The Pennsylvania State University
The Graduate School
Eberly College of Science

ANALYSIS OF RNA POLYMERASE II ON THE DROSOPHILA GENOME
AND THE IMPACT OF THE TERMINATION FACTOR PCF11

A Dissertation in
Biochemistry, Microbiology and Molecular Biology
by
Ruopeng Fan

© 2009 Ruopeng Fan

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

May 2009
The dissertation of Ruopeng Fan was reviewed and approved* by the following:

David S. Gilmour  
Professor of Molecular and Cell Biology  
Dissertation Advisor  
Chair of Committee

B. Franklin Pugh  
Willaman Professor in Molecular Biology

Song Tan  
Associate Professor of Biochemistry and Molecular Biology

Paul Babitzke  
Professor of Biochemistry and Molecular Biology

Philip C. Bevilacqua  
Professor of Chemistry

Richard J. Frisque  
Professor of Molecular Virology  
Head of the Department of Biochemistry and Molecular Biology

*Signatures are on file in the Graduate School
ABSTRACT

The protein Pcf11 has been shown to dismantle RNA polymerase II elongation complexes and has been implicated in transcription termination in yeast, Drosophila and human cells. To obtain more insight into the termination process, I analyzed the distribution of Pol II in Drosophila cells at the 3’ end of three heat shock genes using permanganate genomic footprinting and at the 3’ end of all genes using ChIP-chip. The high resolution footprinting analysis at the 3’ end of the heat shock genes identified several regions where Pol II accumulates. The changes in Pol II distribution at the 3’ end of all three heat shock genes and more than 2000 other genes upon depletion of Pcf11 reveal defects in termination, indicating that Pcf11 functions as a general termination factor. In addition, there are changes in the Pol II distribution at the 5’ end of genes that indicate Pcf11 broadly impacts the behavior of Pol II as might be expected for a protein required for recycling Pol II. My analysis also identified a collection of genes whose transcription might be controlled by Pcf11-dependent premature termination. Three distinct distributions of Pol II are observed at the 3’ end of genes, suggesting the existence of alternative termination mechanisms. Finally, there is a striking decrease in the density of Pol II at the polyadenylation site that correlates with the paucity of nucleosomes in this region. Thus chromatin structure appears to impact the behavior of Pol II at the 3’ end of genes.
TABLE OF CONTENTS

LIST OF FIGURES ................................................................................................. viii
LIST OF TABLES .................................................................................................... xi
LIST OF ABBREVIATIONS .................................................................................... xii
ACKNOWLEDGEMENTS ....................................................................................... xiii

Chapter 1 Literature Review .............................................................................. 1

1.1 Transcription termination is important in eukaryotes ......................... 1
1.2 Transcription termination and mRNA 3'-end processing are linked processes .......................................................... 3
1.3 Working models of transcription termination in eukaryotes ........... 5
  1.3.1 Torpedo model ..................................................................................... 6
  1.3.2 Allosteric model ................................................................................ 7
  1.3.3 Allosteric-Torpedo model ................................................................. 8
1.4 Transcription termination and pausing ................................................ 10
1.5 Transcription termination and chromatin structure ......................... 12
1.6 Transcription termination at snRNA and snoRNA genes ............ 13
1.7 Overview of Pcf11 as a termination factor ........................................ 15
1.8 Thesis summary ......................................................................................... 17

Chapter 2 Materials and Methods ................................................................. 19

2.1 Preparation of antiserum against the Pol II subunit, Rpb3 ................ 19
2.2 RNAi depletion of Pcf11 in Drosophila S2R+ cells and Kc167 cells .... 20
2.3 Western blot analyses and antibodies .................................................. 22
2.4 Chromatin immunoprecipitation ......................................................... 23
2.5 ChIP-chip using Affymetrix Tiling Array ............................................ 26
2.6 Permanganate footprinting using Drosophila S2R+ cells ............... 29
2.7 Bioinformatics Analysis ......................................................................... 32
  2.7.1 Pol II and nucleosome distribution profile ................................ 33
  2.7.2 Polyadenylation signal and GAGA element identification .......... 34
  2.7.3 Clustering analysis .......................................................................... 35
  2.7.4 GO (Gene Ontology) analysis ......................................................... 35
2.8 Reverse transcription real time PCR ..................................................... 36

Chapter 3 Drosophila Pcf11 and Transcription Termination ....................... 37
3.1 Introduction..........................................................................................................................37

3.2 Results.....................................................................................................................................40

3.2.1 Depletion of Pcf11 alters the distribution of Pol II at the 3’ end of the hsp70 and hsp23 genes .........................................................................................................................40

3.2.2 High resolution analyses of the Pol II distribution at the 3’ end of the hsp70 and hsp23 genes reveal uneven patterns of elongation.................................................................46

3.2.3 Genome wide analysis reveals three distinct Pol II behaviors at the 3’ end of genes ..................................................................................................................................................51

3.2.4 Clustering analysis identifies 2611 genes whose termination is affected by the depletion of Pcf11 ..........................................................................................................................55

3.2.5 Evidence that Pcf11 represses transcription of some cellular genes by causing premature termination .......................................................................................................................58

3.2.6 Decreases in Pol II density at the 3’ end of genes correlates with apparent decreases in transcription at the promoters ..................................................................................................60

3.2.7 Depletion of Pcf11 does not cause termination defects at the U1_95Ca gene, an snRNA gene that lacks a polyadenylation signal .................................................60

3.2.8 The behavior of Pol II in the 3’ region appears to be linked to nucleosome positions .................................................................................................................................62

3.3 Discussion....................................................................................................................................65

3.3.1 Pcf11 functions in termination at many protein-encoding genes ......65

3.3.2 Alternate termination mechanisms are implicated by differences in 3’ Pol II distributions.................................................................................................................................65

3.3.3 Termination-coupled pausing downstream from the polyadenylation site ........................................................................................................................................68

3.3.4 Pcf11 depletion decreases transcription initiation on many genes ......72

3.3.5 Some Drosophila genes might be controlled by Pcf11-mediated premature termination .........................................................................................................................73

3.3.6 The function of Drosophila Pcf11 in termination may require a polyadenylation signal .................................................................................................................................74

Chapter 4 Genome Wide Distribution of Pol II at the 5’ End of Genes ....................76

4.1 Introduction..................................................................................................................................76

4.2 Results........................................................................................................................................78

4.2.1 Pol II is concentrated near transcription start sites (TSS) ......................78

4.2.2 Three distribution patterns of Pol II are evident at the 5’ end of genes .................................................................................................................................79

4.2.3 Comparison of NELF and GAGA factor distribution for each category of genes.................................................................................................................................82

4.2.4 Expression levels of category I, II, and III genes .................................85

4.2.5 Bulk and H2A.Z nucleosome distributions for category I, II and III genes .................................................................................................................................87
4.2.6 Nucleosome locations on individual genes with paused Pol II ........... 91
4.3 Discussion .................................................................................................. 93
  4.3.1 Pol II is concentrated at the promoter of thousands of genes .......... 93
  4.3.2 Promoter proximal pausing is a transient intermediate associated with active transcription ................................................................. 95
  4.3.3 GAGA factor might travel with elongating Pol II ......................... 95
  4.3.4 Pol II distribution on category I genes ............................................. 97
  4.3.5 Chromatin structure and promoter proximal pausing ................. 98

Bibliography ............................................................................................... 101

Appendix A  List of Primers Used For Real Time PCR .................................................. 115
Appendix B  List of Primers Used For LM-PCR ......................................................... 118
Appendix C  List of Other Primers ........................................................................... 120
  C.1 List of primers used for making the DNA templates of RNAi ............ 120
  C.2 List of primers used for making the tailed DNA templates (see Appendix D) ........................................................................................................ 121

Appendix D RNAi-Mediated Depletion of Pcf11 .............................................. 122
  D.1 RNAi mediated depletion of Pcf11 in Drosophila S2R+ and Kc167 cells .. 122
  D.2 Pcf11 depletion did not result in a decrease in Pcf11 ChIP signals ........ 125
  D.3 Termination defects of heat shock genes upon depletion of Pcf11 by Pcf11 RNAi 1 or Pcf11 RNAi 2 .............................................................. 127

Appendix E Termination Analysis of the U4_39B Gene ..................................... 130

Appendix F Biological Replicate of Quantified Permanganate Footprinting Experiments ................................................................................................. 133

Appendix G Verification of Changes in Pol II Density Detected by ChIP-chip Analysis ................................................................. 139

Appendix H Definitions of Statistical Terms ..................................................... 143
  H.1 Box plot definition .................................................................................. 143
  H.2 Wilcoxon-Mann-Whitney test ............................................................... 144

Appendix I In Vitro Transcription Assay ............................................................ 145
  I.1 Introduction ............................................................................................. 145
  I.2 Materials and Methods ........................................................................... 146
I.3 Results ............................................................................................................................. 149

Appendix J Permanganate Footprinting of the hsp70B and hsp23 genes............. 153

J.1 Permanganate footprinting of the hsp70B genes ........................................... 153
J.2 Permanganate footprinting of the hsp23 gene................................................. 155

Appendix K Changes in Pol II Density on Genes Controlled by Pcf11-mediated
Premature Termination ............................................................................................ 158

Appendix L Features Associated with the Genes in Category 1, 2 and 3............. 161

L.1 Position of the polyadenylation signal ............................................................ 161
L.2 Chromatin structure ...................................................................................... 162
L.3 GO analysis ...................................................................................................... 163
L.4 Changes in Pol II density upon Pcf11 depletion ........................................... 164
LIST OF FIGURES

Figure 1-1: Illustration of the allosteric model and the torpedo model of eukaryotic transcription termination............................................................6

Figure 1-2: Coupling the allosteric model with the torpedo model..................10

Figure 2-1: Correlation between fold enrichment of Pol II occupancy and matscore..............................................................................................28

Figure 3-1: Depletion of Pcf11 disrupts transcription termination for three heat shock genes..................................................................................42

Figure 3-2: Depleting Pcf11 increases read-through transcription for three heat shock genes and transcript uncleaved ratio at the polyadenylation site for hsp23........................................................................................................45

Figure 3-3: Genomic footprinting with permanganate detects Pol II at the 3’ end of hsp23, hsp70Aa and hsp70Ab genes......................................................47

Figure 3-4: Permanganate footprinting of heat shock genes detects termination defects upon Pcf11 depletion...............................................................49

Figure 3-5: Treeview displays of the permanganate reactivity of T residues of heat shock genes in heat shocked cells relative to non-heat shocked cells ....50

Figure 3-6: Clustering analysis of the Pol II distributions at the 3’ end of genes reveals three distinct Pol II behaviors.........................................................53

Figure 3-7: Different features associated with category 1, 2 and 3 genes .........54

Figure 3-8: Depletion of Pcf11 causes termination defects at many genes ....57

Figure 3-9: Changes in expression of selected genes in response to Pcf11 depletion ...............................................................................................59

Figure 3-10: Pcf11 depletion does not disrupt termination of the U1_95Ca gene ....61

Figure 3-11: The linkage between Pol II and nucleosome distributions at the 3’ end of genes .....................................................................................63

Figure 4-1: Pol II is concentrated at the 5’ end of genes ....................................79

Figure 4-2: Three distribution patterns of Pol II are evident at the 5’ end of genes...81
Figure 4-3: NELF-B distribution for three categories of genes

Figure 4-4: GAGA factor distribution for three categories of genes

Figure 4-5: Comparison of the expression levels of category I, II and III genes
and genes with Pol II levels below 5% FDR threshold within +/-1 kb of the
transcription start site

Figure 4-6: Bulk and H2A.Z nucleosome distributions are similar for category I,
II and III genes

Figure 4-7: The +1 nucleosome is not located at +145 for 6 individual genes that
have promoter proximal pausing

Figure D-1: Experimental design of RNAi-mediated depletion of Pcf11 in tissue
culture cells

Figure D-2: RNAi-mediated depletion of Pcf11 (I)

Figure D-3: RNAi-mediated depletion of Pcf11 (II)

Figure D-4: ChIP does not detect decreases in Pcf11 ChIP signals on the hsp23
gene after Pcf11 RNAi treatment

Figure D-5: Depletion of Pcf11 disrupts transcription termination for three heat
shock genes

Figure E-1: Pcf11 depletion does not disrupt termination of the U4_39B gene

Figure F-1: Results of a biological replicate of permanganate footprinting the
hsp23, hsp70Aa and hsp70Ab genes

Figure F-2: Results of a biological replicate of permanganate footprinting the
U1_95Ca and U4_39B genes

Figure F-3: Comparison of raw data from quantification of permanganate
footprinting experiments to smoothed curves

Figure G-1: Display (UCSC browser) of Pol II ChIP-chip data for 7 regions
where Pol II occupancy increases upon Pcf11 depletion

Figure G-2: Display (UCSC browser) of Pol II ChIP-chip data for 9 regions
where Pol II occupancy decreases upon Pcf11 depletion
Figure G-3: Real time PCR verifies the changes in Pol II occupancy detected by the genome wide analysis upon Pcf11 depletion..................................................142

Figure I-1: Schematic diagram of an in vitro transcription assay............................146

Figure I-2: Optimization of the in vitro transcription reaction..................................151

Figure I-3: Test of spontaneous release of transcripts from the DNA template by sequences downstream of the U1_95Ca and U4_39B genes........................................152

Figure J-1: Permanganate footprinting of the hsp70B genes ................................155

Figure J-2: Permanganate footprinting of the hsp23 gene........................................157

Figure K-1: Display (UCSC browser) of Pol II ChIP-chip data for 9 genes where Pol II occupancy increased in the regions upstream from the polyadenylation site upon Pcf11 depletion.................................................................160

Figure K-2: Real time PCR verified the increases of Pol II density upon Pcf11 depletion at the genes where premature termination possibly occurs ..................160

Figure L-1: The position of the polyadenylation signal is the same for category 1, 2 and 3 genes .................................................................162

Figure L-2: Nucleosome distributions at the 3’ end of category 1, 2 and 3 genes .....163

Figure L-3: Changes in Pol II density at the 3’ end of category 1, 2 and 3 genes......166
LIST OF TABLES

Table 4-1: GO analysis of category I, II and III genes ............................................... 87

Table 4-2: Comparison of the positions of the +1 nucleosomes determined by my analysis and Mavrich’s analysis .............................................................................. 91

Table L-1: GO analysis of category 1, 2, and 3 genes..................................................... 164
# LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>bp</td>
<td>base pair(s)</td>
</tr>
<tr>
<td>CF I/II</td>
<td>cleavage factor I/II</td>
</tr>
<tr>
<td>CF IA</td>
<td>cleavage factor IA</td>
</tr>
<tr>
<td>CF IB</td>
<td>cleavage factor IB</td>
</tr>
<tr>
<td>ChIP</td>
<td>chromatin immunoprecipitation</td>
</tr>
<tr>
<td>CID</td>
<td>CTD interaction domain</td>
</tr>
<tr>
<td>CPF</td>
<td>cleavage-polyadenylation factor</td>
</tr>
<tr>
<td>CPSF</td>
<td>cleavage-polyadenylation specificity factor</td>
</tr>
<tr>
<td>CstF</td>
<td>cleavage stimulatory factor</td>
</tr>
<tr>
<td>CTD</td>
<td>C-terminal domain</td>
</tr>
<tr>
<td>DRB</td>
<td>5,6-dichloro-1-β-D-ribofuranosyl benzimidazole</td>
</tr>
<tr>
<td>DSIF</td>
<td>DRB sensitivity inducing factor</td>
</tr>
<tr>
<td>FDR</td>
<td>false discovery rate</td>
</tr>
<tr>
<td>GO</td>
<td>gene ontology</td>
</tr>
<tr>
<td>HS</td>
<td>heat shock</td>
</tr>
<tr>
<td>LM-PCR</td>
<td>ligation-mediated PCR</td>
</tr>
<tr>
<td>min</td>
<td>minute(s)</td>
</tr>
<tr>
<td>MPE•Fe(II)</td>
<td>methidiumpropyl•Fe(II)</td>
</tr>
<tr>
<td>M/mM/µM</td>
<td>molar/millimolar/micromolar</td>
</tr>
<tr>
<td>mg/ng/µg</td>
<td>milligram(s)/nanograms(s)/micromgram(s)</td>
</tr>
<tr>
<td>ml/µl</td>
<td>milliliter(s)/microliter(s)</td>
</tr>
<tr>
<td>NELF</td>
<td>negative elongation factor</td>
</tr>
<tr>
<td>NHS</td>
<td>non-heat shock</td>
</tr>
<tr>
<td>PAP</td>
<td>polyadenylation polymerase</td>
</tr>
<tr>
<td>PCR</td>
<td>polymerase chain reaction</td>
</tr>
<tr>
<td>Pol II</td>
<td>RNA polymerase II</td>
</tr>
<tr>
<td>poly(A)</td>
<td>polyadenylation</td>
</tr>
<tr>
<td>P-TEFb</td>
<td>positive transcription elongation factor b</td>
</tr>
<tr>
<td>RACE</td>
<td>rapid amplification of cDNA ends</td>
</tr>
<tr>
<td>RNAi</td>
<td>RNA interference</td>
</tr>
<tr>
<td>RTPCR</td>
<td>reverse transcription PCR</td>
</tr>
<tr>
<td>sec</td>
<td>second(s)</td>
</tr>
<tr>
<td>snoRNA</td>
<td>small nucleolar RNA</td>
</tr>
<tr>
<td>snRNA</td>
<td>small nuclear RNA</td>
</tr>
<tr>
<td>TSS</td>
<td>transcription start site</td>
</tr>
</tbody>
</table>
ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. David Gilmour, for his guidance, patience and constant support throughout my Ph.D. studies. I would also like to thank my committee members for their suggestions to my research.

I thank Dr. Bryan Venters in Dr. Frank Pugh’s laboratory for helping me carry out the Pol II ChIP-chip experiments. I also thank Dr. Cizhong Jiang in Dr. Frank Pugh’s laboratory for writing the programs that I used to do my bioinformatics analyses.

I appreciate all the helps and support that I have received from the past and current members of the Gilmour lab. I had a great time working with them.

I would like to thank Ju-Chieh Wung particularly, for always being there no matter whether life is easy or difficult for me. I feel so lucky that I had her affection and continuous support during the last two years of my Ph.D. studies.

Lastly, I would like to thank my parents for their encouragement and endless love through all these years.
1.1 Transcription termination is important in eukaryotes

Transcription in eukaryotes can be regulated at three different stages: initiation, elongation and termination. Initiation involves the binding of RNA polymerase II (Pol II) to a gene’s promoter and the beginning of RNA synthesis. Elongation involves productive synthesis of nascent transcript by adding nucleotides to the 3’ end of the RNA in the active site of Pol II. The last stage of transcription, termination, involves stopping RNA synthesis, release of nascent transcript and dissociation of the Pol II from the DNA template. Tremendous progress has been made towards understanding the mechanism of transcription initiation and elongation, but the mechanism of termination is still poorly understood. Nevertheless, transcription termination is an essential step in regulation of gene expression because termination is required to prevent interference between adjacent transcription units (Proudfoot et al., 2002). The GAL10 gene in yeast provides an example of such interference. Either cis mutations in the polyadenylation signal or trans mutations of mRNA 3’-end cleavage factors resulted in defective termination of the GAL10 gene. The read-through transcripts of the GAL10 gene displaced Gal4p transcription factors from the promoter of the adjacent GAL7 gene and repressed its expression (Greger et al., 2000). Transcription interference has also been shown to repress transcription of a gene involved in serine biosynthesis in yeast (Martens et al.,
2004). In addition, termination is essential to recycle Pol II after Pol II has completed transcription of a gene (Cho et al., 1999; Kobor et al., 1999; Singh and Hampsey, 2007).

Pol II can dissociate from the DNA template not only at the end of a gene, but also during elongation through the body of a gene. The process whereby Pol II terminates transcription before full length transcripts are formed is called premature termination. In bacteria, attenuation and antitermination are related processes in which mutually exclusive patterns of RNA folding dictate whether RNA polymerase will continue transcription or prematurely terminate before completing transcription of an operon (Gollnick and Babitzke, 2002). These processes are widely used to control transcription of genes involved in amino acid biosynthesis in prokaryotes. It is less clear what role premature termination plays in regulating gene expression in eukaryotes and how prevalent this phenomenon is. Thousands of genes in human cells were found to possess a much greater density of Pol II at their 5’ than at their 3’ end (Guenther et al., 2007; Kim et al., 2005; Muse et al., 2007). In many cases, the elongation complex may be stalled when it is associated with the negative elongation factors NELF and DSIF (Aida et al., 2006; Wu et al., 2003). This stalling can be overcome by activator-dependent phosphorylation of Pol II and possibly NELF and DSIF by positive transcription elongation factor b (P-TEFb) (Peterlin and Price, 2006).

One example in eukaryotic cells in which premature termination regulates transcription is the HIV provirus. The Pol II density in the 5’ region of the provirus is much higher than that in the body of the gene. Short transcripts accumulate in the cytoplasm of the cells, indicating that the concentration of Pol II at the 5’ end of the provirus is partially due to premature termination (Cullen, 1991). It was recently shown
that Pcf11, a protein previously implicated in termination at the end of protein-coding genes, could be responsible for premature termination of HIV (Zhang et al., 2007a).

Another example of premature termination in eukaryotes is the negative feedback mechanism that Nrd1, a yeast protein, uses to attenuate its own expression. Nrd1, together with another yeast protein Nab3, directs transcription termination of sn/snoRNA transcripts, some mRNA transcripts, and some intergenic and anti-sense transcripts (Arigo et al., 2006b; Thiebaut et al., 2006). Nrd1 binds specific RNA sequences derived from a snoRNA terminator (Carroll et al., 2007). Corden and co-workers observed that mutations in the Nrd1 protein caused overexpression of the NRD1 gene. There are numerous Nrd1 binding sites in the 5’ region of the Nrd1 transcript. Corden’s group further showed that mutation of these binding sites leads to an increase in the amount of NRD1 message even in the presence of normal Nrd1 protein (Arigo et al., 2006a).

1.2 Transcription termination and mRNA 3’-end processing are linked processes

Transcription termination and mRNA 3’-end processing are tightly linked. In eukaryotes, all protein-encoding mRNAs contain a 3’-end extension of 100-250 adenosine nucleotides, with the exceptions of histone genes (Proudfoot et al., 2002). The formation of a polyadenylation tail relies on a conserved polyadenylation signal (AAUAAA) and a downstream sequence element (DSE), which is a U- or GU-rich motif. The cleavage of pre-mRNA occurs between the polyadenylation signal and the DSE, with a distance of 10-30 nucleotides from each of these two cis elements. Pol II continues transcribing for another 200-2000 nucleotides beyond the cleavage site and finally
dissociates from the DNA (Bauren et al., 1998; Dye and Proudfoot, 1999; Osheim et al., 1999). The upstream cleavage product is immediately polyadenylated (Colgan and Manley, 1997; Zhao et al., 1999). The downstream cleavage product has an uncapped phosphate at its 5’ end, and is highly unstable and rapidly degraded (Manley et al., 1982).

Several multisubunit proteins are involved in recognizing the polyadenylation signal and DSE, and catalyzing the cleavage and polyadenylation reactions. In mammals, the cleavage-polyadenylation specificity factor (CPSF) interacts with AAUAAA, while the cleavage stimulatory factor (CstF) interacts with DSE. The binding of CPSF and CstF appears to be cooperative (Colgan and Manley, 1997; Zhao et al., 1999). Two additional factors, cleavage factor I (CF I) and cleavage factor II (CF II), are essential for cleavage of pre-mRNA. The polyadenylation polymerase (PAP) catalyzes the addition of the polyadenylation tail. Poly(A) binding protein (PABP II) binds the growing polyadenylation tail and enhances the processivity of PAP (Wahle and Ruegsegger, 1999). Most of the components of these complexes have homologs in the yeast cleavage/polyadenylation machinery, comprised of the cleavage-polyadenylation factor (CPF), and cleavage factor IA and IB (CFIA and CFIB).

Analysis of yeast mutants provides direct evidence that transcription termination is functionally linked to cleavage/polyadenylation. Run-on analysis of the CYC1 gene in S. cerevisiae shows that mutation in CFIA (including Rna14, Rna 15 and Pcf11) but not mutation in polyadenylation factors disrupts transcription termination (Birse et al., 1998). The CPF subunits, Yhh1 (Dichtl et al., 2002) and Ssu72 (Ganem et al., 2003) have also been shown to be required for proper termination. The association of these factors with a gene is dependent on the polyadenylation signal in the nascent transcript (Kim et al.,
In vertebrates, some cleavage and polyadenylation factors can be recruited to the gene upstream from the polyadenylation signal and become active after the polyadenylation signal is transcribed (Venkataraman et al., 2005). Direct support for the connection between termination and cleavage/polyadenylation came from studies demonstrating that the polyadenylation signal and DSE are also the cis elements required for normal termination in mammals (Connelly and Manley, 1988; Logan et al., 1987; Whitelaw and Proudfoot, 1986). Several cleavage-polyadenylation-directing signals have also been shown to function in transcription termination in yeast (Russo, 1995). Consistent with these data, it was also demonstrated that polyadenylation signal strength, as measured by the efficiency of RNA cleavage, is directly correlated to termination efficiency (Osheim et al., 1999).

### 1.3 Working models of transcription termination in eukaryotes

In eukaryotes, a discrete signal in the DNA that specifies the location of the termination site has not been identified (Rosonina et al., 2006). Instead, transcription termination seems to be a stochastic process whose probability increases greatly once Pol II has transcribed past the polyadenylation signal. Two models have been offered to explain the connection between termination and cleavage/polyadenylation: the torpedo model and allosteric model (Figure 1-1).
1.3.1 Torpedo model

The torpedo model proposes that the cleavage of the nascent transcript at the polyadenylation site generates an unprotected 5’ end of the 3’ RNA product which is still tethered to the elongation complex. This 5’ end provides an entry point for a 5’→3’ exonuclease to degrade the RNA. When the nuclease reaches the polymerase, it somehow destabilizes the elongation complex, causing termination (Connelly and Manley, 1988). Evidence in support of this model came from the identification of yeast Rat1 and human Xrn2, both 5’ →3’ RNA exonucleases, as necessary for efficient termination (Kim et al., 2004b; West et al., 2004). Mutation in Rat1, as well as RNAi against Xrn2, did not affect
the cleavage efficiency of the nascent transcript. However, disruption of these proteins in yeast or mammalian cells resulted in Pol II transcribing beyond regions where it normally terminated.

1.3.2 Allosteric model

The allosteric model, also known as the anti-terminator model, posits that transcription of the polyadenylation signal causes a conformational change in the elongation complex, which renders Pol II more prone to termination. The conformational change could be either release of an anti-termination factor or recruitment of a termination factor (Logan et al., 1987).

Recent work by Zhang et al. provides support for the allosteric model. Yeast and *Drosophila* Pcf11 were found to cause Pol II to dissociate from a DNA template by associating with the C-terminal domain of the largest subunit of Pol II (CTD) and the nascent transcript (Zhang et al., 2005; Zhang and Gilmour, 2006). The CTD is composed of repeated heptads with the consensus sequence YSPTSPS (25 or 26 tandem repeats in yeast, 42 in *Drosophila*, 52 in mammals) (Corden, 1990). This sequence goes through cycles of phosphorylation and dephosphorylation throughout transcription and is thought to coordinate events such as transcription initiation, elongation and pre-mRNA processing (Hirose and Manley, 2000; Proudfoot and O'Sullivan, 2002; Zorio and Bentley, 2004). It was previously found that the CTD is required for pre-mRNA cleavage and transcription termination in animal cells (McCracken et al., 1997). Pcf11 alone dismantles the elongation complex by a mechanism that is dependent on the CTD but
independent of nucleotide hydrolysis. Because Pcf11 is concentrated at the 3’ end of genes, it might be recruited by the polyadenylation signal in the nascent transcript or other cleavage/polyadenylation factors (Gross and Moore, 2001; Kim et al., 2004a; Zhang and Gilmour, 2006).

In addition to Pcf11, other proteins have been implicated for the allosteric model. Sub1 and Npl3 in yeast have anti-terminator activity. The former interacts with Rna15, a subunit of CFIA, and the latter competes for nascent transcript binding with cleavage/polyadenylation factors. Release of Sub1 and Npl3 might increase the termination efficiency (Bucheli and Buratowski, 2005; Calvo and Manley, 2001). Factors with positive roles in termination include Mbp1 and Grs1, both identified through screens for 3’-end formation defects in yeast (Aranda and Proudfoot, 2001; Magrath and Hyman, 1999).

