual termination of transcription (Proudfoot, 2004). Cleavage of the pre-mRNA at this point would create a new 5'-end of the nascent RNA that is still being extended by Pol II at the time of nuclei isolation. These new 5'-ends are expected to be detected at higher levels relative to the random ends produced by base hydrolysis of continuous sequences. It is also possible that these peaks also represent sites of initiation at the 3'-end of genes (Carninci et al., 2005; Lian et al., 2008).



**Figure 3.19. Alignment of GRO-seq hits to TSSs and 3' ends.** GRO-seq reads flanking the 3'-ends of genes. The sharp peak coincides with the new 5'-end created after cleavage at the polyA site. Polymerase density extends considerably downstream prior to termination (often to ~10kb).

### **3.2.10 Antisense transcription in gene regions**

A number of studies have reported that gene regions are transcribed in the reverse orientation with unanticipated high frequency. Transcript pairs have been identified that overlap at the 5'-ends, 3'-ends, or with full overlap (Kapranov et al., 2007b; Katayama et al., 2005). Although antisense reads in gene regions account for only 6% of the total reads, ~14,545 genes (58.7%) have antisense transcription significantly above background ( $P < 0.01$ ). Of these genes, 273 are accounted for by active annotated genes that overlap at the 5'-end, 4,407 by active convergent genes with a maximum separation of 10kb, and 242 by active annotated genes with full overlap (Figure 3.20).

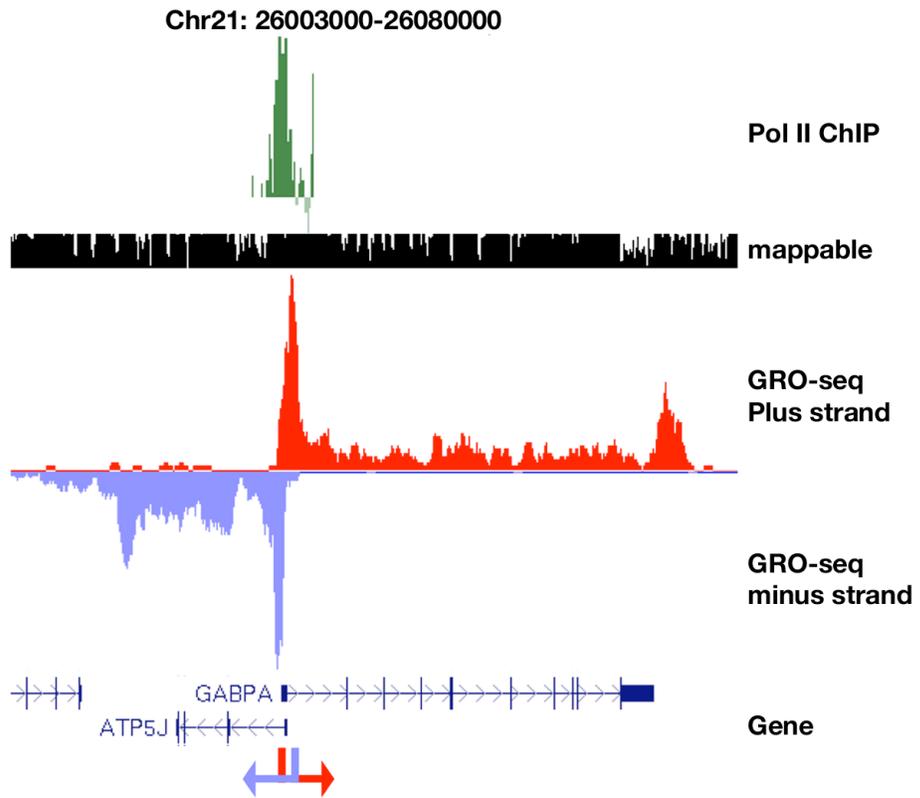
## **3.3 Discussion and perspectives**

### **3.3.2 Pausing, termination, or both?**

Whereas we have clarified our use of terminology here, we are, however, uncertain whether the engaged complexes that we detect at the 5'-end of genes will actually proceed to transcribe to the end of the associated gene given the proper signal. It is possible that some of these polymerases will eventually terminate prematurely in a manner that has been observed for transcription of HIV genes in the absence of the transactivator Tat (Kao et al., 1987; Marciniak and Sharp, 1991). For instance, the presence of promoter-proximal engaged polymerase peaks could also be observed if a promoter experienced high rates of initiation but also high rates of premature termination relative to the amount of polymerases that escape into productive elongation. Under these circumstances, one could expect to detect high levels of engaged polymerases immediately prior to the point of termination. Further

**Figure 3.20. Examples of antisense transcription in the genome.** Three representative loci that show three types of antisense transcription identified previously by others, and presently in this study. The number of occurrences of (A) 5'-overlapping, (B) 3'-overlapping (convergent), and (C) fully overlapping antisense transcription is 273, 4,407, and 242, respectively.

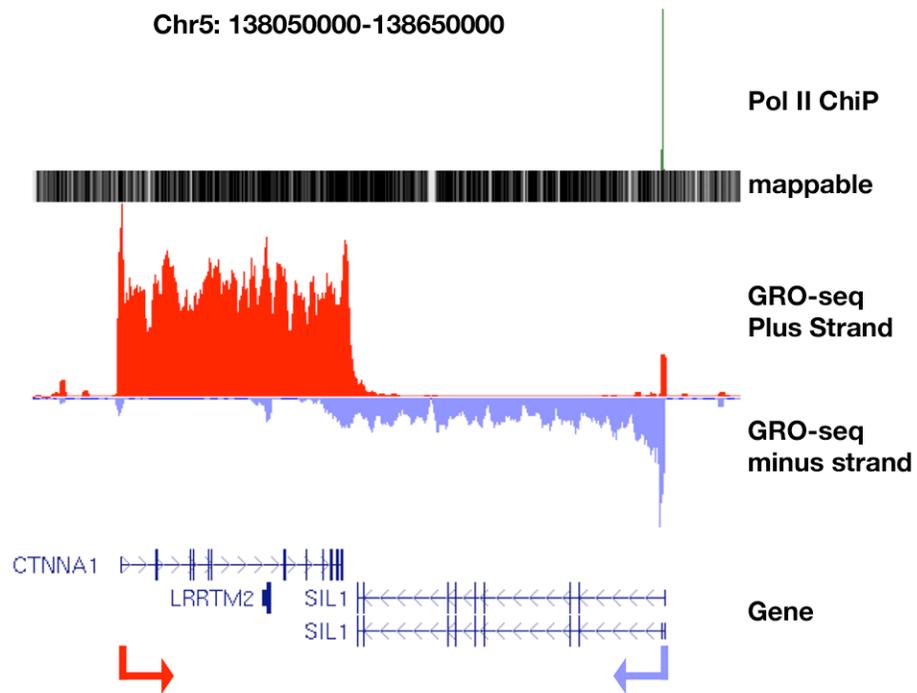
A



5' antisense overlap: 273 genes

Figure 3.20 (continued)

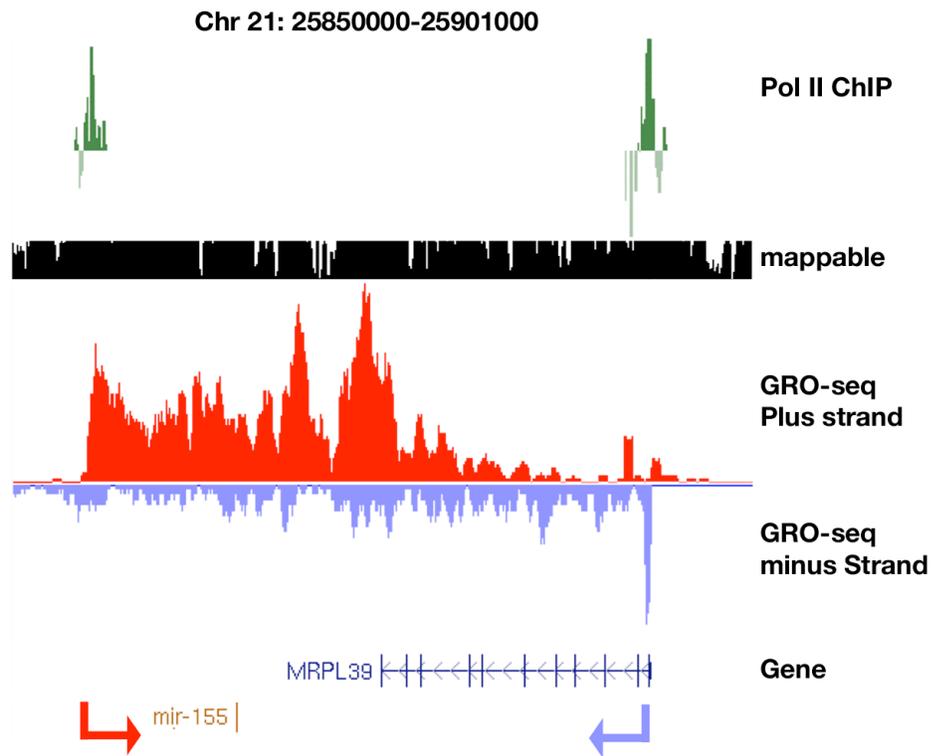
B



**3' antisense overlap: 4,407 genes**

Figure 3.20 (Continued)

C



Full antisense overlap: 242 genes

experimentation and development of new methodologies are required to distinguish between these possibilities *in vivo*. However, our results do show that the transition from initiation to elongation can be rate limiting to gene transcription, whether or not it occurs through holding back a polymerase and causing it to pause or by causing premature termination, or through a combination of pausing and termination.

### **3.3.3 Divergent transcription and histone modifications**

The histone modifications (H3K4me3, H3K4me2, H3/H4ac) that mark active promoters generally occur at the +1 and -1 nucleosomes relative to the transcription start site (TSS). The +1 nucleosome is downstream of the TSS, and is thus associated with initiation, but the modification of the upstream (-1) nucleosome is generally assumed to occur due to the simple proximity of the -1 nucleosome to the control sequences of the promoter. This would suggest that the mechanism by which these modifications are laid down has no strict directionality. Based on our GRO-seq results, and the ubiquity of divergent transcription, an alternative explanation could be that the -1 nucleosome has these modifications either as a consequence of, or perhaps, to allow formation of divergently-engaged polymerase.

### **3.3.4 Possible functions for divergent transcription**

We envision several possible functions for divergent transcription. First, the act of transcription itself could be crucial for granting access of transcription factors to control elements that reside upstream of core promoters, possibly by creating a barrier that prevents nucleosomes from obstructing transcription factor binding sites (Core and Lis, 2008; Gilchrist et

al., 2008). Second, as proposed by Seila et al., negative supercoiling produced in the wake of transcribing polymerases could facilitate initiation in these regions. Third, while the majority of divergent promoters we have found do not appear to produce mRNAs, the nascent RNAs could themselves be functional. A recent study has shown that overlapping promoter-associated transcription produced in low abundances by use of upstream TSSs can produce double-stranded RNA that regulate transcription through an Argonaute dependent pathway that directs machinery associated with repressing transcription to promoter (Han et al., 2007). Another recent study has shown that non coding RNAs produced upstream of the *CCND1* gene can allosterically regulate transcription factors at its promoter and thus regulate gene activity (Wang et al., 2008). Upcoming challenges will be to decipher whether the widespread transcriptional activity that lies upstream but divergent from the direction of coding genes positively or negatively regulates transcription output, and how promoter or unknown DNA elements are designed to distinguish between productive elongation in one direction versus the other.

### **3.4 Concluding remarks**

I have presented here a new methodology for documenting transcribed regions in the human genome by isolation and large-scale sequencing of nascent RNAs. GRO-seq is efficient, requiring only  $\sim 5 \times 10^6$  cells/library, and the resulting NRO-cDNA library is highly enriched relative to total RNA. I have shown that this technology can map RNA polymerase locations with precision, and that this allows the identification of active promoters and their directionality. The distribution of transcriptionally engaged polymerases around gene regions can identify interesting characteristics of promoters and

gene regions such as promoter-proximal pausing, internal pausing, co-transcriptional cleavage of the nascent RNA, the distance Pol II travels beyond annotated 3' ends before termination, and the level antisense transcription within genes. While the present analyses of the assay document many interesting features about how the human genome is transcribed, future analyses are still needed to identify transcription factor binding sites or sequence elements that contribute to pausing, divergent transcription and termination.

## CHAPTER 4

### FUTURE DIRECTIONS

#### **4.1 Further Adaptations of GRO-seq**

While we estimate that the GRO-seq method presented in this dissertation maps polymerase locations to within 40-50 bases, identification of potential pausing elements can benefit from nucleotide resolution mapping of the polymerase with respect to the transcription start site. I have outlined two methods below that will: 1) map the active site of polymerases with near nucleotide resolution, and 2) efficiently identify transcription start sites.

##### **4.1.1 Mapping engaged RNA polymerase with near nucleotide resolution.**

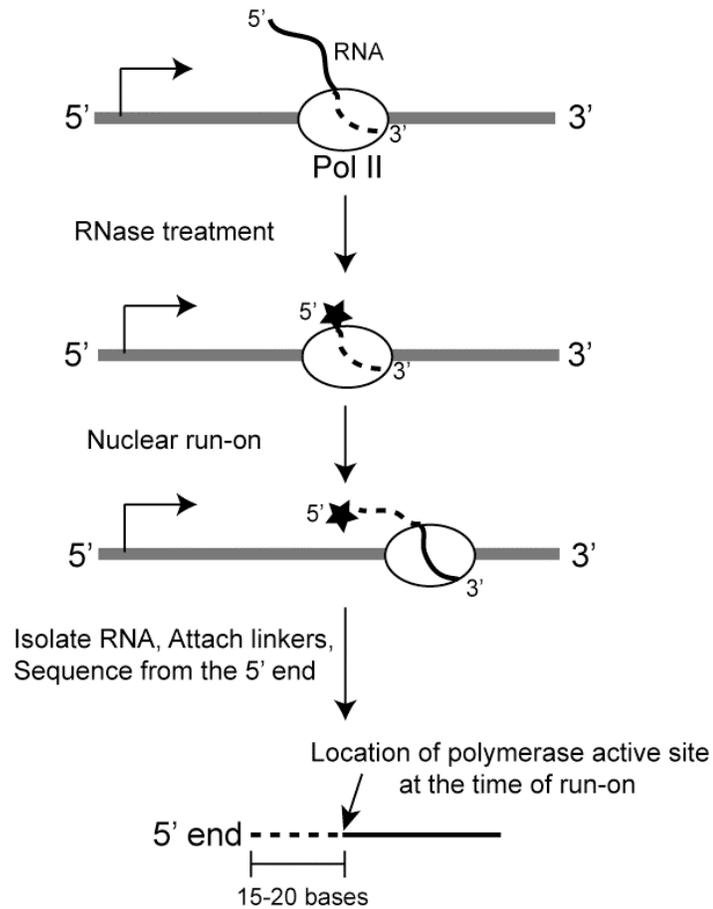
Ideally for this experiment, one would use a nucleotide analog that contains 1) an affinity tag does not inhibit incorporation of the nucleotide by Pol II, but should have a strong enough interaction with some substrate such that one incorporated tag allows efficient purification of the RNA, and 2) a blocked 3'-end that can be chemically or otherwise converted into a hydroxyl group so that an adapter can be ligated. Sequencing from the 3'-end of this 'reversibly -terminated' NRO will map the active site of polymerases with single nucleotide resolution. However, a suitable analog does not currently exist, so an alternative strategy is outlined below. Perhaps we should employ a chemist.

This alternative adaptation involves RNase treatment of isolated nuclei prior to the run-on step. Pol II protects 15-20 bases of nascent RNA upstream

of the active site from RNase treatment, and is capable of resuming transcription when nucleotides are added (Gu et al., 1996; Kireeva et al., 2005). Analysis of sequences produced by the GRO-seq procedure using nuclei that have been RNase-treated prior to the run-on reaction will locate the positions of the active site of the polymerases. The Pol II active site, which is defined as the 3' end of the RNA associated with transcriptionally-engaged polymerase (Rudd et al., 1994), can be extrapolated to reside 15-20 bases downstream of the observed 5'-ends of the sequences determined by this adaptation of GRO-seq (Figure 4.1).

Before making GRO-seq libraries by this method, RNase protection of the expected 15-20 nucleotides by transcriptionally engaged polymerases should be confirmed directly by treating nuclei with various RNase cocktails, followed by a run-on reaction that includes three radiolabeled nucleotides and one chain terminating nucleotide. The average size of the run-on RNAs detected in this experiment will be on average about 3 nucleotides longer than the size of the protected RNA prior to the run-on, and will allow estimation of the average size and range of RNase protection. These types of experiments were presented in Chapter 2 of this dissertation. When this data is combined with the cap analysis described below, we should be able to determine precisely where Pol II is located relative to TSSs, which will be critical in evaluating transcription regulatory mechanisms associated with early elongation.

