



January 2019

# Chromatin Associated Small RNA Show Evidence Of Processing After Promoter Proximal RNA Polymerase II Pausing

Nii Koney-Kwaku Koney

Follow this and additional works at: <https://commons.und.edu/theses>

---

## Recommended Citation

Koney, Nii Koney-Kwaku, "Chromatin Associated Small RNA Show Evidence Of Processing After Promoter Proximal RNA Polymerase II Pausing" (2019). *Theses and Dissertations*. 2567.  
<https://commons.und.edu/theses/2567>

This Dissertation is brought to you for free and open access by the Theses, Dissertations, and Senior Projects at UND Scholarly Commons. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of UND Scholarly Commons. For more information, please contact [zeinebyousif@library.und.edu](mailto:zeinebyousif@library.und.edu).

CHROMATIN ASSOCIATED SMALL RNA SHOW EVIDENCE  
OF PROCESSING AFTER PROMOTER PROXIMAL  
RNA POLYMERASE II PAUSING

by

Nii Koney-Kwaku Koney

Bachelor of Science Zoology, University of Ghana, 2005

Master of Philosophy Human Anatomy, University of Ghana 2009

A Dissertation

Submitted to the Graduate Faculty

of the

University of North Dakota

in partial fulfillment of the requirements

for the degree of Doctor of Philosophy in Anatomy and Cell Biology

Grand Forks, North Dakota

August

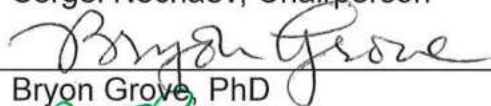
2019

Copyright 2019 Nii Koney-Kwaku Koney

This dissertation, submitted by Nii Koney-Kwaku Koney in partial fulfillment of the requirements for the Degree of Doctor of Philosophy from the University of North Dakota, has been read by the Faculty Advisory Committee under whom the work has been done and is hereby approved.



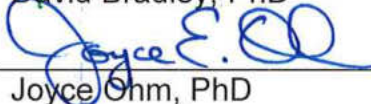
\_\_\_\_\_  
Sergei Nechaev, Chairperson



\_\_\_\_\_  
Bryon Grove, PhD



\_\_\_\_\_  
David Bradley, PhD



\_\_\_\_\_  
Joyce Ohm, PhD



\_\_\_\_\_  
Diane Darland, PhD, Member at Large

This dissertation is being submitted by the appointed advisory committee as having met all of the requirements of the School of Graduate Studies at the University of North Dakota and is hereby approved.



\_\_\_\_\_  
Chris Nelson, PhD  
Assoc. Dean of the School of Graduate Studies

7/30/19  
Date

## PERMISSION

Title: CHROMATIN ASSOCIATED SMALL RNA SHOW EVIDENCE OF  
PROCESSING AFTER PROMOTER PROXIMAL RNA POLYMERASE  
II PAUSING

Department: Anatomy and Cell Biology

Degree: Doctor of Philosophy

In presenting this dissertation in partial fulfillment of the requirements for a graduate degree from the University of North Dakota, I agree that the library of this University shall make it freely available for inspection. I further agree that permission for extensive copying for scholarly purposes may be granted by the professor who supervised my dissertation work, or in his absence, by the Chairperson of the department or the Dean of the School of Graduate Studies. It is understood that any copying or publication or other use of this dissertation or part thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of North Dakota in any scholarly use which may be made of any material in my dissertation.

Nii Koney-Kwaku Koney  
August, 2019

# TABLE OF CONTENTS

TABLE OF CONTENTS .....	v
List of Figures .....	ix
List of Tables .....	xv
ACKNOWLEDGMENTS .....	xvi
ABSTRACT.....	xviii
<b>CHAPTER 1 .....</b>	<b>1</b>
BACKGROUND.....	1
Eukaryotic Transcription .....	2
Pol II Pausing.....	3
Requirements of Early Elongation.....	4
Products of Transcription.....	5
Discovery of Non-canonical Short RNA Species.....	5
Premature Transcription Termination at the Site of Promoter-Proximal Pausing.....	9
Pol II Pausing and Histones .....	12
Products of Paused Pol II.....	13
Proposed Mechanism for Generation of Some Promoter Associated RNAs.....	14
Models of Premature Termination.....	15
Methods Used to Decipher Gene Regulation at the Transcription Level .....	17
Gap in the Field.....	19
<b>CHAPTER 2 .....</b>	<b>23</b>
MATERIALS AND METHODS .....	23
Cell Culture.....	23

Western Blotting.....	23
siRNA Depletion.....	25
Cell Fractionation.....	25
Run-On Permanganate footprinting detects positional change in Pol II complex....	26
Nuclear Run-On shows that isolated nuclei have RNAs that can be extended .....	28
Nuclei Fractionation.....	30
Radioactive Ligation Mediated (LM) PCR detects small RNA .....	34
Testing enzymes to map the 5' End Status of the Observed LM PCR RNA	
Products.....	42
RNA Preparation.....	44
Total RNA isolation.....	44
qPCR.....	44
Preparation of short-capped RNAs .....	46
Short Capped RNA library preparation .....	48
Short uncapped RNA Sequencing .....	49
Precision Run On sequencing (PRO-Seq) .....	52
Sequencing and Bioinformatics .....	53
Bioinformatics.....	54
Gene List and Curation .....	56
<b>CHAPTER 3 .....</b>	<b>57</b>
RESULTS: DEVELOPMENT OF PROTOCOL TO INVESTIGATE PREMATURE TERMINATION	
DURING POL II PAUSING .....	57
Introduction.....	57

LM PCR detects short RNA Products at the promoter region.....	59
Mapping the 5' End Status of the Observed LM PCR RNA Products .....	64
<b>CHAPTER 4.....</b>	<b>71</b>
RESULTS: PROCESSED RNAs ARE THE INTERMEDIATE PRODUCT BETWEEN PAUSED RNAs	
AND PREMATURE TERMINATION PRODUCTS.....	71
Genome-wide approach to characterizing RNAs at the 5' ends of genes.....	71
Characteristics of short capped RNA.....	73
Characteristics of short uncapped RNAs from chromatin fractions (Processed RNA)	
.....	81
Short uncapped processed RNAs which are found at the 5' ends of genes are	
generated from paused complexes .....	84
Processed RNAs which are enriched in highly paused genes.....	92
Role of XRN2 in generation of processed RNA.....	97
<b>CHAPTER 5.....</b>	<b>107</b>
RESULTS: MODIFIED PRO-SEQ REVEALS PROCESSING OF RNAs AT THE 5' ENDS OF	
GENES FROM PAUSED COMPLEXES .....	107
Introduction.....	107
Characteristics of PRO-Start and processed PRO-Seq.....	108
Using modified PRO-Seq, most genes do not show backtracking at the 3' ends of	
genes .....	117
Modified PRO-Seq shows that processed RNAs are strongly correlated to pausing,	
but weakly correlated to gene expression. ....	117



Modified PRO-Seq is comparable to the small RNA sequencing approach to studying generation of processed RNA on chromatin. ....	120
<b>CHAPTER 6.....</b>	<b>126</b>
DISCUSSION AND CONCLUSION.....	126
Discussion.....	126
Radioactive LM PCR.....	127
Capped and uncapped RNA sequencing.....	129
Backtracking.....	131
PRO-Seq.....	133
Model of premature termination.....	134
Conclusion.....	139

## List of Figures

FIGURE 1-1. NON CANONICAL RNA SPECIES AROUND A HYPOTHETICAL PROTEIN CODING GENE. ....	7
FIGURE 1-2. CHARACTERISTICS OF NON CANONICAL RNAs. ....	8
FIGURE 1-3. A SCHEME SHOWING PROMOTER PROXIMAL RNA POL II PAUSING. ....	11
FIGURE 1-4. MODELS OF PREMATURE TERMINATION AT PROMOTER PROXIMAL RNA POL II PAUSING REGION.....	16
FIGURE 1-5. A SCHEME OF SHOWING THE CELL FRACTIONS THAT CAN HAVE PRODUCTS OF PREMATURE TERMINATION .....	22
FIGURE 2-1: PERMANGANATE FOOTPRINTING OF THE SNAIL1 GENE ON RUN-ON NUCLEI IN HELA CELLS SHOWS CHANGES IN POL II COMPLEXES ON SNAIL1.....	27
FIGURE 2-2: SPECIFICITY OF A-BRDU BEADS AFTER ONE ROUND OF BEAD BINDING. ....	29
FIGURE 2-3. A. WESTERN BLOT VALIDATION OF CELL FRACTIONATION IN MCF7 CELLS. .	32
FIGURE 2-4: POL II COMPLEXES ARE STABLE IN HIGH SALT WASHES. ....	33
FIGURE 2-5: LIGATION MEDIATED PCR SCHEMATIC DIAGRAM FOR RADIOACTIVE LIGATION MEDIATED PCR (RADIOACTIVE LM PCR). ....	35
FIGURE 2-6: VALIDATION OF LIGATION MEDIATED PCR.....	38
FIGURE 2-7: OPTIMIZATION OF AMPLIFICATION CYCLE FOR LM RADIOACTIVE PCR.....	41
FIGURE 2-8: POLYACRYLAMIDE GEL IMAGES OF ENZYME ACTIVITY. ....	43
FIGURE 2-9: TESTING DIFFERENT RNA COLUMNS FROM DIFFERENT VENDORS FOR EXTRACTION OF SMALL RNA USING A RADIOLABELLED DNA OLIGONUCLEOTIDE (32P LABELLED HCDH1). ....	47
FIGURE 2-10: SYSTEMATIC APPROACH FOR STUDYING SHORT RNA DISTRIBUTIONS. ....	51