1.3.3 Allosteric-Torpedo model

The torpedo model and the allosteric model are not mutually exclusive. In fact, several pieces of evidence have challenged the torpedo model as the only explanation for termination. First, alleles of Ssu72 and Pcf11 that inhibit cleavage do not disrupt termination (Dichtl et al., 2002; He et al., 2003; Sadowski et al., 2003). Second, electron microscopy failed to detect elongation complexes downstream from genes that would represent elongation complexes with cleaved transcripts (Bauren et al., 1998; Osheim et al., 1999; Osheim et al., 2002). It was argued that cleavage does not precede termination, but it is also possible that the degradation of the nascent transcript occurs so rapidly that
intermediates with cleaved transcripts escape detection. In addition, polyadenylation-dependent termination has been reconstituted in a crude nuclear extract and such termination occurs without transcript cleavage (Tran et al., 2001). That purified elongation complexes can survive extensive digestion by RNase without dissociation from the DNA template (Gu et al., 1996) also questions the possibility that Rat1 destabilizes Pol II solely through its exonuclease activity.

Recent results from Bentley and colleagues suggest a new hybrid model, the allosteric-torpedo model (Figure 1-2), in which Rat1 is an integral component of the cleavage/polyadenylation apparatus and its association with other 3’-end processing factors, such as Pcf11, facilitates termination (Luo et al., 2006). In this study, Luo et al. found that the recruitment of Rat1 to the 3’ end of the ADH4 gene was reduced by a Pcf11 mutation. Reciprocally, mutation of Rat1 also reduced the recruitment of Pcf11 and Rna15 to the 3’ end of ADH4 gene. Luo et al. also demonstrated that the exonuclease activity was not sufficient for termination since a cytoplasmic yeast exonuclease, Xrn1, could degrade the 3’ product of transcript cleavage when directed into the nucleus by fusion with a nuclear localization signal, but could not rescue the termination defect in rat1-1 cells. This puts Pcf11, with its activity to dismantle the elongation complex in vitro, at the center of the transcript release process. Rat1 and Pcf11 might be bridged to each other by Rtt103, a protein that binds both of these proteins (Kim et al., 2004b). It remains unclear how Pcf11 causes Pol II to dissociate from the DNA template.
1.4 Transcription termination and pausing

Pausing has been implicated in transcription termination of Pol II (Enriquez-Harris et al., 1991). Transcriptional pause sites located downstream of the polyadenylation site might enhance the recognition of the polyadenylation signal by proteins or facilitate cleavage/polyadenylation (Yonaha and Proudfoot, 1999) thereby promoting termination. In *S. pombe*, the DSE is important for efficient 3’-end processing and transcription termination of the *ura4* and *nmt2* genes. Nuclear run-on analysis showed increased Pol II density over these DSEs, suggesting they were pause sites. However, no sequence homology among the DSEs was found (Aranda and Proudfoot, 1999; Birse et al., 1997). A G-rich element is located between two closely spaced human complement genes, C2 and factor B (Ashfield et al., 1991). The G-rich element binds to a zinc finger protein, MAZ. This DNA binding protein was first thought to cause Pol II to
pause, thus facilitating termination. However, later studies showed that the G-rich
element by itself causes Pol II to pause (Yonaha and Proudfoot, 1999). Notably pausing
is also important for termination of transcription by Pol I and Pol III. Yeast Reb1p
associates with a region downstream of a T-rich sequence and causes Pol I to pause
(Lang and Reeder, 1995). The lac repressor can substitute for Reb1p in vitro to block
elongation and cause termination (Jeong et al., 1995). Two RNA Pol III-specific subunits,
C53 and C37, slow the rate of elongation, so they could cause Pol III to pause long
enough at particular sites for efficient dissociation of Pol III (Landrieux et al., 2006).

Although pausing increases the termination efficiency, a consensus pause site has
not been identified at the 3’ end of genes. The only sequence element generally required
for termination is the polyadenylation signal itself. Available evidence indicates that the
pausing activity required for termination is exclusively a function of the polyadenylation
signal and does not involve additional nucleic acid sequences or the CTD of Pol II (Nag
et al., 2006; Orozco et al., 2002; Park et al., 2004; Tran et al., 2001).

Nag et al. further explored the mechanism by which the hexamer AAUAAA in
the nascent transcript causes pausing (Nag et al., 2007). Their data showed that CPSF
interacts with the body of the Rpb1 subunit of Pol II through its 30 kDa subunit, CPSF30.
This interaction, together with the recognition of the polyadenylation signal by another
subunit of CPSF, CPSF160, is sufficient to cause Pol II to pause, even in the absence of
the CTD and CstF. Hence, pausing is functionally a distinct step from 3’-end processing
and termination, both of which require the CTD. However, the pause induced by CPSF
and AAUAAA could facilitate assembly of the cleavage/polyadenylation machinery and
recruitment of termination factors (Nag et al., 2007).
It is unclear how the binding of CPSF30 to Rpb1 and of CPSF160 to the polyadenylation signal cause pausing. One possibility is that such binding leads to a conformational change of Pol II since the inhibitory effect of the polyadenylation signal on Pol II elongation can last over 1kb downstream from the hexamer (Nag et al., 2006); (Orozco et al., 2002). Another possibility is that CPSF triggers the release of positive elongation factors. In yeast, ChIP experiments indicate that several positive elongation factors dissociate from the gene downstream from the polyadenylation signal (Kim et al., 2004a).

1.5 Transcription termination and chromatin structure

The chromatin structure across the termination region can affect the termination efficiency. Cells lacking the Chd1 chromatin remodeling factor showed defects in both chromatin structure at the 3’ end of genes and termination itself (Alen et al., 2002). Since Chd1 has been implicated to impede transcription elongation and to enhance transcription pausing (Woodage et al., 1997), it is possible that chromatin remodeling at sites downstream of polyadenylation site might induce or facilitate pausing or template release by Pol II. Transcription by Pol II in vitro can be blocked by a single nucleosome at low ionic strength. Passage of Pol II causes a quantitative loss of one H2A/H2B dimer without altering the location of the nucleosome (Kireeva et al., 2002). The barrier formed by nucleosomes has polarity, which is determined by both the strength of the histone-DNA interaction and the location of the high-affinity DNA region within the nucleosome (Bondarenko et al., 2006). Observations arising from recent high-resolution genome wide
H2A.Z nucleosome mapping (Mavrich et al., 2008b), as well as some other studies (Brown et al., 1996; Brown and Kingston, 1997; Carey et al., 2006), indicate that the +1 nucleosome near the transcript start site (TSS) might contribute to promoter proximal pausing. However, to what extent nucleosomes can cause transient pausing at the 3’ end and facilitate termination is yet to be answered.

1.6 Transcription termination at snRNA and snoRNA genes

In addition to protein-coding genes, Pol II also transcribes some non-coding RNAs, including small nuclear RNAs (snRNAs) and small nucleolar RNAs (snoRNAs). snRNAs are involved in a variety of important processes such as RNA splicing, regulation of transcription factors and Pol II (Blencowe, 2002; Espinoza et al., 2004; Lacadie and Rosbash, 2005; Wassarman and Steitz, 1991). snRNA genes differ from protein-coding genes in that their promoters contain a gene-specific proximal sequence element (PSE), their transcripts are not spliced, and their transcripts are not polyadenylated (Egloff et al., 2008). A 3’ box with the sequence GTTTTN₀.

₃AAAPuNNAGA is located 9-19 nucleotides downstream of the end of the human snRNA transcript (Ciliberto et al., 1986; Hernandez, 1985). Although initial studies suggested that the 3’ box is the functional element causing termination on human U1 snRNA genes (Kunkel and Pederson, 1985), Cuello et al. subsequently provided evidence that the 3’ box was not an efficient transcriptional terminator but a processing element (Cuello et al., 1999). Cuello et al. further suggested that the abrupt termination at the human U1 snRNA gene near the 3’ box resulted from the concerted action of the 3’ box
and a downstream sequence element that bound a protein which could be detected by genomic footprinting (Cuello et al., 1999). snoRNAs, a class of snRNAs, are required for maturation of pre-rRNAs (Kim et al., 2006). The 3’ end of snoRNA is cleaved by the CFIA complex without being subsequently polyadenylated (Fatica et al., 2000; Morlando et al., 2002).

Several protein factors have been implicated in the 3’-end processing and termination of snRNA and snoRNA genes. Mutation of the yeast proteins Ssu72 or Sen1 (a helicase) led to termination defects for snoRNA genes, as well as some protein-coding genes (Ganem et al., 2003; Kim et al., 2006; Rasmussen and Culbertson, 1998; Steinmetz et al., 2001; Ursic et al., 1997). The sequence-specific RNA binding proteins, Nrd1 and Nab3, also function as termination factors for snRNA and snoRNA genes in yeast (Steinmetz et al., 2001). These two proteins form a complex with Sen1 and the nuclear exosome, a collection of RNases capable of trimming the 3’ end of snRNA precursors (Butler, 2002; Mitchell et al., 1997). The CTD of Pol II plays an important role in transcription termination of snRNA genes like in termination of protein-encoding genes. Nrd1 has an amino terminus that interacts with the CTD of Pol II and this interaction is essential for its termination function (Steinmetz et al., 2001). These results suggest similarity between the termination mechanisms of snoRNA, snRNA genes, and protein-encoding genes.
1.7 Overview of Pcf11 as a termination factor

Pcf11 is a 3’-end processing factor that has been characterized in yeast, *Drosophila*, and human. Yeast Pcf11 has a molecular weight of 70 kDa and is a component of the CF I complex (Amrani et al., 1997). Sadowski et al. observed that a mutation in Pcf11 which disrupts Pcf11’s CTD binding activity but not its cleavage activity caused a defect in termination of the *CYC1* gene (Sadowski et al., 2003). However later evidence suggests that the same Pcf11 mutation previously demonstrated to affect only cleavage in fact caused both cleavage and termination defects (Kim et al., 2006; Luo et al., 2006). This discrepancy, though, could be due to different model genes used by these studies. Additional evidence indicates that the CID (CTD interaction domain) of yeast and *Drosophila* Pcf11 can dissociate Pol II from the DNA template *in vitro* by bridging the CTD with the nascent transcript (Zhang et al., 2005; Zhang and Gilmour, 2006). Human Pcf11, first characterized by de Vries et al. in 2000, (de Vries et al., 2000) is a subunit of the CF II complex. Human Pcf11 and another subunit of CF II, Clp1, bind to CstF and CPSF separately (de Vries et al., 2000; Qu et al., 2007). During the purification of the CF II complex, Pcf11 corresponded to a group of polypeptides with molecular weights ranging from 140-200 kDa, possibly due to degradation of a large protein. Depletion of Pcf11 from human cells impaired polyadenylation signal-dependent termination, and degradation of the nascent transcript after cleavage at the polyadenylation site (West and Proudfoot, 2008). The latter result is consistent with Luo’s finding in yeast that Pcf11 is involved in recruiting the RNA exonuclease, Rat1, to the 3’ end of a gene (Luo et al., 2006).
The investigation of the role of *Drosophila* Pcf11 in termination was initiated by us. In addition to our finding that *Drosophila* Pcf11 (dPcf11) causes termination both *in vivo* and *in vitro*, our results also suggest that pausing may be required for Pcf11-dependent termination since Pcf11 was only able to dismantle the elongation complex at a low nucleotide concentration *in vitro*. We proposed that the low nucleotide concentration promoted pausing (Zhang *et al.*, 2005; Zhang and Gilmour, 2006). The ability of Pcf11 to bind to Pol II and terminate transcription makes it an excellent candidate for the allosteric model.

Many questions remain. First, is Pcf11 a universal termination factor for protein-encoding genes and non-coding genes, given that these two types of genes share some but not all 3’-end processing factors (Rosonina *et al.*, 2006). Second, what is the role of pausing in Pcf11-dependent termination and how prevalent is such pausing in the *Drosophila* genome? Third, if Pcf11 dismantles the elongation complex in an allosteric way, what are the conformational changes it brings to Pol II? Forth, how important is the phosphorylation cycle of the CTD in transcription termination? Fifth, is Pcf11 involved in other transcription regulatory processes such as premature termination? Sixth, what additional proteins assist in the process of termination?

dPcf11, unlike its homolog in yeast, has several isoforms in *Drosophila* tissue culture cells and adult flies. The initial characterization of dPcf11 by Zhiqiang Zhang focused on a 60 kDa isoform. I subsequently discovered another isoform of dPcf11 by western analysis that was larger than the Rpb1 subunit of Pol II (200 kDa). It is likely that Zhang missed the large form of dPcf11 because of the acrylamide gel system he was using. The antibody we used to do western analysis was raised against the first 283 amino
acids at the N-terminus of dPcf11 predicted by Flybase (http://flybase.org/). This 283 amino acid region includes the Pcf11 CTD interaction domain, which is highly conserved among all eukaryotes. 5’ RACE and 3’ RACE confirms that both dPcf11 isoforms exist in Drosophila S2 cells (Ju-Chieh Wung, unpublished data) and that each isoform is encoded by multiple RNAs with different polyadenylation sites. The larger isoform of Pcf11 has a predicted molecular weight of 200 kDa. It is generally thought that the choice of polyadenylation site is related to tissue types and developmental stages (Beaudoing and Gautheret, 2001). In fact, our Pcf11 antibody can only detect 200 kDa dPcf11 but not 60 kDa Pcf11 in a Drosophila embryo extract (Ju-Chieh Wung, unpublished). Except for the 59 amino acids at its C-terminus, the 60 kDa dPcf11 is mostly identical to the N-terminal portion of the 200 kDa dPcf11. Regions close to the C-terminal end of 200 kDa dPcf11 have some similarity to its homologs in human and mouse. Our understanding of dPcf11 function will be greatly advanced by fully characterizing the functions of different isoforms of the protein.

1.8 Thesis summary

My thesis focuses on the functional analysis of Drosophila Pcf11 in regulating transcription termination. I describe a high resolution permanganate footprinting analysis of the behavior of Pol II at the 3’ end of several heat shock genes and a genome wide study of Pol II distribution in the region where termination occurs. Pol II was found to accumulate at several locations downstream from the polyadenylation site of the heat shock genes, suggesting that termination is preceded by the pausing of Pol II. Three
distinct Pol II distribution patterns were discovered at the 3’ end of protein-encoding genes, indicating the existence of alternate termination mechanisms in *Drosophila*. By using RNAi-mediated depletion of Pcf11 in *Drosophila* cells, I demonstrated that *Drosophila* Pcf11 is important for proper transcription termination on more than 2500 genes. In addition, I provide some evidence that *Drosophila* Pcf11 might be involved in the recycling of Pol II and premature termination. Lastly, a comparison of the Pol II distribution to the nucleosome distribution at the 3’ end of genes indicates a correlation between the behavior of Pol II and chromatin structure in this region.
2.1 Preparation of antiserum against the Pol II subunit, Rpb3

A pilot purification of His-Rpb3 under native conditions suggested that most of the protein was insoluble when expressed in *E. Coli*, so a different strategy employing denaturing conditions was used to obtain purified His-Rpb3. BL21 (DE3) *E. coli* cells containing His-Rpb3 plasmid were grown overnight at 37°C in LB (Ampicillin, 100 µg/ml) media and diluted into 500 ml of fresh media with the OD<sub>600</sub> equal to 0.1. The culture was grown at 37°C with vigorous shaking until the OD<sub>600</sub> reached 0.6. IPTG was then added to the culture to a final concentration of 0.4 mM. The culture was further grown at 37°C for 3 hr to induce expression of His-Rpb3 protein. After induction, the cell culture was chilled on ice for 10 min, followed by centrifugation at 6000 rpm (GSA rotor) for 15 min. The cell pellet was resuspended in 125 ml of pre-chilled 20 mM Tris-Cl, pH 8.0, and then spun down again at 6000 rpm (GSA rotor) for 15 min. The cell pellet was quick frozen with liquid nitrogen and stored at -80°C overnight. Frozen cells were thawed on ice for 15 min, and resuspended in equilibration buffer (8 M urea, 50 mM sodium phosphate, pH 8.0, 0.3 M sodium chloride and 10 mM imidazole). The cell suspension was transferred to a beaker and was sonicated for 3 x 1.5 min (output: 2.8, duty cycle: 70%). After sonication, cell debris was removed by centrifugation at 15,000 rpm (SS-34 rotor) at 4 °C for 30 min. The cell lysate was loaded onto a pre-equilibrated HIS-Select
Cartridge (Sigma, product # H-8286). The cartridge was washed with 10 ml of wash buffer (8 M urea, 50 mM sodium phosphate, pH 8.0, 0.3 M sodium chloride and 20 mM imidazole). His-Rpb3 protein was eluted with 6-7 ml of elution buffer (8 M urea, 50 mM sodium phosphate, pH 8.0, 0.3 M sodium chloride and 250 mM imidazole). Eluted fractions were collected in 1.5 ml tubes. The concentration of His-Rpb3 protein in each fraction was estimated by the Bradford Assay (Biorad). Eluted fractions were dialyzed against phosphate-buffered saline (PBS, pH 7.4, Molecular Cloning) at 4°C for overnight, fast frozen with liquid nitrogen and stored in -80°C. His-Rpb3 was soluble after dialysis. Purified His-Rpb3 was sent to Pocono farms for antibody production. Antiserums from two injected rabbits (#20578 and #20579) were tested in a western blot with whole Drosophila S2 cell lysate. Both rabbits generated antiserums with high specificity and sensitivity. The 8th bleed of antiserum from rabbit #20578 was chosen for the chromatin immunoprecipitation experiments.

2.2 RNAi depletion of Pcf11 in Drosophila S2R+ cells and Kc167 cells

DNA template for making double strand RNA was amplified by PCR using genomic DNA extracted from Drosophila S2R+ cells and primers containing the T7 promoter sequence: 5’-GAATTAATACGACTCACT ATAGGGAGA-3’. LacZ RNAi, generated from a template amplified from the plasmid 5’LacZ.1, served as a negative control for Pcf11 RNAi since there is no LacZ gene in the Drosophila genome. Spt5 RNAi served as a positive control for Pcf11 RNAi since previously we had succeeded in knocking down Spt5 in Drosophila S2R+ cells (primers targeting specific genes,
including Spt5, LacZ and Pcf11, are listed in Appendix C). The PCR product was phenol/chloroform extracted and ethanol precipitated. A 20 µl in vitro transcription reaction contained: 350 mM HEPES, pH 7.5, 32 mM MgAcetate, 40 mM DTT, 2 mM spermidine, 28 mM each NTP, 500 ng template DNA (amplified PCR product using primers containing T7 promoter sequences) and 500 ng T7 RNA polymerase. The reaction mixture was pre-incubated at 37°C for 5 min prior to the addition of T7 RNA polymerase. After T7 RNA polymerase was added, the reaction was incubated at 37°C for overnight. The reaction was then centrifuged at maximum speed in a microcentrifuge for 2 min to pellet magnesium pyrophosphate. The supernatant was incubated with 2 units of DNase I at 37°C for 30 min. The reaction was stopped by adding 15 µl of 5 M ammonium acetate, 100 mM EDTA and diluted with DEPC treated water to a total volume of 150 µl. The reaction mixture was extracted with an equal volume of phenol/chloroform followed by an equal volume of chloroform. RNA was precipitated by adding 1 volume of isopropanol and dissolved in DEPC treated water. One reaction normally yields 100-150 µg of double strand RNA.

RNAi treatment of *Drosophila* S2R+ cells for subsequent permanganate footprinting: 1 x 10^6 *Drosophila* S2R+ cells (*Drosophila* Genomics Resource Center) in 1 ml of Schneider’s *Drosophila* media (Gibco BRL) containing 10% fetal bovine serum (FBS, Sigma) were seeded in a tissue culture dish (35 mm). One hour after seeding, the media was replaced with 1 ml of fresh media without FBS, together with RNAi of the desired amount (25 µg under standard condition). After 2 hr, the cells were supplemented with 2 ml of 10% FBS containing media and grown for 4.5 days.
RNAi treatment of *Drosophila* S2R+ cells for chromatin immunoprecipitation:

$1.6 \times 10^7$ *Drosophila* S2R+ cells in 8 ml of Schneider’s *Drosophila* media (Gibco BRL) containing 10% FBS were seeded in a tissue culture T-flask (75 cm$^2$). One hour after seeding, the media was replaced with 8 ml of fresh media without FBS, together with 200 µg of double strand RNA. After 2 hr, the cells were supplemented with 17 ml of 10% FBS containing media and grown for 4.5 days.

RNAi treatment of *Drosophila* Kc167 cells (*Drosophila* Genomics Resource Center) was done in dishes (35 mm) in the same way as RNAi treatment of *Drosophila* S2R+ cells, but followed only by western blot analyses.

### 2.3 Western blot analyses and antibodies

Whole cell extracts were prepared by directly lysing cell pellets in 1 x SDS-PAGE loading buffer (2% SDS, 100 mM Tris-HCl, pH 6.8, 10% Glycerol and 0.1% Bromophenol Blue). Proteins were separated on a 6% SDS-PAGE gel and transferred using a semi-dry blotting device (Biorad) to nitrocellulose membrane. Spt 5 was detected by anti-Spt5 antiserum (1:5000). Pol II was detected by ARNA3 antiserum (1:10000), which does not bind the CTD and detects all forms of Pol II. Pcf11 was detected by anti-Pcf11 antiserum (1:1000), which recognizes a 200 kDa isoform and a 60 kDa isoform of Pcf11.
2.4 Chromatin immunoprecipitation

The ChIP assay was done as described previously with some modifications (Boehm et al., 2003; Park et al., 2001; Wu et al., 2003). Twenty-five milliliters of S2R+ cells (from *Drosophila* Genome Resource Center) in a 75 cm² T-flask were grown to a density of $3 \times 10^6$-5 x $10^6$ cells/ml. For heat shock conditions, the T-flask containing cells was partially submerged in a 37°C water bath for 15 min. For crosslinking, 2.5 ml of 11% formaldehyde solution (prepared by diluting 37% formaldehyde (Fisher #F79-500) in 0.1 M NaCl, 1 mM EDTA, 0.5 mM EGTA and 50 mM Tris-Cl, pH 8.0) was added to the 25 ml of media and cells in the T-flask. After 10 min incubation at room temperature, the crosslinking was quenched by adding 3.75 ml of 2 M glycine to a final concentration of 240 mM. Cells were collected by centrifugation at 700g for 10 min. Cell pellets were resuspended and washed with 10 ml cold PBS buffer. Cells were collected again by centrifugation at 700g for 10 min and then resuspended in 1 ml of sonication buffer (10 mM Tris-Cl, pH 8.0, 1 mM EDTA, 0.5 mM EGTA and supplemented with protease inhibitors as follows: 0.5 mM PMSF, 1 μg/ml aprotinin and 1 μg/ml pepstatin A). The cell suspension was transferred to a Corex tube (3 ml) and was sonicated in a dry ice/ethanol bath (0°C-10°C) with the sonicator for 25 min (output: 2, duty cycle: 25%). The size of sheared chromatin DNA, as visualized on a 1% agarose gel, ranges from 100 bp to 600 bp with an average of 300 bp. Cell debris was removed by centrifugation at 13,000 rpm at 4 °C for 10 min. The supernatant was mixed with 1 ml of 6 M urea, and dialyzed in ChIP buffer (10 mM Tris-Cl, pH 8.0, 1 mM EDTA, 0.5 mM EGTA, 0.5 mM PMSF, 10% glycerol, 0.1% sodium deoxycholate and 1.0% Triton X-100) at 4 °C
overnight. Cell debris was removed by centrifugation. The entire preparation of chromatin solution from one 75cm² T-flask was precleared by adding 80 μl of 50% protein A sepharose beads (Amersham Biosciences, suspended in 10 mM Tris-Cl, pH 8.0, 1 mM EDTA and 1 mg/ml BSA) and mixing the solution gently for 30 min at 4 °C. Beads were removed by centrifugation at 13,000 rpm for 5 min. The DNA concentration of the chromatin solution was determined using the fluorometer. A total of 100-150 μl of chromatin solution, which corresponds to 5 μg of DNA, was used for one immunoprecipitation. Antibody was added to the solution and incubated overnight at 4°C. Two microliters of antiserum against Rpb3 (rabbit Ab) and 5 μl of antiserum against Pcf11 (rabbit Ab) were used. When available, the same amount of each corresponding preimmune serum was used for controls. Thirty-five microliters of 50% protein A sepharose were added and incubated for 2 hr at 4 °C. The protein A sepharose beads were collected by brief centrifugation in a microcentrifuge at 3000 rpm for 5 min. The immunoprecipitates were washed 6 times with 1 ml of low salt buffer (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris-Cl, pH 8.1 and 150 mM NaCl), three times with 1 ml of high salt buffer (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris-Cl, pH 8.1 and 500 mM NaCl), twice with 1 ml of lithium wash buffer (0.25 M LiCl, 1% NP-40, 1% sodium deoxycholate, 1 mM EDTA and 10 mM Tris-Cl, pH 8.1) and twice with 1 ml of TE (10 mM Tris-Cl, pH 8.0 and 1 mM EDTA). Each wash was incubated at 4 °C for at least 5 min and centrifuged at 3000 rpm for 5 min. For the lithium wash step, the immunoprecipitates were incubated with the wash solution for 10 min for the first wash and overnight for the second wash. After the first wash with TE buffer, the beads were moved to a new 1.5 ml siliconized tube. After all the washes, the protein–DNA
complexes were eluted by adding 200 µl of elution buffer (1% SDS and 0.1 M NaHCO3) and incubating the solution at room temperature for 15 min. The beads were collected by centrifugation at 3000 rpm and the supernatants were transferred to new tubes. The beads were resuspended in 200 µl of elution buffer. The second eluate was combined with the first eluate. For the preparation of input DNA, the elution buffer was added to 20% amount of the precleared chromatin solution used in one immunoprecipitation to make a total volume of 400 µl. This represents 20% input DNA. Formaldehyde crosslinking was reversed by adding 16 µl of 5 M NaCl to each eluate and input, and incubating at 65°C for 4 hr. After reversal of crosslinking, each solution was incubated with 8 µl of 0.5 M EDTA, 16 µl of 1 M Tris-Cl, pH 6.5 and 1 µl of 20 mg/ml proteinase K at 45 °C for 1 hr. Immunoprecipitated DNA and input DNA were purified by phenol/chloroform extraction and ethanol precipitation in the presence of 25 µg of linear polyacrylamide (Sigma #56575). The DNA pellet was dissolved in 30 µl of distilled water.

A series of dilutions were made from the 20% input DNA to generate 10%, 2%, 1%, 0.2% and 0.1% input. Each input DNA and immunoprecipitated DNA was further diluted 8-fold. One microliter of each 8-fold diluted input DNA and 8-fold diluted immunoprecipitated DNA was used for a real time PCR reaction in a final volume of 20 µl. Each real time PCR reaction contained 1.25 picomoles of each primer, 10 µl of Power SYBR Green PCR Master Mix (Applied Biosystems). Real time PCR reactions were carried out in 96-well plates using Applied Biosystems 7300 PCR machine. The reactions were run in absolute quantification mode and 50 amplification cycles were used. The amplification cycles were as follows: 1 cycle of 95°C for 10 min, 50 cycles of 95°C for 15 sec followed by 60°C for 1 min. Dissociation stage were as follows: 95°C for 15 sec,
then 60°C for 30 sec, followed by 95°C for 15 sec. Duplicate reactions were performed for each immunoprecipitated DNA. PCR primers are listed in Appendix A. C\textsubscript{i} number for each reaction was exported to Excel using 7300 System SDS Software (Applied Biosystems) after the reactions were done. The percentage of immunoprecipitated DNA was calculated by comparison of C\textsubscript{i} numbers for input DNA and immunoprecipitated DNA, and linear interpolation taken between two input data points.

### 2.5 ChIP-chip using Affymetrix Tiling Array

The Rpb3-antiserum-immunoprecipitated chromatin (Pol II ChIP) DNA was analyzed by the Penn State Cartography project. The Pol II ChIP DNA samples from LacZ RNAi (control) and Pcf11 RNAi treated S2R+ cells were amplified by ligation-mediated PCR (Harbison et al., 2004). The signals generated from mock ChIP (pre-immune GAGA factor serum) DNA samples (Lee et al., 2008) was used as background against which the Pol II ChIP signals were normalized. Using the GeneChip WT (Whole Transcript) Double-Stranded DNA Terminal Labeling Kit (Affymetrix 900812), two biological replicates of each amplified DNA containing dUTP was fragmented with uracil DNA glycosylase (DGE) and apyrimidinic/apurinic endonuclease (APE), then biotinylated and finally hybridized to Affymetrix *Drosophila* tiling 1.0R arrays, each of which contains 3 million oligonucleotide 25 bp probes that cover the euchromatic portion of the genome at an average resolution of about one per 35 bp. Signal analysis, interval analysis and peak calling were performed using Model-based Analysis of Tiling-arrays (MAT) software (Johnson et al., 2006). A bandwidth value of 150 bp was used for the
signal analysis. A max gap value of 150 bp, a minimum probe value of 5 and a 5% regional false discovery rate (1 out of 20 regions may be a false call) were used for interval and peak calling analysis. Pol II ChIP signals were represented in the form of matscores.