It will likely be easiest to perform this adaptation in *Drosophila* S2 cells, since the genome is approximately 1/20 the size of the human genome. This feature should greatly reduce the amount of sequencing needed to obtain



**Figure 4.1. Schematic for mapping the 3'-end of the engaged Pol II.**

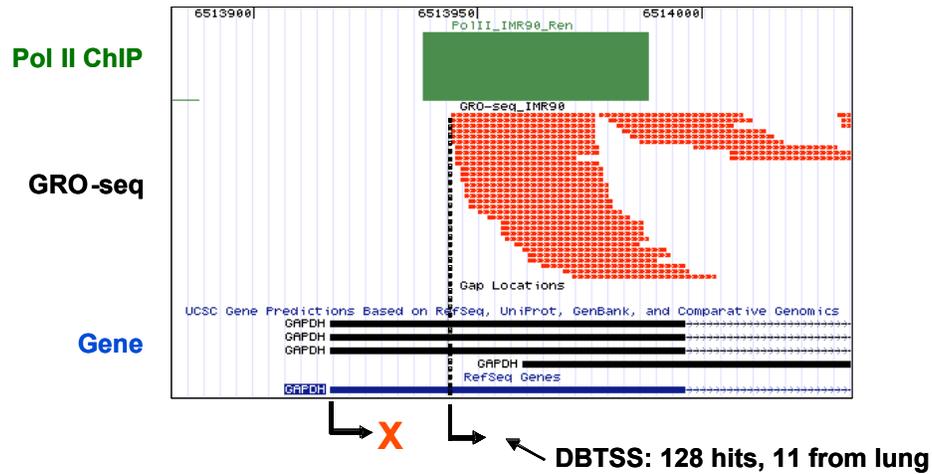
Transcriptionally-engaged Pol II protects 15-20 bp of the nascent transcripts, which could be further transcribed and to produce short run-on transcripts.

Note that the 5' end of the run-on transcript (marked as a star) maps the 3' end of the transcript generated prior to the run-on analysis minus the 15-20 bp Pol II protected site by RNase. Figure courtesy of Irene Min.

sufficient depth of coverage. Also, human promoters largely reside within CpG islands (Sandelin et al., 2007). CpG containing promoters tend to allow initiation at multiple locations (typically with 75 bases). These broad distributions of transcription start sites could complicate a method such as this. The *Drosophila* genome has fewer promoters with high CpG content, and are thus expected to have a higher frequency of focused transcription start sites. In addition, high resolution mapping of Pol II has been recently produced for ~80-90 genes in *Drosophila* S2 cells by potassium permanganate footprinting (Giardina et al., 1992; Lee et al., 2008; Muse et al., 2007; Wang et al., 2007; Zeitlinger et al., 2007a). This gives us the ability to compare our high-resolution global results with a sizable independent set of data; an opportunity that is currently not afforded in any other cell type.

#### **4.1.2 Mapping Transcription start sites with GRO-seq**

In the original design of GRO-seq I reasoned that we should also obtain start site information evidenced by a sharp drop of in the 5'-end of genes. Indeed, we do seem to map some alternative TSSs of highly transcribed genes with this procedure in the original form (Figure 4.2). However, we would need even deeper sequencing to map all TSSs. Therefore, In order to map the relative location of polymerase active sites to the sites where transcription initiated, I propose to map transcription start sites (TSSs) by combining GRO-seq with the oligo-capping method (Wakaguri et al., 2008) (also called RNA-ligase mediated- RACE or RLM-RACE). In this method, total RNA is treated with Calf intestinal alkaline phosphatase (CIAP) to remove the 5'-phosphoryl group. RNAs with a 5-methyl guanosine cap are protected from this treatment. The RNA is then treated with Tobacco acid pyrophosphatase (TAP), an enzyme



**Figure 4.2. GRO-seq can identify TSSs.** A close-up view of GRO-seq data within ~200 bp surrounding the GAPDH TSS. The 5'-end of GRO-seq hits drop off sharply at nucleotides that are not annotated as TSSs in the major databases, but map to the precise location of TSSs found in similar cell lines and that are listed in the Database of Transcription Start Sites.

that removes the cap but leaves a 5'-phosphoryl group. An adapter oligo is then ligated specifically to the once capped RNAs, since the CIAP treated non-capped RNAs can no longer act as donor molecules for the ligation reaction. After reverse transcription, capped RNAs are amplified by using the ligated oligo as a template. The cap site, and hence transcription start site, is then mapped by sequencing of the cDNA. This method is usually performed to map full-length transcripts by using Poly-A RNA as the template for reverse transcription. GRO-seq method already utilizes TAP treatment and ligation of a 5'-adapter ligation before reverse transcription. GRO-seq is easily adaptable to the oligo-capping method by adding a single CIAP treatment step before TAP treatment.

Adaptation of GRO-seq with the oligo-capping method will have two major advantages over current methods for obtaining transcription start site data. First, when mapping TSSs of steady-state RNAs, high abundance mRNAs will predominate in the sequencing reactions. Since, GRO-seq does not measure RNA accumulation, GRO-seq RNAs will represent a smaller fraction of the steady-state RNA pool, and allow the detection of low abundance RNA with less total sequencing. Second, we ligate our own adapter to both the 5'- and 3'-end of the GRO-seq RNA, thus making full-length transcription and RNA processing unnecessary for detection of a capped RNA. This adaptation will also aid our transcript mapping strategies for the first generation GRO-seq, and will allow us to determine whether the sites where we see multiple exact hits do indeed represent TSSs, in addition to allowing us to assess the prevalence of pausing more accurately. If this method proves efficient, it can be applied in combination to all cell lines and conditions for which first generation GRO-seq is performed.

## 4.2 Characterizing the genome-wide transcription response to heat shock in *Drosophila melanogaster*.

*Drosophila melanogaster* is a well-studied model organism and was recently selected for intensive study by the modENCODE consortium. The heat shock response of *Drosophila* has been studied in extreme detail, making it a particularly useful system with which to test the power of GRO-seq. During heat shock, several heat shock genes are rapidly and robustly induced. Changes in transcription at the *Hsp70* gene in particular, occurs within seconds after exposure to heat and results in up to a 200-fold increase in mRNA levels within 20 minutes. The changes in transcription are well documented by examining the distribution of Pol II along the gene by conventional nuclear run-on, ChIP, and potassium permanganate footprinting assays (Boehm et al., 2003; Giardina et al., 1992; O'Brien and Lis, 1993). Conventional run-on or ChIP analyses suggest that transcribing Pol II molecules reach a maximum possible density on *Hsp70* genes during full induction (O'Brien and Lis, 1993). In addition, the majority of transcription from non-heat shock genes is dramatically reduced during heat shock. For instance, H1 gene transcription drops to 50% of non-heat shock levels after 30 seconds, and is reduced to less than 10% after 5 min, of heat shock (O'Brien and Lis, 1993). Shutdown of transcription during heat shock has been documented by redistribution of active Pol II on salivary gland polytene chromosomes from normally active loci to sites of heat shock gene transcription (Greenleaf et al., 1978). In addition, heat shock-mediated transcription shutdown has been shown to be directly mediated by a Pol III transcribed RNA in mammals (Espinoza et al., 2004). It is possible that

performing GRO-seq in S2 cells, we could identify the *Drosophila* equivalent of this RNA.

I plan to do a GRO-seq analysis of *Drosophila* S2 cells during a time course of heat shock induction. This will allow us to assess the sensitivity of the assay to detect changes in transcription rates during the initial stages of both gene activation and repression. Also, since the Pol II density at these genes is at maximum density after 5 min of induction, this system will allow us to determine the dynamic range of GRO-seq. Nuclei will be prepared from S2 cells at 0, 1, 2, 5, and 20 minutes post heat shock. Given that conventional run-ons can detect changes at several heat shock genes in these time frames, I expect GRO-seq will be able to detect these changes in transcription, and will provide us a measure of the necessary sequencing depth needed to detect various levels of transcription.

#### **4.3 Cell cycle control of transcription**

One of the first uses of the nuclear run-on assay was to examine the level of steady state transcription in mitotic versus asynchronously growing cells (Gariglio et al., 1974). It had been assumed that the dense heterochromatin formed during mitosis would prevent transcription during this stage. Indeed, studies have shown that polymerases and sequence specific DNA binding factors can be stripped from the chromatin during in mitotically arrested cells (Martinez-Balbas et al., 1995; Parsons and Spencer, 1997; Shermoen and O'Farrell, 1991). However, nuclear run-ons have shown that mitotic cells do contain a measurable amount of transcription (Gariglio et al., 1974), and that some promoters (ACTG1) can retain Pol II at the promoter, while others cannot (FOS) (Parsons and Spencer, 1997). An interpretation of

these results could be that promoter-proximal pausing plays a role in setting which genes will be transcribed immediately following exit from mitosis. To investigate this possibility, I propose to analyze transcription during in mitotically arrested cells, as well as several timepoints following release from the block. GRO-seq will directly document transcription during these phases with higher dynamicity than arrays, since accumulated RNAs will not interfere with signal. In addition, temporal and spatial resolution that will be produced by GRO-seq cannot be done with any other assay.

#### **4.3 Concluding remarks**

I have outlined here a number of experiments to further adapt GRO-seq to obtain higher resolution of polymerases with respect to the transcription start sites, as well as studies that will investigate the transcription during specific cellular responses. There are conceivably many more adaptations that can be made, and biological pathways to examine. Nonetheless, GRO-seq, combined with our knowledge of transcription factor occupancies, DNA sequence elements, and accumulated RNA levels, will begin to shed new light on how biological processes and signaling pathways work to specify direct transcriptional outcomes.

## APPENDIX A

### ISOLATION OF IN VIVO FORMED PROTEIN/DNA COMPLEXES BY NUCLEOPROTEIN HYBRIDIZATION

#### **A.1 Introduction**

In order to fully understand how a particular gene is regulated during the phases of transcription, it is important to know all the players involved. Our current understanding of the array of protein transcription factors comes from many sources: 1) Biochemical fractionations designed to identify factors that bind specific DNA elements, modulate the activity of Pol II, or physically or chemically alter the chromatin environment. 2) forward and reverse genetic screens that identify genes that important for animal development, or transcription of specific genes. During the last decade, many factors identified by these means have been confirmed to be present at specific genes or regions of genes through the development of in vivo imaging techniques, such as the Chromatin immunoprecipitation (ChIP) assay. After isolating cross-linked protein-DNA complexes, using an antibody to a protein of interest, the location of the protein can be deciphered by determining the DNA sequence(s) that are isolated with it. The DNA is identified at discreet sites by direct hybridization of known DNA (Gilmour and Lis, 1986; Gilmour et al., 1986), quantitative PCR (qPCR)(Boehm et al., 2003), or over large, genomic regions by ligation-mediated PCR followed by hybridization to genomic microarrays (Ren et al., 2000), or massively parallel sequencing (Barski et al., 2007). The ChIP assay was originally developed using UV light to crosslink proteins that directly bind to DNA (Gilmour et al., 1986), and was later adapted by the use

of cell-permeable chemical crosslinkers, such as formaldehyde (Solomon et al., 1988). The benefit of using formaldehyde over UV is that it crosslinks proteins to DNA, as well as other proteins, so that proteins that do not directly contact the DNA can also be assayed. Also, formaldehyde crosslinks are fully reversible, allowing the enriched DNA to be extracted and directly amplified by PCR for quantification as described above. Our lab has successfully used the ChIP assay to document the changes in transcription factor occupancy and distribution along the *Hsp70* gene of *Drosophila melanogaster* (Adelman et al., 2005a; Andrulis et al., 2000; Andrulis et al., 2002; Gilmour and Lis, 1985; Gilmour and Lis, 1986; Gilmour et al., 1986; O'Brien et al., 1994a; O'Brien et al., 1995; Saunders et al., 2003).

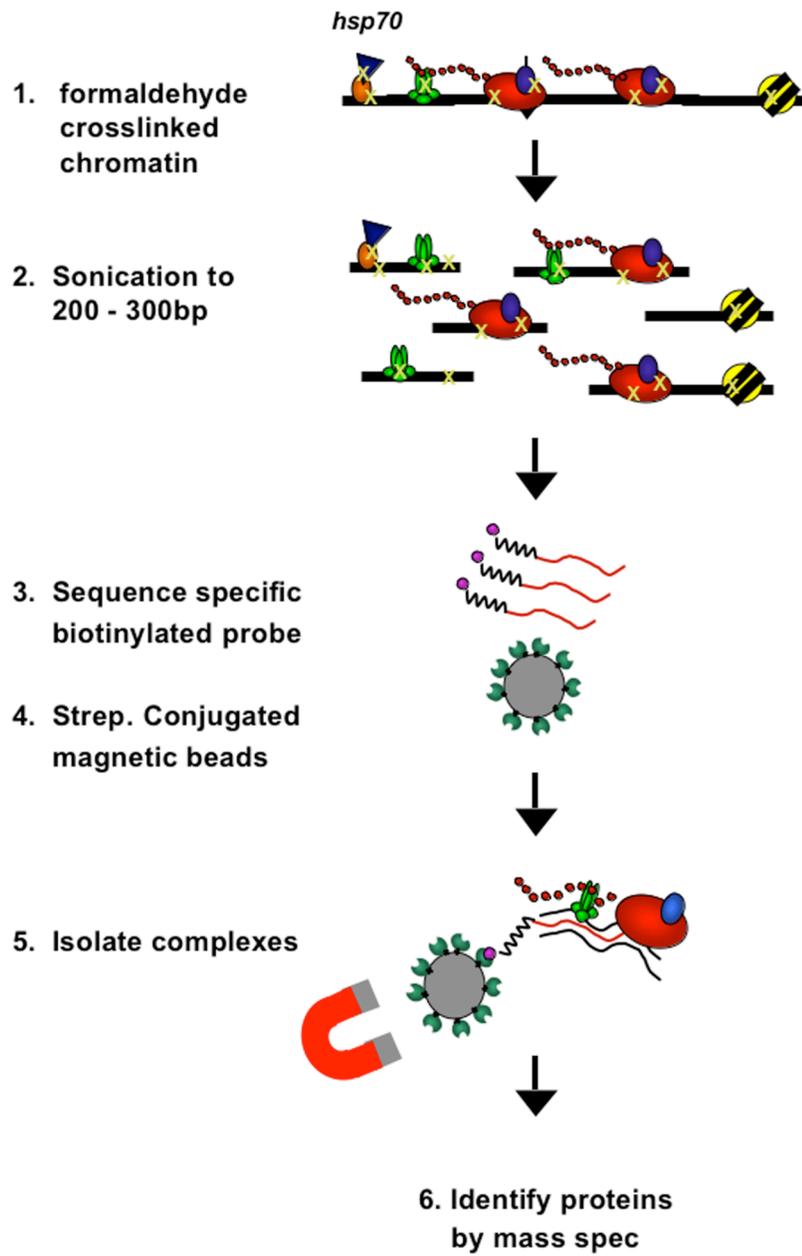
Despite these advances, the major challenges to deciphering the full complement of proteins present at specific genes remains identification of candidate factors, and the production of high-quality antibodies that are suitable for ChIP. Considering the time, expense, and uncertainty of antibody production, I was prompted to develop a general method that would identify all the proteins interacting with specific DNA elements in a single experiment. A simple way to visualize the experiment is to think of it as a 'reverse-ChIP assay'. Instead of identifying the DNA isolated by immunoprecipitation of a single protein, this experiment is designed to isolate a discrete region of the genome, by hybridizing a biotinylated nucleic acid probe to crosslinked chromatin, and identify all the proteins associated with the DNA by mass spectrometry. An outline of this nucleoprotein hybridization experiment designed to target the *Drosophila melanogaster Hsp70* promoter is shown in Figure A.1.

Before embarking on this pursuit, it is important to consider the theoretical yield of protein that can be isolated, the relative enrichment above background, and whether these parameters will yield enough material of sufficient purity to enable detection by mass spectrometry. At the inception of this project, the lower limit of detecting peptides from digested proteins was ~ 10 femtomoles. The theoretical amount of protein that can be associated with any given sequence of the genome depends on the number of molecules that can physically bind a specific region at any given time, the frequency with which the binding site is occupied, and the efficiency with which that protein is crosslinked to the DNA in the presence of formaldehyde. The amount of protein that can then be isolated from a single cell is a combination of the above parameters in conjunction with the number of copies of the targeted sequence in the genome, and the efficiency with which the probe can capture the target and be isolated by bead chromatography. For instance, if a sequence was represented twice in a diploid genome, and a protein bound once to this sequence with full occupancy,  $\sim 3 \times 10^9$  cells (600ml) would be required to even start with 10 femtomoles of material. Considering that the crosslinking efficiency and isolation of the crosslinked protein/DNA complex are not going to be 100% efficient, this is clearly not an acceptable starting point. However, the *Hsp70* gene of *Drosophila* provides a unique opportunity to use this gene promoter as a proof of principle.