FIGURE 3-1. MODEL OF PREMATURE TERMINATION AROUND THE PROMOTER REGION OF GENES. ....	58
FIGURE 3.2. RADIOACTIVE LM PCR OF SNORD49A IN WHOLE CELL, CYTOPLASM, NUCLEI, CHROMATIN AND NUCLEOPLASM FRACTIONS. ....	61
FIGURE 3.3. RADIOACTIVE LM RADIOACTIVE PCR OF SNAI1 IN WHOLE CELL, CYTOPLASM, NUCLEAR, CHROMATIN, AND NUCLEOPLASM FRACTIONS.....	62
FIGURE 3.4. LM PCR OF HSPA1B IN WHOLE CELL, CYTOPLASM, NUCLEI, CHROMATIN, AND NUCLEOPLASM FRACTIONS. ....	63
FIGURE 3-5: MAPPING RNA ENDS IN RADIOACTIVE LM PCR REACTIONS. SNORD49A, HSPA1B, AND SNAI1 GENES WERE TESTED IN NUCLEI FRACTIONS.....	66
FIGURE 3-6: MAPPING RNA ENDS IN RADIOACTIVE LM PCR REACTIONS. ....	67
FIGURE 3-7. RADIOACTIVE LM PCR DETECTION OF SPIKE IN 2S (UNPHOSPHORYLATED) AND SPIKE IN 4S (CAPPED). ....	68
FIGURE 3-8 DIFFERENT SCENARIOS THAT CAN LEAD TO SHORTER RNA FRAGMENTS AROUND THE TSS OF GENES. ....	70
FIGURE 4-1: DIFFERENT TYPES OF RNAs (CAPPED AND UNCAPPED) AROUND THE PROMOTER REGION OF GENES.....	72
FIGURE 4-2: SHORT CAPPED RNA SIZE DISTRIBUTION OF READS.....	75
FIGURE 4-3: COMPARISON IN LENGTH DISTRIBUTION BETWEEN SHORT CAPPED RNAs FROM CHROMATIN AND SHORT CAPPED RNAs PUBLISHED FROM NUCLEI (SAMARAKKODY ET AL., 2015).....	76

FIGURE 4-4: CORRELATION PLOT FOR PROMOTER-PROXIMAL COUNTS BETWEEN SHORT CAPPED RNAs FROM CHROMATIN AND SHORT CAPPED RNAs FROM NUCLEI IN 10,114 GENES. ....	77
FIGURE 4-5: SHORT CAPPED RNAs 5' AND 3' POSITIONS WITHIN 50 NUCLEOTIDES UPSTREAM AND 100 NUCLEOTIDES DOWNSTREAM FROM THE TSS .....	78
FIGURE 4-6: SHORT CAPPED RNAs 3' POSITIONS WITHIN THE REGION 50 NUCLEOTIDES UPSTREAM AND 100 NUCLEOTIDES DOWNSTREAM OF THE TSS.....	79
FIGURE 4-7: A UCSC GENOME BROWSER SHOT OF THE SNAI1 GENE ILLUSTRATING THE 5' AND 3' END POSITIONS OF SHORT CAPPED RNAs FROM CHROMATIN AND PUBLISHED SHORT CAPPED RNA FROM NUCLEI (SAMARAKKODY ET AL., 2015). ....	80
FIGURE 4-8: METAGENE PROFILE OF SHORT UNCAPPED PROCESSED RNA 50 NUCLEOTIDES UPSTREAM AND 100 NUCLEOTIDES DOWNSTREAM OF THE TSS IN A. CYTOPLASMIC, B. NUCLEOPLASMIC, AND C. CHROMATIN FRACTIONS.....	82
FIGURE 4-9: SHORT UNCAPPED PROCESSED RNAs SIZE DISTRIBUTION OF READS .....	83
FIGURE 4-10: SHORT CAPPED AND SMALL UNCAPPED PROCESSED RNA 5' AND 3' POSITIONS 50 NUCLEOTIDES UPSTREAM AND 100 NUCLEOTIDES DOWNSTREAM OF TSS.....	86
FIGURE 4-11: DISTRIBUTION OF AVERAGE DIFFERENCE OF 5' END POSITIONS OF SHORT CAPPED AND SHORT UNCAPPED RNAs. ....	87
FIGURE 4-12: DISTRIBUTION OF AVERAGE DIFFERENCE OF 3' END POSITION OF SHORT CAPPED AND SHORT UNCAPPED RNAs. ....	89
FIGURE 4-13. UCSC GENOME BROWSER SHOT OF SNAI1 GENE SHOWING SHORT CAPPED RNAs (PAUSED RNA) AND UNCAPPED PROCESSED RNAs ON CHROMATIN. ....	90

FIGURE 4-14. SCATTER PLOTS SHOW THE 5' START SITE, 3' POSITION, AND DISTANCE TRANSCRIBED BY POL II IN SHORT CAPPED RNAs AND UNCAPPED PROCESSED RNAs ON THE SNAI1 GENE. ....	91
FIGURE 4-15. PROCESSED RNAs ARE ENRICHED IN HIGHLY PAUSED GENES (SHORT CAPPED RNA).....	94
FIGURE 4-16: PROCESSED RNAs CORRELATE WITH PAUSING BUT NOT WITH GENE EXPRESSION WITHIN THE REGION BETWEEN 100 NUCLEOTIDES UPSTREAM AND DOWNSTREAM OF THE TSS.....	96
FIGURE 4-17: WESTERN BLOT ANALYSIS OF XRN2 siRNA TREATED CELLS. ....	99
FIGURE 4-18: SIZE DISTRIBUTION OF SHORT UNCAPPED PROCESSED RNAs OF XRN2 DEPLETED CELLS.....	100
FIGURE 4-19: METAGENE PLOTS OF 5' AND 3' POSITIONS OF XRN2 PROCESSED RNAs WITHIN THE FIRST 100 NUCLEOTIDES OF GENES .....	101
FIGURE 4-20: NORMALIZED PROCESSED RNA LENGTH READS DISTRIBUTION FOR THE REGION BETWEEN 100 NUCLEOTIDES UPSTREAM AND DOWNSTREAM FROM TSS BETWEEN PROCESSED RNAs AND XRN2-KD PROCESSED RNAs.....	102
FIGURE 4-21: NORMALIZED 5' ENDS OF PROCESSED RNAs AND 5' ENDS OF PROCESSED XRN2 DEPLETED DISTANCE FROM TSS.....	104
FIGURE 4-22. XRN2 DEPLETED PROCESSED RNAs ARE ENRICHED IN HIGHLY PAUSED GENES (SHORT CAPPED RNA). ....	106
FIGURE 5-1: PRO-START RNA SIZE DISTRIBUTION OF READS WITHIN THE FIRST 100 NUCLEOTIDES OF GENES.....	110

FIGURE 5-2: PRO-CAP RNA 5' AND 3' POSITIONS WITHIN THE REGION BETWEEN 100 NUCLEOTIDES UPSTREAM AND DOWNSTREAM FROM THE TSS OF GENES. ....	111
FIGURE 5-3: SIZE DISTRIBUTION OF SHORT UNCAPPED PROCESSED PRO-SEQ RNAs READS WITHIN THE REGION BETWEEN 100 NUCLEOTIDES UPSTREAM AND DOWNSTREAM OF THE TSS OF GENES.....	112
FIGURE 5-4: METAGENE PLOTS OF 5' ENDS AND 3' ENDS OF PROCESSED PRO-SEQ WITHIN THE REGION BETWEEN 100 NUCLEOTIDES UPSTREAM AND DOWNSTREAM OF THE TSS. ....	113
FIGURE 5-5: FIGURE 4-10: PRO-START AND SMALL UNCAPPED PROCESSED RNAs 5' AND 3' POSITIONS WITHIN THE REGION BETWEEN 100 NUCLEOTIDES UPSTREAM AND DOWNSTREAM OF THE TSS.....	115
FIGURE 5-6: SCATTER PLOTS SHOW THE 5' START SITE, 3' POSITION, AND DISTANCE TRANSCRIBED BY POL II IN PRO-START RNAs AND PROCESSED PRO-SEQ RNAs ON THE SNAI1 GENE.....	116
FIGURE 5-7: DISTRIBUTION OF AVERAGE DIFFERENCE OF 5' AND 3' ENDS IN MODIFIED PRO-SEQ DATA. ....	118
FIGURE 5-8: THE RELATIONSHIP BETWEEN PROCESSING OF PAUSED RNAs TO PAUSING AND GENE EXPRESSION. ....	119
FIGURE 5-9: UCSC BROWSER SHOTS COMPARING THE SMALL RNA SEQUENCING TO MODIFIED PRO-SEQ. ....	121
FIGURE 5-10: FIGURE 4-3: COMPARISON IN LENGTH DISTRIBUTION BETWEEN PRO-START RNA FROM SHORT CAPPED RNAs FROM CHROMATIN WITHIN THE FIRST 100 NUCLEOTIDES OF GENES. (1 OF 2 REPLICATES EACH). ....	122

FIGURE 5-11 COMPARISON BETWEEN PROCESSED PRO-START AND SHORT CAPPED RNAs  
(CHROMATIN) AT 3' POSITIONS WITHIN THE REGION BETWEEN 100 NUCLEOTIDES  
UPSTREAM AND DOWNSTREAM OF THE TSS. .... 123

FIGURE 5-12. COMPARISON BETWEEN PROCESSED PRO-SEQ AND PROCESSED RNAs  
(CHROMATIN) AT 5' POSITIONS WITHIN THE REGION BETWEEN 100 NUCLEOTIDES  
UPSTREAM OF AND DOWNSTREAM OF THE TSS..... 124

FIGURE 5-13: COMPARISON BETWEEN PROCESSED PRO-SEQ AND PROCESSED RNAs  
(CHROMATIN) AT 3' POSITIONS WITHIN THE REGION 100 NUCLEOTIDES UPSTREAM AND  
DOWNSTREAM OF THE TSS..... 125