Genome wide ChIP-chip data of GAGA factor and NELF-B were generated by Chanhyo Lee (Gilmour lab) following a similar procedure as described above (Lee et al., 2008). The GAGA factor ChIP DNA was hybridized to Affymetrix Drosophila tiling 1.0R arrays. The NELF-B ChIP DNA was hybridized to Affymetrix Drosophila tiling 1.0F arrays. The GAGA factor ChIP-chip data was processed by Bryan Venters (Pugh lab) using MAT software. A bandwidth value of 150 bp was used for the signal analysis. A max gap value of 150 bp, a minimum probe value of 5 and a 1% regional false discovery rate threshold were used for interval and peak calling analysis. I re-processed the NELF-B ChIP-chip data using MAT software (previously processed by Mark Biggin using TiMAT software). A bandwidth value of 150 bp was used for the signal analysis. A max gap value of 150 bp, a minimum probe value of 5 and a 1% regional false discovery rate threshold were used for interval and peak calling analysis. The 1% FDR applied to GAGA factor and NELF-B ChIP-chip data identified 2907 and 4273 regions enriched for respective proteins (for a comparison, the 5% FDR applied to Pol II ChIP-chip data identified 7689 regions enriched for Pol II in LacZ RNAi treated samples).

Figure 2-1 shows a correlation between the matscores and the fold enrichment of Pol II occupancy in LacZ and Pcf11 RNAi treated samples. Typically, a 6-fold enrichment of Pol II occupancy corresponds to a matscore of 6.5.
To calculate the changes of Pol II density upon Pcf11 depletion, Tiling Analysis Software (TAS, Affymetrix) was used to calculate the ratio of the intensities of Pol II ChIP signal from LacZ RNAi treated S2R+ cells versus those from Pcf11 RNAi treated S2R+ cells. A bandwidth of 150 bp was used. Quantile normalization was applied and the ratio was represented in a log$_2$ transformed format and considered meaningful if the probe is not filtered out in the interval analysis using MAT as described above. This assured that the signal for either the control or Pcf11-depleted cells was above the 5% false discovery rate (the situation where Pol II level is above 5% FDR only in LacZ RNAi treated samples represents a loss of Pol II upon Pcf11 depletion; the situation

---

1 I have tried to use the MAT software to calculate the ratios of the intensities of Pol II ChIP signal from LacZ RNAi treated S2R+ cells versus those from Pcf11 RNAi treated S2R+ cells. However, for some unknown reasons, the calculated ratios are not consistent with the results from direct comparison between individual ChIP-chip data (LacZ RNAi vs. Pcf11 RNAi). The TAS software, on the other hand, does well in calculating the ratios and truly reflecting the differences. The 150 bp bandwidth applied in TAS analysis helps to smooth the calculated ratios across the *Drosophila* genome.
where Pol II level is above 5% FDR only in Pcf11 RNAi treated samples represents a gain of Pol II upon Pcf11 depletion).

2.6 Permanganate footprinting using *Drosophila* S2R+ cells

Three million S2R+ cells in 2-3 ml of Schneider’s *Drosophila* media containing 10% FBS were seeded in a tissue culture dish (35 mm). If the cells were RNAi treated, they were resuspended in the original dish at the end of the 4th day of treatment by pipetting the media up and down and transferred to a new tissue culture dish. Cells were left undisturbed in the dish for overnight at 25°C. This allowed S2R+ cells to attach to the bottom of dishes and recover from stress. The media was then pipetted out to remove dead cells. For non-heat shock conditions, the medium was removed before the addition of permanganate. For heat shock conditions, the medium was removed and cells were heat shocked by adding 2 ml of Schneider’s *Drosophila* media (without FBS) pre-warmed to 37°C. The dish was then immediately placed in a 37°C water bath for 15 min. After heat shock, media was pipetted out again and the dish was placed on ice.

Cells were treated with permanganate by adding 800 µl of ice-cold 10 mM KMnO₄ dissolved in PBS and incubating them for 1 min with gentle shaking. The reaction was stopped by the adding of 200 µl of 5X stop solution (50 mM Tris-Cl, pH 7.5, 50 mM NaCl, 50 mM Na₂EDTA, pH 8.0, 2.5% SDS and 500 mM 2-mercaptoethanol). The cell lysate was transferred to a 1.5 ml tube and then incubated overnight at 37°C with 50 µg of proteinase K. Samples were extracted once with phenol/chloroform/isoamyl alcohol (25:24:1) and once with chloroform. DNA was precipitated by adding 2 volumes
of cold ethanol, incubating at room temperature for 2 min and spinning in a microcentrifuge at room temperature for 10 min (if ethanol precipitation was done in the cold at this stage, a huge white pellet appeared and successive LM-PCR reactions using this precipitated DNA never worked). Precipitated DNA from one dish was dissolved in 40 µl of TE, pH 8.0. The DNA was quantified with the fluorometer. Typically, 3 µg of DNA was recovered from one 35 mm dish of cells. For piperidine cleavage of permanganate treated DNA, 1.5 µg of DNA was diluted to a final volume of 90 µl with H₂O and then incubated with 10 µl of piperidine at 90°C for 30 min. After 30 min, 300 µl of water was added to the DNA solution and then samples were extracted twice with 800 µl of isobutanol and once with 400 µl of isobutanol. DNA was precipitated by adding 2 volumes of cold ethanol, incubating at room temperature for 2 min and spinning in a microcentrifuge at room temperature for 10 min. Precipitated DNA was dissolved in 100 µl of 0.5 x TE.

For purification of genomic DNA to generate naked DNA (as controls) and GA markers (as reference for sequencing mapping), cells were treated with 800 µl of PBS followed by 200 µl of 5x stop solution. DNA was purified as described above. One and a half micrograms of DNA was diluted to a final volume of 100 µl with cold PBS. DNA samples were treated with 100 µl of 20 mM KMnO₄ in PBS on ice for 0, 30, and 60 sec separately. The reactions were stopped by adding 200 µl of 2 x stop solution (20 mM Tris-Cl, pH 7.5, 20 mM NaCl, 20 mM Na₂EDTA, pH 8.0, 1 % SDS and 200 mM 2-mercaptoethanol). DNA was ethanol precipitated at room temperature and dissolved in 90 µl of H₂O. For GA markers, 1 µg of genomic DNA in 20 µl of H₂O was incubated with 50 µl of 95% formic acid at 15°C for 5 min. After 5 min, 200 µl of cold depurination stop
solution (0.3M sodium acetate, pH 7.0 and 50 µg/ml tRNA) was added. Depurinated DNA was ethanol precipitated at 4°C and then dissolved in 90 µl of H$_2$O. Piperidine cleavage was done as described above. After piperidine cleavage, DNA was purified and diluted with 50 µl (G/A marker) or 100 µl (naked DNA control) of 0.5 x TE.

Each 10 µl of KMnO$_4$ treated samples, naked DNA samples, and GA markers were subjected to ligation-mediated PCR (LM-PCR) as previously described (Cartwright et al., 1999). Pfu polymerase generated by Dr. Song Tan was used for the DNA amplifications. Primers and annealing temperatures used in the LM-PCR analyses are listed in Appendix B. The final radiolabelled product from each LM-PCR reaction was dissolved in 15 µl of sequencing gel loading buffer (98% deionized formamide, 10 mM EDTA, pH 8.0, 0.025% xylene cyanol FF and 0.025% bromophenol blue) and one third of the dissolved DNA (5 µl) was run on a 6% or 8% sequencing gel (8M urea, 6% or 8% acrylamide/bis-acrylamide [19:1] and 1 x TBE) at 45W (voltage and current not limited). 6% sequencing gels were run for 4.5 hr and 8% sequencing gels were run for 3 hr. The gel was dried and the radioactivity in the dried gel was detected using a PhosphorImager Screen (Molecular Dynamics).

The intensity of each band within well-resolved regions of each gel was corrected and quantified using SAFA (Semi-Automated Footprinting Analysis) software (https://simtk.org/home/safa). The well-resolved and readable length is about 100-150 bp per gel and 300 bp per LM-PCR primer set. The LM-PCR primer sets were spaced about 250 bp apart so that each gel overlapped with another gel by about 20-30 bp. The gel was first straightened using the aligning function of SAFA software. Each band was then assigned to one of the A, C, G and T residues according to their position relative to the
GA marker and the DNA sequence. The band intensity was given by SAFA software and exported to a .txt file. The intensities of C residues were used as an internal standard for normalizing loading differences between lanes. The intensities of T residues were connected to form a continuous region using the overlapping part among different samples. For heat shock genes, the Pol II intensity at a specific T residue was represented as a ratio over the intensity of the same T residue in the non-heat shock control lane. For non-heat shock genes, the Pol II intensity at a specific T residue was represented as a ratio over the intensity of the same T residue in the naked DNA (60" KMnO₄ treatment) control lane, since it was determined experimentally that 60 sec of treatment using 10 mM KMnO₄ generated similar background on naked DNA as on S2R+ cells. The Pol II distribution was represented in a heatmap (log₂) using Treeview (Eisen et al., 1998) or in a smoothed curve using Kaleidagraph (weighted curve fit, 5-10%).

2.7 Bioinformatics Analysis

The probe intensities from the microarray analysis were obtained in tabulated form. Analysis of tabulated data was done using programs written by Cizhong Jiang (Pugh lab) and tools provided at Penn State Galaxy: http://g2.trac.bx.psu.edu/. Galaxy was also used to prepare and transfer tabulated data for viewing in the UCSC browser located at: http://genome.ucsc.edu/. The annotation of all Drosophila genes were downloaded from the UCSC table browser (track: Flybase Genes; assembly: Apr. 2006), which matches Flybase release 5.2
(ftp://ftp.flybase.net/genomes/dmel_r5.2_FB2007_01/fasta). Alternative transcripts from one gene are regarded as separate genes for my analysis.

2.7.1 Pol II and nucleosome distribution profile

For the composite plot of the Pol II distribution or the nucleosome distribution in reference to the transcription start sites or polyadenylation sites, intensities of Pol II ChIP signals or counts of nucleosomes at distances from the reference gene were binned in 10 bp intervals. To minimize the influence of nearby genes, the following filtering rules were applied: for Pol II, GAGA factor and NELF distributions relative to the transcription start site, regions located within 500 bp of a nearby transcription start site or 1000 bp of a nearby polyadenylation site were removed from the analysis; for Pol II, GAGA factor and NELF distributions relative to the polyadenylation site, regions located 500 bp downstream of the transcription start site of the same gene or within 500 bp of a nearby transcription start site or within 1500 bp of a nearby polyadenylation site were removed from the analysis; for the composite nucleosome distribution relative to the transcription start site, regions located 300 bp upstream of the polyadenylation site of the same gene or within 300 bp of a nearby transcription start site or within 1000 bp of a nearby polyadenylation site were removed from the analysis; for the nucleosome distribution relative to the polyadenylation site, regions located 300 bp downstream of the transcription start site of the same gene or within 300 bp of a nearby transcription start site or within 1500 bp of a nearby polyadenylation site were removed from the analysis. None of these filters substantially affected the distribution of the data or conclusion.
Binned data were then normalized to the number of regions represented in each bin. Plots were smoothed using a moving average with Kaleidagraph software (weighted curve fit). The size of the moving average (typically between 5% and 15% weight) was set to provide the optimal balance between signal and noise for purpose of visual display. The size of the moving average within the chosen range had no effect on the conclusions drawn.

2.7.2 Polyadenylation signal and GAGA element identification

The Drosophila genome was scanned for AATAAA hexamers that resided within +/-1 kb of each annotated polyadenylation site; non-canonical polyadenylation signals (for example, with single nucleotide variance) are not included. The counts of polyadenylation signals were binned in 10 bp intervals based on their distance from the polyadenylation sites and plots were smoothed as aforementioned. Polyadenylation signals located 300 bp downstream of the transcription start site of the same gene or within 300 bp of a nearby transcription start site or within 1000 bp of a nearby polyadenylation site were removed from the analysis (which had little effect on the distribution).

The GAGA element was scanned within +/-1 kb of each transcription start site. The sequence of GAGAGAG was used for scanning. The counts of GAGA elements were binned in 10 bp intervals based on their distance from the transcription start sites and plots were smoothed as aforementioned. GAGA elements located 300 bp upstream of the polyadenylation site of the same gene or within 300 bp of a nearby polyadenylation
site or within 1000 bp of a nearby transcription start site were removed from the analysis (which had little effect on the distribution).

2.7.3 Clustering analysis

Intensities of Pol II ChIP signals or changes in the intensities of Pol II ChIP signals were binned in 36 bp intervals based on their distance from the transcription start sites or the polyadenylation sites. Binned data were mapped to each transcription start site or polyadenylation site if they were within the depicted range (mostly from -1 kb to 1 kb) of the reference point. To minimize the ambiguity caused by neighboring genes, the transcription start sites located within 2 kb of another transcription start site or 2 kb of any polyadenylation site, and the polyadenylation sites located within 2 kb of another polyadenylation site or 1.5 kb of any transcription start site were removed from the analysis (“filtered”). Alternatively, every transcription start site or polyadenylation site is included in the analysis regardless of their genome context (“unfiltered”). K-mean clustering was applied using the Cluster software and the result was displayed in Treeview (Eisen et al., 1998).

2.7.4 GO (Gene Ontology) analysis

G:profile (http://biit.cs.ut.ee/gprofiler/welcome.cgi) was used to convert the annotation of all genes into Flybase_ID format (FBGN). The Database for Annotation,
Visualization and Integrated Discovery (DAVID) project (http://david.abcc.ncifcrf.gov/home.jsp) was used to complete the functional clustering.

2.8 Reverse transcription real time PCR

Total RNA was extracted from Drosophila S2R+ cells (treated with LacZ RNAi or Pcf11 RNAi) under non-heat shock conditions or after 15 min of heat shock at 37°C using RNeasy Mini Kit (Qiagen). The RNA was digested with 10 units of DNase I (RNase free) at 37°C for 15 min to remove genomic DNA. cDNA was generated in a 20 µl reaction containing 5 µg of total RNA, 3 µl of 20 µM Oligo(dT)_{16}, 2 µl of 75 ng/µl random hexamer, 1 µl of 10 mM dNTP mix and sterilized water. The cDNA was diluted in 5-fold series and used as template in a real time PCR reaction in a final volume of 20 µl. Each real time PCR reaction contained 1 µl of the diluted cDNA template, 10 picomoles of each gene-specific primer, 10 µl of Power SYBR Green PCR Master Mix (Applied Biosystems). To control for the RNA recovery from different batches of cells, the mRNA level from the ribosomal protein gene RP49 was used as a normalization standard. Duplicate reactions were performed for each cDNA sample. The data analysis using 7300 System SDS Software (Applied Biosystems) was the same as previously described in ChIP analysis. Primers used in the real time PCR are listed in the Appendix A.
Chapter 3

*Drosophila* Pcf11 and Transcription Termination

3.1 Introduction

Transcription termination is defined as the point during the transcription cycle when RNA polymerase (Pol) releases the transcript and dissociates from the DNA template. Termination has been shown to be important for preventing RNA polymerase from upstream genes from interfering with transcription of downstream genes (Greger *et al.*, 2000; Shearwin *et al.*, 2005) and has been proposed to be essential for recycling the RNA polymerase (Ansari and Hampsey, 2005; Cho *et al.*, 1999; Kobor *et al.*, 1999; Singh and Hampsey, 2007). Pol I, Pol III and bacterial RNA polymerase generally terminate at specific sites on the DNA that are associated with particular nucleic acid sequences or bound proteins which destabilize the elongation complex (Gilmour and Fan, 2008). In contrast, Pol II termination appears to be a stochastic process that occurs after Pol II transcribes the polyadenylation signal (Kim and Martinson, 2003; Orozco *et al.*, 2002). In animal cells, chromatin immunoprecipitation analysis of Pol II and nuclear run-on assays indicate that Pol II pauses after transcribing the polyadenylation signal and then terminates in regions ranging from 200 to 2000 nucleotides farther downstream (Bauren *et al.*, 1998; Dye and Proudfoot, 1999). Since relatively few genes have been analyzed in animal cells, it is not known how varied the behavior of Pol II at the 3’ end of genes might be. Electron microscopic studies of elongation complexes on *Drosophila* genes
provided evidence that termination occurred abruptly on some genes but over an extended region on other genes (Osheim et al., 2002).

Transcription termination and 3’-end processing are tightly linked. Numerous mutations in components of the polyadenylation machinery that impair 3’ end formation also impair termination (Proudfoot, 2004). One of these proteins, Pcf11, has been shown to dissociate a stalled elongation complex from DNA in vitro, indicating that Pcf11 has a key role in termination. Supporting this conclusion are the observations that mutations in Pcf11 in yeast and RNAi depletion of Pcf11 in Drosophila or human cells cause Pol II to transcribe beyond the region where it normally terminates on several genes (Kim et al., 2006; Zhang et al., 2005; Zhang and Gilmour, 2006).

Another protein, a 5’ to 3’ RNA exonuclease, also contributes to termination in yeast and human cells (Kim et al., 2004b; West et al., 2004). The RNA exonuclease, known as RAT1 in yeast and Xrn2 in humans, provides an elegant way to ensure that termination does not occur until after the nascent transcript has been cleaved since the product of this cleavage reaction provides the entry point for the exonuclease. A recent study in yeast indicated that the association of RAT1 and Pcf11 with the 3’ end of a gene was mutually interdependent, thus providing a way to coordinate cleavage and termination (Luo et al., 2006).

Much of our knowledge of transcription termination in animal cells is based on extrapolation of data from yeast and detailed studies of a small number of genes in human cells. No studies in yeast or animal cells have attempted to analyze the behavior of Pol II at the 3’ end of genes with base pair resolution. Moreover, the extent to which Pcf11 affects termination in any cell is not known. Here, I employ permanganate
genomic footprinting to analyze the behavior of Pol II in the region downstream from the polyadenylation site of three heat shock genes before and after RNAi-mediated depletion of Pcf11. Prior to depletion of Pcf11, my footprinting analysis identifies several regions spanning about 100 base pairs where the Pol II level increases. This accumulation appears to precede termination since depletion of Pcf11 results in an increase in the level of Pol II downstream from these pause sites. I also use ChIP-chip to characterize the behavior of Pol II at the 3’ end of thousands of genes in *Drosophila* cells. These analyses reveal 3 different behaviors of Pol II at the 3’ end of genes, indicating that alternate termination mechanisms may exist in *Drosophila*.

Upon depletion of Pcf11, termination at more than 2000 genes is affected. Pol II levels at the promoter of numerous genes were also seen to decrease. The broad impact of Pcf11 depletion on Pol II levels at both ends of genes can be explained by the importance of Pcf11 in both the termination process and the recycling of Pol II. My ChIP-chip analysis identified thousands of gene candidates whose transcription might be controlled by a Pcf11-mediated premature termination mechanism. Finally, a comparison between the distribution of Pol II and nucleosomes, both genome-wide and at the heat shock genes, identifies a strong correlation between pausing of Pol II and positioned nucleosome at the 3’ end of genes.
3.2 Results

3.2.1 Depletion of Pcf11 alters the distribution of Pol II at the 3’ end of the hsp70 and hsp23 genes

Previously, Zhang et al. determined that RNAi-mediated depletion of Pcf11 altered the distribution of Pol II at the 3’ end of the hsp70 genes in a way that was indicative of a defect in termination (Zhang and Gilmour, 2006). Chromatin immunoprecipitation (ChIP) analysis showed that the density of Pol II detected downstream from the polyadenylation site increased with depletion of Pcf11 (Zhang and Gilmour, 2006). However, there are 3 copies of the hsp70B gene and the DNA sequence downstream from the polyadenylation site of each is identical. Hence an unambiguous assessment of the effect of depleting Pcf11 on the behavior of Pol II at a single gene was not possible. To avoid such ambiguity in analyzing the behavior of Pol II at the 3’ end of genes and to more accurately understand the function of *Drosophila* Pcf11 in termination, I sought single copy gene candidates.

The hsp70Aa and hsp70Ab genes are separated by approximately 1.7 kb and divergently transcribed. Each gene has unique sequences downstream from the polyadenylation site where I anticipated Pol II termination would occur. To determine where Pol II terminated at the 3’ end of each gene and whether this termination was influenced by Pcf11, I used ChIP analysis to compare the distribution of Pol II in control
(LacZ RNAi treated) and Pcf11-depleted (Pcf11 RNAi treated) cells. Figure 3-1C shows the specificity of the RNAi treatments. Western blot analysis indicates that Pcf11 was depleted from the cells treated with Pcf11 RNAi but not by RNAi against LacZ or Spt5.

In control cells, the density of Pol II diminished significantly in the region approximately 500-800 bp downstream from the polyadenylation site of each hsp70A gene, suggesting that Pol II terminated transcription somewhere in this region (Figure 3-1A and B). Depletion of Pcf11 (Figure 3-1C) increased the level of Pol II detected in these regions relative to the control cells. Given that previous work demonstrated the activity of Pcf11 in dismantling stalled elongation complexes and releasing nascent transcript in vitro (Zhang and Gilmour, 2006), the elevated Pol II occupancy at the 3’ end of the hsp70A genes most likely represents a defect in transcription termination upon depletion of Pcf11 as opposed to increased pausing of Pol II.

---

2 The depletion of Pcf11 can be achieved by using three RNAi constructs (Pcf11 RNAi 1, 2 and 3) targeting different regions of Pcf11 (Figure D-1 and Figure D-2). Each of the three RNAi construct was able to deplete the 200 kDa isoform of Pcf11 (shown in Figure 3-1C). The 60 kDa isoform of Pcf11, however, did not appear to be depleted as shown by western blot analysis (see Appendix D for detailed result and discussion). All RNAi experiments described in this chapter were carried out using Pcf11 RNAi 1 construct.
Figure 3-1: Depletion of Pcf11 disrupts transcription termination for three heat shock genes. Chromatin immunoprecipitation was used to monitor Pol II occupancy downstream of the polyadenylation site of the A) hsp70Aa, B) hsp70Ab and D) hsp23 genes before and after Pcf11 depletion. Polyadenylation sites (vertical arrows) and PCR amplified regions (horizontal lines) are indicated in a schematic above each bar graph. Pol II signals in downstream regions are normalized against the signal in the most upstream region. Signals of pre-immune serum immunoprecipitated samples are normalized against signals from corresponding samples immunoprecipitated with antibody against the Rpb3 subunit of Pol II. Each bar represents the mean of three independent experiments. A P-value is calculated for comparison between Pol II ChIP signals using the unpaired student t-test. P-values smaller than 0.05 are labeled, indicating that the differences are statistically significant. C) Drosophila S2R+ cells were treated with 25 µg of double strand RNA targeting LacZ (negative control), Pcf11 or Spt5 (positive control). Western blot was done to monitor the levels of Spt5, Pcf11 and Pol II.
I sought another situation where constraints placed on termination might differ from those of the hsp70A genes. The hsp23 and hsp28 genes are heat shock genes that are arranged in tandem and transcribed in the same direction (Figure 3-1D). The polyadenylation site for hsp23 is located 1.3 kb upstream from the transcription start for hsp28, and I reasoned that efficient termination must occur in the space between the genes to prevent Pol II molecules transcribing the hsp23 gene from interfering with the hsp28 promoter. I compared the distribution of Pol II in Pcf11-depleted cells to control cells. In control cells, the Pol II density was observed to peak approximately 300 base pairs downstream from the polyadenylation site, and decrease significantly in the region 350 to 600 base pairs downstream from the polyadenylation site. Within the resolution afforded by ChIP, these decreases were similar to what was observed at the end of each of the hsp70A genes. Depletion of Pcf11 markedly altered the distribution of Pol II at the end of the hsp23 gene. Significantly more Pol II was detected in the region 600 to 1000 base pairs downstream from the polyadenylation site than was detected in control cells (Figure 3-1D). This indicates that Pcf11 is also involved in terminating transcription at the end of the hsp23 gene.

Since transcription termination of the three heat shock genes was disrupted after Pcf11 was reduced in cells, I predicted that more read-through RNA would be generated. To detect the read-through products, I used reverse transcription real time PCR to measure the transcript levels downstream from the polyadenylation site of these genes in Pcf11-depleted cells and control cells. In accordance with a termination defect, Pcf11 depletion resulted in an increase in transcript levels downstream from the
polyadenylation site of all three heat shock genes (Figure 3-2A, B and C). However, further analysis uncovers a potential limitation in the interpretation of the RT-PCR results. A relative measure of the cleavage efficiency at the polyadenylation site of the hsp23 gene was obtained by determining the ratio of PCR product generated by a pair of primers that flank each side of the polyadenylation site to the PCR product generated by a pair of primers located upstream from the polyadenylation site (Figure 3-2D). Depletion of Pcf11 caused an 8-fold increase in this ratio indicating that the efficiency of cleavage was markedly reduced. Since the nascent transcript remaining after cleavage is normally sensitive to degradation (as described by the torpedo model), it is not possible to assess whether the increase in downstream transcripts is due to a defect in termination or an increase in transcript stability. This complication provided additional motivation for monitoring the distribution of Pol II by a method that was independent of the ChIP analysis.
Figure 3-2: Depleting Pcf11 increases read-through transcription for three heat shock genes and transcript uncleaved ratio at the polyadenylation site for hsp23. Reverse transcription real time PCR was used to monitor the levels of RNA upstream or downstream of the polyadenylation site of A) hsp70Aa, B) hsp70Ab and C) hsp23 genes. Polyadenylation sites (arrow) and PCR amplified regions (horizontal lines) are indicated in a schematic above each bar graph. RNA levels detected by each primer pair in Pcf11 RNAi treated cells were normalized against control cells. Each bar represents the mean of three independent experiments, and the error bar represents the standard deviation. D) Reverse transcription real time PCR was used to monitor the levels of total transcripts (697-789) and uncleaved transcripts (697-915) from the hsp23 gene. The uncleaved ratio was calculated by taking the ratio of uncleaved transcript versus total transcript. The uncleaved ratio in Pcf11 RNAi treated cells was normalized against control cells. Each bar represents the mean of three independent experiments. The error bar represents the standard deviation.
3.2.2 High resolution analyses of the Pol II distribution at the 3’ end of the hsp70 and hsp23 genes reveal uneven patterns of elongation

ChIP is widely used to evaluate the behavior of Pol II on genes in living cells, but limitations inherent to the technique prevent high resolution mapping of the distribution of Pol II. Moreover as just shown by my analysis of the transcripts, the instability of the nascent transcript after cleavage at the polyadenylation site limits this approach for mapping the site of termination. High resolution data obtained with permanganate genomic footprinting could reveal novel aspects of the termination process.

Permanganate genomic footprinting detects transcriptionally engaged Pol II by the enhanced reactivity of thymines in transcription bubbles towards oxidation by permanganate (Sasse-Dwight and Gralla, 1989). Lee et al. have used this technique extensively to monitor the behavior of Pol II at the 5’ end of genes in Drosophila (Lee et al., 2008). Figure 3-3 shows that permanganate readily detects Pol II at the 3’ end of the hsp23, hsp70Aa and hsp70Ab genes following heat shock. Each dark band that is in the heat shock lanes (lane 4, 9 and 13) but not in the non-heat shock lanes (lane 3, 8 and 12) corresponds to a T residue that is hyper-reactive to permanganate treatment when the gene is activated. DRB, an inhibitor of Pol II elongation (Wada et al., 1998; Yamaguchi

---

3 Extensive pilot experiments of permanganate genomic footprinting were performed to optimize the conditions of permanganate treatment of Drosophila S2R+ cells. High resolution mapping of Pol II distribution at the 3’ end of the hsp70B genes (hsp70Bbb, hsp70Bb and hsp70Bc) was carried out, but yielded no extraordinary findings about the termination at these genes due to the genome context and technical limitations. Comparison of the Pol II distribution at the hsp23 gene after 5 min and 15 min of heat shock generated incomplete but interesting results, which suggest either the existence of two different termination sites at the hsp23 gene or a delayed induction of the hsp23 gene. For the concision and focus of Chapter 3, these results are omitted here and described in detail in Appendix J.
et al., 1999), eliminates most of the permanganate hyper-reactive sites at the end of the hsp23 gene (lane 5), indicating that this reactivity is due to elongating Pol II.