The *Hsp70* promoter is present in five to six copies/haploid genome, depending on the strain or cell line. Kc cells have been reported to contain six copies. The *Hsp70* promoter contains three heat shock elements (HSEs), each capable of binding a trimer of the Heat Shock Factor (HSF) activator, with full occupancy upon thermal stress. This translates into  $10^8$  HSF

**Figure A.1 Schematic of nucleoprotein hybridization experiment.**

Chromatin is crosslinked and sonicated as in ChIP assays **(1,2)**, however, instead of immunoprecipitating a specific protein, a biotinylated nucleic acid probed is targeted to a specific region of the genome **(3)**. The nucleoprotein complexes are then isolated with streptavidin-coated magnetic beads via the biotin tag **(4,5)**. Proteins are eluted and identified by mass spectrometry **(6)**



molecules bound to an *Hsp70* promoter/cell. Based on the efficiency of ChIP assays, I estimated that the efficiency of crosslinking these HSF molecules to the DNA is ~10%, or 10 isolatable molecules/cell. Thus,  $6 \times 10^8$  (60ml) cells could be used to obtain 10 femtomoles of crosslinked HSF/HSE starting material, assuming 100% efficiency of the isolation procedure. Realistically, I assumed that I would end up with ~10% isolation efficiency, thus requiring  $6 \times 10^9$  cells (600ml) to detect HSF and up to  $1 \times 10^{11}$  cells (10 liters) to detect proteins that are bound at 1 molecule/promoter with 10% crosslinking efficiency. Thus, the initial goal of this project was to determine the isolation efficiency by systematically testing the variables of probe hybridization in small-scale experiments, and establish a framework for scaling up the procedure to obtain enough protein for identification by mass spectrometry.

Perhaps the most challenging aspect of the isolation procedure lies in finding a condition that allows the probe to find and hybridize to its complementary sequence within crosslinked chromatin, while at the same time, does not catalyze the reversal of crosslinks between the proteins and DNA. For standard nucleic acid hybridizations, double-stranded DNA must be denatured to allow hybridization of another DNA molecule or probe. Furthermore, this probe must be in excess concentration relative to the target molecule to prevent rehybridization of the original molecules during the time-frame of the experiment. Considering the situation encountered when hybridizing a probe to crosslinked chromatin, it should be noted that the proteins surrounding the crosslinked DNA could prevent efficient denaturing of the DNA. Even if denaturation is possible, there is still a chance that proteins could form a bridge between the two strands, which would vastly increase the reassociation kinetics between the two strands. Under this condition, probes

present even in excess will likely be displaced by re-zipping of the two DNA strands, through a process known as branch migration.

Formaldehyde forms crosslinks between positively charged primary and secondary amines found in lysine, arginine, N-termini of proteins or within the DNA bases. These crosslinks are fully reversible by heating in the presence of concentrated ionic or amphoteric buffers. The thermal reversibility of formaldehyde crosslinks adds to the challenge of denaturing the target DNA strands while maintaining crosslinks between proteins and DNA. Also, thermal denaturing of the DNA is likely to denature proteins, which could then form aggregates resulting in decreased yield and specificity of the isolation. I therefore set out to test conditions that could reduce the temperature required to denature DNA or otherwise capture the target DNA sequence, while at the same time prevent crosslink reversal and aggregation of proteins.

The first set of experiments describe my attempt to test whether DNA probes are capable of isolating DNA from crosslinked chromatin preparations under various denaturing conditions. I also use the ChIP assay to document whether the formaldehyde crosslinks are preserved under the same conditions. The second set of experiments document the characterization of Peptide Nucleic Acids (PNA) as alternatives to DNA probes for isolation of nucleoprotein complexes. The third set of experiments document the development of an in vitro crosslinking assay, during which HSF is crosslinked to HSEs. This simple system is used to test conditions for target DNA denaturation as well as preservation of crosslinks.

## **A.2 Materials and Methods**

### ***Probe design***

Parameters for the design of a biotinylated nucleic acid probe for pulling down the *Hsp70* promoter included: a region that was central relative to the three HSEs, conserved amongst the copies of *Hsp70*, unique amongst the genome, and not known to directly bind transcription factors such as GAGA factor or HSF. With these parameters in mind, a biotinylated DNA oligo (Probe 1) was designed to hybridize to the bottom strand from -226 to -195 relative to the *Hsp70* transcription start site. Another biotinylated oligo (Probe 2) was designed to hybridize to the top strand from -160 to -195. This second probe is less ideal, since it directly overlaps the central HSE of the *Hsp70* promoter. See main text for description of PNA design.

### ***Probe sequences***

DNA Probe 1: 5'-/Biotin-TEG/-

TCTCCTGGTTATTGTGGTAGGTCATTTGTTTGGC-3'

DNA Probe 2: 5'-/Biotin-TEG/-

TCTCGAATCACGGCCAGAGAAATTTCTCGAGTTTTCTTTG-3'

PNA-U1: 5'-TTTGT TTTGGGATTCT-Lysine-Lysine-3'

PNA-B2: 5'-/Destiobiotin/linker-15/linker-15/AACTGGTTATTGTGG-Lysine-3'

PNA-D2: 5'-AGGTCATTTGTTTGG-Lysine-Lysine-3'

### ***Heat shocking and crosslinking of large volumes of Drosophila cells***

Crosslinked chromatin was prepared from 8-10L of *Drosophila* Kc cells grown in a spinner flask in SFx media (Hyclone), and to a density of 0.5-1x10<sup>7</sup> cells/ml. Spinner flasks were heat shocked in a 20 gallon trash barrel

containing 6L of 48°C water. The temperature was maintained by a closed-pump system running 48°C water through a ¼ inch copper tube that was submerged in the water, and coiled to fit tightly around the spinner flask. The flask was placed within the submerged coil, and the temperature monitored, while constantly stirring the culture. Once the temperature inside the flask reached 36.5°C for 20 minutes, the temperature was dropped to 22-25°C by addition of 2L of ice-cold water to the barrel. Formaldehyde was then added to 0.3% for 3min, followed by addition glycine to a final concentration of 125mM.

***Preparation of crosslinked chromatin.***

Crosslinked cells were pelleted for 10 min, at 4°C, at 2,600xg (~4000rpm) in a Sorvall-GSA rotor. Pellets were washed twice in 50ml cold PBS/liter of culture. Pellets were resuspended in 5 packed cell volumes (PCVs) of buffer A (swelling buffer): 10mM HEPES pH 7.8, 10mM KCl, 1.5mM MgCl<sub>2</sub>, 0.1mM EDTA, 20% glycerol, protease inhibitors. Cells were allowed to swell for 15min at 4°C. Cells were then pelleted in conical tubes, at setting 4 of an IEC clinical centrifuge, and resuspended in 2 PCVs of Buffer B (Buffer A +0.1% Igepal. Cells were then dounced with 50 strokes using a loose pestle, nuclei pelleted, and resuspended in 0.5 packed nuclear volumes (PNVs) of buffer B. 0.4 PNVs of buffer C (10mM HEPES pH 7.8, 1M KCL, 0.1% Igepal) was then added while swirling on ice. The extract was incubated at 4°C on a rocking platform for 45min, and then dounced with 25 strokes of a tight pestle. The homogenate was spun at 4°C at 14,000rpm, for 20min in an SA-600 rotor (Sorval). The chromatin pellet was washed twice with sonication buffer (20mM HEPES pH 8,2mM EDTA, 0.5mM EGTA, 0.5% SDS, 0.5mM PMSF, protease inhibitors), and resuspended in sonication buffer at 300ul/liter of original cell

culture. Chromatin was solubilized by sonication 5 times with the program: 20sec pulse, 20sec off, 5 pulses. The chromatin samples were dialyzed (10mM HEPES pH 8, 1mM EGTA, 0.5mM EGTA, 10% glycerol, 0.1% Sodium deoxycholate, 0.1% triton, 0.5mM PMSF), snap frozen in liquid nitrogen, and stored at -80°C.

***ChIP for HSF under mock denaturing/hybridization conditions.***

ChIP was performed as described (Boehm et al., 2003), with several modifications. Chromatin equating to  $10^7$  cells was used/IP (typically 4ul). ChIP dilution buffer was replaced by a hybridization buffer: 10mM HEPES pH 7.9, 1mM EDTA, 100mM NaCl, 0.1% Igepal, 0.05% SDS, 0.5mM PMSF, and 1mM DTT). Formamide was used at 70% when indicated. The total volume of each sample was 25ul. Samples were heated to indicated temperatures for 10min, followed by 2X dilution in hybridization buffer (without formamide). Samples were incubated at 25°C for 2-5 hours (mock of hybridization). Prior to immunoprecipitation, the reaction was diluted to 500ul in standard ChIP dilution buffer. For HSF ChIP, 1.5ul of rabbit  $\alpha$ -dHSF was used/IP. ChIP samples were then processed as described (Boehm et al., 2003).

***Nucleoprotein hybridization.***

Small-scale nucleoprotein hybridizations were prepared as ChIP samples described above, but processed with several modifications. 50 picomoles of probe 1 and probe 2 were added to each reaction prior to the heat denaturing step. After hybridization, an equal volume (0.1ug/fmol of probe) of preblocked M-280 beads (Invitrogen) were added, and the reaction incubated at 25°C for 1 hour, while rotating. M-280 beads were blocked in

10mM HEPES pH 7.9, 100mM NaCl, 1mM EDTA, 0.1% Igepal, 10mg/ml casein, 5mg/ml sonicated salmon sperm DNA for 1 hour at 25°C. Beads were resuspended in 2X binding buffer (10mM HEPES pH 7.9, 1M NaCl, 1mM EDTA, and 0.1% Igepal) prior to addition to the samples. Beads were washed 3 times at 35°C with 1X binding buffer supplemented with 17.5% formamide, 3 times at 35°C, with hybridization buffer supplemented with 17.5% formamide, and 2 times at 25°C with hybridization buffer. Samples were eluted twice at 85°C in 100ul of hybridization buffer that had reduced NaCl (10mM), or by boiling in 95% formamide for 10min. Large-scale nucleoprotein hybridizations were carried out at 1000X the scale described above.

***Primer sequences for qPCR from ChIP or nucleoprotein hybridization fractions.***

qPCR was carried out as described as in (Boehm et al., 2003).

Hsp70 -200 forward: 5'-tgacagaaagaaaactagagaaa-3'

Hsp70 -108 reverse: 5'gacagagtgagagagcaatagtacagaga-3'

Sens +5960 forward: 5'-cccaaaattggcagctaaacg-3'

Sens +6080 reverse: 5'-gtgggtgatgccatcaataaac-3'

18S +312 forward: 5'-gccctatcaacttttgatggtagta-3'

18S +414 reverse: 5'-ggtagccgtttctcaggct-3'

***Characterization of PNAs.***

PNA were used at the indicated concentrations, with radiolabeled HSEs (described below). Unless otherwise noted the reactions were incubated at 37°C for 2hrs in 5mM Tris-HCl pH 7, 5mM NaCl, 0.5mM EDTA.

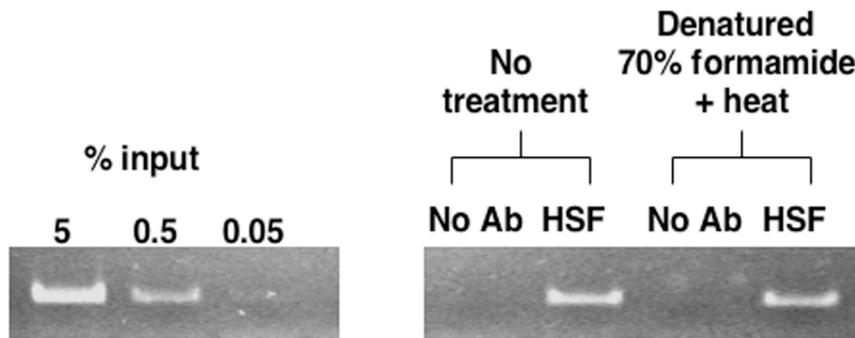
### ***in vitro crosslinking of dHSF to HSEs.***

$\alpha$ -<sup>32</sup>P- labeled HSEs were prepared either by end labeling a PCR product containing -271 to -136 of the Hsp70 promoter, or by klenow fill in of two hybridized oligos (plus strand: -251 to -192; minus strand: -172 to -231). Unless otherwise noted, His-tagged, purified HSF (35nM) was incubated with HSEs in HSF binding buffer (20mM HEPES pH 7.9, 0.1mM EDTA, 4mM DTT, 0.05% Igepal, 50mM NaCl, 0.5mM MgCl<sub>2</sub>, 50ng/ul poly dI/dC, 0.2mg/ml BSA), for 1 hour at 25°C. Crosslinking was performed with 0.3% (w/v) formaldehyde for 5min, unless otherwise noted. Reactions were quenched by addition of glycine to a final concentration of 125mM. Samples were run on a discontinuous SDS-PAGE system, consisting of a 2.5cm, 4% acrylamide stacking gel, and a 10cm, 5% acrylamide separating gel. When noted, samples were buffer exchanged using a micro-Biospin, P-30 column (BioRad), equilibrated in 10mM HEPES pH 7.9, 0.1% Igepal.

## **A.3 Results**

### ***Small-scale reconstruction using hybridization-pulls downs and ChIP for HSF.***

I first tested whether formamide could be used lower the melting temperature (T<sub>m</sub>) of DNA within crosslinked chromatin, without reversing the crosslinking. Formamide reduces the T<sub>m</sub> of DNA by 0.6°C per percent formamide. Thus, 70% formamide should reduce the T<sub>m</sub> of an ~300bp fragment from 95°C to 53°C. To test the effect that the temperature and formamide has on formaldehyde crosslinks, I performed ChIP for HSF at the *Hsp70* promoter using crosslinked chromatin from heat shocked Kc cells.



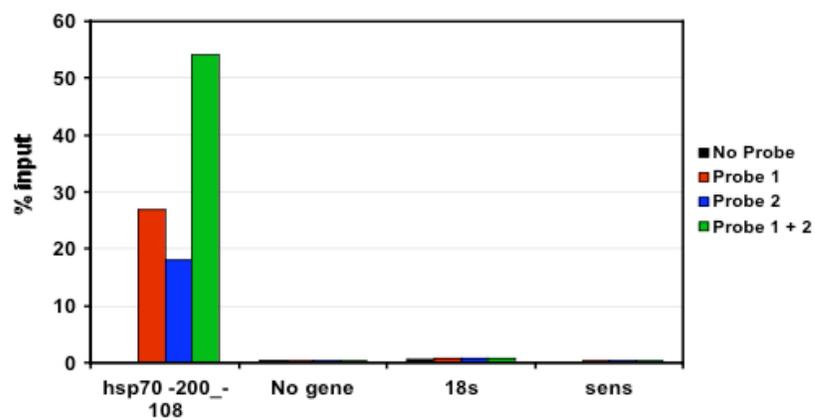
**Figure A.2 ChIP of HSF at Hsp70 promoter under mock denaturing conditions.** Control (No Antibody) and HSF immunoprecipitations (IPs) were performed from untreated, or denatured chromatin prepared from heat shocked Kc cells. Denatured sample was incubated at 55°C for 10 min, followed by 2 hours at 25°C, prior to performing the IP. PCR was performed with primers to the Hsp70 promoter, and compared to input samples (left).