FIGURE 6-1. PROPOSED MODEL OF BIOGENESIS OF PROCESSED RNA BY DXO, DECAPPING  
AND EXORIBONUCLEASE PROTEIN AND XRN2, 5' TO 3' EXORIBONUCLEASE PROTEIN.  
..... 138

## List of Tables

TABLE 1. 1. PROTEINS IMPLICATED IN PREMATURE TERMINATION FROM THREE PROPOSED MODELS OF PREMATURE TERMINATION.....	18
TABLE 2.1. LIST OF ANTIBODIES USED FOR WESTERN BLOTTING. ....	24
TABLE 2-2: CT VALUES OF SMALL RNA ADAPTER LIGATION QPCR TESTED ON SPECIFIC GENES UNDER FOUR CONDITIONS. ....	37
TABLE 2-3: CT VALUES OF SMALL RNA ADAPTER LIGATION QPCR TESTED ON SPECIFIC GENES WITH AND WITHOUT REVERSE TRANSCRIPTASE IN VARIOUS CELL FRACTIONS. ....	37
TABLE 2-4: CT VALUES OF SMALL RNA ADAPTER LIGATION QPCR TESTED ON <i>IN VITRO</i> MADE RNAS WITH AND WITHOUT REVERSE TRANSCRIPTASE.....	39
TABLE 2-5: DETERMINING THE AMOUNT OF STARTING RNA SUFFICIENT FOR ADAPTER LIGATION QPCR -EXTRACTED RNA WITH AND WITHOUT REVERSE TRANSCRIPTASE... ..	39
TABLE 2. 6. LIST OF QPCR PRIMERS AND THEIR SEQUENCES .....	45
TABLE 2-7: ALIGNMENT PERCENTAGES FOR SHORT CAPPED RNA SEQUENCING .....	48
TABLE 2.8. ALIGNMENT STATISTICS FOR SHORT UNCAPPED RNA.....	52
TABLE 2.9. MODIFIED PRO SEQ APPROACH.....	53
TABLE 3.1: APPROACH USED TO DELINEATE 5' END MODIFICATION USING RADIOACTIVE LM PCR.....	64



## ACKNOWLEDGMENTS

It takes a village to raise a kid. My journey at UND would not have been possible without the influence of many individuals:

First, I wish to express my appreciation to Sergei Nechaev (PhD), my advisor, for exposing me to the world of Pol II pausing and to the many techniques used to study transcription. To Bryon Grove (PhD), David Bradley (PhD), Joyce Ohm (PhD), and Diane Darland (PhD), thank you for your time and encouragement.

To my lab mates, I would not have survived without you. You showed me that irrespective of race and religion, love and kindness is the most powerful language. To Sayantani, Jessica, Ata, Atrayee, Smruthi, Atrayee Ray, Damien, Ann, Oscar, Bo, Humaira, Janani, Shawn, Mariia, Kole, Clemence, Maureen, Maria, and Kristina, the lab was always fun with all of you. I cannot express every act of kindness shown here, but I will always remember them. To Menglan, Kai, Junguk, Danielle, and Damien, my work moved forward because of your time and patience. Thank you all for your efforts with the bioinformatic analyses. Damien, your sacrifices pushed me through.

To Patrick and family, Thomas, Anushika, Fredice, Gisele, Leo, Sarmad, and Matt, thank you all for the support. To Kwaku Baryeh and family, you have been amazing. Kwaku Baryeh, I am still waiting for the \$1000 voucher I turned down for you.

To my friends: Grace, Lamisi, Dzabaku and Beryl, Sulley, Ken, Mwendalubi, Aku, Angela, Juliet, and Betty, thank you for the support. Thank you for consistently checking up on me, being there for me, and encouraging me.

To my former Anatomy colleagues at the University of Ghana: Kwame, Benjamin, Ewurama, Richard, Ann, Tanam, Eistine, and Godwin, we keep improving.

To my dad Lt Col.(rtd) E.M.O Koney and my mum Gifty Quaye, thank you for all the sacrifices. You have been amazing parents, and your prayers, faith, and trust have brought us this far. To my sisters Naa Atswei, Naa Ayorkor, and Naa Kai, thank you for being there for my kids. Thank you for all the support.

To my wife Dede Hesse, thank you for being part of this crazy journey. Thank you for your support. Thank you to your parents Prof Adukwei Hesse and Prof Afua Hesse for the wonderful support.

To my daughters Naa Adzeley and Naa Adzorkor, thank you for your sacrifices. We did it.

It has been an amazing time in Grand Forks, a city that will live in my heart for years to come for all of the experiences, the wonderful people, the amazing weather, and the growth as a person. There were challenges, but everything is about perspectives and how you see life.

To my grandmother, Naa Odoley (deceased), my mum Naa Adzeley, my daughters Naa Adzeley and Naa Adzorkor,

## ABSTRACT

Paused Pol II has been implicated in the generation of short non-coding RNAs, such as transcription start site RNAs (TSSa RNAs) and transcription initiation RNAs (tiRNAs), which are found within or around the 5' ends of gene promoters. The generation of these RNAs (which I termed "processed RNAs") is known to occur in the nucleus; however, whether their biogenesis is co-transcriptional or post-transcriptional is unknown. These RNAs have been proposed to be remnants of 5' to 3' processing and protected by paused Pol II. Processed RNAs may represent a novel class of small RNAs derived from promoter-proximal transcription termination.

I hypothesized that processed RNAs are co-transcriptionally generated from RNAs associated with paused Pol II. In order to distinguish between a co-transcriptional or post-transcriptional event, I fractionated the nucleus into the chromatin fraction and the nucleoplasmic fraction. If these processed RNAs were generated co-transcriptionally, they would be present in the chromatin fraction along with the chromatin-bound paused Pol II complex, and if they were generated post-transcriptionally, they would be present in the nucleoplasmic fraction. I demonstrated that processed RNAs associate with the chromatin fraction, indicating that they are co-transcriptionally generated. I further developed a modified version of PRO-Seq (PRO-Start and processed PRO-Seq) to confirm that processed RNAs are generated from paused Pol II.

In the literature, it is unclear whether the generation of these RNAs is independent of backtracking. To test whether backtracking was involved in the biogenesis of

processed RNAs, I compared the 3' end locations of the processed RNAs and short capped RNAs and found that most genes do not demonstrate backtracking. Alternatively, to study processing from the 5' ends, I investigated the role of XRN2, a nuclear 5' to 3' exoribonuclease, in the generation of processed RNAs. I discovered that XRN2 depletion resulted in an enrichment of these processed RNAs, but the results indicated other proteins may also be involved.

Currently, these processed RNAs are proposed to play a role in gene repression. Future experiments will focus on identifying and demonstrating the functional roles of these transcripts.

## **CHAPTER 1**

### **BACKGROUND**

Regulation of gene expression at the transcriptional level is a highly complex process which includes transcriptional and epigenetic control. It is not surprising, therefore, that this process can go amiss and lead to a broad range of diseases such as cancer, autoimmunity, diabetes, obesity, cardiovascular, and congenital diseases (Lee & Young, 2013). Disruption of epigenetic processes can lead to an altered gene regulation. Global epigenetic changes including DNA methylation, histone modifications, nucleosome positioning, and changes in non-coding mRNAs such as microRNAs have become hallmarks of cancer (Sharma et al., 2009). Therefore, regulation of gene expression is important for every aspect of cellular function. To understand how developmental and homeostatic programmes function requires that we know the transcription factors involved, their targets, and regulatory points (Adelman & Lis, 2012). RNA polymerase II (Pol II) is the main protein responsible for transcription of all protein-coding genes and has been a focus of transcriptional research; thus, it is imperative to understand every aspect of its regulation and the importance of the subsequent outcomes in the contribution to human health and disease.

## Eukaryotic Transcription

Eukaryotic organisms have developed sophisticated mechanisms of gene regulation to respond to developmental, environmental, and nutritional cues (Adelman & Lis, 2012). Regulation of gene expression is achieved in large part by controlling transcription (synthesis of mRNAs) (Margaritis & Holstege, 2008). Traditionally, transcription is divided into three main phases: initiation, elongation, and termination (Orphanides & Reinberg, 2002). In eukaryotes, transcription begins with the assembly of the pre-initiation complex, which includes general transcription factors TFIIA, TFIIB, TFIIC, TFIID, TFIIE, TFIIF (GTFs), and Pol II (Fujita & Schlegel, 2010). With the assembly of the pre-initiation complex, there are alterations in the manner in which DNA is packaged into chromatin (Margaritis & Holstege, 2008).

One important feature of Pol II in eukaryotes is its 52 tandem copies of the consensus repeat heptad  $Y_1S_2P_3T_4S_5P_6S_7$  (Corden, 1990; Phatnani & Greenleaf, 2006). In the early phase of transcription, TFIIH phosphorylates the serine 5 residue within the carboxy-terminal heptapeptide repeat (CTD) of the largest Pol II subunit (Komarnitsky et al., 2000; Schroeder et al., 2000). CDK9 of the pTEFb complex is known to phosphorylate serine 2 (Marshall et al., 1996).

The classical paradigm of transcriptional regulation focuses on the recruitment of Pol II to the promoter region of the gene as the main step – if not the only step – which is important in deciding which promoters, and genes, would be active. This observation may be mostly true for bacteria and yeast. However, the advent of technology has enabled numerous studies in eukaryotes, including *Drosophila* (Muse et al., 2007; Zeitlinger et al., 2007), human embryonic stem cells (Guenther et al., 2007), and other

human cell lines, resulting in a paradigm shift from transcriptional regulation occurring exclusively during initiation to regulation occurring post-initiation as well.