Figure 3-3: Genomic footprinting with permanganate detects Pol II at the 3’ end of hsp23, hsp70Aa and hsp70Ab genes. G/A ladders (lanes 6 and 10) generated by partial depurination with formic acid served as landmarks for identifying nucleotide positions. Purified DNA (lanes 1, 2, 7, and 11) were treated for the indicated times with permanganate on ice. Non-heat shocked (NHS) and heat shocked (HS) cells (lanes 3, 4, 8, 9, 12 and 13) were treated with 10 mM potassium permanganate for 1 min on ice.
Isolated DNA was cleaved at oxidized thymines with piperidine and the resulting breaks in the non-template strand were mapped by LM-PCR. Treatment of cells with the transcription elongation inhibitor 200 µM DRB prior to heat shock greatly reduces the permanganate reactivity detected in the 3’ region (lane 5).

Through judicious use of primers, I mapped the distribution of Pol II across a 1 kb region downstream of each of the heat shock genes. Biological replicates of both control and Pcf11-depleted cells were analyzed resulting in 56 sequencing gels, each containing an average of 6 samples. To analyze and present these data, I quantified the relative intensities of all bands on the gels using the Semi-Automatic Footprinting Analysis (SAFA) program. Figure 3-4 presents smoothed traces of the ratio of band intensity at thymines detected in heat shocked cells verses non-heat shocked cells. As a visual check of the smoothed traces, I also displayed the unsmoothed raw data from quantification in log2 transformed formats (Figure 3-5) with the Treeview program (Eisen et al., 1998)4.

The impact of depleting Pcf11 is striking. There is both an increase in the density of Pol II detected downstream from the polyadenylation site and a downstream shift in the Pol II distribution (Figure 3-4). Both changes suggest that Pcf11 plays a role in terminating transcription in the 3’ region of these three genes. In addition, permanganate footprinting reveals an uneven distribution of Pol II in the 1 kb region downstream from the polyadenylation site. The peaks of Pol II identify possible places where Pol II molecules accumulate because of a pause. Since these peaks span regions of 50-100 nucleotides and the permanganate footprint associated with a single Pol II molecule is

4 Results represented in Figure 3-4 and Figure 3-5 were generated from one of the biological replicate experiments. Results generated from the other set of experiments are presented in Figure F-1.
less than 20 nucleotides, these peaks can be regions where several Pol II molecules queue up on the DNA.

Figure 3-4: Permanganate footprinting of heat shock genes detects termination defects upon Pcf11 depletion. Drosophila S2R+ cells were treated with 10 mM potassium permanganate for 1 min on ice. The intensity of the permanganate reactivity of each thymine in heat shocked cells was represented as a ratio over the reactivity of the same thymine in non-heat shocked cells for: A) hsp70Aa, B) hsp70Ab and C) hsp23 genes. The red line represents the smoothed Pol II distribution in LacZ RNAi treated samples. The blue line represents the smoothed Pol II distribution in Pcf11 RNAi treated samples. The orange dots represent all T residues. The polyadenylation sites are indicated by the vertical dotted lines. Pause sites are labeled “P”. P** represents a new pause site on the hsp23 gene formed after Pcf11 depletion. Shaded areas indicate positioned nucleosomes derived from genome wide nucleosome mapping data (Mavrich et al., 2008b). Hypersensitive sites to MPE•Fe(II) (+1290, +1470, +1655) on the hsp23 gene (Cartwright et al., 1999) are indicated by arrows in panel C. The lengths of arrows correlate to the hypersensitivity.
Figure 3-5: Treeview displays of the permanganate reactivity of T residues of heat shock genes in heat shocked cells (LacZ RNAi HS (control cells) and Pcf11 RNAi HS (Pcf11 depleted cells)) relative to non-heat shocked cells (NHS). The intensity of the permanganate reactivity at the T residues in Figure 3-4A, B and C was log2 transformed and directly displayed using Treeview without smoothing. The top to bottom direction of each panel (A, hsp70Aa; B, hsp70Ab; C, hsp23) corresponds to the 5’ to 3’ direction of transcription. Black indicates the same intensity as the NHS lane. Red and green reflect stronger or weaker signals compared to the NHS lane, respectively.
3.2.3 Genome wide analysis reveals three distinct Pol II behaviors at the 3’ end of genes

The analysis of the Pol II distribution at the 3’ end of three heat shock genes provides evidence for several places where Pol II concentrated downstream from the polyadenylation site. This concentration, or pausing, is similar to that observed for a small number of other genes (Aranda and Proudfoot, 1999; Ashfield et al., 1991; Birse et al., 1997; Connelly and Manley, 1989; Enriquez-Harris et al., 1991). It has been proposed that transcription of the polyadenylation signal reduces the processivity of Pol II (Orozco et al., 2002). As a result, there is speculation that this increases the tendency of Pol II to pause and then terminate transcription.

To investigate the behavior of Pol II at the 3’ end of a broad spectrum of genes, I used chromatin immunoprecipitation and high density microarrays to analyze the distribution of Pol II in the vicinity of the 3’ end of all annotated genes in Drosophila cells. I detected significant levels of Pol II in a 2 kb region encompassing the polyadenylation site of 4672 genes (Figure 3-6). Using K-means clustering (Eisen et al., 1998), I divided the genes based on the Pol II distribution in the 3’ region into 3 categories (Figure 3-6). Category 1 had peaks in the Pol II distribution that flanked each side of the polyadenylation site. This category of genes is enriched in highly expressed genes (Figure 3-7A). Hence, the high density of Pol II upstream from the polyadenylation site could be a result of the high transcription level and the peak downstream from the polyadenylation site might be formed due to a pause that precedes termination.
Category 2 genes have Pol II molecules concentrated upstream from the polyadenylation site but not further downstream. This distribution of Pol II is surprising given the results of past studies in higher eukaryotes that have highlighted the role of pausing downstream from the polyadenylation site in termination. For this category, Pol II appears to be terminating transcription at the polyadenylation site.

Category 3 genes have Pol II concentrated in the region downstream from the polyadenylation site. This distribution is consistent with the pausing of Pol II downstream from the polyadenylation site prior to termination. Category 3 is enriched in genes that are weakly expressed (Figure 3-7A) and have low Pol II level at their promoters (Figure 3-7C). These genes also tend to be longer than the other genes in our clustering analysis (Figure 3-7B), which could account for the failure to detect Pol II upstream from the polyadenylation site. For these genes, I propose that only upon pausing downstream from the polyadenylation site does the density of Pol II become high enough to be significant in the ChIP-chip analysis.
Figure 3-6: Clustering analysis of the Pol II distributions at the 3’ end of genes reveals three distinct Pol II behaviors. A) Pol II ChIP signals within +/-1 kb of the polyadenylation site of 4672 genes are designated with color (the intensity of red color corresponds to the Pol II level; black indicates a Pol II level below 5% FDR threshold). Each row represents an individual gene. All genes are aligned at the polyadenylation site (indicated by the black dot on top of the Treeview panel). B) Composite Pol II distribution from -1.5 kb to 1.5 kb of polyadenylation site for the genes in category 1 (purple), 2 (orange) and 3 (green).
Figure 3-7: Different features associated with category 1, 2 and 3 genes. A) Comparison of the expression levels of the genes in category 1 (purple), 2 (orange) and 3 (green) and genes with Pol II level below the 5% FDR threshold within +/-1 kb of the polyadenylation site (brown). Steady state mRNA values were extracted from published expression microarray data for *Drosophila* S2 cells (Rehwinkel et al., 2004). The median expression level is indicated by the black center line, and the first and third quartiles are the edges of the boxed area (inter-quartile range). The extreme values (within 1.5 times the inter-quartile range from the first or third quartile) are at the end of the lines extending from the box. Points at a greater distance than the extreme values from the median are plotted individually. The *P*-value is calculated for each pairwise comparison.
among the four categories in panel A using the Wilcoxon-Mann-Whitney test. A $P$-value smaller than 0.05 indicates that the differences of media values (black lines) are statistically significant. B) Comparison of the gene length among genes in category 1 (purple), 2 (orange) and 3 (green) and all genes (brown). Box plot is created as in panel A. The $P$-value is calculated for comparison between category 1 and 3 and for comparison between category 2 and 3 using Wilcoxon-Mann-Whitney test. C) Composite distribution of Pol II from -1 kb to 1 kb of the transcription start site for the genes in category 1 (purple), 2 (orange) and 3 (green) in LacZ RNAi treated cells. D) Cluster analysis of the Pol II distribution at 3’ end were done in the same way as in Figure 3-6A, with the exception that analyzed candidates are filtered to remove any gene whose polyadenylation site is either within 2 kb of another polyadenylation site or within 1.5 kb of any transcription start site.

Inspection of the *Drosophila* genome raised the concern that one or more of the three Pol II distribution patterns observed in Figure 3-6 might be an artifact due to nearby promoters. For example, Pol II is concentrated at the 5’ end of thousands of genes (see Chapter 4) and any 5’ peaks of Pol II within +/-1 kb of a polyadenylation site would have been included in the current analysis. To address this concern, I filtered out all genes whose polyadenylation site is within 1.5 kb of a promoter or 2 kb of another polyadenylation site. Although this greatly reduced the number of genes analyzed, the same three distribution patterns of Pol II were detected (Figure 3-7D).

3.2.4 Clustering analysis identifies 2611 genes whose termination is affected by the depletion of Pcf11

To investigate the impact of Pcf11 on Pol II at the 3’ end of all known genes, I compared the distribution of Pol II in the vicinity of the polyadenylation site in control and Pcf11-depleted cells. The ratios were displayed as a map in which red and green
represent increases and decreases in Pol II density respectively upon depletion of Pcf11. Only those genes exhibiting significant Pol II level in either control or Pcf11-depleted cells were shown to insure that the ratios were based on Pol II measurements that were above background. 2611 genes were observed to have an increase in the density of Pol II downstream from the polyadenylation site upon depletion of Pcf11 (Figure 3-8A). The effect of depleting Pcf11 on these genes was similar to that observed on the heat shock genes, making these genes clear examples of where Pcf11 contributes to transcription termination.

5 Real time PCR reactions were performed to verify the increases and decreases in Pol II occupancy detected by ChIP-chip. Six out of 7 randomly selected regions where ChIP-chip detected increases in Pol II occupancy were verified by real time PCR (Figure G-1 and Figure G-3A). Seven out of 9 randomly selected regions where ChIP-chip detected decreases in Pol II occupancy were verified by real time PCR (Figure G-2 and Figure G-3B).
Figure 3-8: Depletion of Pcf11 causes termination defects at many genes. A) Pol II occupancy changes that occurred at the 3’ end upon Pcf11 depletion were clustered using K-mean clustering. Pol II occupancy changes at the 5’ end upon Pcf11 depletion were aligned to those at the 3’ end according to the gene names. A total of 7498 genes are included. Each row represents an individual gene. Increase in Pol II occupancy is indicated with red; decrease in Pol II occupancy is indicated with green. Regions displayed in black represent locations where the Pol II ChIP signals for both the control and Pcf11-depleted samples were below a 5% FDR or the change in Pol II levels between the two samples was insignificant. The 2611 genes exhibiting an increase in Pol II levels downstream from the polyadenylation site are presumed to correspond to termination defects caused by the depletion of Pcf11. B) Genes other than those shown in panel A but with changes in Pol II occupancy at the 5’ end were clustered using K-mean clustering.
based on the Pol II occupancy changes near the promoter. A total of 4375 genes are included. The 786 genes exhibiting an increase in Pol II levels only at the 5’ end and the 1902 genes exhibiting an increase in Pol II levels at both the 5’ and 3’ ends (panel A) are cases where Pcf11 causes premature termination (see discussion).

3.2.5 Evidence that Pcf11 represses transcription of some cellular genes by causing premature termination

I also observed an increase in Pol II density on 1902 genes in the region immediately upstream from the polyadenylation site (Figure 3-8A). Further analysis of the Pol II distribution upstream of these regions suggests that many of these increases correlate with increases in Pol II density near the promoters or in the body of these genes. This raises the possibility that Pcf11 is inhibiting elongation on these genes, perhaps by causing premature termination, as has recently been demonstrated for an HIV provirus (Zhang et al., 2007b). In order to test this possibility, I chose nine genes that exhibited increased levels of Pol II in the body of the gene upon depletion of Pcf11 (Figure 3-9A and Figure K-1) and measured their expression levels using reverse transcription real time PCR. Four out of nine genes showed increased expression levels after reduction of Pcf11, three showed decreased expression levels and the remaining two were unaffected (Figure 3-9B). This indicates that, besides controlling transcription termination, Pcf11 is also involved in inhibiting transcription elongation on some genes. The five genes that displayed decreased or unchanged levels of transcripts could be cases where the defects in 3’-end processing offset the increases in transcription elongation implicated by the ChIP-chip data.
Figure 3-9: Changes in expression of selected genes in response to Pcf11 depletion. A) Treeview display of changes in Pol II density at the 5’ end of nine genes upon depletion of Pcf11. Numbers above the panel indicate distance from the transcription start site (+1) of each gene. Increased Pol II occupancy is indicated with red; decreased Pol II occupancy is indicated with green. No change in Pol II occupancy or Pol II ChIP signals below 5% FDR (in both control and Pcf11 RNAi samples) is indicated with black. B) Relative mRNA levels of nine genes in control and Pcf11 RNAi treated cells. To control for the RNA recovery from different batches of cells, the mRNA level of each gene was normalized to the mRNA level of the RP49 gene in the same cells. mRNA level in Pcf11 RNAi treated cells was normalized against control cells. Each bar represents the mean of three independent experiments. The error bar represents the standard deviation. Genes labeled in red showed an increase in their expression levels. Genes labeled in blue showed a decrease in expression levels. Genes labeled in black showed no difference.
3.2.6 Decreases in Pol II density at the 3’ end of genes correlates with apparent decreases in transcription at the promoters

It was surprising to observe that depletion of Pcf11 caused a decrease in the density of Pol II at the 3’ end of 2985 genes (Figure 3-8), since the biochemical activity associated with Pcf11 is dissociation of Pol II from DNA. Almost all of these decreases correlate with decreases in Pol II density detected in the promoter regions of the genes. This raises the possibility that Pcf11 somehow contributes to the level of transcription initiation at many genes.

3.2.7 Depletion of Pcf11 does not cause termination defects at the U1_95Ca gene, an snRNA gene that lacks a polyadenylation signal

Several small nuclear RNA genes are transcribed by Pol II but lack polyadenylation signals. Studies in yeast indicate that Pcf11 is involved in terminating transcription at the 3’ end of some small nuclear RNA genes (Kim et al., 2006). I tested if this was the case for the snRNA gene, U1_95Ca, in Drosophila cells by analyzing the distribution of Pol II in the 3’ region of this gene in control and Pcf11-depleted cells. Both permanganate genomic footprinting⁶ and ChIP (or ChIP-chip) indicate that there are no significant changes in the distribution of Pol II at the 3’ end of this gene upon depletion of Pcf11 (Figure 3-10).

⁶ Results represented in Figure 3-10B and C were generated from one of the biological replicate experiments. Results generated from the other set of experiments are presented in Figure F-2A and C.
Figure 3-10: Pcf11 depletion does not disrupt termination of the U1_95Ca gene. A) Chromatin immunoprecipitation assay was used to detect Pol II occupancy at the U1-95Ca gene. PCR amplified regions (horizontal lines) in ChIP analysis and the annotated 3’ end of the U1_95Ca gene (vertical arrow) are indicated in a schematic above the bar graph. Pol II signals in downstream regions are normalized against the signal in the most upstream region. Signals of pre-immune serum immunoprecipitated samples are
normalized against signals from corresponding samples immunoprecipitated with antibody against the Rpb3 subunit of Pol II. Each bar represents the mean of three independent experiments. A *P*-value is calculated for comparison between unequal Pol II ChIP signals using the unpaired student *t*-test. Both calculated *P*-values are larger than 0.05, indicating that the differences are not statistically significant. B) Permanganate footprinting of the U1_95Ca gene did not detect termination defects upon Pcf11 depletion. *Drosophila* S2R+ cells were treated with 10 mM potassium permanganate for 1 min on ice. The intensity of the permanganate reactivity of each thymine was represented as a ratio over the reactivity of the same thymine in the naked DNA control. The red line represents the smoothed Pol II distribution in LacZ RNAi treated samples. The blue line represents the smoothed Pol II distribution in Pcf11 RNAi treated samples. The orange dots represent all T residues. The annotated 3’ end of the U1_95Ca gene is indicated by the vertical dotted line. C) The intensity of the permanganate reactivity at the T residues in B was log2 transformed and displayed using Treeview. The top to bottom direction corresponds to the 5’ to 3’ direction of transcription. Black indicates the same intensity as the naked DNA lane. Red and green reflect stronger or weaker signals compared to the naked DNA lane, respectively. D) Display (UCSC browser) of Pol II ChIP-chip data for the U1_95Ca gene before and after Pcf11 depletion. The height of vertical bars corresponds to the intensity of Pol II ChIP signals. The black arrow indicates the transcription direction of the U1_95Ca gene.

### 3.2.8 The behavior of Pol II in the 3’ region appears to be linked to nucleosome positions

A striking feature of the Pol II distributions at the 3’ end of genes is the paucity of Pol II in the vicinity of the polyadenylation site (Figure 3-6 and Figure 3-11). This is reminiscent of the paucity of nucleosomes detected at the end of the open reading frames in *Drosophila* embryos (Mavrich *et al.*, 2008b). Overlaying the genome wide nucleosome distribution data (from Mavrich *et al.*, 2008b, replotted using the polyadenylation site as the reference point) with Pol II distribution data showed that the minima for both distributions coincide (Figure 3-11A). One explanation for this alignment is that the absence of a nucleosome at the polyadenylation site would allow Pol II to rapidly transit this region, thus preventing an accumulation of paused Pol II. Accordingly, nucleosomes
have been shown to inhibit transcription elongation by Pol II in vitro (Bondarenko et al., 2006; Brown et al., 1996; Kireeva et al., 2005).

Figure 3-11: The linkage between Pol II and nucleosome distributions at the 3’ end of genes. Composite genome wide distribution of nucleosomes (gray background) and of Pol II in LacZ RNAi treated cells (red curve) and Pcf11 RNAi treated cells (blue curve) within +/-1.5 kb of the polyadenylation site. Nucleosome ChIP-chip data was generated by Mavrich et al. (Mavrich et al., 2008b).

A corollary of this model is that the pausing of Pol II detected downstream from the polyadenylation site might be linked to the presence of nucleosomes. Indeed, Figure 3-11 showed coincident increases in both Pol II and nucleosome occupancies in a region from the polyadenylation site to 500 nucleotides further downstream. To test the hypothesis that nucleosomes downstream of the polyadenylation site form a barrier to moving Pol II molecules, I examined the positions of nucleosomes in the 3’ region of the hsp70Aa, hsp70Ab, and hsp23 genes. For both hsp70A genes, several peaks of Pol II occupancy map adjacent to or in the promoter proximal half of positioned nucleosomes in
the control cells (Figure 3-4A and B), consistent with the co-elevation of Pol II and nucleosome occupancies downstream from the polyadenylation site in the genome wide analysis (Figure 3-11A and B). Upon depleting Pcf11, peaks in the Pol II distribution distal to the polyadenylation site (P2 for both hsp70A genes, Figure 3-4A and B) were seen to increase but remained located adjacent to nucleosomes, further supporting the conclusion that nucleosomes are blocking elongation of Pol II. On the other hand, nucleosomes were not detected at or immediately downstream of the peak of Pol II occupancy at the end of the hsp23 gene in control cells (Figure 3-4C), suggesting that something other than a nucleosome is inducing this pause (see discussion). Upon Pcf11 depletion, a new peak in the Pol II distribution (P\*; Figure 3-4C) was formed around 200 nucleotides downstream from where Pol II was observed to accumulate in control cells. The genome wide nucleosome mapping data did not detect a nucleosome immediately downstream of the newly formed peak (P\*). However, Cartwright and Elgin (Cartwright and Elgin, 1986) detected a complex pattern of micrococcal nuclease cuts and MPE cuts that were proposed to be caused by a nucleosome array in the region just downstream of P\*. The basis for the discrepancy between the whole genome mapping of nucleosomes and the early results of Cartwright and Elgin are not known, but it is possible that the region is organized with non-histone proteins rather than nucleosomes.
3.3 Discussion

3.3.1 Pcf11 functions in termination at many protein-encoding genes

Pcf11 is the only protein among the cadre of factors implicated in termination and polyadenylation that has been shown to dismantle a Pol II elongation complex in vitro (Zhang et al., 2005; Zhang and Gilmour, 2006). I found that depletion of Pcf11 caused the level of Pol II detected at the 3’ end of thousands of genes to increase and the distribution of Pol II to shift downstream from the original location. This was not simply because of an increase in transcription initiation since the density of Pol II detected at the beginning of most of these genes actually decreased (Figure 3-8). ChIP and permanganate genomic footprinting of three heat shock genes also showed that the Pol II level increased and peaks of Pol II density shifted downstream following depletion of Pcf11. The broad impact of Pcf11 on Pol II at the 3’ end of genes indicates that Pcf11 has a central role in transcription termination.

3.3.2 Alternate termination mechanisms are implicated by differences in 3’ Pol II distributions

Cluster analysis of the Pol II distribution detected in the 3’ regions of over 4000 genes uncovered three distinct distribution patterns. One distribution (category 3) matches what has been observed previously on a small number of genes (Aranda and Proudfoot, 1999; Birse et al., 1997; Enriquez-Harris et al., 1991; Yonaha and Proudfoot, 1999). Peaks of Pol II were detected downstream from the polyadenylation site,
consistent with pausing of Pol II at specific sites after transcribing the polyadenylation signal. Another distribution (category 1) exhibited relatively high density of Pol II on each side of the polyadenylation site with an obvious decrease in Pol II density at the polyadenylation site. Both category 1 and 3 showed a gradual decrease in Pol II occupancy after 750 nucleotides downstream from the polyadenylation site (Figure 3-6B), representing the occurrence of termination.

A third distribution (category 2) of Pol II indicates that Pol II terminates in close proximity to the polyadenylation site without concomitant pausing since Pol II was detected upstream but not downstream from the polyadenylation site (Figure 3-6). Based on the bulk of information found in the literature, abrupt termination at the polyadenylation signal would seem to be highly unusual. However, there are two studies supporting the existence of abrupt transcription termination. Beyer and colleagues used electron microscopy to examine the distribution of Pol II on 114 Drosophila genes and found that chains of Pol II molecules ended abruptly at the ends of genes in 69% of investigated cases (Osheim et al., 2002). Gullerova et al. demonstrated that, in S. pombe, the binding of cohesin protein to the intergenic regions of two convergent genes prevented the elongating polymerase molecules from transcribing beyond the polyadenylation site and therefore led to polyadenylation site proximal termination (Gullerova and Proudfoot, 2008). However, the binding of cohesin between convergent genes might not be the cause of the abrupt transcription termination on category 2 genes. Unlike yeast where cohesin binds almost exclusively between genes, Drosophila cohesin preferentially binds transcribed regions (Misulovin et al., 2008). In addition, in my
clustering analysis where filtering was applied to remove neighboring genes, the category 2 Pol II distribution was still present (Figure 3-7D) indicating that convergent genes are not required for transcription to terminate abruptly at the polyadenylation site in *Drosophila*.

I do not know why the genes represented in category 2 terminate transcription terminated at the polyadenylation site, but one possibility involves gene looping. Proudfoot and Hempsey have provided evidence that the promoter and terminator of some gene are in close proximity (Ansari and Hampsey, 2005; O'Sullivan *et al.*, 2004). It was proposed that the juxtaposition of the termination region to the promoter could enhance the efficiency of termination and recycling of Pol II (Singh and Hampsey, 2007). Although there is no evidence indicating that gene looping occurs in *Drosophila*, it is still an intriguing idea that gene looping may result in an abrupt transcription termination on some genes (for example, category 2 genes in my ChIP-chip analysis) through a direct blocking of elongating Pol II by the initiation complex.

The three distinct distribution patterns of Pol II at the 3’ end of genes indicate that alternate termination mechanisms might exist for protein-coding genes. In support of this possibility, approximately 30% of the genes with Pol II terminating downstream from the polyadenylation signal exhibited an increase in Pol II levels upon depletion of Pcf11 whereas only 13% of the genes with polyadenylation site proximal termination were affected (*P*-value < 10^-38). As most studies of termination have focused on cases where Pol II terminates downstream from the polyadenylation site, it will be important to analyze this new type of termination to fully understand termination.
A GO analysis was carried out in order to identify common biological functions of genes represented in each category (Table 1). However, nothing extraordinary was found. Category 1 was modestly enriched in ribosomal genes than the other two categories, a finding consistent with the higher expression levels of category 1 genes (Figure 3-7A) since ribosomal genes are known to be highly transcribed.

Although I favor the idea that the Pol II ChIP signals detected upstream from the polyadenylation site in category 2 represent transcription that terminates at the polyadenylation site, another explanation for these Pol II ChIP signals is anti-sense transcription initiated at the 3’ end of a gene. Mavrich et al. detected some anti-sense transcripts initiating in the vicinity of the 3’ end of genes in yeast (Mavrich et al., 2008a). However, those genes make up only a small fraction of the yeast genome while the number of category 2 genes in my clustering analysis is quite large. Recent results from a genome-wide analysis of nuclear run-on products in human cells also failed to detect significant amounts of anti-sense transcription originated from the 3’ end of human genes (Core et al., 2008). Therefore, chances are small that the Pol II ChIP signals on category 2 genes represent anti-sense transcription.

3.3.3 Termination-coupled pausing downstream from the polyadenylation site

A peak in the Pol II distribution centered in a region approximately 500 nucleotides downstream from the polyadenylation site was observed for 3116 Drosophila genes (category 1 and category 3, Figure 3-6B). Recent results from a genome-wide analysis of nuclear run-on products also shows a peak of Pol II in the composite plots
downstream from the polyadenylation site in human cells, although this peak is located approximately 1.5 kb downstream from the polyadenylation site (Core et al., 2008). Pol II tends to pause closer to the polyadenylation site in *Drosophila* than in human cells. The basis for this difference is unclear, but the reason for having Pol II terminate closer to the polyadenylation site in *Drosophila* than human cells could be due to the compact organization of genes in *Drosophila*.

Pausing has always been viewed as an obligate step in elongation that precedes termination (Aranda and Proudfoot, 1999; Birse et al., 1997; Enriquez-Harris et al., 1991; Yonaha and Proudfoot, 1999) and different studies have suggested various causes for such pausing. In yeast, several elongation factors appear to dissociate from Pol II after Pol II transcribes the polyadenylation signal (Kim et al., 2004a), which in turn could render Pol II more susceptible to pausing. Other evidence suggests that the polyadenylation signal by itself is sufficient to pause Pol II (Nag et al., 2006; Nag et al., 2007; Orozco et al., 2002; Park et al., 2004; Tran et al., 2001).

My comparison of the genome wide nucleosome distribution and the Pol II distribution and my high resolution analysis of the behavior of Pol II at the 3’ end of two hsp70A genes provide evidence that some pauses are linked to the chromatin structure. It was proposed that the termination function of Pcf11 requires the Pol II molecule to be in a transiently paused or slowly elongating state, since Pcf11 could dismantle an elongating complex *in vitro* only when transcription was carried out at a low nucleotide concentration (Zhang and Gilmour, 2006). Therefore, nucleosome-induced pauses downstream from the polyadenylation site might render the elongation complex susceptible to dissociation by Pcf11.
Recently, Lis and colleagues showed that there was a rapidly loss of canonical nucleosomes at the 3’ end of the two hsp70A genes when the genes were subjected to heat shock induction (Petesch and Lis, 2008). In this study, nucleosomes were detected by monitoring the level of protection against extensive micrococcal nuclease digestion afforded by the chromatin structure in 100 bp regions (Petesch and Lis, 2008). While the results indicate that the canonical nucleosome structure is disrupted upon heat shock, they do not establish that nucleosomes are absent. It is possible that the nucleosome structure is altered in a way that increases micrococcal nuclease cutting within the 147 base pair region associated with the histone octamer. Two results favor the hypothesis that the structure of the nucleosomes is altered rather than the nucleosomes being ejected. First, analysis of the chromatin structure of heat shock genes with nucleases indicates that the DNA is still packaged into a repeating structure (Cartwright and Elgin, 1986; Udvardy et al., 1985). Second, ChIP analyses show that core histones are still associated with the heat shock genes under heat shocked conditions (Petesch and Lis, 2008; Solomon et al., 1988). Hence, the nucleosomes that cause Pol II to pause in the 3’ region could have a structure that differs from that of bulk chromatin.