Initial results showed that heating to 55°C in the presence of 70% formamide had no appreciable effect on the amount of the *Hsp70* promoter that could be immunoprecipitated with the HSF antibody (Figure A.2). Under these same conditions, I was also able to isolate the promoter by hybridizing biotinylated oligos targeted to both strands of the *Hsp70* promoter. The promoter was isolated with the same efficiency if crosslinked chromatin or fully denatured naked DNA was used in the procedure (Figure A.3 and data not shown), indicating that the DNA within chromatin was being denatured under these conditions. The hybridization appeared to be specific, since the *Hsp70* promoter was isolated with ~200 fold more efficiency than a region of the senseless gene, which has the highest similarity to the target site; and with ~100 fold more efficiency than the highly abundant 18S rDNA repeat. However, due to the generally low efficiency of crosslinking and the incomplete capture of all the promoters, this experiment could not determine whether the oligo-captured promoters were actually crosslinked to proteins. I therefore scaled up the procedure and performed western blots to determine whether HSF or GAGA factor could be detected in the eluted fractions. Unfortunately, HSF and GAGA factor were not detected in the elution from a nucleoprotein hybridization, even though the expected yield was within the linear range of detection of these antibodies (data not shown). Several scenarios, and combinations thereof, could account for this result: 1) the sensitivity of the procedure is low; 2) the oligo probes preferentially hybridized to DNA that was not crosslinked to protein in the vicinity of the target site; or 3) the crosslinks were indeed being reversed. Repeated experiments revealed that isolation of

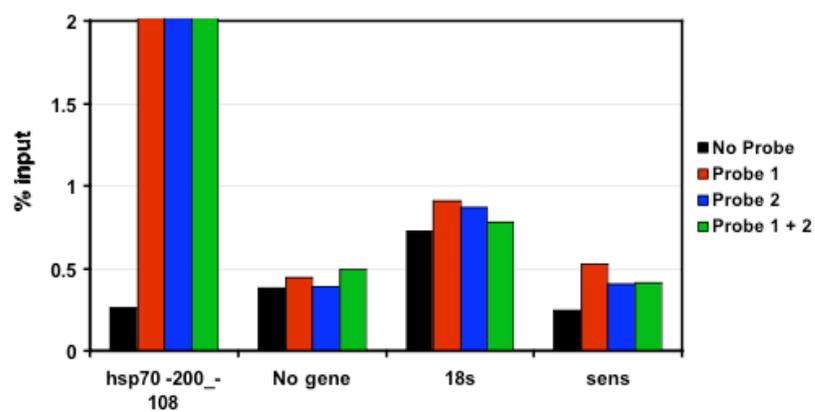
**Figure A.3 Isolation of Hsp70 promoter with biotinylated DNA oligos.**

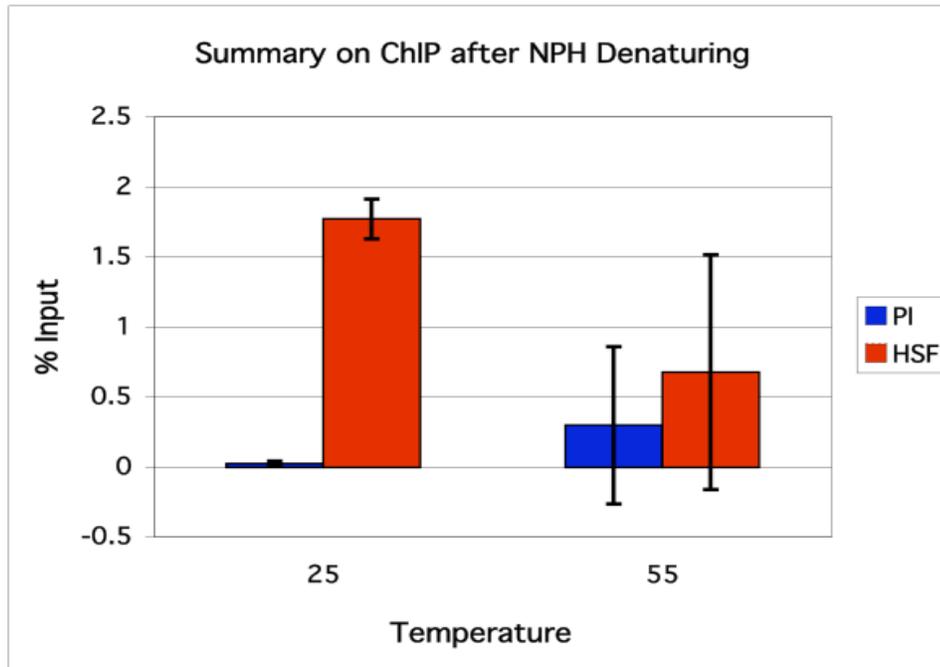
Both strands of the Hsp70 promoter were isolated by hybridizing biotinylated DNA oligos to crosslinked chromatin (see methods). Efficiency and enrichment over background (no gene, 18s), or a region of similar sequence to the target was assayed by qPCR . Top panel is full scale, bottom panel has a reduced y-axis for viewing the controls

Pull down of both strands of hsp70 promoter



Pull down of both strands of hsp70 promoter





**Figure A.4 Summary of HSF ChIP after DNA denaturing protocol.**

Y axis shows the percent input for HSF (Red bars) IPs either before (25°C) or after (55°C) thermal denaturing. Blue bars represent Pre-immune controls. Error bars represent the standard error of the mean for five separate experiments.

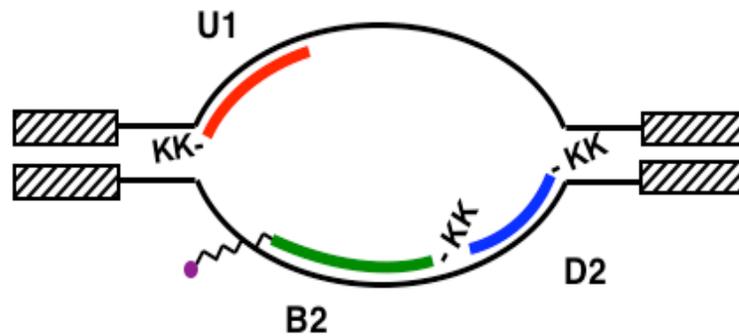
the *Hsp70* promoter via immunoprecipitation of HSF was variable under the denaturing conditions described above, ranging from non-denatured control levels to background (Figure A.4). Also, the signal from the heat-denatured no antibody control was more variable than the non-denatured sample; sometimes as high as the HSF antibody signal. I hypothesized that the variability in the CHIP experiments was being caused by the formation of aggregates during the denaturing step. Attempts to add combinations of ionic and non-ionic detergents did not appear to prevent the variability (data not shown).

***Attempts to use peptide nucleic acids as probes for nucleoprotein hybridizations.***

In order to circumvent the apparent problems with thermal reversal of crosslinks, and possible aggregate formation, I began seeking alternative probe designs. The use of locked nucleic acids (LNAs) or peptide nucleic acids (PNAs) seemed to provide specific advantages for this project relative to DNA oligos. LNAs have a methylene bridge between the 2' and 4' carbon of the ribose moiety that 'locks' the nucleic acid in the A-form, which in turn, increases the base stacking properties between bases (Reviewed in Vester and Wengel, 2004)). This organization leads to greater thermal stability of LNAs that are hybridized with DNA or RNA. PNAs have a pseudo-peptide backbone in place of the sugar-phosphate backbone found in DNA or RNA (Reviewed in Nielsen, 2004). Since the PNA backbone contains no charged phosphates, PNAs are extremely flexible and stable in complex with other nucleic acids. PNAs can form triplex complexes with dsDNA, displace one strand of the DNA in a process known as invasion, and can also form triplex

invasion complexes whereby two PNAs invade dsDNA and hybridize to one strand by both Watson-Crick and Hoogsteen base-pairing. Nucleic acid structure can greatly affect the invasion rate of PNAs, but once bound, PNAs have been shown to have extremely slow off rates - on the order of days to weeks. The most powerful reagent for creating triplex invasion complexes is the bis-PNA, whereby two PNAs targeting the same sequence are arranged in a head to head orientation separated by a flexible linker. Given the invasion properties and extreme stability of PNA/DNA hybrids, I chose to use PNAs as oligo-capture probes over LNAs.

The PNAs used in this study were synthesized by Muris Kobasliga from Tyler McQuade's lab, here at Cornell. Due to considerations of yield, I was not given the option to design a bis-PNA. I therefore designed three overlapping PNAs, with the intent that they would act cooperatively and increase the specificity of the isolation. The PNAs were designed to the same region of the *Hsp70* promoter as the biotinylated DNA oligos. A 15-mer, PNA-B2, was designed as the 'capture' PNA. It was modified on the 5' end by two flexible linkers that each contained 15 bond lengths. These spacers separate the PNA sequence from a desthiobiotin moiety, which is the tag that can be isolated with streptavidin magnetic beads. The spacers are intended increase the yield by preventing the protein/DNA/PNA complexes from sterically hindering the desthiobiotin/streptavidin interaction. The desthiobiotin/streptavidin interaction has a dissociation constant ( $K_d$ ) several orders of magnitude greater than that of biotin/streptavidin. This allows complexes that are isolated by the PNA to be eluted with biotin rather than harsh thermal or chemical elution methods that can cause the release of the



**Figure A.5 Schematic of PNA design.**

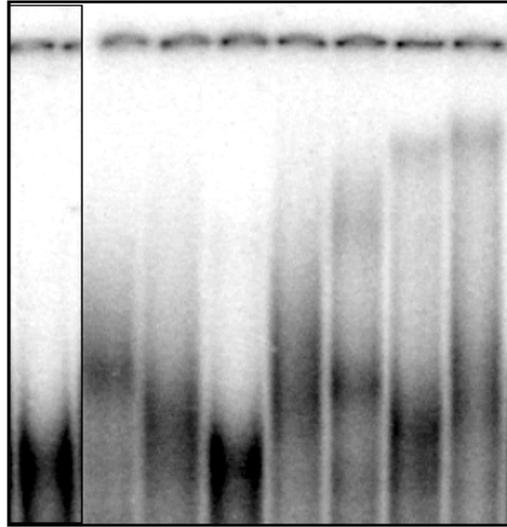
PNAs designed to isolate the Hsp70 promoter from crosslinked chromatin. All three PNAs are targeted between HSEs (hashed boxes). PNAs U1 and D2 are designed to aid in the binding and specificity of PNA-B2. Stable binding of PNA B2 is critical for isolation, since it contains the desthiobiotin moiety (purple). See text for further description.

intended target as well as non-specifically bound complexes. The 3'-end of PNA-B2 was modified with a lysine amino acid. The added positive charge provided by lysines has been shown to increase the rate of PNA/target formation by increasing the local concentration of the PNA around the negatively charged backbone of nucleic acids (Nielsen, 2004). Two more PNAs, PNA-U1 and PNA-D2, were designed to bind immediately upstream and downstream of PNA-B2, respectively. PNA-U1 is 16 bases, hybridizes to the opposite strand with respect to PNA-B2, and overlaps PNA-B2 by two bases. PNA-D1 is 15 bases and hybridizes to the same strand as PNA-B2, with a single base separating the target sites. Neither PNA-U1, or D2 have desthiobiotin modifications, but both have two lysines at the 3'-end. The purpose of these two PNAs is to increase the rate and specificity of PNA-B2 for the *Hsp70* promoter. A schematic showing how the PNAs are intended to bind to dsDNA is shown in Figure A.5.

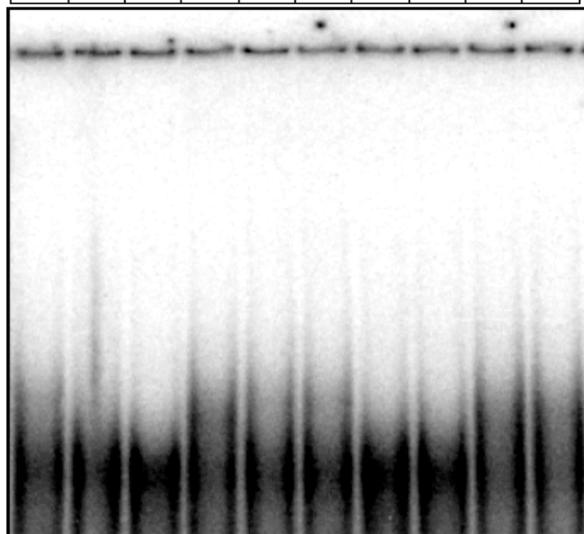
I first tested the general characteristics of the PNAs by determining their ability to bind non-crosslinked PCR-products containing the target site. In summary, I found that the PNA binding to dsDNA was slow (requiring hours) and the rate was greatly affected by the ionic strength of the solution, as well as the presence non-specific DNA. On the plus side, the PNAs were specific for their targets, and appear to have a very slow off-rate that is not affected by addition of salts or non-specific DNA. In regards to the goal of this project, I found that the PNAs appear to be cooperative at high concentrations (>5 $\mu$ M), but show no detectable binding or cooperativity at concentrations feasible for a large-scale experiment (low nanomolar range) (Figure A.6). The PNAs could bind to single-stranded DNA at low nanomolar or sub-nanomolar concentrations (0.5-5 $\mu$ M) (data not shown). Thus, under the current design,

**Figure A.6. PNAs do not bind at concentrations feasible for large-scale experiments.** Top panel shows the cooperativity of PNA binding to dsDNA in the micromolar range. Bottom band is the free dsDNA, and slower migrating bands are in complex with the PNAs. Note that PNA-B2 does not bind to the DNA on its own (lane 4), but only when the other two PNAs (lanes 5,6) or both (lane7) are present. Bottom panel shows that the PNAs do not bind dsDNA at concentrations (low nanomolar) that are feasible for large-scale nucleoprotein hybridizations.

[U1] $\mu\text{M}$	0	5	0	0	5	5	0	5
[D2] $\mu\text{M}$	0	0	7.5	0	7.5	0	7.5	7.5
[B2] $\mu\text{M}$	0	0	0	7.5	0	7.5	7.5	7.5



[U1] nM	50	0	0	50	50	50	0	0	50	50
[D2] nM	0	50	0	50	0	0	50	50	50	50
[B2] nM	0	0	50	0	5	50	5	50	5	50

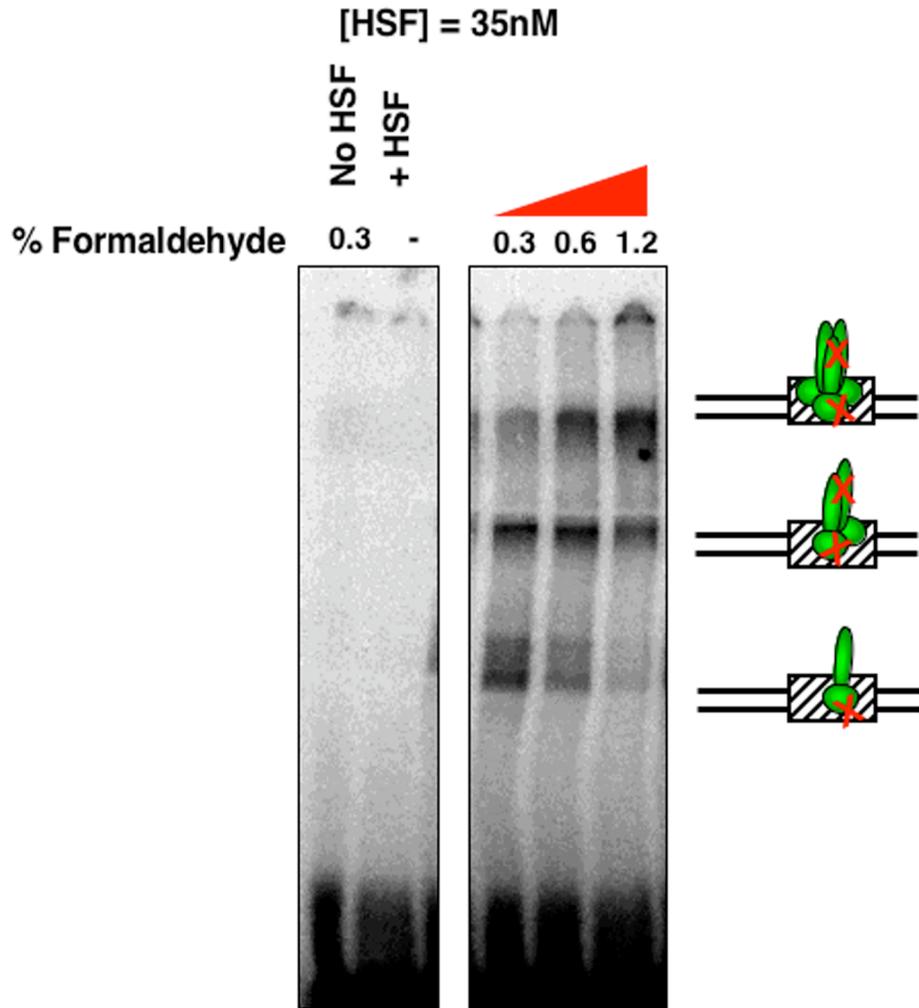


the PNAs still require some degree of denaturing of the DNA for efficient capture in a nucleoprotein hybridization experiment. In order to fully understand what conditions would permit partial or full denaturation of crosslinked complexes, I developed an in vitro assay to test this directly.

**In vitro crosslinking of HSF to HSEs as a direct readout of formaldehyde crosslink stability.**

In parallel with the above experiments, I developed an in vitro crosslinking method whereby purified HSF is crosslinked directly to a radiolabeled PCR product containing an HSE. The goal was to have a more direct readout of the stability of formaldehyde crosslinks under any tested denaturing condition. When the complexes are run on through SDS-PAGE, the crosslinked complex can be tracked by observing the slower migrating HSF/HSE complex by autoradiography. In these experiments HSF is added at sufficient concentration to occupy all HSEs prior to crosslinking (data not shown). Titration of formaldehyde in the reaction revealed that 0.3% formaldehyde results 5% crosslinking efficiency (Figure A.7 and data not shown). Interestingly, three bands were apparent on the gel, which I hypothesize represent monomer, dimer, and trimer HSF molecules in complex with the HSE (Figure A.7). Note that the dimer and trimer may very well be a result of additional protein/protein crosslinks rather than each individual HSF molecule directly crosslinking to the DNA.