For example, at the genome-wide level, many studies using metazoan model organisms have revealed a high density of transcriptionally engaged Pol II paused near the promoter region. This indicates an additional layer of regulation which occurs after transcription initiation, but before the Pol II (complex) has cleared the promoter (Core & Lis, 2008; Guenther et al., 2007; Kim et al., 2005; Muse et al., 2007; Zeitlinger et al., 2007). This new finding was termed Pol II pausing.

### **Pol II Pausing**

In prokaryotes, once Pol II is recruited, the full-length mRNA transcript is made. However, studies observed an accumulation of Pol at the 5' end of the transcribed region of the *Drosophila* hsp70 gene, despite the gene being inactive (Giardina et al., 1992; D S Gilmour & Lis, 1986; Rasmussen & Lis, 1993; Rougvie & Lis, 1988). This phenomenon was thought to be a strategy that permitted robust and rapid activation of genes (Core & Lis, 2008; Muse et al., 2007; Nechaev & Adelman, 2008). Further findings that Pol II carried out transcription in isolated nuclei suggested that paused Pol II had initiated transcription on inactive genes (Law et al., 1998; Rougvie & Lis, 1988). In addition to findings in *Drosophila*, viral genes (HIV) and mammalian genes (Myc, Junb, and Fos) that exhibited this phenomenon were initially seen as exceptions to a universal rule (Kao et al., 1987; Krumm et al., 1992; Strobl & Eick, 1992). However, recent studies showed that in eukaryotes, this accumulation of Pol II (pausing) reflects a regulatory mechanism (Scheidegger et al., 2019).



Pausing may seem to function as a repressive mechanism. However, a study in human embryonic stem cells (ESCs) revealed that certain chromatin signatures accompanied accumulation of Pol II at the promoter region, which was indicative of gene activity. Surprisingly, only a subset of these genes with such active signatures showed detectable full length transcripts (Guenther et al., 2007). This contradiction suggested that both active and seemingly inactive genes can accumulate Pol II at promoter (proximal) regions.

Further studies that look at transcriptionally engaged Pol II indicated that few paused genes are transcriptionally inactive (Core et al., 2008). Most models have revealed that genes which are enriched in signal responsive pathways such as development, cell proliferation, and stress or damage responses exhibit paused Pol II (Adelman & Lis, 2012). Based on Pol II occupancy at the promoter regions, epigenetic marks, and gene expression data, Pol II pausing appears to be a new regulatory mechanism that can be observed on both active and inactive genes.

### **Requirements of Early Elongation**

Many different proteins have been proposed to modulate activities of pausing and its release into elongation. Factors such as DSIF (DRB Sensitivity Inducing Factor) and NELF (Negative ELongation Factor), as well as transcription termination factors (Dcp1 a, Xrn2, and TTF2) (Brannan et al., 2012), are enriched at promoters of metazoan genes and have been implicated in pausing. Pause release is mediated by the kinase activity of positive Transcription Elongation Factor b (pTEFb). pTEFb phosphorylates the DSIF/NELF complex (Cheng & Price, 2007; Wada et al., 1998; Yamaguchi et al., 1999;

Zhu et al., 1997). This leads to the dissociation of NELF from Pol II and leads to elongation.

### **Products of Transcription**

Technological advancements have facilitated whole genome sequencing and characterization of transcriptomes (Chaitankar et al., 2016; Reuter et al., 2015). In addition to mRNA, these advancements have revealed a broad range of poorly understood transcripts (Pertea, 2012). These new findings have disproven previous hypothesis that the principal purpose of Pol II transcription is protein production; in fact, only about 2% of transcription initiation events lead to translation (Lander et al., 2001). Genome sequencing projects have shed further light on the complexities of the genome to the effect that RNAs are equally as, if not more, important than proteins for controlling cellular function and phenotype (long non-coding RNAs, miRNA, siRNA, piwi RNA, etc).

### **Discovery of Non-canonical Short RNA Species**

Currently, there are a plethora of non-coding RNAs, such as the capped Promoter Associated RNAs (PASRs) (Kapranov et al., 2007; Project, 2009), the uncapped transcription start site RNAs (TSSa RNAs) (Seila et al., 2008; Valen et al., 2011), and the uncapped transcription initiation RNAs (tiRNAs) (Taft et al., 2009, 2010), which are within or around 5' ends of genes with different characteristics and unknown functions (Figure 1-1). These RNAs are either short (PASR and TSSa RNAs) or long non-coding RNAs with varying degrees of stability (Figure 1-2). Other small capped RNAs (Xie et al., 2013; Zamudio et al., 2014) are substrates for miRNA generation. Similarly, other

smaller groups of RNAs have been observed around splice sites, within the gene body, and at the 3' ends of genes. The role and importance of the generation of these RNAs in the cell is still unknown.

Some of these RNAs have arisen out of divergent transcription, a phenomenon in mammals and yeast promoter regions where Pol II is initiated in both the sense and antisense directions. Divergent transcription has been observed in most promoters: products of divergent transcription generate these RNAs (Transcription Start Site-associated RNAs TSSa-RNAs) in both the sense and antisense directions (Seila et al., 2008). A study showed that certain short RNAs associated with the promoter were involved in gene silencing (Kanhere et al., 2010).

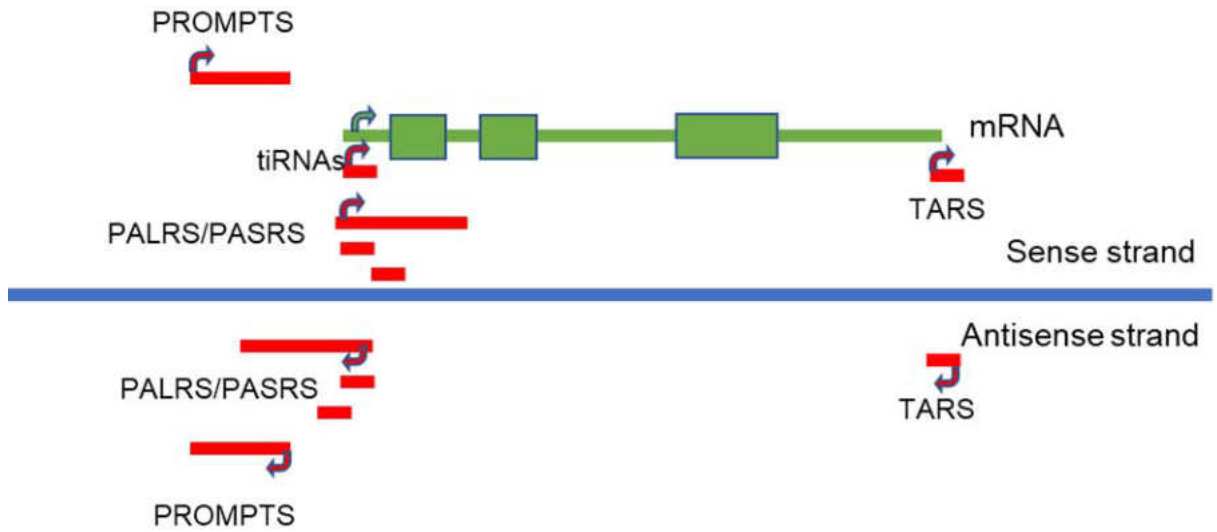


Figure 1-1. Non canonical RNA species around a hypothetical protein coding gene. DNA is shown in blue and standard RNA are shown in green. Non canonical RNA species are shown in red. These transcripts are found both in the sense and antisense direction. Some of the transcripts are long (PROMPTS and PALRS) and others are short (PARS and tiRNAs). TARS are located at the 3' ends of genes. Adapted from (Clark et al., 2013).

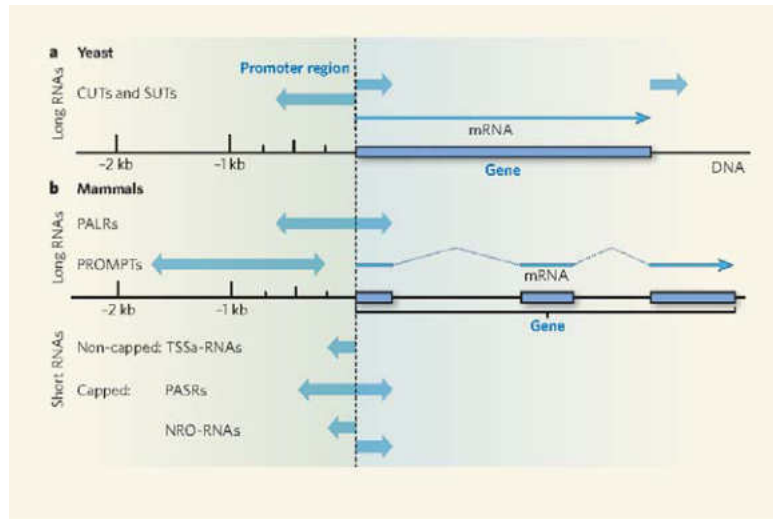


Figure 1-2. Characteristics of non canonical RNAs. A. In yeast, CUTS and SUTS are long noncoding RNA sequences. Non-coding RNAs are capped, Non-capped long and short. Single-headed arrows indicate the direction of transcription, and their position indicates the genomic region from which the ncRNA is transcribed. Mammalian genome show more complexities. There are long noncoding RNAs such as PARLS and PROMPTS, short non-capped (TSSa-RNAs) and capped (PASRs and NRO-RNAs). Used with permission (Carninci, 2009).