Previous work in yeast provides links between chromatin structure and termination (Alen et al., 2002; Morillon et al., 2003). Mutations in the chromatin remodeling proteins Chd1 and Isw1 impaired termination and perturbed the chromatin structure at the 3’ end of several genes. These results along with mine indicate that understanding termination at many genes will likely require an assessment of the contribution of chromatin. It will be interesting to investigate whether perturbing the
chromatin structure at the 3’ end of genes in higher eukaryotes alters termination and whether the presence of a nucleosome facilitates Pcf11-mediated dismantling in vitro.

My result for the hsp23 gene indicates that there are likely to be nucleosome-independent pauses that are involved in termination at some genes. One possible cause for such nucleosome-independent pauses is a sequence that inhibits elongation. In yeast, ChIP experiments indicate that several positive elongation factors dissociate from the gene downstream from the polyadenylation signal (Kim et al., 2004a). Such dissociation might enhance the probability of pausing downstream from the polyadenylation site. Another possible cause for nucleosome-independent pauses is a protein bound to the DNA. Using electron microscopy, Beyer and colleagues observed cases where trains of Drosophila polymerases appeared to terminate at sites associated with a particle whose size exceeded that of a Pol II molecule or a nucleosome (Osheim et al., 2002). In human, a G-rich element located between two closely spaced genes, C2 and factor B, binds to a zinc finger protein, MAZ (Ashfield et al., 1991). This protein-DNA binding is able to cause Pol II to terminate transcription in vivo (Ashfield et al., 1994). Later studies showed that the G-rich element by itself causes Pol II to pause in vitro (Yonaha and Proudfoot, 1999). It was proposed that the binding of MAZ enhances the specificity of this pausing effect.

The Pol II distribution detected in my ChIP-chip analysis exhibited a striking paucity of Pol II at the polyadenylation site, correlating with a low density of nucleosomes that map in the same region (Figure 3-11). This low density of nucleosomes may be linked to both the absence of nucleosome positioning sequences, such as the repeating 10 bp AA/TT dinucleotides, and the presence of DNA sequences
that antagonize nucleosome assembly, such as the polyadenylation signal (Mavrich et al., 2008a; Mavrich et al., 2008b). The nucleosome depleted region at the 3’ end of genes could be a zone that is rapidly transited by Pol II, thus resulting in a low density of Pol II in this region.

### 3.3.4 Pcf11 depletion decreases transcription initiation on many genes

Besides the increase in Pol II occupancy downstream from the polyadenylation site, I also observed a decrease in Pol II occupancy at the 3’ end of many genes upon Pcf11 depletion. Most decreases in Pol II occupancy at the 3’ end were accompanied by decreases in Pol II occupancy at the 5’ end of the gene (Figure 3-8A), suggesting that the depletion of Pcf11 was impairing transcription initiation. Although one postulated role for termination is that it is required to recycle Pol II, no direct evidence for this has ever been provided. I propose that the decreases I observed in Pol II occupancy were due to a depletion of the pool of free Pol II. Kimura et al. determined that approximately 34% of the 320,000 molecules of Pol II in HeLa cells are transcriptionally engaged (Kimura et al., 1999), so even a modest impairment in the dissociation of Pol II could affect the free pool. In addition, Steinmetz et al. provided evidence that a mutation in yeast Sen1 protein, a Pol II termination factor for non-coding RNA genes and short protein-encoding genes, caused a decrease in Pol II occupancy on many genes (Steinmetz et al., 2006). Therefore, whether the Pol II occupancy downstream from a polyadenylation site will increase upon depletion of Pcf11 may dependent on a balance between a defect in termination and a decrease in transcription initiation caused by loss of Pcf11. The actual number of
Drosophila genes whose termination is dependent on Pcf11 may surpass the number (2611) indicated in Figure 3-8A.

Another possible cause for the decrease in Pol II density at the promoters is transcription interference. Pol II molecules that fail to terminate upstream could interfere with transcription of downstream genes by displacing activators and other components of the transcriptional machinery from downstream genes (Greger et al., 2000; Shearwin et al., 2005). Although I did not detect an increase in Pol II levels upstream from the promoters of the genes whose transcription initiation appeared to be impaired after Pcf11 depletion, it remains possible that small amounts of Pol II are enough to cause transcription interference while escaping the detection of ChIP-chip analysis.

3.3.5 Some Drosophila genes might be controlled by Pcf11-mediated premature termination

I was surprised to find that depletion of Pcf11 caused the level of Pol II associated with many genes to increase. On 1902 genes, the increase in Pol II occupancy at the 5’ end extended to the 3’ end (Figure 3-8A). On 786 genes, the increase was limited to the 5’ end (Figure 3-8B). There are also some genes that showed an increase in Pol II occupancy upstream of their promoters, possibly representing cases where anti-sense transcription was enhanced (Figure 3-8B). Some of the increases in Pol II occupancy in the body of genes were accompanied by increases in mRNA levels (Figure 3-9).

Recently, Pcf11 was found to repress transcription of a latent HIV provirus by causing premature termination (Zhang et al., 2007b). My ChIP-chip data and RNA measurements
suggest that Pcf11-mediated premature termination is also repressing expression of some cellular genes. One gene, CG13075-RA, exhibited more than 100 fold increase in mRNA level after depletion of Pcf11. This gene is normally weakly expressed in *Drosophila S2* cells (Rehwinkel *et al.*, 2006).

Recently, several studies have shown that Pol II is concentrated at the 5’ end of thousands of genes in human and *Drosophila* cells (Gilchrist *et al.*, 2008; Guenther *et al.*, 2007; Muse *et al.*, 2007; Zeitlinger *et al.*, 2007). In *Drosophila*, this has been ascribed to the action of NELF, a protein that inhibits transcription elongation (Lee *et al.*, 2008; Wu *et al.*, 2005; Wu *et al.*, 2003). The results presented here identify premature termination by Pcf11 as another mechanism by which Pol II can be concentrated at the 5’ end of genes. Further investigation of the participation of Pcf11 in the elongation process will likely enhance our understanding of transcription regulation.

### 3.3.6 The function of *Drosophila* Pcf11 in termination may require a polyadenylation signal

Termination of the snRNA gene, U1_95Ca, was not affected by depletion of Pcf11, implicating the necessity of a polyadenylation signal for the termination function of *Drosophila* Pcf11. The AATAAA hexamer is thought to be essential for recruiting cleavage and polyadenylation factors to the elongation complex (Kim *et al.*, 2004a). The U1_95Ca gene, which lacks the polyadenylation signal, may not be able to recruit Pcf11 to the 3’ end of the gene. It was also shown that the polyadenylation signal is important to cause pausing of Pol II (Nag *et al.*, 2007), which could facilitate the dismantling activity
of Pcf11. Neither the ChIP analysis nor the high resolution permanganate assay detected an accumulation of Pol II molecules at the 3’ end of the U1_95Ca gene, providing another explanation for the failure to detect termination defects on this gene upon Pcf11 depletion. Thus, transcription termination of such non-polyadenylated RNAs as snRNAs and snoRNAs may involve a Pcf11-independent termination mechanism in *Drosophila*.

My result is different from what has been shown in yeast. Kim *et al.* showed that certain mutation alleles of yeast Pcf11 caused termination defects at snoRNA genes (Kim *et al.*, 2006). The discrepancy between my result and Kim’s result might reflect different termination mechanisms employed by snRNA genes in different organisms. In yeast, Nrd1 and Nab3 proteins have been shown to regulate termination of snRNA genes (Steinmetz *et al.*, 2001). Mutation of Ssu72 and Sen1 proteins also impairs termination of snoRNA genes (Ganem *et al.*, 2003; Kim *et al.*, 2006; Rasmussen and Culbertson, 1998; Ursic *et al.*, 1997). It will be interesting to investigate the effects of disrupting the counterparts of these proteins in *Drosophila* on transcription termination. In human, the phosphorylation of the ser7 residue of the CTD heptad facilitates the interaction of CTD with the snRNA gene-specific Integrator complex, a 12 subunit complex mediating the 3’-end processing of snRNA genes (Baillat *et al.*, 2005). It was proposed that the Integrator complex might participate in termination of human snRNA genes. However, a blast search did not identify any *Drosophila* proteins homologous to subunits of the human Integrator complex.
Chapter 4

Genome Wide Distribution of Pol II at the 5’ End of Genes

4.1 Introduction

Several recent analyses of the distribution of Pol II across the genome reveal that Pol II is concentrated at the beginning of thousands of genes in Drosophila and human cells (Guenther et al., 2007; Muse et al., 2007; Steinmetz et al., 2006; Zeitlinger et al., 2007). The level of Pol II concentrated at the promoter tends to increase with the level of expression, suggesting that there is a rate-limiting step in the transcription process that occurs after Pol II associates with the promoter of a gene (Saunders et al., 2006). Other work indicates that this rate-limiting step is often a pause in elongation that occurs 20 to 50 nucleotides downstream from the transcription start site (Giardina et al., 1992; Lee et al., 2008; Lis, 1998). Approximately 12% of all known Drosophila genes were categorized as being stalled based on the ratio of the levels of Pol II detected at the promoter to the body of a gene (Muse et al., 2007; Zeitlinger et al., 2007). Many of these genes are stress-response genes and developmental-control genes. Genes with Pol II poised at their 5’ end could have the potential to be quickly induced so that rapid changes in transcription during development can be achieved (Wang et al., 2007).
Several proteins appear to be important for pausing Pol II in the promoter proximal region. Both negative elongation factor (NELF) and DRB\textsuperscript{7} sensitivity inducing factor (DSIF) (Wada \textit{et al.}, 1998; Yamaguchi \textit{et al.}, 1999) have been shown to be involved in promoter proximal pausing at the hsp70 gene \textit{in vivo} in \textit{Drosophila} (Wu \textit{et al.}, 2005; Wu \textit{et al.}, 2003). NELF and DSIF associate with the elongation complexes and inhibit elongation. GAGA factor, a DNA binding protein specifically recognizing GA sequences, is also involved in the pausing. GAGA factor associates with the general transcription factor, TFIIID, and the chromatin remodeling factor, NURF (Giot \textit{et al.}, 2003; Nakayama \textit{et al.}, 2007; Xiao \textit{et al.}, 2001). These interactions could be involved in recruiting Pol II to the promoter, which would be a prerequisite for establishing a paused Pol II. However, nothing directly links GAGA factor to pausing per se except for the finding that the removal or addition of a GAGA element to transgenes in \textit{Drosophila} inhibits or promotes, respectively, the formation of paused Pol II (Lee \textit{et al.}, 1992; Wang \textit{et al.}, 2005).

Although substantial information has been generated from the aforementioned ChIP-chip analyses, the resolution of these studies is not as high as that afforded by the Affymetrix tiling arrays. In Chapter 3, I used the Affymetrix \textit{Drosophila} tiling 1.0R arrays to analyze Pol II at the 3’ end of genes. I found my data to be far more informative about Pol II in \textit{Drosophila} than data of lower resolution reported by other investigators. In this chapter, I use data from high resolution arrays to analyze Pol II distributions near the promoter regions. In addition, I have compared the Pol II distribution to the

\textsuperscript{7} DRB: 5,6-dichloro-1-β-D-ribofuranosyl benzimidazole, a nucleoside analogue, inhibitor of a kinase P-TEFb (positive transcription elongation factor b).
distributions of NELF and GAGA factor to better understand the relationship between pausing and these two factors. The results of my analysis identify thousands of genes with Pol II concentrated at the 5’ end as has been described in previous studies (Muse et al., 2007; Zeitlinger et al., 2007). In addition, the resolution afforded by the high density arrays provides new information. I have identified a collection of genes that have Pol II concentrated upstream from the annotated transcription start site. I have also obtained evidence that GAGA factor accompanies Pol II during elongation on some genes.

4.2 Results

4.2.1 Pol II is concentrated near transcription start sites (TSS)

A plot of the composite distribution of Pol II at the 5’ end of genes reveals a strong peak of Pol II shifted slightly downstream from the transcription start site (Figure 4-1). This high level of Pol II near the 5’ end of genes has been seen in other genome wide studies of Pol II distributions (Guenther et al., 2007; Muse et al., 2007; Steinmetz et al., 2006; Zeitlinger et al., 2007), suggesting that a rate-limiting step in transcription regulation occurs in this region on many genes.
4.2.2 Three distribution patterns of Pol II are evident at the 5’ end of genes

The apex in the composite distribution of Pol II at the 5’ end of genes is located at +100. It seemed likely that this reflected Pol II paused in the promoter proximal region of numerous genes. However this was shifted slightly downstream from the +20 to +50 region where Pol II has been found to pause when mapped by permanganate genomic footprinting. One explanation for this could be that the composite distribution of all genes with Pol II was hiding distinct categories of genes that skewed the overall distribution. As was seen in chapter 3 for the analysis of Pol II at the 3’ end, K-mean clustering analyses can deconvolute the composite distribution into distinct categories of genes.

Figure 4-1: Pol II is concentrated at the 5’ end of genes. Plot of composite genome wide distribution of Pol II ChIP signals (Affymetrix *Drosophila* tiling arrays) from -1 kb to 1 kb of all annotated transcription start sites for which Pol II could be detected. Transcription start sites that lacked Pol II in this region were not included. The ChIP experiment was performed on *Drosophila* S2R+ cells treated with LacZ RNAi.
Using K-mean clustering, I was able to divide *Drosophila* genes into three categories and separate genes that had Pol II concentrated in the promoter proximal region from the others. Category II in Figure 4-2 represents 3794 genes whose peak in the composite distribution of Pol II is centered at +85. The Pol II distribution at the 5’ end of category II genes suggests that they are likely to have paused Pol II in the promoter proximal region. Category III consists of 2099 genes that have Pol II ChIP signals extended into the body of the gene, indicating that these genes are probably actively transcribed.

Category I genes seem unusual because Pol II resides mainly upstream from the transcription start site. It is possible that these Pol II ChIP signals result from unannotated or mis-annotated promoters, or are due to transcription occurring at neighboring genes. To test the latter possibility, I removed from the clustering analysis the genes that have a transcription start site within 2 kb of another transcription start site or a polyadenylation site. Even after filtering, the fraction of genes with Pol II located upstream from the promoter was not significantly altered, indicating that this category is not generated by neighboring genes (Figure 4-2B).
Figure 4-2: Three distribution patterns of Pol II are evident at the 5’ end of genes. A) K-mean clustering analysis of Pol II distributions detected at the 5’ end of genes. Pol II ChIP signals within +/-1 kb of the transcription start sites of 8378 genes are designated with color (intensity of red color corresponds to the Pol II level; black color indicates Pol II level below 5% FDR threshold). Each row represents an individual gene. All genes are aligned at the transcription start site (indicated by the black dot on top of the Treeview panel). The genes separate into 3 categories. B) Clustering analysis of the Pol II distributions at the 5’ end were done in the same way as in panel A, except that analyzed candidates are filtered to remove any gene whose transcription start site is within 2 kb of another transcription start site or any polyadenylation site. C) Composite Pol II distribution from -1 kb to 1 kb of transcription start site for the genes in category I (orange), II (purple) and III (green).
4.2.3 Comparison of NELF and GAGA factor distribution for each category of genes

The distinct distribution patterns for the three categories of genes are striking and are possibly caused by association with different transcription factors. It was shown previously that there is a strong correlation between NELF, GAGA factor and promoter proximal pausing (Lee et al., 2008). GAGA factor is also involved in transcription initiation and chromatin remodeling at promoters (Nakayama et al., 2007; Wilkins and Lis, 1997). To investigate whether there is any correlation between the behavior of Pol II at the 5’ end of genes and the distributions of NELF and GAGA factor, I took the available ChIP-chip data for NELF-B (a subunit of NELF complex) and GAGA factor generated from Drosophila S2 cells (Lee et al., 2008) and plotted the composite distributions of these two factors for category I, II and III genes.

NELF was enriched near the promoters of category II genes. The peak of the NELF distribution aligned well with the peak of the Pol II distribution for category II genes (Figure 4-3 and Figure 4-2C), indicating that NELF could be involved in causing Pol II to concentrate at the 5’ end of these genes. NELF was also detected at category III genes (Figure 4-3). However, unlike the Pol II distribution which extends into the body of category III genes, the NELF distribution is concentrated at the promoter. NELF could be transiently associating with Pol II as Pol II transcribes the promoter proximal region but dissociating as the Pol II progresses into the body of the gene. Finally, NELF was observed to occupy regions upstream from the transcription start site of category I genes (Figure 4-3). Since NELF is concentrated at the 5’ region of both category II and III
genes, the distribution of NELF for category I genes is consistent with the hypothesis that unannotated or mis-annotated genes reside in these regions.

Figure 4-3: NELF-B distribution for three categories of genes. Composite NELF-B distribution from -1 kb to 1 kb of the transcription start site for genes in category I (orange), II (purple) and III (green) is plotted. NELF-B ChIP-chip data was generated by Lee et al. (Lee et al., 2008).

GAGA factor was concentrated in the region upstream from the transcription start site of category II genes, consistent with what one would anticipate for a typical transcriptional regulator (Figure 4-4A). Category III genes had GAGA factor much more broadly distributed with GAGA factor evident at comparable levels upstream and downstream from the transcription start site. The presence of GAGA factor downstream from the transcription start raises the possibility that GAGA factor could be co-localizing with Pol II. Category I genes have GAGA factor broadly distributed but in the upstream direction relative to the transcription start site.
I analyzed the distribution of a GAGA element, GAGAGAG, to see if this could account for the disparate distributions of GAGA factor detected on my three categories of genes. A peak in the GAGA element distribution was readily apparent upstream from the transcription start site of category II genes that aligned with the peak in the GAGA factor distribution (Figure 4-4B). It seems, however, that proximity to the region upstream from the transcription start is also a determinant for the binding of GAGA factor since GAGA factor was not enriched in the vicinity of +500 for category II genes where there is another peak in the GAGA element distribution.

Category III genes had a clear enrichment of GAGA elements downstream from the transcription start site but lacked the prominent peak found upstream from the transcription start site. This enrichment of GAGA elements in the downstream region could contribute to the presence of GAGA factor detected downstream from the transcription start of category III genes. However, the distribution of the element does not align precisely with the protein, raising the possibility that sequence alone does not dictate GAGA factor binding.

In contrast to category II and III genes, GAGA elements are randomly distributed for category I genes. This would be consistent with the hypothesis that this category arises because of deficiencies in the genome annotation, since such deficiencies would have obscured the reference points used to define where Pol II was located.
4.2.4 Expression levels of category I, II, and III genes

I compared the expression levels of the three categories of genes to determine whether the presence of Pol II in the body of a gene corresponded to active transcription and whether those cases with Pol II concentrated near the promoter might be transcriptionally inactive. As shown in Figure 4-5, category III genes tend to have higher expression than the other two categories of genes. In accordance with this, an analysis of the types of genes found in each category using a gene ontology algorithm revealed that category III genes are highly enriched with genes encoding ribosomal proteins (Table 4-1), which are known to be highly transcribed. Hence the elevated levels of Pol II detected
in the body of category III genes are indicative of Pol II molecules undergoing active elongation.

The analysis of expression levels also indicated that all three categories associated with Pol II had an average expression level that was approximately 16 times higher than the genes where Pol II was not detected. This serves as validation of my Pol II ChIP-chip analysis.

Figure 4-5: Comparison of the expression levels of category I, II and III genes and genes with Pol II levels below 5% FDR threshold within +/-1 kb of the transcription start site. Steady state mRNA values were extracted from published expression microarray data for Drosophila S2 cells (Rehwinkel et al., 2004). The median expression level is indicated by the black center line, and the first and third quartiles are the edges of the boxed area (inter-quartile range). The extreme values (within 1.5 times the inter-quartile range from the first or third quartile) are the ends of the lines extending from the box. Points at a greater distance than the extreme values from the median are plotted individually. The \( P \)-value is calculated for each pair wise comparison among the four categories of genes using Wilcoxon-Mann-Whitney test. A \( P \)-value smaller than 0.05 was obtained for all pair wise comparisons indicating that the median values (black lines) are significantly different from each other.
4.2.5 Bulk and H2A.Z nucleosome distributions for category I, II and III genes

Recent genome-wide mapping of canonical nucleosomes and H2A.Z-containing nucleosomes in *Drosophila* embryos reveal that chromatin structure at the 5’ end of genes is organized in a distinct way. Results from yeast indicate that H2A.Z is incorporated into the nucleosome during transcription and serves to mark genes that are being or have been transcribed (Brickner *et al.*, 2007; Santisteban *et al.*, 2000). Based on the composite distribution of both canonical and H2A.Z nucleosomes, the region

---

**Table 4-1: GO analysis of category I, II and III genes**

<table>
<thead>
<tr>
<th>Category I (1882)*</th>
<th>Category II (2601)*</th>
<th>Category III (1424)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zinc-finger protein/Metal-binding</td>
<td>Zinc-finger protein/Metal-binding</td>
<td>Ribosomal protein</td>
</tr>
<tr>
<td>21.62 (136)</td>
<td>21.84 (196)</td>
<td>36.4 (111)</td>
</tr>
<tr>
<td>Transcription regulation/ Nucleus</td>
<td>Transcription regulation/ Nucleus</td>
<td>Zinc-finger protein/Metal-binding</td>
</tr>
<tr>
<td>17.52 (167)</td>
<td>19.19 (206)</td>
<td>7.12 (91)</td>
</tr>
<tr>
<td>Kinase</td>
<td>Kinase</td>
<td>Kinase/ Transferase</td>
</tr>
<tr>
<td>4.94 (69)</td>
<td>14.84 (185)</td>
<td>6.68 (107)</td>
</tr>
<tr>
<td>mRNA splicing</td>
<td>Cell cycle/Mitosis</td>
<td>Cell cycle/Mitosis</td>
</tr>
<tr>
<td>2.77 (10)</td>
<td>5.03 (28)</td>
<td>5.15 (20)</td>
</tr>
<tr>
<td>Cell cycle/Mitosis</td>
<td>Stress response</td>
<td>Isomerase</td>
</tr>
<tr>
<td>1.4 (13)</td>
<td>3.79 (15)</td>
<td>4.68 (33)</td>
</tr>
<tr>
<td></td>
<td>ATP synthesis/Ion transport</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.44 (25)</td>
<td></td>
</tr>
</tbody>
</table>

Each biological process cluster is associated with an enrichment score. The higher the score, more significantly enriched are the genes associated with such biological process within the input population. Enrichment score is the geometric mean (in -log scale) of several calculated *P*-values (Huang da *et al.*, 2007). An enrichment score of 1.3 corresponds to a *P*-value of 0.05. The number of genes within each biological process cluster is listed in parenthesis.

*The number indicates the total number of genes with associated GO terms.
encompassing the transcription start site is nucleosome-free and the first nucleosome within a gene, known as the +1 nucleosome, is centered at +135 for genes that lacked paused Pol II and at +145 for genes that had paused Pol II (Mavrich et al., 2008b). I wanted to determine the composite nucleosome distributions on my three categories of genes to determine how these patterns compared to those previously described. I anticipated that my class II genes that appeared to have paused Pol II would have the +1 nucleosome centered at +145. I wondered what the nucleosome distribution for category I genes would look like since the distributions of Pol II, GAGA factor and NELF did not fit any currently characterized models of genes.

The composite distributions of both H2A.Z and bulk nucleosomes for all three categories of genes follow an oscillating pattern similar to the pattern previously described by Mavrich and colleagues (Mavrich et al., 2008b) (Figure 4-6A and B). In addition, the 5’ region of all three categories of genes tends to be nucleosome free. The similarity among the nucleosome distributions over the region encompassing the transcription start site and in the body of the gene argues that the unusual characteristics of category I genes is not simply a consequence of the transcription start site, which served as the reference point for the composite plots, being mis-localized.

Table 4-2 summarizes the location of the +1 nucleosomes exhibited by my composite distributions and those reported by Mavrich et al., 2008b. All of the distributions associated with genes that have Pol II locate the first nucleosome in the region between +115 and +140. There is a notable deviation between my results and those of Mavrich et al., 2008b. Mavrich et al. determined that the +1 nucleosome
associated with genes that had paused Pol II was centered at +145 whereas my three categories are centered at +125 or +130. In particular, my category II genes, which have the strongest similarity to genes with paused Pol II, has the +1 H2A.Z containing nucleosome centered at +125 and the bulk nucleosome centered at +110. Although this discrepancy appears small, it could significantly impact the conclusion that a nucleosome contributes to the pause by blocking elongation. In support of this conclusion, Mavrich et al. observed that the +1 nucleosome was centered at +135 for genes without paused Pol II and at +145 for genes with paused Pol II leading to the idea that an interaction between Pol II and the nucleosome influence the position of the +1 nucleosome.
Figure 4-6: Bulk and H2A.Z nucleosome distributions are similar for category I, II and III genes. A) Composite H2A.Z nucleosome distribution from -1 kb to 1 kb of transcription start site for genes in category I (orange), II (purple) and III (green). Genome wide H2A.Z nucleosome mapping data was generated by Mavrich et al. (Mavrich et al., 2008b). B) Composite bulk nucleosome distribution from -1 kb to 1 kb of transcription start site for genes in category I (orange), II (purple) and III (green). Genome wide bulk nucleosome mapping data was generated by Mavrich et al. (Mavrich et al., 2008b). C) Composite H2A.Z nucleosome distribution from -1 kb to 1 kb of transcription start site for genes with (red) and without (blue) significant Pol II ChIP signals (above 5% FDR threshold). D) Composite bulk nucleosome distribution from -1 kb to 1 kb of the transcription start site for genes with (red) and without (blue) significant Pol II ChIP signals (above 5% FDR threshold) in this region. The genes with significant Pol II ChIP signals are the same 8378 genes included in Figure 4-2A. All the other Drosophila genes other than the 8378 genes are counted as genes without significant Pol II ChIP signals.
4.2.6 Nucleosome locations on individual genes with paused Pol II

In light of the discrepancy between my placement of the +1 nucleosome and that of Mavrich et al., I inspected results for individual genes known to have paused Pol II (Figure 4-7). These genes were selected based on the result of previous permanganate footprinting analyses that had detected paused Pol II on these genes (Lee et al., 2008). These genes are also included in category II of my clustering analysis and in the collection of “stalled” genes determined by Zeitlinger et al. based on a calculated “stalling index” (Zeitlinger et al., 2007). A high level of NELF is detected at the 5’ end of these genes (Figure 4-7). However, the +1 nucleosome mapped by Mavrich et al. is not

<table>
<thead>
<tr>
<th>My result</th>
<th>Mavrich’s result</th>
</tr>
</thead>
<tbody>
<tr>
<td>+1 H2A.Z nucleosomes</td>
<td>+1 Bulk nucleosomes</td>
</tr>
<tr>
<td>Category I</td>
<td>+130</td>
</tr>
<tr>
<td>Category II</td>
<td>+125</td>
</tr>
<tr>
<td>Category III</td>
<td>+130</td>
</tr>
<tr>
<td>Genes with Pol II</td>
<td>+140</td>
</tr>
<tr>
<td>Genes without Pol II</td>
<td>+130</td>
</tr>
</tbody>
</table>

Table 4-2: Comparison of the positions of the +1 nucleosomes determined by my analysis and Mavrich’s analysis. The position of the +1 nucleosome in each panel in Figure 4-6 was determined by visually locating the maximum data point in the smoothed trace from +50 to +200. The positions of the +1 nucleosome for “paused” and “not paused” genes were determined by Mavrich et al. (Mavrich et al., 2008b).
located at +145 as shown in their genome wide nucleosome mapping (Mavrich et al., 2008b). Instead, the location of the +1 nucleosome ranges from +180 to +600.

Figure 4-7: The +1 nucleosome is not located at +145 for 6 individual genes that have promoter proximal pausing. The location of bulk nucleosomes and the occupancies of Pol II and NELF-B are displayed in UCSC browser for 6 genes that have promoter proximal
pausing. The gene name is listed on top of each panel. The height of vertical bars corresponds to the intensity of Pol II or NELF-B ChIP signals. The red rectangular boxes overlaid on Pol II ChIP-chip data indicate the estimated size of the permanganate footprints detected by Lee et al. (Lee et al., 2008). Each nucleosome is indicated by a black box covering a region of 146 base pairs. The black arrow indicates the transcription direction of the gene. The position of +1 nucleosome, measured by the distance from the center of +1 nucleosome to the transcription start site, is indicated for each gene. Pol II ChIP-chip data was generated in Drosophila S2R+ cells treated with LacZ RNAi. NELF-B ChIP-chip data was generated by Lee et al. (Lee et al., 2008). The genome wide bulk nucleosome mapping data was generated by Mavrich et al. (Mavrich et al., 2008b).