I then used this system to test various methods of denaturing the DNA within the crosslinked complexes. In summary, I found that any buffer or detergent that carries an ionic charge can catalyze the reversal of

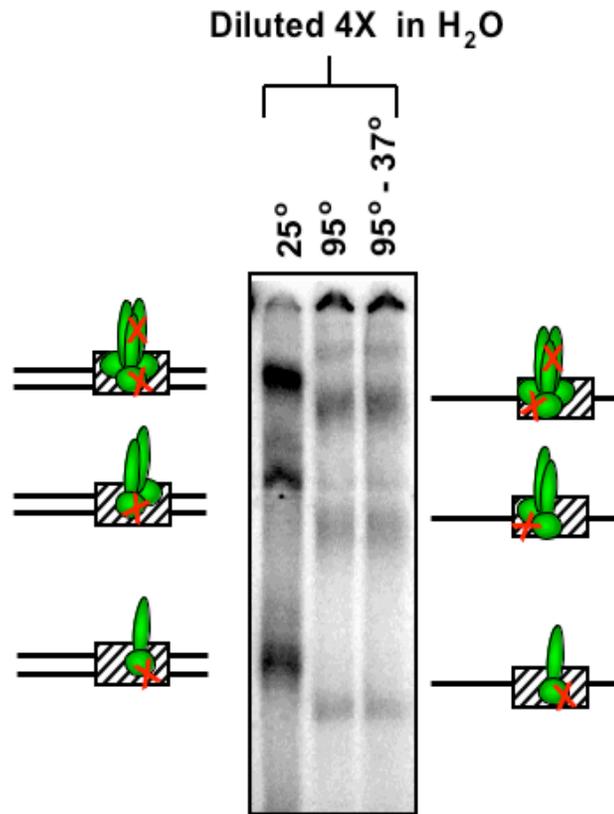


**Figure A.7. In vitro crosslinking of HSF to HSEs with formaldehyde.**

SDS-PAGE gel showing HSF crosslinked to a radiolabeled HSE. Free HSEs (bottom) are not shifted if either HSF or formaldehyde is left out (left panel). Right panel shows the extent of HSF crosslinking in response to increasing levels of formaldehyde. Bands, from top to bottom, are thought to represent an HSE crosslinked to a HSF trimer, dimer, or monomer, respectively (illustrated to the right).

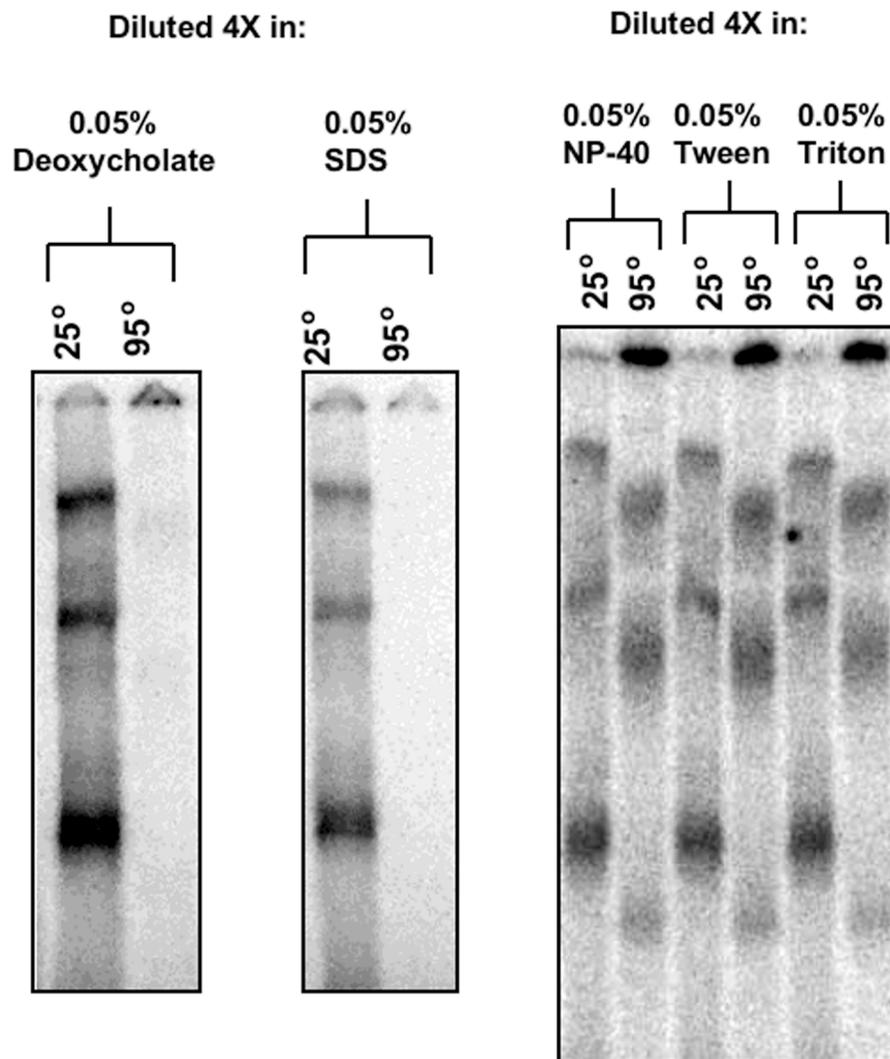
formaldehyde crosslinks (Figure A.8, A.9 and data not shown). I also discussed these results with Tadhg Begely, and after pushing some electrons around on a dry-erase board, he came to the conclusion that any charged buffer should indeed serve as a catalyst for crosslink reversal. Interestingly, if the crosslinked complexes were diluted in H<sub>2</sub>O, or solutions of nonionic detergents instead of the hybridization buffer, the crosslinks appeared to be maintained. The resulting complexes migrate faster in the gel, and are less susceptible to double-strand specific DNases, indicating that they are indeed single-stranded (Figure A.8 and data not shown). However, these complexes could not be shifted on a gel or isolated by the biotinylated DNA or PNA probes described above (data not shown).

Since the only denaturing conditions that appear to preserve the crosslinked complex is low ionic strength and nonionic detergents, I added a step to completely remove the potential affects from the HSF binding buffer, and formaldehyde that is quenched by glycine after the crosslinking reaction. Furthermore, the buffer conditions in the in vitro experiments do not completely mirror the situation encountered when preparing crosslinked chromatin from cells. For instance, the in vitro assay contains HSF binding buffer, as well as formaldehyde that has been quenched by 125mM Glycine. The former is not present in the chromatin preparations, and pelleting and washing the cells and nuclei after crosslinking largely remove the latter. I therefore exchanged the buffer after in vitro crosslinking by running the complexes through a desalting column. I then tested whether the complexes were stable under heat denaturing conditions if exchanged into a buffer with low ionic strength. In contrast to the experiments where crosslinked complexes are diluted in H<sub>2</sub>O,



**Figure A.8. Denaturing of dsDNA within HSF/HSE complexes.**

Diluted, crosslinked reactions were incubated at 95°C for 5min (lane 2), or 95°C for 5min followed by 37°C for two hours (lane 3). The faster migrating bands relative to the control (lane a) are proposed to be the single stranded HSF/HSE complexes (illustrated to the right).

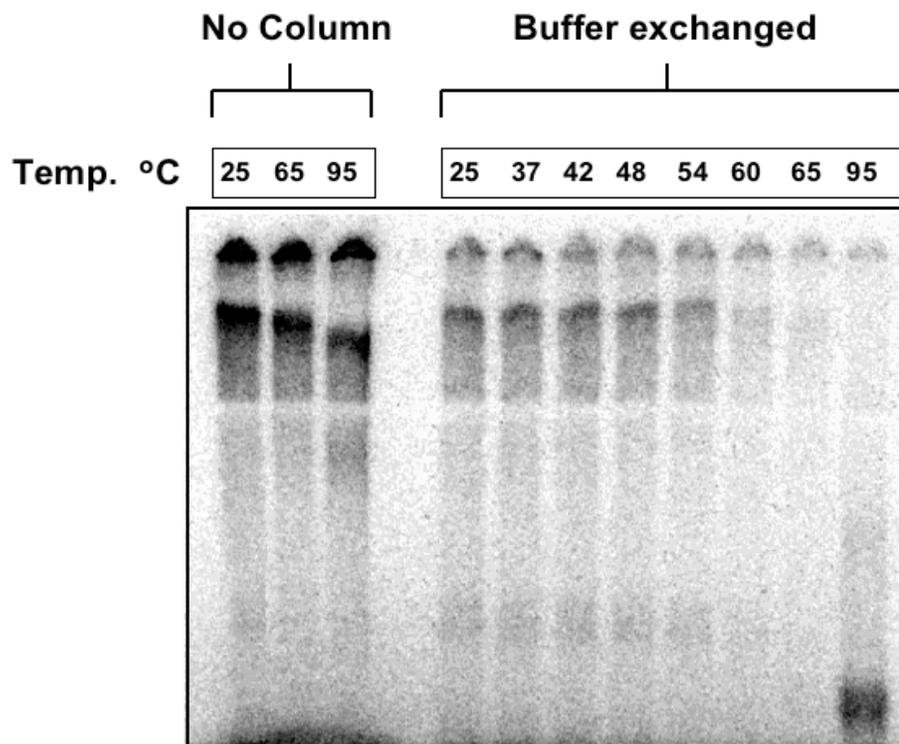


**Figure A.9. Formaldehyde crosslink stability with ionic or non-ionic detergents.** Left panel shows that thermal denaturing in the presence of ionic detergents, SDS or sodium deoxycholate, results in reversal of crosslinks. Right panel shows that non-ionic detergent, NP-40, Tween, and Triton X-100, do not reverse formaldehyde crosslinks under the same conditions.

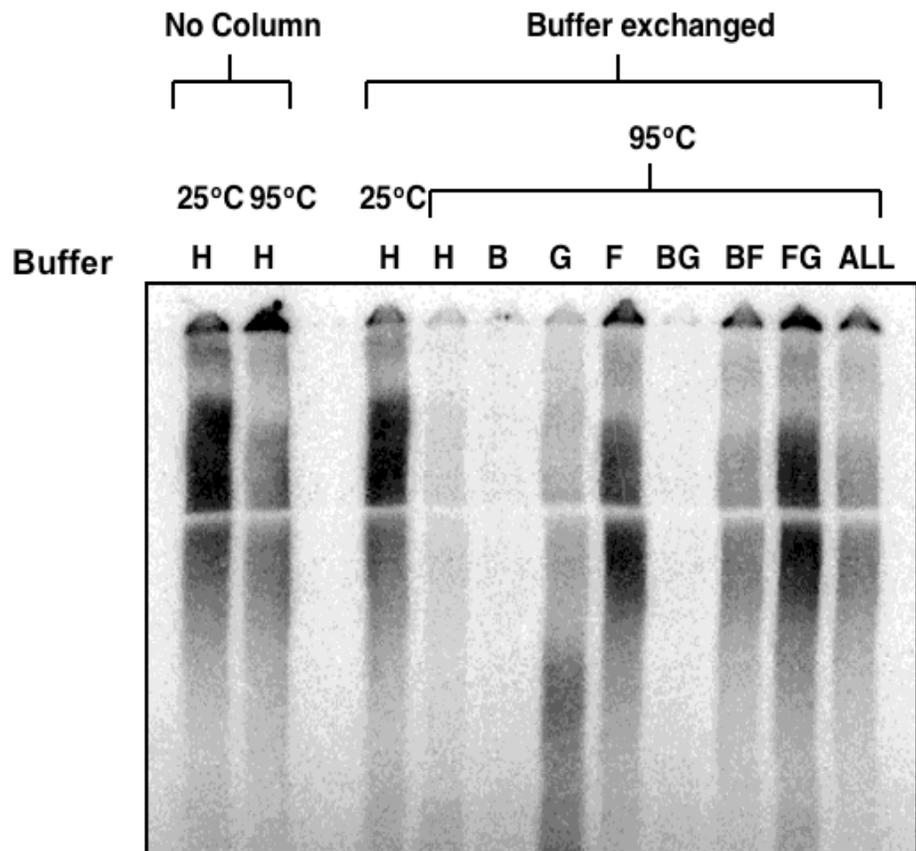
crosslinks in buffer exchanged complexes are not as stable, and appear to be completely reversed at 65°C (Figure A.10). I then tested what component of the in vitro crosslinked reaction accounted for the apparent stabilization of the complexes by adding back each component or combinations thereof, after buffer exchange. Surprisingly, the presence of formaldehyde or quenched formaldehyde led to the stabilization of crosslinks after buffer exchange (Figure A.11). This is especially unexpected for the quenched formaldehyde condition, since the glycine, nonspecific DNA, excess HSF, and BSA is expected to absorb any crosslinks that are reversed during the thermal denaturing step. However, the results from the add back of quenched formaldehyde after the buffer exchange suggest that the crosslinks are in equilibrium between reversing and reforming during the denaturing step. This is problematic, since formation of new crosslinks could cause artifacts when isolating complexes from nuclear extracts. Even if the artifacts produced as such were not significant, the quenched formaldehyde is not present in the chromatin isolated from nuclei, and would have to be added back to the preparations, creating unpredictable results.

Most of the experiments described above used denaturing conditions of 95°C, whereas crosslinked complexes appear to be held intact at up to 48 - 54°C after buffer exchange (Figure A.10). It is therefore possible that lower temperatures could destabilize dsDNA enough to allow displacement of one strands by a more stable nucleic acid probe, particularly PNAs. However, I found that increasing reaction temperatures from 37°C to 47°C or 57°C, did not enhance PNA binding to naked dsDNA (data not shown).

Because of my inability to nail down specific conditions that could preserve formaldehyde crosslinks, and at the same time efficiently introduce an



**Figure A.10. Thermal stability of HSF/HSE crosslinked complexes after buffer exchange.** Crosslinked complexes that exchanged into a low ionic strength buffer (lanes 4-11) (see methods), are more labile in response to heat the reactions before buffer exchange (lanes 1-3).



**Figure A.11. Quenched formaldehyde maintains HSF/HSE complexes during thermal denaturing.** Crosslinked reactions were buffer exchanged (lanes 3-11), and then the indicated buffer were added back prior to thermal denaturing. Buffers: H: H<sub>2</sub>O, B: HSF binding buffer, G: 125mM Glycine, and F: formaldehyde. Results show that formaldehyde is required during the denaturing step if the quenched formaldehyde is removed by buffer exchange.

isolatable probe to the target sequence, this project was deemed unreasonable at this time. There are several more options to try for this project. For example, the region between two HSEs may be especially unavailable for probe binding due to steric hindrance caused by HSF molecules in the immediate vicinity. Alternative probe designs, or placement of probes could alleviate this problem. Also, as a proof of principle, one could consider targeting a highly abundant sequence, such as ribosomal promoter (present at ~250 copies in *Drosophila*), to increase the likelihood of success. One could also increase the amount of crosslinking, such that after thermal denaturing, enough crosslinking persists. However, the additional amount of time expected to pursue all options, along with the uncertainty of success, was considered too great and reckless for the timeline of a normal Ph.D. candidate.

#### **A.4 Concluding remarks.**

The results presented in this appendix describe my attempt to isolate protein/DNA complexes with DNA sequence-specific nucleic acid probes. Several problems were encountered in the course of experiments, including: 1) inconsistency of ChIP results following mock denaturation treatments, 2) insufficient binding of PNA probes to dsDNA, and 3) the apparent rearrangements of formaldehyde crosslinks during any thermal denaturing step. The last point is the most problematic, due to the potential of crosslinks being formed between proteins and DNA that did not exist at the time of original crosslinking in cells.

As I am writing this appendix, Robert Kingston's group apparently had success with this project, by using the human telomeres as a proof of principle (Dejardin and Kingston, 2009). Telomeres are present at ~100 copies per cell,

making them an attractive target over single copy genes. They used a 25 base nucleic acid probe that contained an LNA base at every other position. They also used a desthiobiotin tag, however, it was separated from the probe by a spacer of 108 bond lengths rather than the 30 that I used. Starting with  $\sim 2 \times 10^9$  HeLa S3 cells, they crosslinked with 3% formaldehyde, which is 10 times the amount that I used. For the nucleoprotein hybridizations, they used a buffer very similar to the one I used, but the denaturing condition was different. Instead of a one-time denaturing step that I used, they heated and cooled the sample in three cycles, with short incubation times  $\sim 2$ -6 minutes at the high temperature, followed by prolonged (1-2 hour) incubations at the 38°C. The first cycle was heated to 70°C, and the following two were heated to 60°C. This would apparently allow the DNA to breathe several times, thus increasing the chances of the LNA to find the target DNA and displace the strand of DNA. There are no results in the paper or supplemental data that describe any tests of the efficiency of the capture. Even so, the procedure appears to work since they identified many known telomere-binding proteins, and were able to validate novel ones.