Preker et al., (2008) depleted the exosome machinery in their study, and this led to the discovery of small RNAs (PROMoter upstream Transcripts - PROMPTs) which are polyadenylated and unstable, suggesting that some of these RNAs are short lived. These PROMPTs transcripts also occur both in the sense and antisense directions (Preker et al., 2008). Preker et al., (2008) suggested that these transcripts correlate with gene activity and may have a regulatory role. Recently, a study by Zamudio et al., (2014) discovered that some RNA Pol II genes produce hairpin RNAs that feed into the miRNA pathway to produce transcription start site miRNA.

### **Premature Transcription Termination at the Site of Promoter-Proximal Pausing**

The difference between the density of Pol II at the promoter region and the gene body has intrigued the transcriptional field. This disconnect can occur as a result of Pol IIs which are not distributed evenly along the DNA (gene), but rather favouring certain locations such as the promoter-proximal sites, where they spend more time before resuming elongation. Potentially, the fraction of time Pol II spends on a given gene relative to release into elongation can contribute to the differences among gene expression.

Alternatively, the same pattern can be observed if these Pol IIs are prematurely terminating (Buratowski, 2009) (Figure 1-3). Subsequent studies have hypothesized that premature termination at the promoter region can explain the relatively lower amount of Pol II on gene bodies. Core et al., (2008) suggest premature termination may be a reason for the high densities of Pol II at the promoter region; this may be due to a promoter

experiencing high initiation rates as well as high rates of premature termination relative to the number of Pol II that escape into productive elongation (Figure 1-3). Hence, there is detection of high levels of engaged Pol II immediately prior to the point of termination.

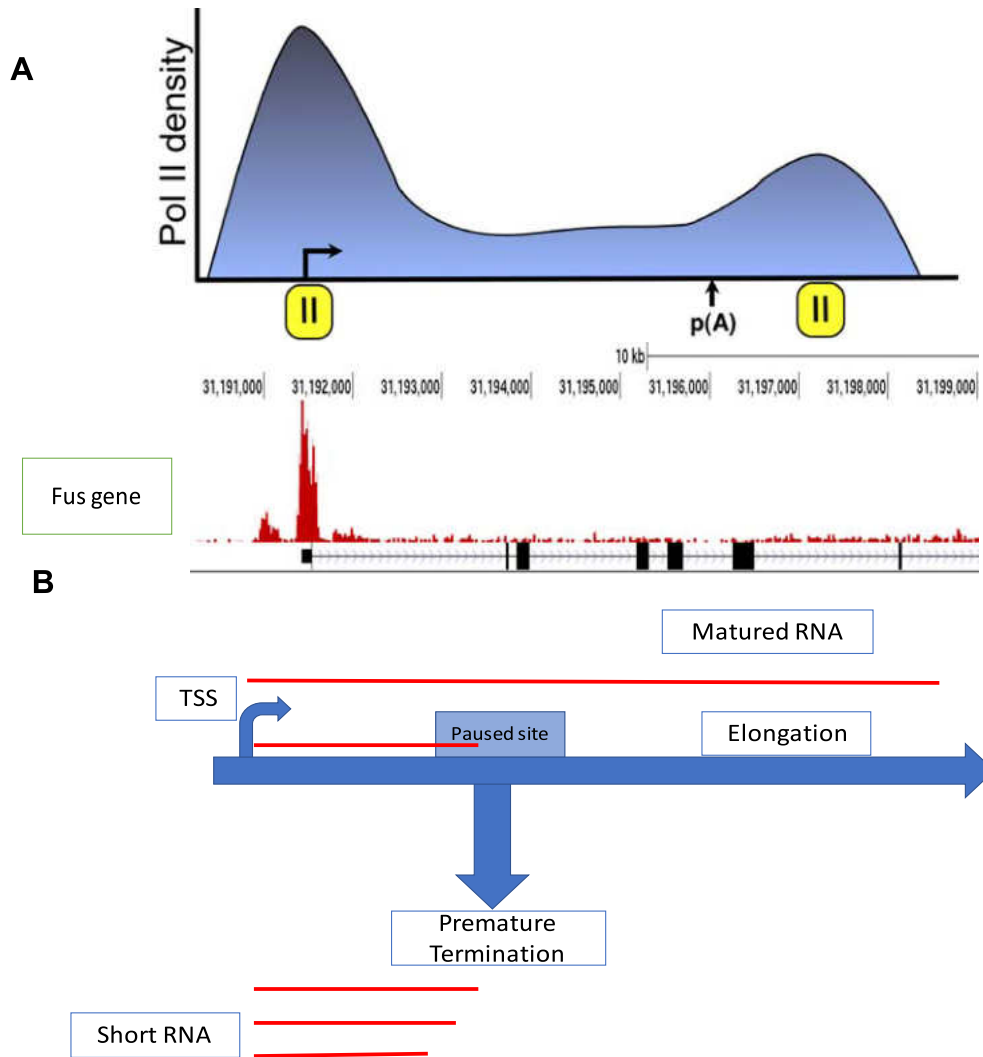


Figure 1-3. A scheme showing Promoter Proximal RNA Pol II pausing. **A** Pol II ChIP of Fus gene (Abbas unpublished) showing the discrepancy in signal at the promoter region compared to gene body and **B** with two possible outcomes. Premature termination to produce a short non-coding RNA and elongation to produce a matured RNA.



Few findings support the idea that once at the paused site, Pol II can either elongate to synthesize the mRNA or undergo premature termination by producing a short RNA (Buckley et al., 2014; Krebs et al., 2017). Buckley et al., (2014), using Hsp 70 as a model system in *Drosophila*, attempted to address the gap in knowledge related to the contribution of premature termination to accumulation of Pol II at the promoter. They observed that heat shock, which activates Hsp70 genes, dramatically increased the incidence of elongation of Pol II without decreasing the incidence of premature termination.

Previously, studies have shown that the HIV gene produces an abundant 59 nucleotide RNA, a product of premature termination of the early elongation complex. These short transcripts play a transcriptional repressive role in the regulation of the HIV gene in humans. This strategy resembles transcription anti-termination in bacteria, wherein transcription begins at the promoter region but then soon terminates. However, there is no current evidence that suggests high levels of promoter-proximal termination by Pol II at endogenous genes in eukaryotes performs similar functions to anti-termination in prokaryotes (Adelman & Lis, 2012).

It is an established fact that many genes exhibit paused Pol II. It is important to investigate how the elongation machinery is manipulated to give rise to gene activation or repression.

### **Pol II Pausing and Histones**

Studies have shown that combinations of histone modifications can determine gene activity; that is, either activation or repression. A study showed that most genes

associated with H3K4me3, H3k9ac and H3k14ac modifications in the promoter region. About half of these genes showed inactivity. Genes that showed activity by proceeding to elongation to produce matured transcripts showed enrichment of additional histone marks H3k36me3 and H3k79me2. Genes that lacked these marks still exhibited Pol II binding to the promoter region and produced 5' transcripts of 70 or less nucleotides (Guenther et al., 2007). This study showed that most genes support initiation, but do not complete transcription. This, together with other studies, suggests that premature termination may be a source of all the various RNAs seen which are associated with the promoter regions.

A review (Adelman & Lis, 2012) highlighted the absence of contribution of premature termination to the promoter proximal Pol II signal and elongation signal in the field of transcriptional regulation. Premature termination at the paused site remains enigmatic, even though in recent years there have been studies (Kanhare et al., 2010; Project, 2009) to suggest that these terminated RNAs may not be merely by-products of transcription. These suggest a repressive regulation of gene expression in the cell.

### **Products of Paused Pol II**

There is mounting evidence that the promoter is not only a hub for transcription initiation to produce mature mRNA. Additionally, many non-coding RNAs appear to be generated within the promoter region. Consistent with the short RNAs produced by Pol II pausing, recent studies have discovered short RNAs associated with the promoter regions of genes.

Even though different studies support the notion that Pol II pausing may have an alternative non-canonical pathway by premature termination, which generates small

RNAs, there is an absence of conclusive studies. For example, transcription termination factors XRN2 and TTF2 (Brannan et al., 2012), which are known to be responsible for transcription termination at the 3' end of genes, have been recently found to be enriched at promoters of human genes and may be responsible for premature termination of paused RNA pol-II complex. These findings support the notion that Pol II pausing is a decision point at which production of full length mRNA transcripts and short RNA transcripts can take place from the same gene promoter, which is an indication that the interplay between the mRNA and short RNA produced from the same gene may be regulatory. However, because this is a novel finding, the mechanisms and regulatory roles of premature termination remain unknown.

### **Proposed Mechanism for Generation of Some Promoter Associated RNAs**

Some promoter associated RNAs were proposed to be by-products of backtracking (Taft et al., 2009), a process where after pausing, Pol II retreats to a site with high thermodynamic stability when it encounters an obstacle (Nechaev et al., 2010). Studies by Valen et al., (2011) proposed that these RNAs are remnants of RNA processing and are protected by paused RNA Pol II. When XRN1/2 was depleted in the same study, they observed an increase in the lengths of these RNAs towards the 5' end, suggesting a role of XRN1/2 in post transcriptional processing of transcriptional start site associated RNAs and supporting the study by Brannan et al. (2012).

These findings point to a novel, previously unappreciated regulatory mechanism that involves Pol II pausing producing both mRNA and microRNA-like molecules from the same promoter, leading to a paradigm shift in our understanding of transcriptional regulation.

## Models of Premature Termination

Brannan et al., (2012) observed that XRN2, a nuclear 5'-3' exonuclease, and TTF2 are responsible for transcription termination at the 3' ends of genes. XRN2 and TTF2 were enriched at promoters of metazoan genes and colocalized with decapping proteins DCP1a and DCP2. The location of these proteins coincided with the occupancy of paused Pol II. Knockdown of XRN2 and TTF2 led to the redistribution of paused Pol II away from the start site either upstream or downstream. This finding suggests that under certain conditions, paused Pol II can undergo premature termination (Figure 1-4A).