4.3 Discussion

4.3.1 Pol II is concentrated at the promoter of thousands of genes

Clustering analysis of Pol II ChIP-chip data revealed three different Pol II distribution patterns at the 5’ end of Drosophila genes. The resolution of Pol II mapping described here is higher than what have been reported (Muse et al., 2007; Zeitlinger et al., 2007), but all of these genome wide studies identified thousands of genes that have Pol II concentrated at the 5’ end. Over 60 individual cases in Drosophila have been analyzed by permanganate genomic footprinting, and all show the presence of transcriptionally engaged Pol II in the region from +20 to +50 when peaks of both NELF and Pol II are detected at the 5’ end (Lee et al., 2008). Because both NELF and Pol II are most concentrated at the 5’ end of my category II genes, I conclude that these are the genes with paused Pol II, and that there are at least 3000 of these genes. In support of this, 80% of the genes on which Lee et al. has detected paused Pol II by permanganate footprinting (Lee et al., 2008) are present in category II of my clustering analysis.
Zeitlinger et al. identified genes with stalled Pol II by calculating the ratio of the maximum Pol II occupancy within 300 bp around a transcription start site over the median Pol II occupancy in the body of the gene (from +600 to the 3’ end) (Zeitlinger et al., 2007). This ratio was called a “stalling index”. Genes with unusually high “stalling index” (outside a fitted normal distribution curve plotted using the “stalling index” values for all genes) were judged to have “stalled” Pol II in the promoter proximal region. The advantage of this approach is that it established a mathematical measurement of the strength of “stalling”. Based on these criteria, Zeitlinger et al. identified approximately 1000 genes that had stalled Pol II. 61% of these genes fell within my category II genes thus providing further evidence that my category II genes represent genes with paused Pol II.

My category II genes number 2 to 3 times more than the number of paused or stalled genes identified by Zeitlinger et al. and Lee et al. This difference can be ascribed to the methods by which the genes were categorized. As mentioned above, Zeitlinger et al. used strict mathematical criteria. Inspection of their data, however, reveals that the mean value for the ratio of Pol II at the promoter over Pol II in the body of the gene was 2. This indicates that there are several thousand genes associated with Pol II that have Pol II concentrated in the promoter region. Lee et al. based their identification of genes with paused Pol II on a ranking of genes based on the levels of NELF detected in the 5’ region and a sampling of genes in the ranking with permanganate footprinting. My identification of category II genes was based on the cluster algorithm that looks for similar patterns in the distribution of Pol II seen in the promoter region. Genes with Pol II concentrated at
the 5’ end but with an intermediate “stalling index” in Zeitlinger’s analysis are also included in category II.

4.3.2 Promoter proximal pausing is a transient intermediate associated with active transcription

It is interesting that NELF is also enriched at the promoters of class III genes (Figure 4-3), suggesting that highly transcribed genes could also have paused Pol II. How can we correlate the pausing of Pol II with the high expression levels of category III genes? One explanation is that NELF may transiently associate with Pol II on a category III gene and then dissociate from the elongation complex to allow productive transcription. Consistent with this possibility, there is evidence showing that NELF-dependent pausing may enhance the expression of some genes by maintaining the accessibility of the promoter region to the transcriptional machinery (Gilchrist et al., 2008). Pausing, as well as NELF binding, is not necessarily associated with gene repression (Aiyar et al., 2007; Core et al., 2008; Lee et al., 2008).

4.3.3 GAGA factor might travel with elongating Pol II

GAGA factor was found to localize upstream from the transcription start site of category II genes. This is in good agreement with previous work showing that GAGA factor is involved in setting up paused Pol II (Lee et al., 1992; Wang et al., 2005). The distribution of GAGA factor on category III genes raises the intriguing possibility that GAGA factor travels with elongating Pol II molecules. The composite distribution of
GAGA factor extends downstream from the transcription start site in accordance with Pol II but in contrast with NELF (Figure 4-4A and Figure 4-3).

Previous studies have yielded conflicting results concerning the relationship between GAGA factor and elongation. O’Brien et al. observed that UV light crosslinked GAGA factor to the body of the Drosophila hsp70 and hsp26 genes in a transcription-dependent manner (O’Brien et al., 1995). They were also able to demonstrate that the recruitment of GAGA factor to the hsp70 gene followed the kinetics of heat shock induction. However, the tracking of GAGA factor with Pol II on the hsp70 gene was not seen in a later ChIP experiment employing formaldehyde crosslinking (Wu et al., 2005), nor has this possibility appeared in any recent models of heat shock induction from the Lis lab where the UV crosslinking experiments were done. Using a GAGA factor-DAM methylase fusing protein for mapping, van Steensel et al. provided evidence that GAGA factor is enriched on introns but not exons (van Steensel et al., 2003).

My data now provides the first support for formaldehyde-dependent ChIP that GAGA factor does indeed associate with the body of genes. The composite distribution of GAGA factor in the body of category III genes does not correlate with the frequency distribution of the GAGA element, indicating the association may not be directed by sequence specific binding. A caveat to this could be the sequence I used to search for possible binding sites. This was done by examining the frequency of GAGAGAG DNA motifs from -1 kb to 1 kb of the transcription start site. The GAGAGAG motif was chosen because the double sites of GAGAG (the optimal binding site of GAGA factor) tend to co-occur (Mavrich et al., 2008b; van Steensel et al., 2003). However, a number of other motifs containing (GA)$_n$ repeats are able to bind GAGA factor and the binding
affinity is dependent on genome context (van Steensel et al., 2003). These GAGA element variants are preferentially localized within introns but not exons. Therefore I cannot rule out the possibility that the occupancy of GAGA factor detected in the body of a gene is due to the recognition of GAGA elements other than the GAGAGAG motif.

My results raise the possibility that GAGA factor associates with elongation complexes. Depending on the context, such an interaction might promote or inhibit elongation. To further investigate whether GAGA factor associates with elongation complexes, a gel shift assay using reconstituted elongation complexes could be done. I developed a facile way of generating tailed templates that can be used to generate elongation complexes with purified Drosophila Pol II (Appendix I).

4.3.4 Pol II distribution on category I genes

The high density of Pol II upstream from the transcription start site of category I genes is unexpected. The intermediate occupancy of NELF and GAGA factor in the same regions suggests that these regions may encompass some un- or mis- annotated promoters. Arguing against this possibility is my finding that a nucleosome free region followed by a +1 nucleosome is evident in the composite nucleosome distribution generated for these genes.

Recently published work suggests an interesting explanation for the presence of Pol II upstream from the promoters of category I genes. High throughput sequencing techniques have identified antisense transcripts initiating in the region upstream from many promoters in human and mouse cells (Core et al., 2008; Seila et al., 2008).
However, there are three differences between the distribution pattern of Pol II engaged in anti-sense transcription in mammals and the distribution pattern of Pol II observed at category I genes upstream from the transcription start site. First, the peak in the Pol II distribution associated with the anti-sense orientated transcripts in mammals is 250 bp upstream from the transcription start site, but the peak in the Pol II distribution for my category I genes is more than 500 bp upstream from the transcription start site. Second, the Pol II engaged in anti-sense transcription in mammals tends to pause and therefore generate short transcripts (Core et al., 2008; Seila et al., 2008), but the Pol II distribution for category I genes spans a broad region (Figure 4-2C). Third, the anti-sense transcripts in mammals correlate with high levels of sense transcription, but the Pol II density at or downstream from the transcription start site of category I genes is very low.

Clearly additional experiments are needed to understand the basis for Pol II located upstream from transcription start site. Nuclear run-on assays which provided the basis for Lis and colleagues’ (Core et al., 2008) genome wide analysis could be employed to verify that Pol II is present upstream from genes and to establish whether these Pol II molecules transcribe RNA in the sense or antisense direction. If antisense transcripts are observed, *Drosophila* could serve as an ideal system for investigating the function of these antisense transcripts in regulating gene expression during development.

4.3.5 Chromatin structure and promoter proximal pausing

Early studies of the human hsp70 gene indicated that promoter proximal pausing on this gene depends on chromatin structure (Brown et al., 1996). The importance of
chromatin structure in pausing was further supported by the finding that HSF, an activator of the hsp70 gene, associated with the chromatin remodeling factor, Swi/Snf. Early results for the *Drosophila* hsp70 gene suggested that chromatin structure was not involved in pausing Pol II. Paused Pol II could be reconstituted in *Drosophila* nuclear extracts in the absence of chromatin (Benjamin and Gilmour, 1998). Moreover, induction of the hsp70 gene in *Drosophila* occurred independently of Swi/Snf (Armstrong *et al.*, 2002).

Recent results for mapping of nucleosomes across the entire *Drosophila* genome have reignited our interest in the possibility that nucleosomes play a role in promoter proximal pausing. Mavrich *et al.* observed that the +1 nucleosome associated with paused genes was centered 10 nucleotides downstream from that of genes without paused Pol II (Mavrich *et al.*, 2008b). A connection between pausing and chromatin has also been made by studying a gene called TepII whose transcription is repressed upon RNAi-mediated depletion of NELF from cells (Gilchrist *et al.*, 2008). Depletion of NELF causes a decrease in transcription that is counter to expectations given that NELF is an inhibitor of transcription elongation. Further analysis showed that the presence of the paused Pol II prevented a nucleosome from encroaching on the promoter region.

My analysis of category II genes revealed a nucleosome distribution which was similar but not identical to the distribution described by Mavrich *et al.* (Mavrich *et al.*, 2008b). However, both the exact location and the 10 base pair difference between the distributions on paused and not paused genes that were noted by Mavrich *et al.* were not evident in my analysis. This prompted me to examine individual genes that had been shown previously to be associated with paused Pol II. Inspection of these genes indicated
that while Pol II is restricted to a narrow region spanning from +20 to +50 (Lee et al., 2008), the location of the +1 nucleosome was much more variable. In agreement with these results, recent mapping of nucleosomes on the hsp70 gene reveals that the +1 nucleosome is centered at +330 (Petesch and Lis, 2008). These observations argue against a simple model in which the collision of Pol II with a nucleosome causes promoter proximal pausing.
Bibliography


Blencowe, B.J. (2002). Transcription: surprising role for an elusive small nuclear RNA. Curr Biol 12, R147-149.


MacDonald, C.C., and Redondo, J.L. (2002). Reexamining the polyadenylation signal: were we wrong about AAUAAA? Mol Cell Endocrinol 190, 1-8.


Whitelaw, E., and Proudfoot, N. (1986). Alpha-thalassaemia caused by a poly(A) site mutation reveals that transcriptional termination is linked to 3' end processing in the human alpha 2 globin gene. EMBO J 5, 2915-2922.


Appendix A

List of Primers Used For Real Time PCR

All the primers with a name starting with “RealT” were used to do ChIP analysis. Some of them were also used to measure the RNA level at different regions of heat shock genes (Figure 3-2).

All the primers with a name starting with “ArrayTest” were used to verify the changes of Pol II ChIP signals upon Pcf11 depletion detected by the ChIP-chip experiments (Figure G-3).

All the primers with a name containing “RTRT” were used to measure the mRNA levels of the genes that are possibly controlled by Pcf11-mediated premature termination (Figure 3-9). They were also used to verify the increases of Pol II ChIP signals on corresponding genes upon Pcf11 depletion detected by the ChIP-chip (Figure K-2).

Tm is provided by the manufacturer of the primers (IDT).
<table>
<thead>
<tr>
<th>Gene</th>
<th>Primer</th>
<th>Sequence</th>
<th>Tm</th>
</tr>
</thead>
<tbody>
<tr>
<td>β-tubulin</td>
<td>β-tubulin+5F</td>
<td>5'-GGCTTCCAAAGCGAATGCCTA</td>
<td>57.6</td>
</tr>
<tr>
<td>β-tubulin</td>
<td>β-tubulin+61R</td>
<td>5'-AACAAACTGTCAGGTCCTACA</td>
<td>57.4</td>
</tr>
<tr>
<td>hsp23</td>
<td>RealT Hsp2+100R</td>
<td>5'-TGCTCTCAAAGCAGATCTCTCCTC</td>
<td>54.4</td>
</tr>
<tr>
<td>hsp23</td>
<td>RealT Hsp2+1143F</td>
<td>5'-ATGTAAGACTGCATGAACTGCATGTA</td>
<td>55.0</td>
</tr>
<tr>
<td>hsp23</td>
<td>RealT Hsp2+1239R</td>
<td>5'-GAGATGACAGAAAATTCATAAAAAGAAG</td>
<td>53.2</td>
</tr>
<tr>
<td>hsp23</td>
<td>RealT Hsp2+1494F</td>
<td>5'-GCAACAGACAGTGGCAAATGATTG</td>
<td>55.2</td>
</tr>
<tr>
<td>hsp23</td>
<td>RealT Hsp2+1569R</td>
<td>5'-AGTGTCCTGGCAGCTCTTT</td>
<td>58.0</td>
</tr>
<tr>
<td>hsp23</td>
<td>RealT Hsp2+1637F</td>
<td>5'-CAAATGCCTGGTCAATGAAATAC</td>
<td>53.5</td>
</tr>
<tr>
<td>hsp23</td>
<td>RealT Hsp2+1723R</td>
<td>5'-TAAAGCATTTTCTCTCTCTGCTTT</td>
<td>54.9</td>
</tr>
<tr>
<td>hsp23</td>
<td>RealT Hsp2+17F</td>
<td>5'-AAAGCCAAGGCAGAAACAGCTA</td>
<td>55.2</td>
</tr>
<tr>
<td>hsp23</td>
<td>RealT Hsp2+1882F</td>
<td>5'-GCTGTCAGGAGAGAAGCAGAG</td>
<td>57.8</td>
</tr>
<tr>
<td>hsp23</td>
<td>RealT Hsp2+1949R</td>
<td>5'-CCTTTGCATAACGGCCCGTAA</td>
<td>55.0</td>
</tr>
<tr>
<td>hsp23</td>
<td>RealT Hsp2+697F</td>
<td>5'-CCTTGCAATATTTAAACATCGATT</td>
<td>53.1</td>
</tr>
<tr>
<td>hsp23</td>
<td>RealT Hsp2+789R</td>
<td>5'-CCTTAAATGCATATTGTTAGCTGAGA</td>
<td>55.0</td>
</tr>
<tr>
<td>hsp23</td>
<td>RealT Hsp2+900F</td>
<td>5'-TTTCTCTTGTGTTATGGGTTACTATT</td>
<td>54.7</td>
</tr>
<tr>
<td>hsp23</td>
<td>RealT Hsp2+982R</td>
<td>5'-AACCAGGTATCAGAAGAAAG</td>
<td>55.3</td>
</tr>
<tr>
<td>hsp23</td>
<td>RealT Hsp2+1122R</td>
<td>5'-TTTTCTTCATCTGCAACAGTCTC</td>
<td>54.9</td>
</tr>
<tr>
<td>hsp23</td>
<td>RealT Hsp2+20F</td>
<td>5'-TGAAAGCGAAACAGTGGTGAAGCA</td>
<td>55.6</td>
</tr>
<tr>
<td>hsp23</td>
<td>RealT Hsp2+2659F</td>
<td>5'-GACTTTAAATTTCCGTATTTCCGATG</td>
<td>55.4</td>
</tr>
<tr>
<td>hsp23</td>
<td>RealT Hsp2+2737R</td>
<td>5'-TCATGTGGTTAATCAGGAAATCTCTCATAA</td>
<td>54.4</td>
</tr>
<tr>
<td>hsp23</td>
<td>RealT Hsp2+2934F</td>
<td>5'-GGTTGATGTGGCTTTGGACTTG</td>
<td>58.9</td>
</tr>
<tr>
<td>hsp23</td>
<td>RealT Hsp2+3021R</td>
<td>5'-AGCAACCGTATCAGTACAC</td>
<td>57.5</td>
</tr>
<tr>
<td>hsp23</td>
<td>RealT Hsp2+3302F</td>
<td>5'-CGCTGGTCTGAAGTTTTGG</td>
<td>57.2</td>
</tr>
<tr>
<td>U1_95Ca</td>
<td>RealT U1_95Ca+270F</td>
<td>5'-TGTTGTTGTCTGGCTGATGGAT</td>
<td>55.5</td>
</tr>
<tr>
<td>U1_95Ca</td>
<td>RealT U1_95Ca+355R</td>
<td>5'-ATAATATATCTATCTCTATATATTGATG</td>
<td>55.5</td>
</tr>
<tr>
<td>U1_95Ca</td>
<td>RealT U1_95Ca+638R</td>
<td>5'-CAATCGTATGAGATCCGCAAA</td>
<td>53.1</td>
</tr>
<tr>
<td>U1_95Ca</td>
<td>RealT U1_95Ca+70R</td>
<td>5'-GCCATGAGCTTCTCCTCTGATA</td>
<td>55.2</td>
</tr>
<tr>
<td>U1_95Ca</td>
<td>RealT U1_95Ca-18F</td>
<td>5'-TCCCGAGCAGTGGGAAACAGC</td>
<td>56.3</td>
</tr>
<tr>
<td>U4_39B</td>
<td>RealT U4_39B+233F</td>
<td>5'-TCTTTTTACTGTTATTGAGGATAGTT</td>
<td>55.2</td>
</tr>
<tr>
<td>U4_39B</td>
<td>RealT U4_39B+335R</td>
<td>5'-TTTTGTAAGGGCGTGTCATT</td>
<td>55.6</td>
</tr>
<tr>
<td>U4_39B</td>
<td>RealT U4_39B+38R</td>
<td>5'-GGACCTTATGGTACGGTACTGG</td>
<td>56.1</td>
</tr>
<tr>
<td>U4_39B</td>
<td>RealT U4_39B+527F</td>
<td>5'-GACAGTTCCTGCTTGGGTGGCT</td>
<td>59.6</td>
</tr>
<tr>
<td>U4_39B</td>
<td>RealT U4_39B+611R</td>
<td>5'-TGGCGAATATCTTCTCCACCTT</td>
<td>57.8</td>
</tr>
<tr>
<td>U4_39B</td>
<td>RealT U4_39B-38F</td>
<td>5'-GCCCGATGGTGATATATGTAAGAA</td>
<td>55.1</td>
</tr>
<tr>
<td>Pcf11</td>
<td>ArrayTest_1+418F</td>
<td>5'-GGCAACACTGTCCTGCTTGGCT</td>
<td>60.3</td>
</tr>
<tr>
<td>CG9148-RA</td>
<td>ArrayTest_2+200F</td>
<td>5'-ACAATCGAGACCGGAGAAT</td>
<td>57.3</td>
</tr>
<tr>
<td>CG9148-RA</td>
<td>ArrayTest_2+368R</td>
<td>5'-TAGGTCTCGAGCAGTGTTGG</td>
<td>58.4</td>
</tr>
<tr>
<td>GenBank Accession</td>
<td>Array Test Type</td>
<td>Primer 1</td>
<td>Primer 2</td>
</tr>
<tr>
<td>-------------------</td>
<td>------------------</td>
<td>--------------</td>
<td>--------------</td>
</tr>
<tr>
<td>CG4394-RB</td>
<td>ArrayTest_3+199R</td>
<td>5'-TTGCCCACTCGGTAATGAG 55.1</td>
<td></td>
</tr>
<tr>
<td>CG4394-RB</td>
<td>ArrayTest_3+72F</td>
<td>5'-AGCATGGATCGCTGGATAAG 56.8</td>
<td></td>
</tr>
<tr>
<td>CG30038-RA</td>
<td>ArrayTest_4+165R</td>
<td>5'-TGGCTTTATGGAGGCCCTTCT 56.1</td>
<td></td>
</tr>
<tr>
<td>CG30038-RA</td>
<td>ArrayTest_4+38F</td>
<td>5'-CGATGATGACTGGGCTTTGA 54.8</td>
<td></td>
</tr>
<tr>
<td>CG8036-RB</td>
<td>ArrayTest_5+202R</td>
<td>5'-GTCGCGATGAACTGGGATAACC 55.6</td>
<td></td>
</tr>
<tr>
<td>CG8036-RB</td>
<td>ArrayTest_5+76F</td>
<td>5'-TGGAACCGACCGAATTGGA 56.0</td>
<td></td>
</tr>
<tr>
<td>CG14780-RA</td>
<td>ArrayTest_6+22F</td>
<td>5'-ATGCCAGACCAATCGAGAG 56.9</td>
<td></td>
</tr>
<tr>
<td>CG14780-RA</td>
<td>ArrayTest_6+88R</td>
<td>5'-AGCTGGCCAGGGACGACTAC 59.9</td>
<td></td>
</tr>
<tr>
<td>CG30456-RA</td>
<td>ArrayTest_7+131R</td>
<td>5'-TGACTCCGTTGTGGATCTGCT 56.7</td>
<td></td>
</tr>
<tr>
<td>CG30456-RA</td>
<td>ArrayTest_7+14F</td>
<td>5'-CCAGCTCTCTGCTACCCGATGAA 58.5</td>
<td></td>
</tr>
<tr>
<td>CG1903-RB</td>
<td>ArrayTest_8+114F</td>
<td>5'-AATGATGACCGGGGATTTG 55.3</td>
<td></td>
</tr>
<tr>
<td>CG1903-RB</td>
<td>ArrayTest_8+182R</td>
<td>5'-AGTCGGGCTGGATTCCTTATC 56.4</td>
<td></td>
</tr>
<tr>
<td>CG17018-RD</td>
<td>ArrayTest_9+169F</td>
<td>5'-GGGAGGCTGAATCCGAGTCT 57.6</td>
<td></td>
</tr>
<tr>
<td>CG17018-RD</td>
<td>ArrayTest_9+230R</td>
<td>5'-TCAAATGCTATCCGATCCTAA 57.6</td>
<td></td>
</tr>
<tr>
<td>CG5254-RA</td>
<td>ArrayTest_10+355F</td>
<td>5'-TGGAGGTCTCGATCAGTCGAA 56.9</td>
<td></td>
</tr>
<tr>
<td>CG5254-RA</td>
<td>ArrayTest_10+420R</td>
<td>5'-CGGGAGGTTGGCCTGATCT 59.5</td>
<td></td>
</tr>
<tr>
<td>CG1702-RA</td>
<td>ArrayTest_11+185F</td>
<td>5'-GGCCCAATTTATTTTGGTAAT 52.5</td>
<td></td>
</tr>
<tr>
<td>CG1702-RA</td>
<td>ArrayTest_11+253R</td>
<td>5'-CTTGAGGGGAATCTGAAGAGA 56.8</td>
<td></td>
</tr>
<tr>
<td>CG7144-RA</td>
<td>ArrayTest_12+306F</td>
<td>5'-CATCTTTGTCGCCGAGCT 56.6</td>
<td></td>
</tr>
<tr>
<td>CG7144-RA</td>
<td>ArrayTest_12+370R</td>
<td>5'-CGGAGGGTTCCTCGATATCTCA 56.2</td>
<td></td>
</tr>
<tr>
<td>CG17209-RA</td>
<td>ArrayTest_13+161F</td>
<td>5'-AACGCTTTTGGAGACCTGTGACT 57.0</td>
<td></td>
</tr>
<tr>
<td>CG17209-RA</td>
<td>ArrayTest_13+226R</td>
<td>5'-ACTCACCGCAGATCCTCAATTTGA 54.6</td>
<td></td>
</tr>
<tr>
<td>CG7737-RA</td>
<td>ArrayTest_14+408F</td>
<td>5'-TGTGATGTTTTCGGGTTCG 54.0</td>
<td></td>
</tr>
<tr>
<td>CG7737-RA</td>
<td>ArrayTest_14+482R</td>
<td>5'-GCCAACTGTGGGAATACTC 56.9</td>
<td></td>
</tr>
<tr>
<td>CG3399-RA</td>
<td>ArrayTest_15+211F</td>
<td>5'-TGCCCCCTCTTGTCTATTTTTGA 55.0</td>
<td></td>
</tr>
<tr>
<td>CG3399-RA</td>
<td>ArrayTest_15+284R</td>
<td>5'-CGGTCCATGTTGGAAATCGGA 53.2</td>
<td></td>
</tr>
<tr>
<td>CG10123-RA</td>
<td>ArrayTest_16+293F</td>
<td>5'-AATGAGACACGGCAGCTGCAAT 57.2</td>
<td></td>
</tr>
<tr>
<td>CG10123-RA</td>
<td>ArrayTest_16+355R</td>
<td>5'-GTCGACTTTGGAGCCTTTCCCATCA 57.2</td>
<td></td>
</tr>
<tr>
<td>CG17904-RA</td>
<td>CG17904-RA RTRT_F</td>
<td>5'-TGTAAGAAAGCAAAATATCCCATTG 52.6</td>
<td></td>
</tr>
<tr>
<td>CG17904-RA</td>
<td>CG17904-RA RTRT_R</td>
<td>5'-TGCCCGCAGAGGAACCT 58.1</td>
<td></td>
</tr>
<tr>
<td>CG17035-RA</td>
<td>CG17035-RA RTRT_F</td>
<td>5'-GCCAGGAGTCTGATGAGA 55.1</td>
<td></td>
</tr>
<tr>
<td>CG17035-RA</td>
<td>CG17035-RA RTRT_R</td>
<td>5'-AATGCTCCGTTGGAATAGCATCT 56.9</td>
<td></td>
</tr>
<tr>
<td>CG5134-RC</td>
<td>CG5134-RC RTRT_F</td>
<td>5'-GCATGCCGGCCGGAGCTAC 58.2</td>
<td></td>
</tr>
<tr>
<td>CG5134-RC</td>
<td>CG5134-RC RTRT_R</td>
<td>5'-GCCGGTACATCGGAAAAGCA 57.2</td>
<td></td>
</tr>
<tr>
<td>CG7959-RA</td>
<td>CG7959-RA RTRT_F</td>
<td>5'-TGACACGAGGACGCACTC 57.5</td>
<td></td>
</tr>
<tr>
<td>CG7959-RA</td>
<td>CG7959-RA RTRT_R</td>
<td>5'-CAGCGACGCTGGGTGTGTAG 58.6</td>
<td></td>
</tr>
<tr>
<td>CG3781-RA</td>
<td>CG3781-RA RTRT_F</td>
<td>5'-AGTTTTCGCCACCAAGCTC 57.9</td>
<td></td>
</tr>
<tr>
<td>CG3781-RA</td>
<td>CG3781-RA RTRT_R</td>
<td>5'-TCCGCTCCTCGGCTCAGACA 58.9</td>
<td></td>
</tr>
<tr>
<td>CG4274-RA</td>
<td>CG4274-RA RTRT_F</td>
<td>5'-AGAGGAGGTTTGGGCGCTAA 56.9</td>
<td></td>
</tr>
<tr>
<td>CG4274-RA</td>
<td>CG4274-RA RTRT_R</td>
<td>5'-TGGCTCTTGAGCTTGAAT 57.1</td>
<td></td>
</tr>
<tr>
<td>CG2948-RA</td>
<td>CG2948-RA RTRT_F</td>
<td>5'-CCTGCCCTGTCAGCCCAAT 55.6</td>
<td></td>
</tr>
<tr>
<td>CG2948-RA</td>
<td>CG2948-RA RTRT_R</td>
<td>5'-GCCCAAAGGGAAGGACTAA 56.0</td>
<td></td>
</tr>
<tr>
<td>CG8617-RA</td>
<td>CG8617-RA RTRT_F</td>
<td>5'-GCACGGATTTGACGCACTAG 54.4</td>
<td></td>
</tr>
<tr>
<td>CG8617-RA</td>
<td>CG8617-RA RTRT_R</td>
<td>5'-AGCTCATGCTGATCGAA 55.3</td>
<td></td>
</tr>
<tr>
<td>CG13075-RA</td>
<td>CG13075-RA RTRT_F</td>
<td>5'-CAGCTACCTGTACTGCCAGAATACT 57.3</td>
<td></td>
</tr>
<tr>
<td>CG13075-RA</td>
<td>CG13075-RA RTRT_R</td>
<td>5'-GCCGCCGCAAGTCTATAGA 57.4</td>
<td></td>
</tr>
</tbody>
</table>
Appendix B

List of Primers Used For LM-PCR

LM1, LM2, and LM3 indicate the order in which each primer is used of a 3 primer set. Tm is provided by the manufacturer of the primers (IDT).