Why did my plan not work? The standard oligonucleotide probe design was probably doomed from the start since the oligo would only be efficient if the dsDNA was completely denatured from the crosslinked complex. Complete denaturation seemed unlikely given that a network of crosslinks were likely to keep the strands in close proximity. The PNAs were promising, but I used a non-standard design that did not provide the cooperativity needed for efficient isolation. It is quite possible that the linker used for my probes was not long enough to relieve steric hindrance that blocks the interaction between the biotin tag and streptavidin beads. Finally, my results with ChIP and the in

in vitro crosslinking assay suggest that crosslinks are reversed or rearranged during thermal denaturing. More crosslinking, as mentioned above and performed by the Kingston group, could alleviate this problem by maintaining some of the original crosslinks. Finally, cyclic heating and cooling might be a more efficient method for probe invasion.

## REFERENCES

- Adelman, K., Marr, M. T., Werner, J., Saunders, A., Ni, Z., Andrulis, E. D. and Lis, J. T. (2005a).** Efficient Release from Promoter-Proximal Stall Sites Requires Transcript Cleavage Factor TFIIS. *Mol. Cell* **17**, 103-112.
- Adelman, K., Marr, M. T., Werner, J., Saunders, A., Ni, Z., Andrulis, E. D. and Lis, J. T. (2005b).** Efficient Release from Promoter-Proximal Stall Sites Requires Transcript Cleavage Factor TFIIS. *Mol Cell* **17**, 103-12.
- Aida, M., Chen, Y., Nakajima, K., Yamaguchi, Y., Wada, T. and Handa, H. (2006).** Transcriptional Pausing Caused by NELF Plays a Dual Role in Regulating Immediate-Early Expression of the junB Gene. *Mol. Cell. Biol.* **26**, 6094-6104.
- Aiyar, S. E., Sun, J. L., Blair, A. L., Moskaluk, C. A., Lu, Y. Z., Ye, Q. N., Yamaguchi, Y., Mukherjee, A., Ren, D. M., Handa, H. et al. (2004).** Attenuation of Estrogen Receptor Alpha-Mediated Transcription through Estrogen-Stimulated Recruitment of a Negative Elongation Factor. *Genes Dev* **18**, 2134-46.
- Akoulitchev, S., Makela, T. P., Weinberg, R. A. and Reinberg, D. (1995).** Requirement for TFIIH Kinase Activity in Transcription by RNA Polymerase II. *Nature* **377**, 557-60.
- Andrulis, E. D., Guzman, E., Doring, P., Werner, J. and Lis, J. T. (2000).** High-Resolution Localization of Drosophila Spt5 and Spt6 at Heat Shock

Genes in Vivo: Roles in Promoter Proximal Pausing and Transcription  
Elongation. *Genes Dev.* **14**, 2635-2649.

**Andrulis, E. D., Werner, J., Nazarian, A., Erdjument-Bromage, H., Tempst, P. and Lis, J. T.** (2002). The RNA Processing Exosome is Linked to Elongating RNA Polymerase II in *Drosophila*. *Nature* **420**, 837-841.

**Asturias, F. J.** (2004). RNA Polymerase II Structure, and Organization of the Preinitiation Complex. *Curr Opin Struct Biol* **14**, 121-9.

**Barberis, A., Pearlberg, J., Simkovich, N., Farrell, S., Reinagel, P., Bamdad, C., Sigal, G. and Ptashne, M.** (1995). Contact with a Component of the Polymerase II Holoenzyme Suffices for Gene Activation. *Cell* **81**, 359-368.

**Barski, A., Cuddapah, S., Cui, K., Roh, T. Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I. and Zhao, K.** (2007). High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell* **129**, 823-837.

**Bentley, D. L.** (1995). Regulation of Transcriptional Elongation by RNA Polymerase II. *Curr. Opin. Genet. Dev.* **5**, 210-216.

**Bertone, P., Stolc, V., Royce, T. E., Rozowsky, J. S., Urban, A. E., Zhu, X., Rinn, J. L., Tongprasit, W., Samanta, M., Weissman, S. et al.** (2004). Global Identification of Human Transcribed Sequences with Genome Tiling Arrays. *Science* **306**, 2242-2246.

**Blau, J., Xiao, H., McCracken, S., O'Hare, P., Greenblatt, J. and Bentley, D.** (1996). Three Functional Classes of Transcriptional Activation Domain. *Mol Cell Biol* **16**, 2044-55.

**Boeger, H., Bushnell, D. A., Davis, R., Griesenbeck, J., Lorch, Y., Strattan, J. S., Westover, K. D. and Kornberg, R. D.** (2005). Structural Basis of Eukaryotic Gene Transcription. *FEBS Lett* **579**, 899-903.

**Boehm, A. K., Saunders, A., Werner, J. and Lis, J. T.** (2003). Transcription Factor and Polymerase Recruitment, Modification, and Movement on dhsp70 in Vivo in the Minutes Following Heat Shock. *Mol. Cell. Biol.* **23**, 7628-7637.

**Brower-Toland, B., Wacker, D. A., Fulbright, R. M., Lis, J. T., Kraus, W. L. and Wang, M. D.** (2005). Specific Contributions of Histone Tails and their Acetylation to the Mechanical Stability of Nucleosomes. *J. Mol. Biol.* **346**, 135-146.

**Brown, S. A., Imbalzano, A. N. and Kingston, R. E.** (1996). Activator-Dependent Regulation of Transcriptional Pausing on Nucleosomal Templates. *Genes Dev* **10**, 1479-90.

**Bushnell, D. A., Westover, K. D., Davis, R. E. and Kornberg, R. D.** (2004). Structural Basis of Transcription: An RNA Polymerase II-TFIIB Cocrystal at 4.5 Angstroms. *Science* **303**, 983-8.

**Cai, H. and Luse, D. S.** (1987). Transcription Initiation by RNA Polymerase II in Vitro. Properties of Preinitiation, Initiation, and Elongation Complexes. *J Biol Chem* **262**, 298-304.

**Cameron, V. and Uhlenbeck, O. C.** (1977). 3'-Phosphatase Activity in T4 Polynucleotide Kinase. *Biochemistry* **16**, 5120-5126.

**Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M. C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C. et al. (2005).** The Transcriptional Landscape of the Mammalian Genome. *Science* **309**, 1559-1563.

**Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C. A., Taylor, M. S., Engstrom, P. G., Frith, M. C. et al. (2006).** Genome-Wide Analysis of Mammalian Promoter Architecture and Evolution. *Nat. Genet.* **38**, 626-635.

**Chapman, R. D., Heidemann, M., Albert, T. K., Mailhammer, R., Flatley, A., Meisterernst, M., Kremmer, E. and Eick, D. (2007).** Transcribing RNA Polymerase II is Phosphorylated at CTD Residue Serine-7. *Science* **318**, 1780-1782.

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

**Chen, B. S. and Hampsey, M. (2004).** Functional Interaction between TFIIB and the Rpb2 Subunit of RNA Polymerase II: Implications for the Mechanism of Transcription Initiation. *Mol Cell Biol* **24**, 3983-91.

**Chen, H. T., Warfield, L. and Hahn, S. (2007).** The Positions of TFIIF and TFIIE in the RNA Polymerase II Transcription Preinitiation Complex. *Nat. Struct. Mol. Biol.* **14**, 696-703.

**Cheng, C. and Sharp, P. A.** (2003). RNA Polymerase II Accumulation in the Promoter-Proximal Region of the Dihydrofolate Reductase and Gamma-Actin Genes. *Mol Cell Biol* **23**, 1961-7.

**Core, L. J. and Lis, J. T.** (2008). Transcription Regulation through Promoter-Proximal Pausing of RNA Polymerase II. *Science* **319**, 1791-1792.

**Core, L. J., Waterfall, J. and Lis, J.** (2008). Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. *Science*.

**Cramer, P.** (2004). RNA Polymerase II Structure: From Core to Functional Complexes. *Curr Opin Genet Dev* **14**, 218-26.

**Dejardin, J. and Kingston, R. E.** (2009). Purification of Proteins Associated with Specific Genomic Loci. *Cell* **136**, 175-186.

**Dvir, A.** (2002). Promoter Escape by RNA Polymerase II. *Biochim Biophys Acta* **1577**, 208-223.

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

**Egloff, S., O'Reilly, D., Chapman, R. D., Taylor, A., Tanzhaus, K., Pitts, L., Eick, D. and Murphy, S.** (2007). Serine-7 of the RNA Polymerase II CTD is Specifically Required for snRNA Gene Expression. *Science* **318**, 1777-1779.

**ENCODE Project Consortium, Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigo, R., Gingeras, T. R., Margulies, E. H., Weng, Z., Snyder, M., Dermitzakis, E. T. et al.** (2007). Identification and Analysis of Functional

Elements in 1% of the Human Genome by the ENCODE Pilot Project. *Nature* **447**, 799-816.

**Espinoza, C. A., Allen, T. A., Hieb, A. R., Kugel, J. F. and Goodrich, J. A.** (2004). B2 RNA Binds Directly to RNA Polymerase II to Repress Transcript Synthesis. *Nat. Struct. Mol. Biol.* **11**, 822-829.

**Faro-Trindade, I. and Cook, P. R.** (2006). Transcription Factories: Structures Conserved during Differentiation and Evolution. *Biochem. Soc. Trans.* **34**, 1133-1137.

**Fish, R. N. and Kane, C. M.** (2002). Promoting Elongation with Transcript Cleavage Stimulatory Factors. *Biochim Biophys Acta* **1577**, 287-307.

**Fivaz, J., Bassi, M. C., Pinaud, S. and Mirkovitch, J.** (2000). RNA Polymerase II Promoter-Proximal Pausing Upregulates c-Fos Gene Expression. *Gene* **255**, 185-194.

**Garcia-Martinez, J., Aranda, A. and Perez-Ortin, J. E.** (2004). Genomic Run-on Evaluates Transcription Rates for all Yeast Genes and Identifies Gene Regulatory Mechanisms. *Mol. Cell* **15**, 303-313.

**Gariglio, P., Bellard, M. and Chambon, P.** (1981a). Clustering of RNA Polymerase B Molecules in the 5' Moiety of the Adult Beta-Globin Gene of Hen Erythrocytes. *Nucleic Acids Res* **9**, 2589-98.

**Gariglio, P., Bellard, M. and Chambon, P.** (1981b). Clustering of RNA Polymerase B Molecules in the 5' Moiety of the Adult Beta-Globin Gene of Hen Erythrocytes. *Nucleic Acids Res.* **9**, 2589-2598.

- Gariglio, P., Buss, J. and Green, M. H.** (1974). Sarkosyl Activation of RNA Polymerase Activity in Mitotic Mouse Cells. *FEBS Lett.* **44**, 330-333.
- Giardina, C. and Lis, J. T.** (1993). Polymerase Processivity and Termination on Drosophila Heat Shock Genes. *J. Biol. Chem.* **268**, 23806-23811.
- Giardina, C., Perez-Riba, M. and Lis, J. T.** (1992). Promoter Melting and TFIID Complexes on Drosophila Genes in Vivo. *Genes Dev.* **6**, 2190-2200.
- Gilchrist, D. A., Nechaev, S., Lee, C., Ghosh, S. K., Collins, J. B., Li, L., Gilmour, D. S. and Adelman, K.** (2008). NELF-Mediated Stalling of Pol II can Enhance Gene Expression by Blocking Promoter-Proximal Nucleosome Assembly. *Genes Dev.* **22**, 1921-1933.
- Gilmour, D. S.** (2008). Promoter Proximal Pausing on Genes in Metazoans. *Chromosoma.*
- Gilmour, D. S. and Lis, J. T.** (1985). In Vivo Interactions of RNA Polymerase II with Genes of Drosophila Melanogaster. *Mol. Cell. Biol.* **5**, 2009-2018.
- Gilmour, D. S. and Lis, J. T.** (1986). RNA Polymerase II Interacts with the Promoter Region of the Noninduced hsp70 Gene in Drosophila Melanogaster Cells. *Mol. Cell. Biol.* **6**, 3984-3989.
- Gilmour, D. S., Pflugfelder, G., Wang, J. C. and Lis, J. T.** (1986). Topoisomerase I Interacts with Transcribed Regions in Drosophila Cells. *Cell* **44**, 401-407.

- Gnatt, A. L., Cramer, P., Fu, J., Bushnell, D. A. and Kornberg, R. D.** (2001). Structural Basis of Transcription: An RNA Polymerase II Elongation Complex at 3.3 Å Resolution. *Science* **292**, 1876-82.
- Greenleaf, A. L., Plagens, U., Jamrich, M. and Bautz, E. K.** (1978). RNA Polymerase B (Or II) in Heat Induced Puffs of *Drosophila* Polytene Chromosomes. *Chromosoma* **65**, 127-136.
- Gromak, N., West, S. and Proudfoot, N. J.** (2006). Pause Sites Promote Transcriptional Termination of Mammalian RNA Polymerase II. *Mol. Cell. Biol.* **26**, 3986-3996.
- Gu, W., Wind, M. and Reines, D.** (1996). Increased Accommodation of Nascent RNA in a Product Site on RNA Polymerase II during Arrest. *Proc Natl Acad Sci U S A* **93**, 6935-40.
- Guenther, M. G., Levine, S. S., Boyer, L. A., Jaenisch, R. and Young, R. A.** (2007). A Chromatin Landmark and Transcription Initiation at most Promoters in Human Cells. *Cell* **130**, 77-88.
- Guiguen, A., Soutourina, J., Dewez, M., Tafforeau, L., Dieu, M., Raes, M., Vandenhoute, J., Werner, M. and Hermand, D.** (2007). Recruitment of P-TEFb (Cdk9-Pch1) to Chromatin by the Cap-Methyl Transferase Pcm1 in Fission Yeast. *EMBO J.* **26**, 1552-1559.
- Hahn, S.** (2004). Structure and Mechanism of the RNA Polymerase II Transcription Machinery. *Nat Struct Mol Biol* **11**, 394-403.

- Han, J., Kim, D. and Morris, K. V.** (2007). Promoter-Associated RNA is Required for RNA-Directed Transcriptional Gene Silencing in Human Cells. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 12422-12427.
- Hassan, A. H., Prochasson, P., Neely, K. E., Galasinski, S. C., Chandy, M., Carrozza, M. J. and Workman, J. L.** (2002). Function and Selectivity of Bromodomains in Anchoring Chromatin-Modifying Complexes to Promoter Nucleosomes. *Cell* **111**, 369-379.
- Hawley, D. K. and Roeder, R. G.** (1985). Separation and Partial Characterization of Three Functional Steps in Transcription Initiation by Human RNA Polymerase II. *J. Biol. Chem.* **260**, 8163-8172.
- Hendrix, D. A., Hong, J. W., Zeitlinger, J., Rokhsar, D. S. and Levine, M. S.** (2008). Promoter Elements Associated with RNA Pol II Stalling in the Drosophila Embryo. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 7762-7767.
- Hengartner, C. J., Myer, V. E., Liao, S. M., Wilson, C. J., Koh, S. S. and Young, R. A.** (1998). Temporal Regulation of RNA Polymerase II by Srb10 and Kin28 Cyclin-Dependent Kinases. *Mol Cell* **2**, 43-53.
- Hieb, A. R., Baran, S., Goodrich, J. A. and Kugel, J. F.** (2006). An 8 Nt RNA Triggers a Rate-Limiting Shift of RNA Polymerase II Complexes into Elongation. *EMBO J.* **25**, 3100-3109.
- Holstege, F. C., Fiedler, U. and Timmers, H. T.** (1997). Three Transitions in the RNA Polymerase II Transcription Complex during Initiation. *Embo J* **16**, 7468-80.

- Howe, K. J.** (2002). RNA Polymerase II Conducts a Symphony of Pre-mRNA Processing Activities. *Biochim Biophys Acta* **1577**, 308-24.
- Huisinga, K. L. and Pugh, B. F.** (2004). A Genome-Wide Housekeeping Role for TFIID and a Highly Regulated Stress-Related Role for SAGA in *Saccharomyces Cerevisiae*. *Mol. Cell* **13**, 573-585.
- Iborra, F. J., Pombo, A., Jackson, D. A. and Cook, P. R.** (1996). Active RNA Polymerases are Localized within Discrete Transcription 'Factories' in Human Nuclei. *J. Cell. Sci.* **109 ( Pt 6)**, 1427-1436.
- Ioshikhes, I. P., Albert, I., Zanton, S. J. and Pugh, B. F.** (2006). Nucleosome Positions Predicted through Comparative Genomics. *Nat. Genet.* **38**, 1210-1215.
- Izban, M. G. and Luse, D. S.** (1992). The RNA Polymerase II Ternary Complex Cleaves the Nascent Transcript in a 3'----5' Direction in the Presence of Elongation Factor SII. *Genes Dev* **6**, 1342-56.
- Jackson, D. A., Iborra, F. J., Manders, E. M. and Cook, P. R.** (1998). Numbers and Organization of RNA Polymerases, Nascent Transcripts, and Transcription Units in HeLa Nuclei. *Mol. Biol. Cell* **9**, 1523-1536.
- Jang, M. K., Mochizuki, K., Zhou, M., Jeong, H. S., Brady, J. N. and Ozato, K.** (2005). The Bromodomain Protein Brd4 is a Positive Regulatory Component of P-TEFb and Stimulates RNA Polymerase II-Dependent Transcription. *Mol Cell* **19**, 523-34.