In another model, the microprocessor complex (Drosha and Dgcr8) recruits XRN2 and SetX to the promoter of HIV1 in HeLa cells to initiate premature termination. This involves a stem loop RNA which is cleaved by Drosha. The cleaved RNA is further processed by Rps6, generating a small RNA (Wagschal et al., 2012) (Figure 1-4B).

Austena et al., (2015) also demonstrated that WDR82, SET1 H3K4 methyltransferase, and PP1 phosphatase abolished Pol II termination when depleted. They observed that enhancers and promoters produce Pol II dependent short ncRNA and these were elongated when these proteins were depleted. These studies suggest that different proteins (Table 1-1) play a role in premature termination and may be context dependent. This implies that pausing is a regulatory decision point in transcription that can direct the production of two different types of RNA from the same gene promoter. Generating the mRNA and pause release can also result in premature transcription termination to produce a short noncoding RNA. The possible release of these short RNAs has been difficult to characterize, partly because of exosome activity that rapidly degrades short transcripts that are generated (Preker et al., 2008).

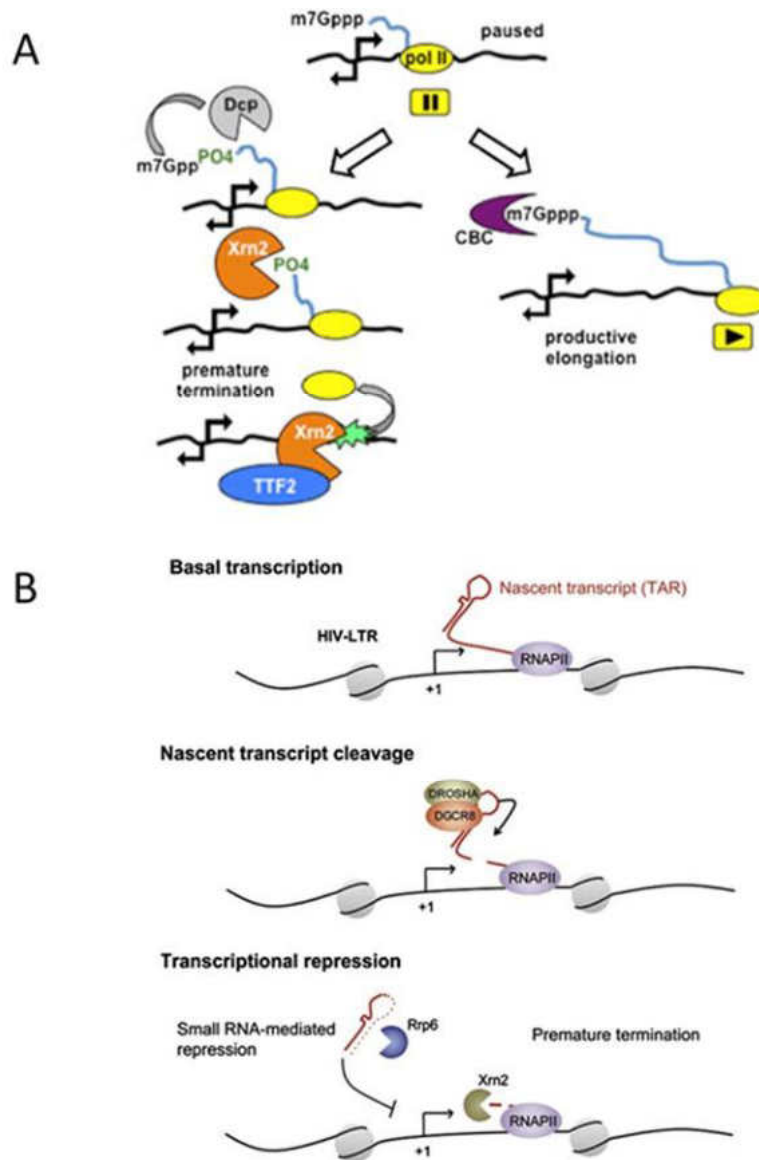


Figure 1-4. Models of premature termination at Promoter Proximal RNA Pol II pausing region. **A** Decapping model of premature termination RNAs at the 5' ends of genes are decapped by decapping proteins DCP leaving the ends with monophosphates. The monophosphate ends serve as a substrate for the 5' to 3' exonuclease to degrade the RNA. Together with transcription factor 2 (TTF2) they terminate Pol II. Used with permission (Brannan et al., 2012) and **B** Microprocessor model of premature termination. Used with permission (Wagschal et al., 2012). The microprocessor complex DGCR8 and Drosha are recruited to the promoter region. The microprocessor complex internally cleaves the RNA and the RNA associated with Pol II is terminated by XRN2. The free-floating RNA from the internal cleavage is further processed by Rrp6 to form a miRNA-like RNA.

## **Methods Used to Decipher Gene Regulation at the Transcription Level**

The balance between production and degradation of RNA from a given locus determines its steady state in a cell. The change in gene transcription (in response to stimuli) can be a result of altered RNA synthesis, stability or both (Paulsen et al., 2013). Current genome-wide approaches used in gene expression studies include Global Run On sequencing (GRO-Seq) (Core et al., 2008), Native elongating transcript sequencing (Net-Seq) (Churchman & Weissman, 2011; Mayer et al., 2015; Nojima et al., 2015), Short capped RNA sequencing (scRNA-Seq) (Nechaev et al., 2010; Samarakkody et al., 2015) and Bromouridine sequencing (Bru-Seq) (Paulsen et al., 2014). In global run-on and sequencing (GRO-Seq), nuclei from cells are isolated and initiated RNA are allowed to run-on *in vitro* in the presence of 5' bromouridine 5'triphosphate (Br-UTP). RNAs with Br-U incorporated are immunoprecipitated and sequenced. In Native elongating transcript sequencing (Net-Seq), nascent RNA is isolated by immunoprecipitation

Table 1.1. Proteins implicated in premature termination from three proposed models of premature termination.

<b><i>Gene Symbol</i></b>	<b>Description</b>		<b>Reference</b>
<b><i>XRN2</i></b>	Nuclear 5'-3' exonuclease	Hela, HIV Model	Brannan et al, 2012, Wagschal et al., 2012
<b><i>TTF2</i></b>	Transcription Termination factor 2	Hela, HIV Model	Brannan et al., 2012
<b><i>EDC3</i></b>	Enhancer of mRNA Decapping 3	Hela	Brannan et al., 2012
<b><i>DCP1a</i></b>	Decapping mRNA 1a	Hela	Brannan et al., 2012
<b><i>DCP2</i></b>	Decapping mRNA 2	Hela	
<b><i>Drosha</i></b>	Class 2 ribonuclease III enzyme	HIV model	Wagschal et al., 2012
<b><i>DGCR8</i></b>	DiGeorge syndrome critical region 8	HIV model	Wagschal et al., 2012
<b><i>Rrp6</i></b>	3'-5' exoribonuclease		
<b><i>SetX</i></b>	Senataxin	HIV model	Wagschal et al., 2012
<b><i>WDR82</i></b>	WD repeat domain 82	Mouse macrophage	Austena et al., 2015
<b><i>SET1</i></b>	H3k4 methyl transferase	Mouse macrophage	Austena et al., 2015
<b><i>PPI phosphatase</i></b>	Nuclear protein phosphatase 1	Mouse macrophage	Austena et al., 2015

of Pol II elongation complex followed by deep sequencing of the 3' ends of the nascent transcript associated with Pol II. Short-capped RNA sequencing (scRNA-Seq) involves isolation of nuclei, size selection of short RNA species, and enzymatic degradation of RNAs that lack a 5' cap before sequencing. scRNA-Seq identifies the start site of genes and the promoter proximal paused site of genes. Bromouridine sequencing (Bru-Seq) involves the metabolic pulse-chase sequencing of nascent RNA with bromouridine in cells to study stability of RNAs by monitoring its synthesis and degradation.

Most genome-wide techniques such as Global run on and sequencing (GRO-Seq) (Core et al., 2008), Native elongating transcript sequencing (Net-Seq) (Churchman & Weissman, 2011), Short capped RNA sequencing (scRNA-Seq) (Nechaev et al., 2010) and Bromouridine sequencing (Bru-Seq) (Paulsen et al., 2013) have not taken the possible role of termination into account. Some of these methods have conflated the extracted RNAs, which makes it difficult to distinguish between RNAs associated with chromatin and premature terminated RNAs. Secondly, some of these approaches (Net-Seq and scRNA-Seq) do not distinguish between newly formed RNAs and previously made RNAs. Thus, these methods do not fully capture the complexities associated with regulation and the contribution of nascent RNA synthesis or RNA decay to steady state RNA changes (Paulsen et al., 2013)

### **Gap in the Field**

Many small RNAs around the promoter region have been ascribed different names but may essentially be the same. The reason for this difference in nomenclature stems from different approaches used and the technological challenges at the time of



discovery. A subset of these RNAs, tiRNAs (Taft et al., 2009) and TSSa RNAs (Valen et al., 2011), are likely to be generated by the same paused Pol II and may be a pathway that leads to premature termination.

The biogenesis of these RNAs has not been fully established, with processes such as backtracking (Taft et al., 2009) and 5' end processing being implicated but not proven (Brannan et al., 2012; Nojima et al., 2015; Valen et al., 2011). The limitation in these studies was the approach used to compare the location of the 3' ends of these RNAs. The 3' ends were compared to the location of Pol II in ChIP experiments. Pol II ChIP has low resolution, which makes it difficult to rule out the processes of backtracking, since it is difficult to obtain the exact location of the Pol II at single nucleotide resolution. Additionally, these experiments were carried out in nuclei or whole cell fractions, even though the results obtained suggest that their biogenesis is co-transcriptional and not post-transcriptional.