<table>
<thead>
<tr>
<th>Gene</th>
<th>Primer</th>
<th>Sequence</th>
<th>Tm</th>
<th>Annealing Temp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>hsp23</td>
<td>Hsp23+1124LM1</td>
<td>5'-GAAGTTCTTACATTACATGCTT</td>
<td>50.6</td>
<td>51.0</td>
</tr>
<tr>
<td>hsp23</td>
<td>Hsp23+1124LM2</td>
<td>5'-CATTCTAGCTTCAGTTTACC</td>
<td>53.7</td>
<td>55.0</td>
</tr>
<tr>
<td>hsp23</td>
<td>Hsp23+1124LM3</td>
<td>5'-GCTTCAAGTCTTACATTACATGCTT</td>
<td>56.9</td>
<td>58.0</td>
</tr>
<tr>
<td>hsp23</td>
<td>Hsp23+1150LM1</td>
<td>5'-CTTACATAGTTAATTACCTCGT</td>
<td>47.4</td>
<td>47.0</td>
</tr>
<tr>
<td>hsp23</td>
<td>Hsp23+1150LM2</td>
<td>5'-CAGATTAAACCTTGACTGAAGTT</td>
<td>51.7</td>
<td>51.0</td>
</tr>
<tr>
<td>hsp23</td>
<td>Hsp23+1150LM3</td>
<td>5'-CTTGAAGTTAAGTCTTCATATCCA</td>
<td>54.9</td>
<td>55.5</td>
</tr>
<tr>
<td>hsp23</td>
<td>Hsp23+1348LM1</td>
<td>5'-ATCATATGTAATGTGATGTC</td>
<td>50.4</td>
<td>46.2</td>
</tr>
<tr>
<td>hsp23</td>
<td>Hsp23+1348LM2</td>
<td>5'-GATATGTAATGTGATGTC</td>
<td>53.4</td>
<td>49.5</td>
</tr>
<tr>
<td>hsp23</td>
<td>Hsp23+1348LM3</td>
<td>5'-GCTTCAATACCTTGACTGAAGTT</td>
<td>53.3</td>
<td>55.6</td>
</tr>
<tr>
<td>hsp23</td>
<td>Hsp23+1629LM1</td>
<td>5'-GTCATCTGCTTCGTTCTCGT</td>
<td>55.9</td>
<td>54.0</td>
</tr>
<tr>
<td>hsp23</td>
<td>Hsp23+1629LM2</td>
<td>5'-CAGGTTCGTCTTCATGTTTGC</td>
<td>56.7</td>
<td>58.0</td>
</tr>
<tr>
<td>hsp23</td>
<td>Hsp23+1629LM3</td>
<td>5'-GCTTCAATACCTTGACTGAAGTT</td>
<td>59.0</td>
<td>62.0</td>
</tr>
<tr>
<td>hsp23</td>
<td>Hsp23+1863LM1</td>
<td>5'-GACATTTGCTTTCTGGAGAGTTTC</td>
<td>55.9</td>
<td>58.7</td>
</tr>
<tr>
<td>hsp23</td>
<td>Hsp23+1863LM2</td>
<td>5'-GAGTTTCTGTCCCGTCTTCG</td>
<td>55.6</td>
<td>54.0</td>
</tr>
<tr>
<td>hsp23</td>
<td>Hsp23+1863LM3</td>
<td>5'-AAGTTGCTTGATGTTCTTTCG</td>
<td>56.9</td>
<td>58.0</td>
</tr>
<tr>
<td>hsp23</td>
<td>Hsp23+140LM1</td>
<td>5'-GAGATTTTGTTGCCCCCTAGCA</td>
<td>60.0</td>
<td>62.0</td>
</tr>
<tr>
<td>hsp23</td>
<td>Hsp23+140LM2</td>
<td>5'-AGTTGCTGAGTGTTCTTTCG</td>
<td>59.1</td>
<td>54.0</td>
</tr>
<tr>
<td>hsp23</td>
<td>Hsp23+140LM3</td>
<td>5'-GCTTCAATACCTTGACTGAAGTT</td>
<td>53.8</td>
<td>55.0</td>
</tr>
<tr>
<td>hsp23</td>
<td>Hsp23+140LM1</td>
<td>5'-GACATTTGCTTTCTGGAGAGTTTC</td>
<td>62.9</td>
<td>58.0</td>
</tr>
<tr>
<td>hsp23</td>
<td>Hsp23+140LM2</td>
<td>5'-GACGCTGTAAGCAAGCTGCTAGC</td>
<td>59.1</td>
<td>54.0</td>
</tr>
<tr>
<td>hsp23</td>
<td>Hsp23+140LM3</td>
<td>5'-GACATTTGCTTTCTGGAGAGTTTC</td>
<td>62.9</td>
<td>58.0</td>
</tr>
<tr>
<td>hsp23</td>
<td>Hsp23+140LM1</td>
<td>5'-GCTTCAATACCTTGACTGAAGTT</td>
<td>57.2</td>
<td>55.0</td>
</tr>
<tr>
<td>hsp23</td>
<td>Hsp23+140LM2</td>
<td>5'-GCTTCAATACCTTGACTGAAGTT</td>
<td>57.2</td>
<td>55.0</td>
</tr>
<tr>
<td>hsp23</td>
<td>Hsp23+140LM3</td>
<td>5'-GCTTCAATACCTTGACTGAAGTT</td>
<td>57.2</td>
<td>55.0</td>
</tr>
<tr>
<td>hsp70Ab</td>
<td>Hsp70Ab+3352LM3</td>
<td>5'-AGGACACACTCCCGCGACA</td>
<td>61.3</td>
<td>62.0</td>
</tr>
<tr>
<td>hsp70Ab</td>
<td>Hsp70Ab+3354LM1</td>
<td>5'-CAGCTGGGGAGTGGTTT</td>
<td>55.7</td>
<td>54.0</td>
</tr>
<tr>
<td>hsp70Ab</td>
<td>Hsp70Ab+3354LM2</td>
<td>5'-GCAGGCGTTGACACACCTTG</td>
<td>54.7</td>
<td>58.0</td>
</tr>
<tr>
<td>hsp70Ab</td>
<td>Hsp70Ab+3354LM3</td>
<td>5'-TTGAACACCTCCCGCGACTCAT</td>
<td>57.4</td>
<td>62.0</td>
</tr>
<tr>
<td>hsp70B</td>
<td>87c+2741 LM1</td>
<td>5'-ATAGCCCTTTTGCTTTTC</td>
<td>48.6</td>
<td>48.0</td>
</tr>
<tr>
<td>hsp70B</td>
<td>87c+2741 LM2</td>
<td>5'-ATAGCCCTTTTGCTTTTC</td>
<td>52.2</td>
<td>55.0</td>
</tr>
<tr>
<td>hsp70B</td>
<td>87c+2741 LM3</td>
<td>5'-ATAGCCCTTTTGCTTTTC</td>
<td>43.3</td>
<td>42.0</td>
</tr>
<tr>
<td>hsp70B</td>
<td>87c+2741 LM4</td>
<td>5'-ATAGCCCTTTTGCTTTTC</td>
<td>47.7</td>
<td>48.0</td>
</tr>
<tr>
<td>hsp70B</td>
<td>87c+2741 LM5</td>
<td>5'-ATAGCCCTTTTGCTTTTC</td>
<td>51.8</td>
<td>54.0</td>
</tr>
<tr>
<td>hsp70B</td>
<td>87c+2741 LM6</td>
<td>5'-ATAGCCCTTTTGCTTTTC</td>
<td>55.5</td>
<td>54.0</td>
</tr>
<tr>
<td>hsp70B</td>
<td>87c+2741 LM7</td>
<td>5'-ATAGCCCTTTTGCTTTTC</td>
<td>59.1</td>
<td>58.0</td>
</tr>
<tr>
<td>hsp70B</td>
<td>87c+2741 LM8</td>
<td>5'-ATAGCCCTTTTGCTTTTC</td>
<td>62.9</td>
<td>62.0</td>
</tr>
<tr>
<td>hsp70B</td>
<td>87c+2741 LM9</td>
<td>5'-ATAGCCCTTTTGCTTTTC</td>
<td>58.2</td>
<td>54.0</td>
</tr>
<tr>
<td>hsp70B</td>
<td>87c+2741 LM10</td>
<td>5'-ATAGCCCTTTTGCTTTTC</td>
<td>61.8</td>
<td>58.0</td>
</tr>
<tr>
<td>hsp70B</td>
<td>87c+2741 LM11</td>
<td>5'-ATAGCCCTTTTGCTTTTC</td>
<td>65.8</td>
<td>62.0</td>
</tr>
<tr>
<td>U1 95Ca</td>
<td>U1 95Ca+394LM1</td>
<td>5'-TGGGAGCTCAGCACGAATGG</td>
<td>54.6</td>
<td>54.0</td>
</tr>
<tr>
<td>U1 95Ca</td>
<td>U1 95Ca+394LM2</td>
<td>5'-ATACACACACGACAGTCCCAA</td>
<td>56.2</td>
<td>58.0</td>
</tr>
<tr>
<td>U1 95Ca</td>
<td>U1 95Ca+394LM3</td>
<td>5'-GGTGGCCACAAAGACACATTCAGATGGA</td>
<td>57.8</td>
<td>62.0</td>
</tr>
<tr>
<td>U1 95Ca</td>
<td>U1 95Ca+653LM1</td>
<td>5'-GGTGGCCACAAAGACACATTCAGATGGA</td>
<td>54.1</td>
<td>54.0</td>
</tr>
<tr>
<td>U1 95Ca</td>
<td>U1 95Ca+653LM2</td>
<td>5'-GGTGGCCACAAAGACACATTCAGATGGA</td>
<td>55.3</td>
<td>58.0</td>
</tr>
<tr>
<td>U1 95Ca</td>
<td>U1 95Ca+653LM3</td>
<td>5'-GGTGGCCACAAAGACACATTCAGATGGA</td>
<td>56.4</td>
<td>62.0</td>
</tr>
<tr>
<td>U4 39B</td>
<td>U4 39B+393LM1</td>
<td>5'-GCTGCTGTTTAGTTAGTTAGT</td>
<td>52.9</td>
<td>51.0</td>
</tr>
<tr>
<td>U4 39B</td>
<td>U4 39B+393LM2</td>
<td>5'-GCTGCTGTTTAGTTAGTTAGT</td>
<td>51.5</td>
<td>55.0</td>
</tr>
<tr>
<td>U4 39B</td>
<td>U4 39B+393LM3</td>
<td>5'-GCTGCTGTTTAGTTAGTTAGT</td>
<td>54.3</td>
<td>58.0</td>
</tr>
<tr>
<td>U4 39B</td>
<td>U4 39B+644LM1</td>
<td>5'-GCTGCTGTTTAGTTAGTTAGT</td>
<td>52.7</td>
<td>54.0</td>
</tr>
<tr>
<td>U4 39B</td>
<td>U4 39B+644LM2</td>
<td>5'-GCTGCTGTTTAGTTAGTTAGT</td>
<td>55.8</td>
<td>58.0</td>
</tr>
<tr>
<td>U4 39B</td>
<td>U4 39B+644LM3</td>
<td>5'-GCTGCTGTTTAGTTAGTTAGT</td>
<td>59.4</td>
<td>62.0</td>
</tr>
</tbody>
</table>
## Appendix C

### List of Other Primers

#### C.1 List of primers used for making the DNA templates of RNAi

<table>
<thead>
<tr>
<th>Gene</th>
<th>Primer</th>
<th>Sequence</th>
<th>Tm</th>
<th>Annealing Temp.*</th>
</tr>
</thead>
<tbody>
<tr>
<td>LacZ</td>
<td>LacZ RNAi F</td>
<td>5’-GAATTAATACGACTCACTATAGGAGA TGAAGCTGGCTACAGGA</td>
<td>64.2</td>
<td>50,65</td>
</tr>
<tr>
<td>LacZ</td>
<td>LacZ RNAi R</td>
<td>5’-GAATTAATACGACTCACTATAGGAGA GCAGGCTTCTGTGCCTTAAT</td>
<td>64.6</td>
<td>50,65</td>
</tr>
<tr>
<td>Pcf11 (RNAi 1)</td>
<td>Pcf11 RNAi F +184F</td>
<td>5’-GAATTAATACGACTCACTATAGGAGA GTGGCGCAGCGAATTAA</td>
<td>64.7</td>
<td>55,65</td>
</tr>
<tr>
<td>Pcf11 (RNAi 1)</td>
<td>Pcf11 RNAi +692R</td>
<td>5’-GAATTAATACGACTCACTATAGGAGA GGTATGGGGTGGTTACCC</td>
<td>64.9</td>
<td>55,65</td>
</tr>
<tr>
<td>Pcf11 (RNAi 2)</td>
<td>Pcf11 T7R corrected</td>
<td>5’-GAATTAATACGACTCACTATAGGAGA TCAACCACATACGGTCCC</td>
<td>64.3</td>
<td>54,65</td>
</tr>
<tr>
<td>Pcf11 (RNAi 2)</td>
<td>Pcf11 T7 S1</td>
<td>5’-GAATTAATACGACTCACTATAGGAGA GCAGTGGCGTAAAGATCAC</td>
<td>65.0</td>
<td>54,65</td>
</tr>
<tr>
<td>Pcf11 (RNAi 3)</td>
<td>Pcf11 RNAi DRSC F</td>
<td>5’-GAATTAATACGACTCACTATAGGAGA GGAAGGCCTCAAAACAAAC</td>
<td>64.3</td>
<td>53,65</td>
</tr>
<tr>
<td>Pcf11 (RNAi 3)</td>
<td>Pcf11 RNAi DRSC R</td>
<td>5’-GAATTAATACGACTCACTATAGGAGA AAACGGGTAGCGGTTCAG</td>
<td>64.6</td>
<td>53,65</td>
</tr>
<tr>
<td>Spt5</td>
<td>Spt5 RNAi-A for</td>
<td>5’-GAATTAATACGACTCACTATAGGAGA CGTTCGGCGAGATCAA</td>
<td>65.2</td>
<td>53,65</td>
</tr>
<tr>
<td>Spt5</td>
<td>Spt5 RNAi-A rev</td>
<td>5’-GAATTAATACGACTCACTATAGGAGA CCGGCAGCGGTTAATCT</td>
<td>65.2</td>
<td>53,65</td>
</tr>
</tbody>
</table>

*Annealing Temperature contain two numbers: the first one is the annealing temperature used in the first 2 cycles of PCR amplification; the second one is the annealing temperature used in the following PCR cycles.
C.2 List of primers used for making the tailed DNA templates (see Appendix D)

<table>
<thead>
<tr>
<th>Gene***</th>
<th>Primer</th>
<th>Sequence</th>
<th>Tm</th>
<th>Annealing Temp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>U1_95Ca (+293)</td>
<td>Probe_U1_95Ca*</td>
<td>5'-AAATCCATCAGCCAGACAACAAC</td>
<td>55.9</td>
<td>N/A</td>
</tr>
<tr>
<td>U1_95Ca (+194)</td>
<td>U1_95Ca_BbvCI_F</td>
<td>5'-TAAATTCTTGCTGAGGAAAATGCAAAATGTATGAAACACG</td>
<td>61.1</td>
<td>55</td>
</tr>
<tr>
<td>U1_95Ca (+194)</td>
<td>U1_95Ca_BbvCI_F2</td>
<td>5'-GAATCTGCTGATGGAGAAAATGCAAAATGTATGAAACACG</td>
<td>64.4</td>
<td>55</td>
</tr>
<tr>
<td>U1_95Ca (+194)</td>
<td>U1_95Ca_BsmI_F</td>
<td>5'-TAAATTTTTTTGCTGACATTCGGAAAATGCAAAATGTATGAAACACG</td>
<td>59.7</td>
<td>55</td>
</tr>
<tr>
<td>U1_95Ca (+194)</td>
<td>U1_95Ca_BsmI_F2</td>
<td>5'-GAATCTGCTGATGGAGAAAATGCAAAATGTATGAAACACG</td>
<td>63.9</td>
<td>55</td>
</tr>
<tr>
<td>U1_95Ca (+194)</td>
<td>U1_95Ca_BsmI_F3</td>
<td>5'-GAATCTGCTGATGGAGAAAATGCAAAATGTATGAAACACG</td>
<td>68.0</td>
<td>55</td>
</tr>
<tr>
<td>U1_95Ca (+401)</td>
<td>U1_95Ca_R</td>
<td>5'-AAAAACAGTTGGGCATCACAAC</td>
<td>54.5</td>
<td>55</td>
</tr>
<tr>
<td>U1_95Ca (+154)</td>
<td>U1_95Ca_BsmI_F3+154</td>
<td>5'-GGCCGCGCCTGATCCGATTCCGCCGTCTGCTTATTTTATCGTT</td>
<td>73.4</td>
<td>55</td>
</tr>
<tr>
<td>U1_95Ca (+401)</td>
<td>U1_95Ca_Biotin_R+401**</td>
<td>/5Bio/AAAAACAGTTGGGCATCACAAC</td>
<td>54.5</td>
<td>55</td>
</tr>
<tr>
<td>U4_39B (+275)</td>
<td>U4_39B_BbvCI_F</td>
<td>5'-TAAATTCTTGCTGAGGAAAATGCAAAATGTATGAAACACG</td>
<td>61.1</td>
<td>55</td>
</tr>
<tr>
<td>U4_39B (+275)</td>
<td>U4_39B_BsmI_F</td>
<td>5'-GAATCTGCTGATGGAGAAAATGCAAAATGTATGAAACACG</td>
<td>59.5</td>
<td>55</td>
</tr>
<tr>
<td>U4_39B (+476)</td>
<td>U4_39B_R</td>
<td>5'-GAATCTGCTGATGGAGAAAATGCAAAATGTATGAAACACG</td>
<td>53.0</td>
<td>55</td>
</tr>
<tr>
<td>U4_39B (+158)</td>
<td>U4_39B_BsmI_F3+158</td>
<td>5'-GAATCTGCTGATGGAGAAAATGCAAAATGTATGAAACACG</td>
<td>68.7</td>
<td>55</td>
</tr>
<tr>
<td>U4_39B (+394)</td>
<td>U4_39B_Biotin_R+394**</td>
<td>/5Bio/TGCTGCTAACTGTTACTATTAGTTGGTG</td>
<td>55.1</td>
<td>55</td>
</tr>
</tbody>
</table>

*This probe was used to bind transcript made in vitro and to test the activity of RNase H (Figure D-2C).

**These primers are biotinylated at their 5’-ends.

***The numbers following the gene names indicate the positions where the primers hybridize.
Appendix D

RNAi-Mediated Depletion of Pcf11

D.1 RNAi mediated depletion of Pcf11 in *Drosophila* S2R+ and Kc167 cells

Three different double strand RNAs (RNAi) were used to deplete Pcf11. As illustrated in Figure **D-1A**, Pcf11 RNAi 1 targets the N-terminus of both the 60 kDa and 200 kDa Pcf11; Pcf11 RNAi 2 targets both the C-terminus of the 60 kDa Pcf11 and the middle part of the 200 kDa Pcf11; Pcf11 RNAi 3 targets only the 200 kDa Pcf11.

---

Figure **D-1**: Experimental design of RNAi-mediated depletion of Pcf11 in tissue culture cells. A) Schematic of different Pcf11 mRNAs and RNAi constructs. Two distinct RNAs are transcribed from the Pcf11 gene as a result of use of alternative splice sites and alternative polyadenylation sites. The solid line indicates identical sequences and dotted line indicates different sequences. Numbers represent distances from the 5’ end of mature mRNA. Numbers with asterisk (*) represent distances only for 60 kDa Pcf11 mRNA. B) Comparison of different Pcf11 proteins. Blue denotes different peptide sequences. CID (CTD interaction domain) is within 17-173 aa.
*Drosophila* S2R+ and Kc167 cells were selected for RNAi treatment. Both cell lines have been shown by the *Drosophila* RNAi Screening Consortium (DRSC) to be good candidates for RNAi experiments. Western blot was used to detect specific proteins. Crude antiserum generated in rabbit against the N-terminus (1-283 aa) of Pcf11 was affinity purified by using the same antigen peptide expressed in *E. coli*. Both Pcf11 isoforms (60 kDa and 200 kDa) were detected by affinity purified antibody (Figure D-2). All three RNAi constructs were able to decrease the level of the 200 kDa Pcf11 in S2R+ cells and Kc167 cells. Surprisingly, the 60 kDa protein was unperturbed by the RNAi treatment (Figure D-2). This conflicts with previous observations (Zhang and Gilmour, 2006).

**Figure D-2**: RNAi-mediated depletion of Pcf11 (I). *Drosophila* S2R+ and Kc167 cells were treated with various amounts of Pcf11 RNAi 1, 2 and 3 as indicated above. Cells were also treated with Spt5 RNAi as a positive control. Western blot was done with affinity purified Pcf11 antiserum (rabbit) to monitor the expression level of Pcf11. Molecular weight standards of 62, 83 and 175 kDa are indicated with lines. Asterisk (*) indicates a possible degradation product of the 200 kDa Pcf11. Dot (·) indicates a non-specific band.
Figure D-3 shows another RNAi experiment where Pcf11 RNAi 1 was used. Spt5 RNAi was also used as a positive control and was able to decrease the level of Spt5 in both cell lines. LacZ RNAi, as a negative control, had no effect on the expression of either Pcf11 or Spt5. S2R+ cells were more susceptible to RNAi treatment since the depletion efficiency was higher in S2R+ cells than in Kc167 cells for both Spt5 and 200 kDa Pcf11 (80% vs. 50%).

Zhiqiang Zhang showed a clear depletion of a 60 kDa protein previously using Pcf11 RNAi 2 (Zhang and Gilmour, 2006). However, he used Drosophila S2 cells, whereas I used S2R+ and Kc167 cells. S2R+ cells are derived from S2 cells but selected to attach to flasks better. It is unclear why 60 kDa Pcf11 was not depleted with any of the
double strand RNA I have used. I tried to do Pcf11 RNAi experiments using S2 cells but still failed to deplete the 60 kDa Pcf11. I have also tried to treat the cells for different lengths of time, used different amounts of RNAi and different combinations of RNAi, but the 60 kDa Pcf11 was not depleted. Even though I affinity purified the antibody, there is a chance that the 60 kDa protein detected in the western blot is due to non-specific binding. Nevertheless, successful depletion of 200 kDa Pcf11 still allowed me to study its function as a termination factor on individual genes and globally.

D.2 Pcf11 depletion did not result in a decrease in Pcf11 ChIP signals

Antibody against Pcf11 was used in ChIP experiments to determine whether Pcf11 occupancy decreased on the hsp23 gene after RNAi treatment (I did not analyze the hsp70 gene). The result shows that Pcf11 was enriched at the 3’ end of the hsp23 gene before depletion of Pcf11 (Figure D-4), which agrees with its termination function. Unexpectedly, no decreases in the Pcf11 ChIP signals were detected after RNAi treatment (Figure D-4). The Pcf11/Pol II ChIP signal ratios were unchanged across the hsp23 gene, suggesting that equal amounts of Pcf11 protein were associated with Pol II. In contrast, Zhang and Gilmour reported observing a decrease in Pcf11 ChIP signals on the hsp70 gene upon RNAi treatment (Zhang and Gilmour, 2006). I tried to resolve this issue by using affinity purified antibody to do the ChIP but the same results were obtained. One explanation for why the Pcf11 ChIP signal did not decrease upon RNAi treatment is that the depletion efficiency of Pcf11 was not high enough to significantly diminish the level of Pcf11 associated with elongation complexes. Pcf11 may have strong
affinity for Pol II, so even small amounts of Pcf111 might still be able to bind Pol II. In addition, there are multiple binding sites for Pcf11 on the CTD, so the level of depletion might be insufficient to remove all the Pcf11 from each Pol II molecule.

The 60 kDa Pcf11 adds another level of complexity to the results of the RNAi experiment. Because the 60 kDa Pcf11 was not depleted by RNAi, decrease in the 200 kDa Pcf11 ChIP signals might be masked by the ChIP signals for the 60 kDa Pcf11. If the 60 kDa Pcf11 can substitute for the 200 kDa Pcf11 in binding to Pol II, but is somehow less efficient at causing termination, I would not be able to detect decreased Pcf11 ChIP signals while still being able to observe a termination defect upon Pcf11 RNAi treatment.
D.3 Termination defects of heat shock genes upon depletion of Pcf11 by Pcf11 RNAi 1 or Pcf11 RNAi 2

Figure D-5 is another display of Figure 3-1, showing termination defects of heat shock genes caused by depletion of Pcf11. Depletion of Pcf11 in Figure 3-1 was done by...
using Pcf11 RNAi 1. In Figure D-5, results of Pol II ChIP experiments done on Drosophila S2R+ cells treated with either Pcf11 RNAi 1 or Pcf11 RNAi 2 are presented. Error bars are added to the graph, replacing calculated $P$-value.
Figure D-5: Depletion of Pcf11 disrupts transcription termination for three heat shock genes. Chromatin immunoprecipitation was used to monitor Pol II occupancy downstream of the polyadenylation sites of the A) hsp70Aa, B) hsp23 and C) hsp70Ab genes before and after Pcf11 depletion by Pcf11 RNAi 1 or Pcf11 RNAi 2. Polyadenylation sites (vertical arrows) and PCR amplified regions (horizontal lines) are indicated in a schematic above each bar graph. Pol II signals in downstream regions are normalized against the signal in the most upstream region. Signals of pre-immune serum immunoprecipitated samples are normalized against signals from corresponding samples immunoprecipitated with antibody against the Rpb3 subunit of Pol II. Each bar represents the mean of three independent experiments and the error bar represents the standard deviation. D) Absolute Pol II ChIP signals from the promoter regions of the hsp23 gene and of the β1-tubulin gene did not change upon Pcf11 depletion.
Appendix E

Termination Analysis of the U4_39B Gene

In chapter 3, I provided evidence that Pcf11 depletion by RNAi treatment did not cause termination defects at the 3’ end of the U1_95Ca gene. To determine whether transcription of other snRNA genes also terminates in a Pcf11-independent manner, I examined the effect of Pcf11 depletion on the changes of Pol II density at the 3’ end of the U4_39B gene. Figure E-1 shows that depletion of Pcf11 had no effect on transcription termination of the U4_39B gene when measured by ChIP.

ChIP results also showed that the Pol II level at the 5’ end of the U4_39B gene was unchanged after Pcf11 depletion (Figure E-1A). However, this data conflicted with my permanganate footprinting results. Quantification of Pol II footprints revealed that the Pol II occupancy on the U4_39B gene was much lower after Pcf11 depletion (Figure E-1C). A biological replicate experiment confirmed this result (Figure F-2B and Figure F-2D). This discrepancy possibly came from the inability of ChIP to distinguish transcriptionally engaged Pol II from transcriptionally disengaged Pol II. A similar discrepancy between ChIP and nuclear run-on assays has been previously reported (Morillon et al., 2003). If judged from the permanganate footprinting result, transcription initiation of the U4_39B gene could be impaired by Pcf11-induced defects in recycling of Pol II or transcription interference from the upstream gene CG8678. Nevertheless, termination of the U4_39B gene occurred in the same region in Pcf11-depleted cells as it
did in control cells, suggesting that Pcf11 is not involved in termination of this snRNA gene.