**Juven-Gershon, T., Hsu, J. Y., Theisen, J. W. and Kadonaga, J. T. (2008).** The RNA Polymerase II Core Promoter - the Gateway to Transcription. *Curr. Opin. Cell Biol.* **20**, 253-259.

**Kaneko, S. and Manley, J. L. (2005).** The Mammalian RNA Polymerase II C-Terminal Domain Interacts with RNA to Suppress Transcription-Coupled 3' End Formation. *Mol Cell* **20**, 91-103.

**Kao, S. Y., Calman, A. F., Luciw, P. A. and Peterlin, B. M. (1987).** Anti-Termination of Transcription within the Long Terminal Repeat of HIV-1 by Tat Gene Product. *Nature* **330**, 489-493.

**Kapranov, P., Cheng, J., Dike, S., Nix, D. A., Dutttagupta, R., Willingham, A. T., Stadler, P. F., Hertel, J., Hackermuller, J., Hofacker, I. L. et al. (2007a).** RNA Maps Reveal New RNA Classes and a Possible Function for Pervasive Transcription. *Science* **316**, 1484-1488.

**Kapranov, P., Willingham, A. T. and Gingeras, T. R. (2007b).** Genome-Wide Transcription and the Implications for Genomic Organization. *Nat. Rev. Genet.* **8**, 413-423.

**Katayama, S., Tomaru, Y., Kasukawa, T., Waki, K., Nakanishi, M., Nakamura, M., Nishida, H., Yap, C. C., Suzuki, M., Kawai, J. et al. (2005).** Antisense Transcription in the Mammalian Transcriptome. *Science* **309**, 1564-1566.

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

**Keene, R. G. and Luse, D. S.** (1999). Initially Transcribed Sequences Strongly Affect the Extent of Abortive Initiation by RNA Polymerase II. *J Biol Chem* **274**, 11526-34.

**Khorasanizadeh, S.** (2004). The Nucleosome: From Genomic Organization to Genomic Regulation. *Cell* **116**, 259-72.

**Kim, T. H., Barrera, L. O., Qu, C., Van Calcar, S., Trinklein, N. D., Cooper, S. J., Luna, R. M., Glass, C. K., Rosenfeld, M. G., Myers, R. M. et al.** (2005a). Direct Isolation and Identification of Promoters in the Human Genome. *Genome Res* **15**, 830-9.

**Kim, T. H., Barrera, L. O., Zheng, M., Qu, C., Singer, M. A., Richmond, T. A., Wu, Y., Green, R. D. and Ren, B.** (2005b). A High-Resolution Map of Active Promoters in the Human Genome. *Nature* **436**, 876-80.

**Kimura, K., Wakamatsu, A., Suzuki, Y., Ota, T., Nishikawa, T., Yamashita, R., Yamamoto, J., Sekine, M., Tsuritani, K., Wakaguri, H. et al.** (2006). Diversification of Transcriptional Modulation: Large-Scale Identification and Characterization of Putative Alternative Promoters of Human Genes. *Genome Res.* **16**, 55-65.

**Kininis, M., Chen, B. S., Diehl, A. G., Isaacs, G. D., Zhang, T., Siepel, A. C., Clark, A. G. and Kraus, W. L.** (2007). Genomic Analyses of Transcription

Factor Binding, Histone Acetylation, and Gene Expression Reveal Mechanistically Distinct Classes of Estrogen-Regulated Promoters. *Mol. Cell. Biol.* **27**, 5090-5104.

**Kininis, M., Isaacs, G. D., Core, L. J., Ha, N. and Kraus, W. L.** Post-Recruitment Regulation of RNA Polymerase II Directs Rapid Signaling Responses at the Promoters of Estrogen Target Genes. *Submitted*.

**Kireeva, M. L., Hancock, B., Cremona, G. H., Walter, W., Studitsky, V. M. and Kashlev, M.** (2005). Nature of the Nucleosomal Barrier to RNA Polymerase II. *Mol Cell* **18**, 97-108.

**Komarnitsky, P., Cho, E. J. and Buratowski, S.** (2000). Different Phosphorylated Forms of RNA Polymerase II and Associated mRNA Processing Factors during Transcription. *Genes Dev* **14**, 2452-60.

**Kouzarides, T.** (2007a). Chromatin Modifications and their Function. *Cell* **128**, 693-705.

**Kouzarides, T.** (2007b). SnapShot: Histone-Modifying Enzymes. *Cell* **131**, 822.

**Krumm, A., Hickey, L. B. and Groudine, M.** (1995). Promoter-Proximal Pausing of RNA Polymerase II Defines a General Rate-Limiting Step After Transcription Initiation. *Genes Dev* **9**, 559-72.

**Krumm, A., Meulia, T., Brunvand, M. and Groudine, M.** (1992). The Block to Transcriptional Elongation within the Human c-Myc Gene is Determined in the Promoter-Proximal Region. *Genes Dev* **6**, 2201-13.

**Kugel, J. F. and Goodrich, J. A.** (2000). A Kinetic Model for the Early Steps of RNA Synthesis by Human RNA Polymerase II. *J Biol Chem* **275**, 40483-91.

**Kugel, J. F. and Goodrich, J. A.** (2002). Translocation After Synthesis of a Four-Nucleotide RNA Commits RNA Polymerase II to Promoter Escape. *Mol Cell Biol* **22**, 762-73.

**Laybourn, P. J. and Dahmus, M. E.** (1990). Phosphorylation of RNA Polymerase IIA Occurs Subsequent to Interaction with the Promoter and before the Initiation of Transcription. *J Biol Chem* **265**, 13165-73.

**Lee, C., Li, X., Hechmer, A., Eisen, M., Biggin, M. D., Venters, B. J., Jiang, C., Li, J., Pugh, B. F. and Gilmour, D. S.** (2008). NELF and GAGA Factor are Linked to Promoter-Proximal Pausing at Many Genes in Drosophila. *Mol. Cell Biol.* **28**, 3290-3300.

**Lee, D. Y., Hayes, J. J., Pruss, D. and Wolffe, A. P.** (1993). A Positive Role for Histone Acetylation in Transcription Factor Access to Nucleosomal DNA. *Cell* **72**, 73-84.

**Lee, H., Kraus, K. W., Wolfner, M. F. and Lis, J. T.** (1992). DNA Sequence Requirements for Generating Paused Polymerase at the Start of hsp70. *Genes Dev.* **6**, 284-295.

**Li, B., Carey, M. and Workman, J. L.** (2007). The Role of Chromatin during Transcription. *Cell* **128**, 707-719.

**Li, Y. and Kornberg, R. D.** (1994). Interplay of Positive and Negative Effectors in Function of the C-Terminal Repeat Domain of RNA Polymerase II. *Proc Natl Acad Sci U S A* **91**, 2362-6.

**Li, Z., Van Calcar, S., Qu, C., Cavenee, W. K., Zhang, M. Q. and Ren, B.** (2003). A Global Transcriptional Regulatory Role for c-Myc in Burkitt's Lymphoma Cells. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 8164-8169.

**Lian, Z., Karpikov, A., Lian, J., Mahajan, M. C., Hartman, S., Gerstein, M., Snyder, M. and Weissman, S. M.** (2008). A Genomics Analysis of RNA Polymerase II Modification and Chromatin Architecture Related to 3' End RNA Polyadenylation. *Genome Res.*

**Lis, J.** (1998a). Promoter-Associated Pausing in Promoter Architecture and Postinitiation Transcriptional Regulation. *Cold Spring Harb Symp Quant Biol* **63**, 347-56.

**Lis, J.** (1998b). Promoter-Associated Pausing in Promoter Architecture and Postinitiation Transcriptional Regulation. *Cold Spring Harb. Symp. Quant. Biol.* **63**, 347-356.

**Lis, J. T., Mason, P., Peng, J., Price, D. H. and Werner, J.** (2000). P-TEFb Kinase Recruitment and Function at Heat Shock Loci. *Genes Dev.* **14**, 792-803.

**Liu, Y., Kung, C., Fishburn, J., Ansari, A. Z., Shokat, K. M. and Hahn, S.** (2004). Two Cyclin-Dependent Kinases Promote RNA Polymerase II

Transcription and Formation of the Scaffold Complex. *Mol Cell Biol* **24**, 1721-35.

**Lu, H., Zawel, L., Fisher, L., Egly, J. M. and Reinberg, D.** (1992). Human General Transcription Factor IIH Phosphorylates the C-Terminal Domain of RNA Polymerase II. *Nature* **358**, 641-5.

**Luse, D. S. and Samkurashvili, I.** (1998). The Transition from Initiation to Elongation by RNA Polymerase II. *Cold Spring Harb Symp Quant Biol* **63**, 289-300.

**Majovski, R. C., Khapersky, D. A., Ghazy, M. A. and Ponticelli, A. S.** (2005). A Functional Role for the Switch 2 Region of Yeast RNA Polymerase II in Transcription Start Site Utilization and Abortive Initiation. *J Biol Chem* **280**, 34917-23.

**Makela, T. P., Parvin, J. D., Kim, J., Huber, L. J., Sharp, P. A. and Weinberg, R. A.** (1995). A Kinase-Deficient Transcription Factor TFIIH is Functional in Basal and Activated Transcription. *Proc Natl Acad Sci U S A* **92**, 5174-8.

**Mancebo, H. S., Lee, G., Flygare, J., Tomassini, J., Luu, P., Zhu, Y., Peng, J., Blau, C., Hazuda, D., Price, D. et al.** (1997). P-TEFb Kinase is Required for HIV Tat Transcriptional Activation in Vivo and in Vitro. *Genes Dev* **11**, 2633-44.

**Mandal, S. S., Chu, C., Wada, T., Handa, H., Shatkin, A. J. and Reinberg, D.** (2004). Functional Interactions of RNA-Capping Enzyme with Factors that

Positively and Negatively Regulate Promoter Escape by RNA Polymerase II.

*Proc Natl Acad Sci U S A* **101**, 7572-7.

**Marciniak, R. A. and Sharp, P. A.** (1991). HIV-1 Tat Protein Promotes Formation of More-Processive Elongation Complexes. *EMBO J.* **10**, 4189-4196.

**Marshall, N. F. and Price, D. H.** (1995). Purification of P-TEFb, a Transcription Factor Required for the Transition into Productive Elongation. *J Biol Chem* **270**, 12335-8.

**Martinez-Balbas, M. A., Dey, A., Rabindran, S. K., Ozato, K. and Wu, C.** (1995). Displacement of Sequence-Specific Transcription Factors from Mitotic Chromatin. *Cell* **83**, 29-38.

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

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

**Morris, D. P., Michelotti, G. A. and Schwinn, D. A.** (2005). Evidence that Phosphorylation of the RNA Polymerase II Carboxyl-Terminal Repeats is Similar in Yeast and Humans. *J Biol Chem* **280**, 31368-77.

- Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. and Wold, B.** (2008). Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq. *Nat. Methods*.
- Muse, G. W., Gilchrist, D. A., Nechaev, S., Shah, R., Parker, J. S., Grissom, S. F., Zeitlinger, J. and Adelman, K.** (2007). RNA Polymerase is Poised for Activation Across the Genome. *Nat. Genet.* **39**, 1507-1511.
- Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M. and Snyder, M.** (2008). The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. *Science* **320**, 1344-1349.
- Nechaev, S. and Adelman, K.** (2008). Promoter-Proximal Pol II: When Stalling Speeds Things Up. *Cell. Cycle* **7**,.
- Nevado, J., Gaudreau, L., Adam, M. and Ptashne, M.** (1999). Transcriptional Activation by Artificial Recruitment in Mammalian Cells. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 2674-2677.
- Ng, H. H., Robert, F., Young, R. A. and Struhl, K.** (2003). Targeted Recruitment of Set1 Histone Methylase by Elongating Pol II Provides a Localized Mark and Memory of Recent Transcriptional Activity. *Mol Cell* **11**, 709-19.
- Nielsen, P. E.** (2004). PNA Technology. *Mol. Biotechnol.* **26**, 233-248.
- O'Brien, T., Hardin, S., Greenleaf, A. and Lis, J. T.** (1994a). Phosphorylation of RNA Polymerase II C-Terminal Domain and Transcriptional Elongation. *Nature* **370**, 75-77.

**O'Brien, T., Hardin, S., Greenleaf, A. and Lis, J. T.** (1994b).

Phosphorylation of RNA Polymerase II C-Terminal Domain and Transcriptional Elongation. *Nature* **370**, 75-7.

**O'Brien, T. and Lis, J. T.** (1993). Rapid Changes in Drosophila Transcription After an Instantaneous Heat Shock. *Mol. Cell. Biol.* **13**, 3456-3463.

**O'Brien, T., Wilkins, R. C., Giardina, C. and Lis, J. T.** (1995). Distribution of GAGA Protein on Drosophila Genes in Vivo. *Genes Dev.* **9**, 1098-1110.

**Orlicky, S. M., Tran, P. T., Sayre, M. H. and Edwards, A. M.** (2001).

Dissociable Rpb4-Rpb7 Subassembly of Rna Polymerase II Binds to Single-Strand Nucleic Acid and Mediates a Post-Recruitment Step in Transcription Initiation. *J Biol Chem* **276**, 10097-102.

**Orphanides, G., Lagrange, T. and Reinberg, D.** (1996). The General Transcription Factors of RNA Polymerase II. *Genes Dev* **10**, 2657-83.

**Oven, I., Brdickova, N., Kohoutek, J., Vaupotic, T., Narat, M. and Peterlin, B. M.** (2007). AIRE Recruits P-TEFb for Transcriptional Elongation of Target Genes in Medullary Thymic Epithelial Cells. *Mol. Cell. Biol.* **27**, 8815-8823.

**Pal, M. and Luse, D. S.** (2002). 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.

**Pal, M. and Luse, D. S.** (2003). 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 U S A* **100**, 5700-5.

- Pal, M., McKean, D. and Luse, D. S.** (2001). Promoter Clearance by RNA Polymerase II is an Extended, Multistep Process Strongly Affected by Sequence. *Mol Cell Biol* **21**, 5815-25.
- Pal, M., Ponticelli, A. S. and Luse, D. S.** (2005). The Role of the Transcription Bubble and TFIIB in Promoter Clearance by RNA Polymerase II. *Mol Cell* **19**, 101-10.
- Parsons, G. G. and Spencer, C. A.** (1997). Mitotic Repression of RNA Polymerase II Transcription is Accompanied by Release of Transcription Elongation Complexes. *Mol. Cell. Biol.* **17**, 5791-5802.
- Payne, J. M., Laybourn, P. J. and Dahmus, M. E.** (1989). The Transition of RNA Polymerase II from Initiation to Elongation is Associated with Phosphorylation of the Carboxyl-Terminal Domain of Subunit Ila. *J Biol Chem* **264**, 19621-9.
- Pei, Y., Du, H., Singer, J., Stamour, C., Granitto, S., Shuman, S. and Fisher, R. P.** (2006). Cyclin-Dependent Kinase 9 (Cdk9) of Fission Yeast is Activated by the CDK-Activating Kinase Csk1, Overlaps Functionally with the TFIIF-Associated Kinase Mcs6, and Associates with the mRNA Cap Methyltransferase Pcm1 in Vivo. *Mol Cell Biol* **26**, 777-88.
- Peterlin, B. M. and Price, D. H.** (2006). Controlling the Elongation Phase of Transcription with P-TEFb. *Mol. Cell* **23**, 297-305.
- Prelich, G.** (2002). RNA Polymerase II Carboxy-Terminal Domain Kinases: Emerging Clues to their Function. *Eukaryot Cell* **1**, 153-62.

- Price, D. H.** (2000). P-TEFb, a Cyclin-Dependent Kinase Controlling Elongation by RNA Polymerase II. *Mol Cell Biol* **20**, 2629-34.
- Proudfoot, N.** (2004). New Perspectives on Connecting Messenger RNA 3' End Formation to Transcription. *Curr. Opin. Cell Biol.* **16**, 272-278.
- Proudfoot, N. J.** (1989). How RNA Polymerase II Terminates Transcription in Higher Eukaryotes. *Trends Biochem. Sci.* **14**, 105-110.
- Ptashne, M. and Gann, A.** (1997). Transcriptional Activation by Recruitment. *Nature* **386**, 569-577.
- Purnell, B. A., Emanuel, P. A. and Gilmour, D. S.** (1994). TFIID Sequence Recognition of the Initiator and Sequences Farther Downstream in *Drosophila* Class II Genes. *Genes Dev* **8**, 830-42.
- Rada-Iglesias, A., Ameer, A., Kapranov, P., Enroth, S., Komorowski, J., Gingeras, T. R. and Wadelius, C.** (2008). Whole-Genome Maps of USF1 and USF2 Binding and Histone H3 Acetylation Reveal New Aspects of Promoter Structure and Candidate Genes for Common Human Disorders. *Genome Res.* **18**, 380-392.
- Rasmussen, E. B. and Lis, J. T.** (1993). In Vivo Transcriptional Pausing and Cap Formation on Three *Drosophila* Heat Shock Genes. *Proc. Natl. Acad. Sci. U. S. A.* **90**, 7923-7927.
- Rasmussen, E. B. and Lis, J. T.** (1995). Short Transcripts of the Ternary Complex Provide Insight into RNA Polymerase II Elongational Pausing. *J. Mol. Biol.* **252**, 522-535.

**Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E. et al. (2000).** Genome-Wide Location and Function of DNA Binding Proteins. *Science* **290**, 2306-2309.

**Robert, F., Pokholok, D. K., Hannett, N. M., Rinaldi, N. J., Chandy, M., Rolfe, A., Workman, J. L., Gifford, D. K. and Young, R. A. (2004).** Global Position and Recruitment of HATs and HDACs in the Yeast Genome. *Mol Cell* **16**, 199-209.

**Rodriguez, C. R., Cho, E. J., Keogh, M. C., Moore, C. L., Greenleaf, A. L. and Buratowski, S. (2000).** Kin28, the TFIIF-Associated Carboxy-Terminal Domain Kinase, Facilitates the Recruitment of mRNA Processing Machinery to RNA Polymerase II. *Mol Cell Biol* **20**, 104-112.

**Rougvie, A. E. and Lis, J. T. (1988a).** The RNA Polymerase II Molecule at the 5' End of the Uninduced hsp70 Gene of *D. Melanogaster* is Transcriptionally Engaged. *Cell* **54**, 795-804.

**Rougvie, A. E. and Lis, J. T. (1988b).** The RNA Polymerase II Molecule at the 5' End of the Uninduced hsp70 Gene of *D. Melanogaster* is Transcriptionally Engaged. *Cell* **54**, 795-804.

**Rougvie, A. E. and Lis, J. T. (1990).** Postinitiation Transcriptional Control in *Drosophila Melanogaster*. *Mol. Cell. Biol.* **10**, 6041-6045.

- Rudd, M. D., Izban, M. G. and Luse, D. S.** (1994). The Active Site of RNA Polymerase II Participates in Transcript Cleavage within Arrested Ternary Complexes. *Proc Natl Acad Sci U S A* **91**, 8057-61.
- Sandelin, A., Carninci, P., Lenhard, B., Ponjavic, J., Hayashizaki, Y. and Hume, D. A.** (2007). Mammalian RNA Polymerase II Core Promoters: Insights from Genome-Wide Studies. *Nat. Rev. Genet.* **8**, 424-436.
- Saunders, A., Core, L. J. and Lis, J. T.** (2006). Breaking Barriers to Transcription Elongation. *Nat. Rev. Mol. Cell Biol.* **7**, 557-567.
- Saunders, A., Werner, J., Andrulis, E. D., Nakayama, T., Hirose, S., Reinberg, D. and Lis, J. T.** (2003). Tracking FACT and the RNA Polymerase II Elongation Complex through Chromatin in Vivo. *Science* **301**, 1094-1096.
- Sawado, T., Halow, J., Bender, M. A. and Groudine, M.** (2003). The Beta - Globin Locus Control Region (LCR) Functions Primarily by Enhancing the Transition from Transcription Initiation to Elongation. *Genes Dev.* **17**, 1009-1018.
- Schilling, L. J. and Farnham, P. J.** (1994). Inappropriate Transcription from the 5' End of the Murine Dihydrofolate Reductase Gene Masks Transcriptional Regulation. *Nucleic Acids Res.* **22**, 3061-3068.
- Schones, D. E., Cui, K., Cuddapah, S., Roh, T. Y., Barski, A., Wang, Z., Wei, G. and Zhao, K.** (2008). Dynamic Regulation of Nucleosome Positioning in the Human Genome. *Cell* **132**, 887-898.

**Schuhmacher, M., Kohlhuber, F., Holzel, M., Kaiser, C., Burtscher, H., Jarsch, M., Bornkamm, G. W., Laux, G., Polack, A., Weidle, U. H. et al.** (2001). The Transcriptional Program of a Human B Cell Line in Response to Myc. *Nucleic Acids Res.* **29**, 397-406.

**Schwartz, B. E., Larochele, S., Suter, B. and Lis, J. T.** (2003). Cdk7 is Required for Full Activation of Drosophila Heat Shock Genes and RNA Polymerase II Phosphorylation in Vivo. *Mol. Cell. Biol.* **23**, 6876-6886.

**Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thastrom, A., Field, Y., Moore, I. K., Wang, J. P. and Widom, J.** (2006). A Genomic Code for Nucleosome Positioning. *Nature* **442**, 772-778.

**Seila, A. and et al.** (in press). Divergent Transcription from Active Promoters. *Science*.

**Serizawa, H., Conaway, J. W. and Conaway, R. C.** (1993). Phosphorylation of C-Terminal Domain of RNA Polymerase II is Not Required in Basal Transcription. *Nature* **363**, 371-4.

**Serizawa, H., Conaway, R. C. and Conaway, J. W.** (1992). A Carboxyl-Terminal-Domain Kinase Associated with RNA Polymerase II Transcription Factor Delta from Rat Liver. *Proc Natl Acad Sci U S A* **89**, 7476-80.

**Shendure, J. and Ji, H.** (2008). Next-Generation DNA Sequencing. *Nat. Biotechnol.* **26**, 1135-1145.

**Shermoen, A. W. and O'Farrell, P. H.** (1991). Progression of the Cell Cycle through Mitosis Leads to Abortion of Nascent Transcripts. *Cell* **67**, 303-310.

**Shopland, L. S., Hirayoshi, K., Fernandes, M. and Lis, J. T.** (1995). HSF Access to Heat Shock Elements in Vivo Depends Critically on Promoter Architecture Defined by GAGA Factor, TFIID, and RNA Polymerase II Binding Sites. *Genes Dev.* **9**, 2756-2769.

**Sims, R. J.,3rd, Belotserkovskaya, R. and Reinberg, D.** (2004). Elongation by RNA Polymerase II: The Short and Long of it. *Genes Dev* **18**, 2437-68.

**Sims, R. J.,3rd, Chen, C. F., Santos-Rosa, H., Kouzarides, T., Patel, S. S. and Reinberg, D.** (2005). Human but Not Yeast CHD1 Binds Directly and Selectively to Histone H3 Methylated at Lysine 4 Via its Tandem Chromodomains. *J Biol Chem* **280**, 41789-92.

**Sims, R. J.,3rd, Millhouse, S., Chen, C. F., Lewis, B. A., Erdjument-Bromage, H., Tempst, P., Manley, J. L. and Reinberg, D.** (2007). Recognition of Trimethylated Histone H3 Lysine 4 Facilitates the Recruitment of Transcription Postinitiation Factors and Pre-mRNA Splicing. *Mol. Cell* **28**, 665-676.

**Solomon, M. J., Larsen, P. L. and Varshavsky, A.** (1988). Mapping Protein-DNA Interactions in Vivo with Formaldehyde: Evidence that Histone H4 is Retained on a Highly Transcribed Gene. *Cell* **53**, 937-947.

**Strobl, L. J. and Eick, D.** (1992). Hold Back of RNA Polymerase II at the Transcription Start Site Mediates Down-Regulation of c-Myc in Vivo. *Embo J* **11**, 3307-14.

**Sultan, M., Schulz, M. H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina, T., Soldatov, A., Parkhomchuk, D. et al. (2008).**

A Global View of Gene Activity and Alternative Splicing by Deep Sequencing of the Human Transcriptome. *Science* **321**, 956-960.

**Sypes, M. A. and Gilmour, D. S. (1994).** Protein/DNA Crosslinking of a TFIID Complex Reveals Novel Interactions Downstream of the Transcription Start. *Nucleic Acids Res* **22**, 807-14.

**Trinklein, N. D., Aldred, S. F., Hartman, S. J., Schroeder, D. I., Otilar, R. P. and Myers, R. M. (2004).** An Abundance of Bidirectional Promoters in the Human Genome. *Genome Res.* **14**, 62-66.

**Tsukiyama, T., Becker, P. B. and Wu, C. (1994).** ATP-Dependent Nucleosome Disruption at a Heat-Shock Promoter Mediated by Binding of GAGA Transcription Factor. *Nature* **367**, 525-32.

**Ujvari, A. and Luse, D. S. (2006).** RNA Emerging from the Active Site of RNA Polymerase II Interacts with the Rpb7 Subunit. *Nat Struct Mol Biol* **13**, 49-54.

**Ujvari, A., Pal, M. and Luse, D. S. (2002).** RNA Polymerase II Transcription Complexes may Become Arrested if the Nascent RNA is Shortened to Less than 50 Nucleotides. *J Biol Chem* **277**, 32527-37.

**Vermeulen, M., Mulder, K. W., Denissov, S., Pijnappel, W. W., van Schaik, F. M., Varier, R. A., Baltissen, M. P., Stunnenberg, H. G., Mann, M. and Timmers, H. T. (2007).** Selective Anchoring of TFIID to Nucleosomes by Trimethylation of Histone H3 Lysine 4. *Cell* **131**, 58-69.

**Vester, B. and Wengel, J.** (2004). LNA (Locked Nucleic Acid): High-Affinity Targeting of Complementary RNA and DNA. *Biochemistry* **43**, 13233-13241.

**Wada, T., Takagi, T., Yamaguchi, Y., Ferdous, A., Imai, T., Hirose, S., Sugimoto, S., Yano, K., Hartzog, G. A., Winston, F. et al.** (1998a). DSIF, a Novel Transcription Elongation Factor that Regulates RNA Polymerase II Processivity, is Composed of Human Spt4 and Spt5 Homologs. *Genes Dev* **12**, 343-56.

**Wada, T., Takagi, T., Yamaguchi, Y., Watanabe, D. and Handa, H.** (1998b). Evidence that P-TEFb Alleviates the Negative Effect of DSIF on RNA Polymerase II-Dependent Transcription in Vitro. *Embo J* **17**, 7395-403.

**Wakaguri, H., Yamashita, R., Suzuki, Y., Sugano, S. and Nakai, K.** (2008). DBTSS: Database of Transcription Start Sites, Progress Report 2008. *Nucleic Acids Res.* **36**, D97-101.

**Wang, W., Carey, M. and Gralla, J. D.** (1992). Polymerase II Promoter Activation: Closed Complex Formation and ATP-Driven Start Site Opening. *Science* **255**, 450-3.

**Wang, X., Arai, S., Song, X., Reichart, D., Du, K., Pascual, G., Tempst, P., Rosenfeld, M. G., Glass, C. K. and Kurokawa, R.** (2008). Induced ncRNAs Allosterically Modify RNA-Binding Proteins in Cis to Inhibit Transcription. *Nature*.

- Wang, X., Lee, C., Gilmour, D. S. and Gergen, J. P.** (2007). Transcription Elongation Controls Cell Fate Specification in the Drosophila Embryo. *Genes Dev.* **21**, 1031-1036.
- Wang, Y. V., Tang, H. and Gilmour, D. S.** (2005). Identification in Vivo of Different Rate-Limiting Steps Associated with Transcriptional Activators in the Presence and Absence of a GAGA Element. *Mol Cell Biol* **25**, 3543-52.
- Weaver, J. R., Kugel, J. F. and Goodrich, J. A.** (2005). 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**, 39860-9.
- Weeks, J. R., Hardin, S. E., Shen, J., Lee, J. M. and Greenleaf, A. L.** (1993). Locus-Specific Variation in Phosphorylation State of RNA Polymerase II in Vivo: Correlations with Gene Activity and Transcript Processing. *Genes Dev* **7**, 2329-44.
- Wen, Y. and Shatkin, A. J.** (1999). Transcription Elongation Factor hSPT5 Stimulates mRNA Capping. *Genes Dev* **13**, 1774-9.
- Westover, K. D., Bushnell, D. A. and Kornberg, R. D.** (2004). Structural Basis of Transcription: Separation of RNA from DNA by RNA Polymerase II. *Science* **303**, 1014-6.
- Widom, J.** (2001). Role of DNA Sequence in Nucleosome Stability and Dynamics. *Q. Rev. Biophys.* **34**, 269-324.

**Wilhelm, B. T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I., Penkett, C. J., Rogers, J. and Bahler, J.** (2008). Dynamic Repertoire of a Eukaryotic Transcriptome Surveyed at Single-Nucleotide Resolution. *Nature*.

**Wilkins, R. C. and Lis, J. T.** (1997). Dynamics of Potentiation and Activation: GAGA Factor and its Role in Heat Shock Gene Regulation. *Nucleic Acids Res.* **25**, 3963-3968.

**Wilkins, R. C. and Lis, J. T.** (1998). GAGA Factor Binding to DNA Via a Single Trinucleotide Sequence Element. *Nucleic Acids Res.* **26**, 2672-2678.

**Wold, B. and Myers, R. M.** (2008). Sequence Census Methods for Functional Genomics. *Nat. Methods* **5**, 19-21.

**Wu, C. H., Lee, C., Fan, R., Smith, M. J., Yamaguchi, Y., Handa, H. and Gilmour, D. S.** (2005). Molecular Characterization of Drosophila NELF. *Nucleic Acids Res* **33**, 1269-79.

**Xiao, H., Friesen, J. D. and Lis, J. T.** (1995). Recruiting TATA-Binding Protein to a Promoter: Transcriptional Activation without an Upstream Activator. *Mol. Cell. Biol.* **15**, 5757-5761.

**Yamaguchi, Y., Inukai, N., Narita, T., Wada, T. and Handa, H.** (2002). Evidence that Negative Elongation Factor Represses Transcription Elongation through Binding to a DRB Sensitivity-Inducing factor/RNA Polymerase II Complex and RNA. *Mol Cell Biol* **22**, 2918-27.

- Yamaguchi, Y., Narita, T., Inukai, N., Wada, T. and Handa, H.** (2001). SPT Genes: Key Players in the Regulation of Transcription, Chromatin Structure and Other Cellular Processes. *J Biochem (Tokyo)* **129**, 185-91.
- Yamaguchi, Y., Takagi, T., Wada, T., Yano, K., Furuya, A., Sugimoto, S., Hasegawa, J. and Handa, H.** (1999). NELF, a Multisubunit Complex Containing RD, Cooperates with DSIF to Repress RNA Polymerase II Elongation. *Cell* **97**, 41-51.
- Yan, M. and Gralla, J. D.** (1997). Multiple ATP-Dependent Steps in RNA Polymerase II Promoter Melting and Initiation. *Embo J* **16**, 7457-67.
- Yang, Z., Yik, J. H., Chen, R., He, N., Jang, M. K., Ozato, K. and Zhou, Q.** (2005). Recruitment of P-TEFb for Stimulation of Transcriptional Elongation by the Bromodomain Protein Brd4. *Mol Cell* **19**, 535-45.
- Yankulov, K., Blau, J., Purton, T., Roberts, S. and Bentley, D. L.** (1994). Transcriptional Elongation by RNA Polymerase II is Stimulated by Transactivators. *Cell* **77**, 749-59.
- Young, R. A.** (1991). RNA Polymerase II. *Annu. Rev. Biochem.* **60**, 689-715.
- Zeitlinger, J., Stark, A., Kellis, M., Hong, J. W., Nechaev, S., Adelman, K., Levine, M. and Young, R. A.** (2007a). RNA Polymerase Stalling at Developmental Control Genes in the *Drosophila Melanogaster* Embryo. *Nat. Genet.* **39**, 1512-1516.
- Zeitlinger, J., Stark, A., Kellis, M., Hong, J. W., Nechaev, S., Adelman, K., Levine, M. and Young, R. A.** (2007b). RNA Polymerase Stalling at

Developmental Control Genes in the *Drosophila Melanogaster* Embryo. *Nat. Genet.* **39**, 1512-1516.

**Zhou, M., Lu, H., Park, H., Wilson-Chiru, J., Linton, R. and Brady, J. N.** (2006). Tax Interacts with P-TEFb in a Novel Manner to Stimulate Human T-Lymphotropic Virus Type 1 Transcription. *J. Virol.* **80**, 4781-4791.

**Zhu, Y., Pe'ery, T., Peng, J., Ramanathan, Y., Marshall, N., Marshall, T., Amendt, B., Mathews, M. B. and Price, D. H.** (1997). Transcription Elongation Factor P-TEFb is Required for HIV-1 Tat Transactivation in Vitro. *Genes Dev* **11**, 2622-32.

**Zorio, D. A. and Bentley, D. L.** (2004). The Link between mRNA Processing and Transcription: Communication Works both Ways. *Exp Cell Res* **296**, 91-7.