I addressed these limitations by purifying short RNAs associated with chromatin and preparing libraries by selecting for capped and uncapped RNA. This approach allowed me to determine, with nucleotide resolution, the location of both capped RNA and uncapped RNA within the promoter region. This approach is also important to prove the premise that these short RNAs are generated from capped RNA. Short RNAs in the paused state should be associated with chromatin, while the nucleoplasm or cytoplasm should contain short RNAs generated by termination (Figure 1-5). Thus, fractionation of cells into chromatin, nucleoplasm, and cytoplasm provides a platform to estimate the amount of short RNAs in these fractions.

This study attempts to address this gap in knowledge by examining the distribution of various RNA species across cellular sub fractions.

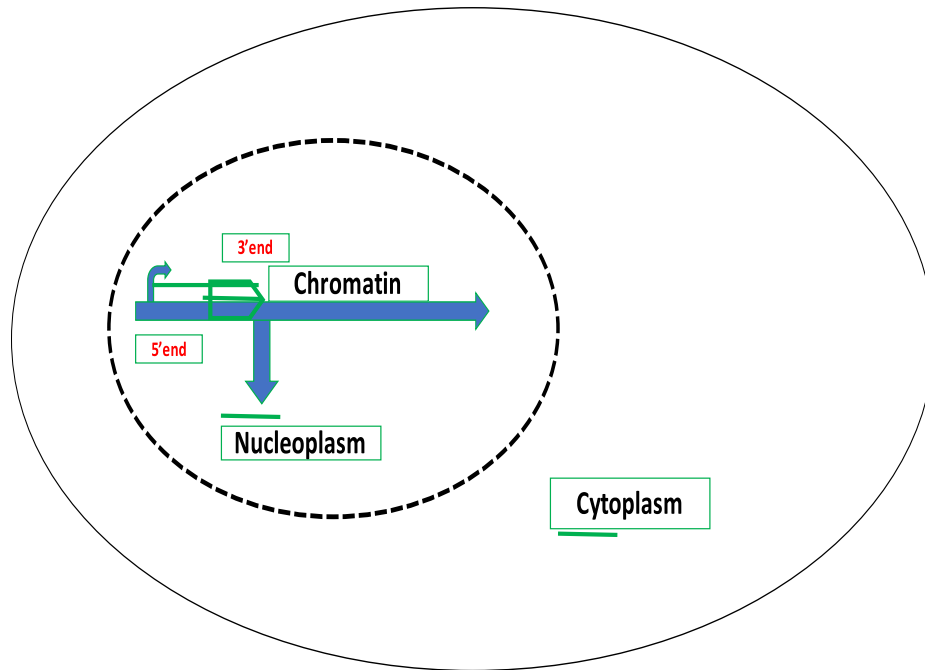


Figure 1-5. A scheme of showing the cell fractions that can have products of premature termination . The short RNAs generated can be found on chromatin, nucleoplasm or cytoplasm.

## **CHAPTER 2**

### **MATERIALS AND METHODS**

#### **Cell Culture**

MCF7, a human breast cancer cell line derived from a metastatic site by pleural effusion, was used in the study. These were obtained from American Type Culture Collection (Manassas, Virginia). Cells were cultured in DMEM/F-12 medium (Gibco, Thermo Fisher Scientific) with 10% fetal bovine serum (Atlanta Biologicals, Georgia). Cells were grown in an Eppendorf New Brunswick Galaxy 170S at 37°C and 5% CO<sub>2</sub> incubator. The growth medium was changed every 3 days.

#### **Western Blotting**

Cells were lysed in Urea lysis buffer (8M Urea, 1% SDS, 126mM Tris pH 6.8), and protein concentrations were determined using a Qubit 1.0 fluorimeter and Qubit protein assay kit (Invitrogen). The cell lysates were run on a Biorad Any kD Mini PROTEAN TGX gel. The ladder used was the ECL Rainbow Marker (Amersham). The gel was transferred onto a PDVF membrane using a wet transfer method for 2 hours. The blots were blocked in 5% milk in TBS with 0.1% Tween 20 (Sigma) for 1 hour. The blots were immunoblotted with anti-XRN2 (Abcam), anti-ExoSC3 (Abcam), anti-ExoSC10 (Abcam) to validate the siRNA depletions and anti-Actin (Millipore) was used as the control protein. Anti-Tubulin (Abcam), a cytoplasmic marker, anti-U1snRNP 70 (Santa Cruz), a nucleoplasmic marker, and anti-histone H3 (Abcam), a chromatin marker, were used to validate the efficiency of the fractionation. These primary antibodies were

incubated overnight at 4°C. The blots were washed with 1x TBST three times for ten minutes each, incubated with horseradish peroxidase (HRP) secondary antibody (GE) at a dilution of 1:10000 for 1 hour, and washed with 1x TBST three times for ten minutes each before being treated with Luminata Forte Western HRP substrate (Millipore) and imaged using an Odyssey imager.

Table 2.1. List of antibodies used for Western blotting.

Antibody	Catalogue Number	Company	Dilution/working concentration used
anti-XRN2	Ab72181	Abcam	1:1000
anti-ExoSC3	Ab190689	Abcam	1:1000
anti-ExoSC10	Ab50558	Abcam	1:1000
anti-Nuclear matrix p84 (5E10)	Ab487	Abcam	1:1000
anti-Tubulin	Ab44928	Abcam	1:200
anti-U1snRNP 70 (c18)	Sc-69571	Santa Cruz	1:100
Anti-Histone h3 (pan)	07-690	Abcam	
Anti-HSF1	H-311	Santa Cruz	1:200
Anti-Ser-2 clone 3E 10	04-1571	Millipore	1:2000
anti-Actin (clone C4)	MAB1501	Millipore	1:5000

### **siRNA Depletion**

Silencer® Select siRNAs (Negative control 1, XRN2, ExoSC3 and ExoSC10) were obtained from Ambion. Using Lipofectamine p3000 (Invitrogen), Opti-MEM (Gibco), 300 pmole of siRNA were transfected into 15cm dishes of MCF7 cells and allowed to grow for 4 days before harvesting.

### **Cell Fractionation**

Cells were grown in a 15cm dish and harvested at a confluency of 80%. Cells were washed with 10ml cold PBS, scraped in PBS and centrifuged at 4°C at 1000xg for a minute. The supernatant was discarded. Cells were resuspended in lysis buffer (10mM Tris-Cl pH7.5, 2mM MgCl<sub>2</sub>, 3mM CaCl<sub>2</sub>, 0.5% IGEPAL (Sigma-Aldrich), 10% glycerol (Sigma), 2 Units/ml SUPERase-In (Invitrogen), Protease Inhibitor Cocktail (Sigma)) and gently pipetted up and down 20 times using a p1000 tip with the end cut off to reduce shearing and incubated on ice for 7 minutes. The nuclei were centrifuged and pelleted for 7mins at 4°C at 1000 – 2000xg. The supernatant (cytoplasm) was aliquoted (by 5% total volume of lysate), and the pellet was resuspended in lysis buffer. The nuclei were centrifuged for 5mins at 4°C at 1000 – 2000xg and the supernatant was discarded. The nuclei were resuspended in 100ul of Freezing buffer (50mM Tris-Cl pH 8.3, 40% glycerol, 5mM MgCl<sub>2</sub>, 0.1mM EDTA) for storage -80°C. This protocol was modified from Core et al., (2008). The cell swelling step was removed and the lysis incubation was shortened to keep the cytoplasmic RNA intact.

### **Run-On Permanganate footprinting detects positional change in Pol II complex**

To determine if Pol II complexes remained stable and supported transcription in isolated frozen nuclei, I performed Run-On and permanganate footprint assays. RNA nucleotides were added to isolated nuclei at 30°C for 5 minutes to allow transcriptionally engaged polymerases to resume transcription. Thymines associated with single stranded DNA were cleaved with piperidine and Ligated-Mediated PCR was performed using radiolabeled primers. PCR products were run on 6% sequencing gels to visualize the Pol II “transcription bubbles” at base pair resolution through mapping of thymines in single stranded DNA regions (Gilmour & Fan, 2009). Figure 2-1 (Purple block) shows the loss of signal downstream from the TSS and an increase in band intensity at +86. The bands signal the location of the transcription bubble. This shift of signal was moderate in lane 8, as expected since the run-on in lane 8 was performed for only 1 minute. In lane 9, where run-on was performed for 5 minutes, the signal intensity of the bands that correspond to the transcription bubble in lane 8 diminished considerably, signifying that Pol II had moved. This indicated that the run-on reaction was successful and that these nuclei were viable.

This result demonstrated that nuclei possess stable complexes which remain intact after isolation. Notably, transcriptional complexes appeared to be stable and were identical in both fresh and frozen nuclei (Figure 2-1, lanes 4 and 5). This suggested that freezing nuclei at -80°C did not affect their integrity. Hence, frozen nuclei were used for subsequent experiments.