Figure E-I: Pcf11 depletion does not disrupt termination of the U4_39B gene. A) Chromatin immunoprecipitation was used to monitor Pol II occupancy on the U4_39B gene.
gene. PCR amplified regions (horizontal lines) in ChIP analysis and the annotated 3’ end of the U4_39B gene (vertical arrow) are indicated in a schematic above the bar graph. Pol II signals in downstream regions are normalized against the signal in the most upstream region. Signals of pre-immune serum immunoprecipitated samples are normalized against signals from corresponding samples immunoprecipitated with antibody against the Rpb3 subunit of Pol II. Each bar represents the mean of three independent experiments and the error bar represents the standard deviation. B) Absolute Pol II ChIP signals from -38--+38 region of the U4_39B gene and the promoter region of β1-tubulin did not change upon Pcf11 depletion. C) Drosophila S2R+ cells were treated with 10 mM potassium permanganate for 1 minute on ice. The intensity of the permanganate reactivity of each thymine was represented as a ratio over the reactivity of the same thymine in naked DNA control. The red line represents the smoothed Pol II distribution in LacZ RNAi treated samples. The blue line represents the smoothed Pol II distribution in Pcf11 RNAi treated samples. The orange dots represent all T residues. The annotated 3’ end of the U4_39B gene is indicated by the vertical dotted line. D) The intensity of the permanganate reactivity at the T residues in B was log2 transformed and directly displayed using Treeview without smoothing. The top to bottom direction corresponds to the 5’ to 3’ direction of transcription. Black indicates the same intensity as the naked DNA lane. Red and green reflect stronger or weaker signals compared to the naked DNA lane, respectively.
Appendix F

Biological Replicate of Quantified Permanganate Footprinting Experiments

A biological replicate experiment was performed for each permanganate footprinting experiment described in my thesis. Here I present the quantified result of these biological replicate experiments. The Pol II distribution was analyzed at a high resolution both before and after Pcf11 depletion. Quantification and plotting were done as described in Chapter 2. Figure F-1 includes data of the hsp23, hsp70Aa and hsp70Ab genes. Figure F-2 includes data of the U1_95Ca and U4_39B genes.
Figure F-1: Results of a biological replicate of permanganate footprinting the hsp23, hsp70Aa and hsp70Ab genes. The intensity of the permanganate reactivity of each thymine in heat shocked cells was represented as a ratio over the reactivity of the same thymine in non-heat shocked cells for: A) hsp23, B) hsp70Aa and C) hsp70Ab genes. The red line represents the smoothed Pol II distribution in LacZ RNAi treated samples. The blue line represents the smoothed Pol II distribution in Pcf11 RNAi treated samples. The orange dots represent all T residues. The polyadenylation sites are indicated by the...
vertical dotted lines. Pause sites are labeled with “P”. P** represents a new pause site on the hsp23 gene formed after Pcf11 depletion. D-F) The intensity of the permanganate reactivity at the T residues in panel A-C was log2 transformed and directly displayed using Treeview without smoothing. The top to bottom direction of each panel (D, hsp23; E, hsp70Aa; F, hsp70Ab) corresponds to the 5’ to 3’ direction of transcription. Black indicates the same intensity as the NHS lane. Red and green reflect stronger or weaker signals compared to the NHS lane, respectively.
Figure F-2: Results of a biological replicate of permanganate footprinting the U1_95Ca and U4_39B genes. The intensity of the permanganate reactivity of each thymine was represented as a ratio over the reactivity of the same thymine in naked DNA control for: A) U1_95Ca and B) U4_39B genes. The red line represents the smoothed Pol II
distribution in LacZ RNAi treated samples. The blue line represents the smoothed Pol II distribution in Pcf11 RNAi treated samples. The orange dots represent all T residues. The annotated 3’ ends of these two genes are indicated by the vertical dotted lines. C and D) The intensity of the permanganate reactivity at the T residues in panel A and B was log2 transformed and directly displayed using Treeview without smoothing. The top to bottom direction of each panel (C, U1_95Ca; D, U4_39B) corresponds to the 5’ to 3’ direction of transcription. Black indicates the same intensity as the NHS lane. Red and green reflect stronger or weaker signals compared to the NHS lane, respectively.

To show the effect of smoothing by KaleidaGraph software, I superimposed the raw data from the quantification of permanganate footprinting experiments on smoothed curves processed by KaleidaGraph software. Figure F-3 includes data for permanganate footprinting experiments performed on LacZ RNAi treated Drosophila S2R+ cells described in Figure 3-4, Figure 3-10 and Figure E-1.
Figure F-3: Comparison of raw data from quantification of permanganate footprinting experiments to smoothed curves. Raw data are presented in pink. Red curves represent smoothed data. Data included in this figure corresponds to permanganate footprinting experiments performed on LacZ RNAi treated Drosophila S2R+ cells described in Figure 3-4, Figure 3-10 and Figure E-1.
Appendix G

Verification of Changes in Pol II Density Detected by ChIP-chip Analysis

To verify the increase and decrease in Pol II occupancy after Pcf11 depletion detected by the genome wide ChIP-chip analysis, I randomly picked 7 regions where the Pol II level was found to increase and 9 regions where the Pol II level was found to decrease and tested increases or decreases of Pol II ChIP signals in these regions by real time PCR. Six out of the 7 increases and 7 out of the 9 decreases of Pol II ChIP signals were confirmed by real time PCR. Here I present the screen shots of the Pol II ChIP-chip data (displayed in UCSC browser) for each region analyzed (Figure G-1 and Figure G-2) together with the real time PCR results (Figure G-3). The primers used for real time PCR are listed in Appendix B.
Figure G-1: Display (UCSC browser) of Pol II ChIP-chip data for 7 regions where Pol II occupancy increases upon Pcf11 depletion. The height of vertical bars corresponds to the intensity of Pol II ChIP signals. The red rectangular box covers the region that was amplified by real time PCR for verifying the ChIP-chip results (Figure G-3). The black arrow indicates the transcription direction of the gene where the amplified region resides.
Figure G-2: Display (UCSC browser) of Pol II ChIP-chip data for 9 regions where Pol II occupancy decreases upon Pcf11 depletion. The height of vertical bars corresponds to the intensity of Pol II ChIP signals. The red rectangular box covers the region that was amplified by real time PCR for verifying the ChIP-chip results (Figure G-3). The black arrow indicates the transcription direction of the gene where the amplified region resides.

Figure G-3: Real time PCR verifies the changes in Pol II occupancy detected by the genome wide analysis upon Pcf11 depletion. A) Pol II density in regions covered by red rectangular boxes in Figure G-1 was examined by real time PCR for both control and Pcf11-depleted samples. Each bar represents the mean of three independent experiments and the error bar represents the standard deviation. Increase of Pol II density on the Pcf11 gene was not verified. Increases of Pol II density on the other 6 genes were verified. B) Pol II density in regions covered by red boxes in Figure G-2 was examined by real time PCR for both control and Pcf11-depleted samples. Each bar represents the mean of three independent experiments and the error bar represents the standard deviation. Decreases of Pol II density on the CG1903-RB and CG17018-RD genes were not verified. Decreases of Pol II density on the other 7 genes were verified.
Appendix H

Definitions of Statistical Terms

H.1 Box plot definition

Each box encloses 50% of the data with the median value of the variable displayed as a line. The top and bottom of the box mark the limits of ± 25% of the variable population. The lines extending from the top and bottom of each box mark the minimum and maximum values within the data set that fall within an acceptable range. Any value outside of this range, called an outlier, is displayed as an individual point.

- Median - The data value located halfway between the smallest and largest values.
- Upper Quartile (UQ) - The data value located halfway between the median and the largest data value.
- Lower Quartile (LQ) - The data value located halfway between the median and the smallest data value.
- Interquartile Distance (IQD) - The distance between the Upper and Lower Quartiles (UQ - LQ).
♦ Outliers - Points whose value is either:

greater than UQ + 1.5 * IQD or less than LQ - 1.5 * IQD

For references, please refer to the manual of the KaleidaGraph software.

H.2 Wilcoxon-Mann-Whitney test

Wilcoxon-Mann-Whitney test is used to see if two different samples have been
drawn from the same population. The two samples can have different numbers of data
points (in my studies, the samples refer to the collection of expression values for
corresponding genes in defined classes), since no relationship is assumed between the
samples. This test assumes the samples are not normally distributed with the same
variances. As part of the results, KaleidaGraph calculates a P value. This value
determines if there is a statistically significant difference between the medians of the two
groups. If the value is below a certain level (usually 0.05), the conclusion is that there is a
difference between the two.

For references, please refer to: 1) Siegel, S. and Castellan, N. J. Jr. Nonparametric
Appendix I

*In Vitro* Transcription Assay

I.1 Introduction

I present a novel way for generating tailed templates that can be used to initiate Pol II transcription at a specific position at the end of the fragment. These templates are easier to generate than the method developed by Zhang *et al.* (Zhang *et al.*, 2004) and could prove useful in future elongation and termination assays. I demonstrate the efficacy of these templates by transcribing two snRNA genes, the U1_95Ca and U4_39B genes. Figure 1-1 illustrates the experiment design.
I.2 Materials and Methods

Genomic DNA extracted from *Drosophila* S2R+ cells was used as template for amplifying a DNA fragment to be transcribed *in vitro*. Primers used for PCR reactions are listed in Appendix C. The forward primers contain the recognition site of a nicking enzyme (either Nb.BsmI or Nb.Bbvcl) for the purpose of nicking the DNA and forming a tailed DNA template. The reverse primers are either labeled with biotin or not labeled.
DNA template generated by using biotin labeled reverse primer can be immobilized through binding to streptavidin beads.

Amplified PCR product was extracted once with phenol/chloroform/isoamyl alcohol (25:24:1) and once with chloroform. DNA was precipitated with ethanol at 4°C. Five micrograms of precipitated DNA, in a total volume of 50 µl, were digested by 10 units of Nb.BsmI (New England Biolabs) at 65°C for 3 hr or by 10 units of Nb.BbvcI (New England Biolabs) at 37°C for 3 hr. After digestion, the enzyme was inactivated by incubation at 80°C for 20 min. The reaction was immediately placed on ice and loaded onto a 1% agarose gel. The digested DNA fragment was purified using Qiagen Gel Purification Kit and dissolved in TE, pH 8.0.

To immobilize the DNA template, the tailed DNA template was bound to streptavidin-coated magnetic beads (Dynabeads M-280 Streptavidin, Novagen). A magnetic Particle Concentrator (Dynal MPC) was used to pull down the beads during washes. Streptavidin-coated magnetic beads were first equilibrated and washed 3 times with B & W buffer (10 mM Tris-HCl, pH 7.5, 1 mM EDTA and 2.0 M NaCl). One hundred nanograms of tailed DNA template were bound to 20 µl of streptavidin-coated magnetic beads following the manufacturer’s instructions. Before being used in transcription reactions, beads bound with tailed DNA template were washed once with 20 µl transcription buffer (200 mM KCl, 50 mM HEPES, pH 7.6, 1.0 mM MnCl₂, 12% glycerol and 0.5 mM dithiothreitol).

In vitro transcription was performed by preparing a 15 µl pre-incubation mixture containing 200 mM KCl, 50 mM HEPES, pH 7.6, 1.0 mM MnCl₂, 12% glycerol, 0.5 mM DTT, 20 U RNasin, 0.25 µl of purified Pol II and 100 ng of tailed DNA template. This
pre-incubation mixture was incubated at room temperature for 5 min to allow Pol II to bind the end of the DNA. A 5 µl cocktail of nucleotides in water was then added to the pre-incubation mixture to yield final nucleotide concentrations of 500 µM ATP, 500 µM GTP, 500 µM CTP, 25 µM UTP, and 1 µCi α-32P UTP/reaction. After the addition of the nucleotide cocktail, the reaction was incubated at room temperature for 25 min. For transcription reactions using G less cassette template (prepared by Zhiqiang Zhang, Gilmour lab) in the absence of GTP (Figure I-2), an extra component, 0.5 mM UpG, was included in the 15 µl pre-incubation mixture (Zhang et al., 2005), and GTP was left out of the 5 µl nucleotide cocktail. When α-32P CTP replaced α-32P UTP in equal amount in the transcription reactions (first two lanes in Figure I-2), the final concentration of CTP and UTP were adjusted to 25 µM and 500 µM respectively.

To isolate total RNA produced by the in vitro transcription, 40 µl of stop solution (20 mM EDTA, pH 8.0, 0.2 M NaCl, 1% SDS, 0.25 mg/ml yeast tRNA and 0.1 mg/ml protease K) was added directly to the reaction mixture. The sample was then incubated at 42°C for 30 min and extracted once with 55 µl of phenol /chloroform/isoamyl alcohol (25:24:1). RNA product was precipitated with ethanol at 4°C and analyzed on 6% denaturing gels (7 M urea, 6% acrylamide/bis-acrylamide [19:1], 1x TBE). Radioactive RNA was detected using PhosphorImager Screen (Molecular Dynamics) and Typhoon 8600.

To distinguish RNA tethered to the elongation complex and RNA released into the solution (Figure I-3), the DNA-bound streptavidin-coated magnetic beads were separated from the supernatant using Dynal MPC before the addition of stop buffer. Forty
microliters of stop buffer was added to the supernatant. The immobilized beads were washed twice with transcription buffer (200 mM KCl, 50 mM HEPES, pH 7.6, 1.0 mM MnCl₂, 12% glycerol and 0.5 mM dithiothreitol) and resuspended in 20 µl of transcription buffer. Forty microliters of stop buffer were then added to the resuspended beads. Samples (supernatant or beads) were incubated at 42°C for 30 min and extracted once with 55 µl of phenol/chloroform/isoamyl alcohol (25:24:1). RNA product was precipitated and detected as previously described.

I.3 Results

Figure I-2 shows the result of optimizing the in vitro transcription reaction without the DNA templates being immobilized. Different tails were tested on a piece of DNA matching +194 to +401 of the U1_95Ca gene. The BsmI_3 tail yielded the most full-length product and generated the least background before the DNA template was nicked (Figure I-2A). The variation in the transcription efficiency using different tailed templates was determined by the efficiency of enzymatic nicking and efficiency of transcription initiation. I speculate that the background in the uncut lanes was caused by the “breathing” of the DNA ends, which generates transient entrance for Pol II.

In some reconstituted in vitro transcription systems, nascent transcripts displace the non-template strand and form a long duplex with the template stand of the DNA (Dedrick and Chamberlin, 1985). To test whether this also occurred in my transcription reaction, RNase digestion assays were carried out. It was observed that the nascent transcripts were sensitive to RNase A but not RNase H (Figure I-2B). When a probe that
matched the sequence in the middle of the transcript was added to the transcription reaction, the full length transcript could then be digested into two halves by RNase H (Figure I-2C). This test indicates that the nascent transcript generated by my transcription reaction does not anneal with the template DNA strand. Hence this *in vitro* transcription system is a good model for studying elongation and termination.
I then tested my *in vitro* transcription system with the DNA sequences downstream from the annotated 3' ends of the U1_95Ca and U4_39B genes. Figure 1-3 shows that full length transcripts could be obtained with both DNA templates. A series of bands corresponding to potential pause sites were also observed. Such pause sites tend to
occur at A or T residues in a AT-rich context, which agrees with previous studies showing that A:T rich sequences tend to pause Pol II (Kireeva et al., 2000; Komissarova et al., 2002). However, no transcripts were found to be released from the elongation complex. Hence, the sequences analyzed here do not cause spontaneous release of nascent transcripts.

Figure I-3: Test of spontaneous release of transcripts from the DNA template by sequences downstream of the U1_95Ca and U4_39B genes. DNA sequences from +154 to +401 of the U1_95Ca gene and from +158 to +394 of the U4_39B gene were amplified by PCR. They partially overlap with the regions where termination of these two genes appears to occur (transcription of the U1_95Ca and U4_39B genes terminates at +164-+300 and +143-+280 respectively). The BsmI_3 tail was used to initiate transcription and the biotinylated DNA template was immobilized on a magnetic bead before adding Pol II and nucleotides. Transcripts of different lengths were found in bound fractions but no transcripts were found in supernatant. Numbers along the gel indicate the positions of potential pause sites relative to the transcription start site of each gene. Nucleotides corresponding to the potential pause sites are labeled in red in the sequences provided on the side.
Appendix J

Permanganate Footprinting of the hsp70B and hsp23 genes

The data presented here is a set of pilot permanganate footprinting experiments examining the Pol II distribution at the 3’ regions of the hsp70B genes (hsp70Bbb, hsp70Bb and hsp70Bc) and of the hsp23 gene.

J.1 Permanganate footprinting of the hsp70B genes

The hsp70Bbb, hsp70Bb and hsp70Bc genes are transcribed in the same direction and the sequences downstream from their polyadenylation sites are identical. Permanganate footprinting indicates that these genes are fully induced within 5 minutes of heat shock, since permanganate footprints in the promoter region and in the region downstream from the polyadenylation signal are identical at 5 minutes and 15 minutes after heat shock (Figure J-1, lane 1-6; polyadenylation signal is located at +2326–+2331). In fact, after 5 minutes of heat shock, Pol II travelled all the way to +2695 (Figure J-1, lane 12). DRB, an inhibitor of Pol II elongation (Wada et al., 1998; Yamaguchi et al., 1999), eliminated most of the permanganate hyper-reactive sites at the 3’ end of the hsp70B genes (Figure J-1, lane 13), indicating that the reactivity was due to elongating Pol II. I had hoped to be able to map Pol II farther downstream but the pattern of bands stopped beyond +2695. Different primers were tested but failed. The problematic region is AT-rich, so the primers matching these DNA sequences have relatively low melting
temperatures. Unfortunately, the DNA sequences at regions further downstream match other parts of the *Drosophila* genome, limiting the use of those sequences for primer design.
Figure J-1: Permanganate footprinting of the hsp70B genes. G/A ladders (lanes 7 and 14) generated by partial depurination with formic acid served as markers for identifying nucleotide positions. Purified DNA (lanes 8, 9, 10, 15, 16 and 11) was treated for the indicated times with permanganate on ice. Non-heat shocked (NHS) and heat shocked (HS) cells (lanes 1, 2, 3, 4, 5, 6, 11, 12, 13, 18, 19 and 20) were treated with 10 mM potassium permanganate for 1 minute on ice. Isolated DNA was cleaved at oxidized thymines with piperidine and the resulting breaks in the non-template strand (the three gels on the left) or in template strand (the gel on the right) were mapped by LM-PCR. Treatment of cells with 200 µM DRB prior to heat shock greatly reduced the permanganate reactivity detected in the 3’ region (lane 13). The number above each gel indicates the primers used in LM-PCR. Numbers along the side of each gel indicate the distance of hyper-reactive thymines from the transcription start site. The polyadenylation signal is at +2326-+2331. The faded regions at the bottom of the third and fourth gels represent failed LM-PCR regions.

J.2 Permanganate footprinting of the hsp23 gene

The hsp23 gene, unlike any of the hsp70 genes, does not seem to have paused Pol II in its promoter proximal region. There are no differences between the permanganate reactivity detected on the template strand of the hsp23 gene under non-heat shock conditions and that detected on the naked DNA control (Figure J-2, compare lanes 1 and 2). I tried to look for paused Pol II by detecting strong permanganate reactivity on the non-template strand of the hsp23 gene but none was detected (data not shown). ChIP-chip analyses detected the presence of GAGA factor but not NELF at the 5’ end of the hsp23 gene under non-heat shock conditions (Lee et al., 2008). It is likely that NELF is required in addition to GAGA factor for Pol II to pause in the promoter proximal region of this gene.

After 5 minutes of heat shock, transcription of the hsp23 gene was induced, as indicated by the appearance of bands from +6 to +96 (lane 3). The sensitivity of
transcription to DRB indicates that the hyper-reactivity was due to elongating Pol II (lane 4). However, in a region 300 bp downstream from the polyadenylation site (+884), the 5’ HS lane (lane 10) looked almost identical to the NHS lane (lane 9) indicating that very little Pol II was present, while the 15’ HS lane (lane 12) showed very strong permanganate reactivity indicating that a high level of Pol II was present.

I did not do further analysis to find out what caused the difference between the hsp23 and hsp70 genes. Apparently, Pol II transcribes more than 2 kb on the hsp70B genes within 5 minutes of heat shock (Figure J-1, lane 12) but less than 1 kb on the hsp23 genes after the gene was induced for the same length of time (Figure J-2, lane 10). One possibility is that the hsp23 gene has two termination sites: one for short induction, the other for prolonged induction. A sequence search identified three AATAAA hexamers at the 3’ end of the hsp23 gene, located at +857-+862, +1418-+1423 and +1724-+1729 respectively. However, Pol II paused at around +1000 after 15 minutes of heat shock and transcription termination occurred after that (Figure 3-4C). Hence, it is unlikely that the hexamers at +1418-+1423 and +1724-+1729 are functional. Alternatively, the hsp23 gene could have a slower induction rate than the hsp70 gene that might have been overlooked by waiting until 5 minutes after heat shock to do the footprinting. For example, if the hsp23 gene took several minutes to become induced by heat shock, Pol II would not be expected to appear in a region more than 1 kb downstream from the transcription start site after 5 minutes of heat shock, since the rate of Pol II transcription is roughly 1.2 kb per minute (Boehm et al., 2003; O'Brien and Lis, 1993). Further kinetic studies are needed to address this possibility.
Figure J-2: Permanganate footprinting of the hsp23 gene. Purified DNA (lanes 1, 7 and 8) was treated for the indicated times with permanganate on ice. Non-heat shocked (NHS) and heat shocked (HS) cells (lanes 2-6 and 9-13) were treated with 10 mM potassium permanganate for 1 minute on ice. Isolated DNA was cleaved at oxidized thymines with piperidine and the resulting breaks in the non-template strand (the gel on the right) or in template strand (the gel on the left) were mapped by LM-PCR. Treatment of cells with 200 µM DRB prior to heat shock greatly reduced the permanganate reactivity detected in both the 5' and 3' region (lane 4, 6, 11 and 13). Number above each gel indicates the primers used in LM-PCR. Numbers along the side of each gel indicate the distance of hyper-reactive thymines from the transcription start site. The polyadenylation signal is at +857-+862.
Appendix K

Changes in Pol II Density on Genes Controlled by Pcf11-mediated Premature Termination

Thousands of genes had increase in Pol II density upstream from their polyadenylation sites upon Pcf11 depletion (Figure 3-8). To test the hypothesis that these genes are controlled by Pcf11-mediated premature termination, 9 genes were chosen and the changes in their mRNA levels caused by the depletion of Pcf11 were examined (Figure 3-9). Here I present the screen shots of the Pol II ChIP-chip data (displayed in UCSC browser) for each of those analyzed genes (Figure K-1). Real time PCR was used to verify the increase in Pol II density on these genes detected by the ChIP-chip analysis. The results of those verifying real time PCR experiments are also presented here (Figure K-2).
Figure K-1: Display (UCSC browser) of Pol II ChIP-chip data for 9 genes where Pol II occupancy increased in the regions upstream from the polyadenylation site upon Pcf11 depletion. The height of vertical bars corresponds to the intensity of Pol II ChIP signals. The red rectangular box covers the region that was amplified by real time PCR for both the mRNA analysis (Figure 3-9) and the verification of ChIP-chip results (Figure K-2). The black arrow indicates the transcription direction of the chosen gene.

Figure K-2: Real time PCR verified the increases of Pol II density upon Pcf11 depletion at the genes where premature termination possibly occurs. Pol II density in regions covered by the red rectangular boxes in Figure K-1 was examined by real time PCR for both control and Pcf11-depleted samples. Each bar represents the mean of three independent experiments and the error bar represents the standard deviation.
Appendix L

Features Associated with the Genes in Category 1, 2 and 3

In Chapter 3, I described the identification of three Pol II distribution patterns, namely category 1, 2 and 3. Here I present the results of some bioinformatics analysis revealing the features associated with each category, including the position of the polyadenylation signal, chromatin structure, biological function and changes in Pol II density upon Pcf11 depletion.

L.1 Position of the polyadenylation signal

The various Pol II distribution patterns at the 3’ end of genes invite the question: is there any correlation between the position of the polyadenylation signal and the Pol II distribution profile? Hence I investigated the position of the polyadenylation signal by plotting the composite frequency of the AATAAA hexamer within +/-1 kb of the annotated polyadenylation site. Figure L-1 shows that the position of AATAAA hexamer is the same for the genes in category 1, 2 and 3. Since the peaks of all three AATAAA frequency distributions are aligned well at around 30 bp upstream from the polyadenylation site, the position of the polyadenylation signal is probably not related to the distinct Pol II distribution patterns at the 3’ end. The difference in the height of the peaks probably reflects heterogeneous composition of canonical and non-canonical polyadenylation signals in each category, since a large portion of eukaryotic genes are
found to possess functional non-canonical polyadenylation signals such as ATTAAA (Beau.doing et al., 2000; Graber et al., 1999; MacDonald and Redondo, 2002).

Figure **L-1**: The position of the polyadenylation signal is the same for category 1, 2 and 3 genes. Composite distribution of the frequency of AATAAA hexamer from -1 kb to 1 kb of the polyadenylation site for the genes in category 1 (purple), 2 (orange) and 3 (green).

**L.2 Chromatin structure**

Figure **L-2** shows that the chromatin structure at the 3’ end is similar for the genes in category 1, 2 and 3. Examination of the composite distributions of H2A.Z, bulk and Pol II-bound nucleosomes revealed a common high-low-high profile. Note that the occupancy of Pol II-bound nucleosomes is intermediate upstream from the polyadenylation site for category 3 genes (Figure **L-2C**), whereas Pol II is present in the same regions at a very low level. This discrepancy was probably caused by the usage of
counts instead of signal intensity in plotting the composite Pol II-bound nucleosome plots (Figure L-2C). See chapter 3 for discussion about the possible relationship between Pol II distribution and nucleosome distribution at the 3’ end of genes.

Figure L-2: Nucleosome distributions at the 3’ end of category 1, 2 and 3 genes. Composite distribution of A) H2A.Z nucleosome, B) bulk nucleosome and C) Pol II-bound nucleosome from -1.5 kb to 1.5 kb of the polyadenylation site for the genes in category 1 (purple), 2 (orange) and 3 (green). Genes with no significant Pol II ChIP signals (above 5% FDR) within +/-1 kb of the polyadenylation site (black) are also considered in panel A and B.

L.3 GO analysis

Gene Ontology (GO) analysis was carried out to identify enriched genes that share similar biological functions within category 1, 2 and 3 (Table L-1). The major difference observed among the three categories is that both category 1 and 2 are enriched with highly transcribed ribosomal protein genes. This is consistent with my expression profile analysis for these categories (Figure 3-7).
L.4 Changes in Pol II density upon Pcf11 depletion

In an attempt to correlate the impact of Pcf11 depletion on transcription termination with the Pol II distribution pattern at the 3’ end of genes before Pcf11 depletion, I investigated the changes in Pol II density in the vicinity of the polyadenylation site for the genes in category 1, 2 and 3 respectively in the following clustering analysis.

As Figure L-3 shows, after the depletion of Pcf11, 28% of category 1 genes, 13% of category 2 genes and 31% of category 3 genes had increased Pol II occupancy downstream from the polyadenylation site, indicating a termination defect. Chi-square tests were used to calculate the statistical significance for the differences between 28%
(category 1) and 13% (category 2) and between 13% (category 2) and 31% (category 3). A $P$-value of 1.99375E-27 was obtained for the comparison between 28% (category 1) and 13% (category 2). A $P$-value of 2.53353E-39 was obtained for the comparison between 13% (category 2) and 31% (category 3). These comparisons indicate that significantly less genes in category 2 than in the other two categories had termination defects upon Pcf11 depletion.

32% of category 1 genes, 20% of category 2 genes and 8% of category 3 genes had increased Pol II occupancy upstream from the polyadenylation site upon depletion of Pcf11, making these genes good candidates for having Pcf11-mediated premature termination (Figure L-3). 40% of category 1 genes, 67% of category 2 genes and 61% of category 3 genes had decreased Pol II occupancy at the 3’ end upon Pcf11 depletion (Figure L-3). These decreases correlate with decreases of Pol II occupancy at the promoter, suggesting that transcription initiation of these genes was strongly affected by the depletion of Pcf11, probably due to a defect in the recycling of Pol II (see the discussion in Chapter 3).
Figure L-3: Changes in Pol II density at the 3’ end of category 1, 2 and 3 genes. For each category of genes, three Treeview panels are displayed. In the left panel, Pol II ChIP signals from the tiling arrays (before Pcf11 depletion) within +/-1 kb of the polyadenylation site of each gene are displayed. The intensity of red color corresponds to the Pol II level; black indicates a Pol II level below 5% FDR threshold. Each row represents an individual gene. All genes are aligned at the polyadenylation site (indicated by the black dot on top of the Treeview panel). In the middle panel, Pol II occupancy
changes occurring at the 3’ end upon Pcf11 depletion are displayed. Genes are arranged in the same order as in the left panel and aligned at the polyadenylation site (indicated by the black dot on top of the Treeview panel). Increase in Pol II occupancy is indicated with red; decrease in Pol II occupancy is indicated with green. No change in Pol II occupancy or Pol II ChIP signals below 5% FDR in both control and Pcf11-depleted samples is indicated with black. In the right panel, Pol II occupancy changes occurring at the 5’ end upon Pcf11 depletion are displayed. Genes are arranged in the same order as in the left panel and aligned at the transcription start site (indicated by the black dot on top of the Treeview panel). Increase in Pol II occupancy is indicated with red; decrease in Pol II occupancy is indicated with green. No change in Pol II occupancy or Pol II ChIP signals below 5% FDR in both control and Pcf11-depleted samples is indicated with black. The percentage of the genes that showed an increase in Pol II density downstream from the polyadenylation site is 28% for category 1, 13% for category 2 and 31% for category 3. The percentage of the genes that showed an increase in Pol II density upstream from the polyadenylation site is 32% for category 1, 20% for category 2 and 8% for category 3. The percentage of the genes that showed a decrease in Pol II density at the 3’ end is 40% for category 1, 67% for category 2 and 61% for category 3.
VITA

Ruopeng Fan

Education

2009  Ph.D. in Biochemistry, Microbiology and Molecular Biology, Department of Biochemistry and Molecular Biology, Pennsylvania State University, USA

2002  B.S. in Biology, Department of Life Sciences, Fudan University, China

Publication