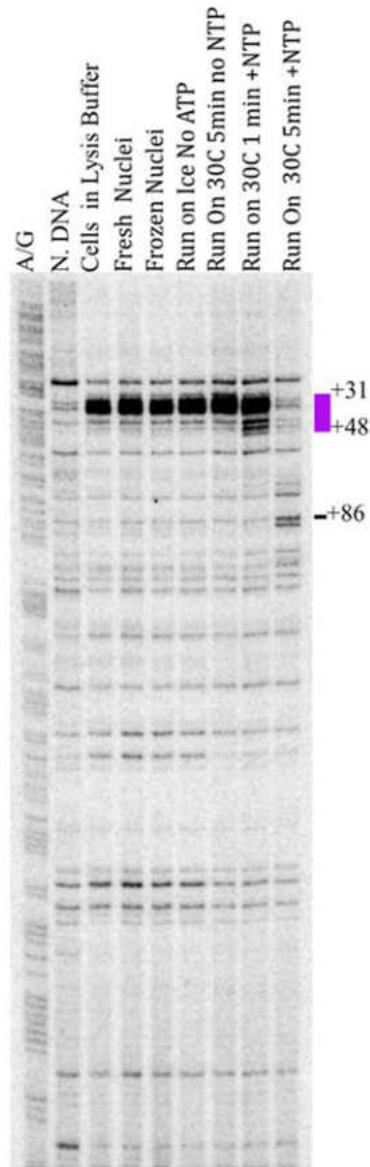


Figure 2-1: Permanganate footprinting of the SNAIL1 gene on run-on nuclei in HeLa cells shows changes in Pol II complexes on SNAIL1. The state of Pol II complexes in nuclei were investigated under different isolation, storage and run-on conditions. Run-on reactions (Lane 6-9) followed by permanganate footprinting treatment (Lane 1-9) underwent ligated mediated PCR. The amplified regions were displayed on a 6% sequencing gel and visualized by autoradiography. Lane 1 is the A/G marker, Lane 2 is naked DNA (N.DNA) treated with permanganate, Lanes 3 is cell in lysis buffer, Lane 4 is fresh nuclei, Lane 5 is frozen nuclei, Lane 6 is run-on ice with no ATP, Lane 7 is run-on at 30°C for 5 min with no ATP, Lane 8 is run-on at 30°C for 1 min with NTP, and Lane 9 is run-on 30°C for 5 min with NTP. The dark bands (purple block) show the location of Pol II.



### **Nuclear Run-On shows that isolated nuclei have RNAs that can be extended**

To test the integrity of the RNA associated with the Pol II complexes during nuclei isolation (Core et al., 2008), the RNA was labeled using a 5-Bromo-UTP (BrUTP) analog and a radioactive nucleotide ( $\alpha^{32}\text{P}$ -CTP), which served as a tracer. The labeled RNA was then pulled down using agarose beads that were conjugated with an antibody specific for  $\alpha$ -BrdU. Figure 2-2 (A and B), lanes 2 and 6 show that unfragmented RNAs were not degraded which suggested that the isolation of nuclei did not affect the integrity of the RNAs.

After one round of bead binding (Figure 2-2 C), I measured both the unbound and eluted RNA for BrU labelled RNA. Run-on was performed using only CTP as a control. Lane 4 shows a 30% enrichment of BrU labelled RNA compared to the control CTP only, lane 2, which is barely detected. RNAs were thus extended in the reaction, which suggested that RNAs remained intact and associated with Pol II during nuclei extraction. These results established that the nuclei isolation procedure could be used for my purposes.

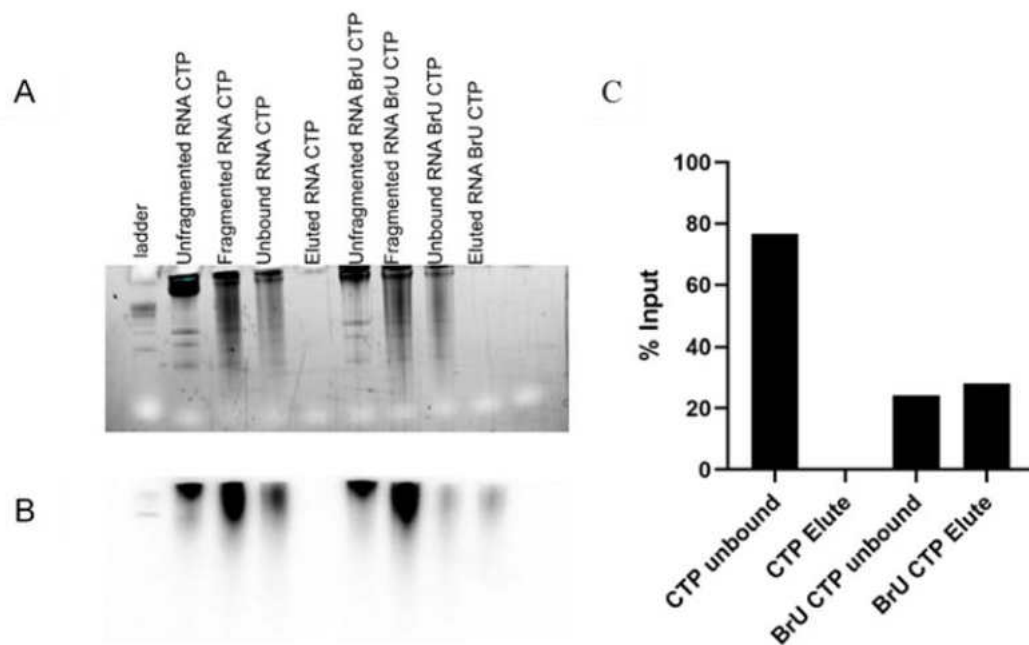


Figure 2-2: Specificity of  $\alpha$ -BrdU beads after one round of bead binding. A 15% gel image (ethidium bromide staining) B. Gel image visualized by autoradiography. Run-On nuclei RNA labelled with Br-UTP and  $\alpha^{32}\text{P}$ -CTP as the radioactive tracer in HeLa nuclei. Lane 1 is ssRNA, lane 2 is unfragmented RNA  $\alpha^{32}\text{P}$ -CTP, Lane 3 is cell in lysis buffer Lane 4: unbound RNA  $\alpha^{32}\text{P}$ -CTP, Lane 5: eluted RNA  $\alpha^{32}\text{P}$ -CTP, Lane 6: unfragmented RNA BrU  $\alpha^{32}\text{P}$ -CTP, Lane 7: fragmented RNA Br-UTP  $\alpha^{32}\text{P}$ -CTP, Lane 8: unbound RNA Br-UTP  $\alpha^{32}\text{P}$ -CTP, Lane 9: eluted RNA Br-UTP CTP  $\alpha^{32}\text{P}$ -CTP. C. Quantification by scintillation counting. Binding and Elution of NaOH-hydrolysed Br-UTP -RNA to  $\alpha$ -BrdU beads were verified by Scintillation counting.

## Nuclei Fractionation

My main goal was to identify and characterize the RNAs associated with Pol II on chromatin and the other fractions. To fractionate the nuclei into the chromatin fraction and the nucleoplasm fraction, the isolated nuclei were resuspended first in NUN1 buffer [20mM Tris-HCl pH 8.0, 75mM NaCl, 0.5mM EDTA, 50% Glycerol, Proteinase Inhibitor 100x Sigma, 2units/ml SUPERase-In (Invitrogen)], vortexed, and then NUN2 buffer [20mM HEPES-KOH pH 7.6, 7.5mM MgCl<sub>2</sub>, 0.2mM EDTA, 300mM NaCl, 1M Urea, 1% IGEPAL(Sigma), Protease Inhibitor cocktail (Sigma), 2units/ml SUPERase-In (Invitrogen)] was added at a ratio of (1:9.6 ul). The nuclei were vortexed for 5 secs every 5 mins at maximum speed and incubated on ice for 15 minutes. The nuclear lysate was spun at 21,000xg at 4°C for 15 minutes to pellet the chromatin. The supernatant was collected as nucleoplasm and the pellet (chromatin) was washed in 500ul of HSB buffer (10mM Tris-HCl, pH 7.5, 500mM NaCl, 10mM MgCl<sub>2</sub>) and spun at 21,000xg for 5 mins. The supernatant was discarded. The fractions were stored in -80°C. The method was adapted from Mayer et al., 2015 and Nojima et al., 2015

To study the distribution of short RNAs in the nucleoplasm and chromatin fractions, I adapted previously published protocols (Mayer et al., 2015; Nojima et al., 2015; Wuarin & Schibler, 1994) to break up nuclei into the chromatin and nucleoplasmic fractions. After I fractionated the nuclei into the chromatin and nucleoplasmic fractions, I blotted for proteins specific to each fraction to check that the isolation approaches were successful and that there was minimal-to-no cross contamination.

Specifically, I selected Nuclear Matrix Protein and SnRNP 70 for my nucleoplasmic markers (Figure 2-3A). I observed that these two proteins were localized

to the whole cell, nuclei, and nucleoplasm as expected, and not to the cytoplasm or chromatin fractions. The cytoplasmic marker Tubulin was specific to the whole cell and the cytoplasmic lysate (Figure 2-3A). Because histone proteins are expected to localize to the chromatin fraction, I used Coomassie staining to visualize these small proteins, which are abundant and less than 20 kDa (Figure 2-3B). The histone proteins localized to the whole cell, nuclei, and chromatin fractions as expected.

To test the effect of salt on Pol II complexes during the fractionation of nuclei, I tested two salt concentrations (300mM and 500 mM). I excluded salt from the nuclei fractionation buffer (300mM) and centrifuged at 1000xg. I collected the supernatant and pellet. I performed this reaction with 300mM and 500mM NaCl. I collected the supernatant and pellet in both instances. Washing with high salt (500 mM NaCl) may remove any RNAs that are not tightly bound to the chromatin. The effectiveness of this step was evaluated by Western blotting. I used Ser 5P and Ser 2P, which are markers of two different phosphorylation states at the C-terminal domain of Pol II. These proteins are expected to localize to the chromatin fraction; however, there is little known about their presence in the nucleoplasm.

I compared the supernatant fractions from the 300 mM NaCl wash protocol and replaced 300 mM NaCl with 500 mM NaCl in another condition. I did not see enrichment of these proteins in the supernatant fractions. snRNP70 was used as a marker for separation of chromatin from the nucleoplasm (Figure 2-4). I found that the complexes remained stable even in high salt concentration and did not dissociate from the chromatin. This confirmed that the DNA, RNA, Pol II ternary complexes were highly stable (Wuarin & Schibler, 1994).

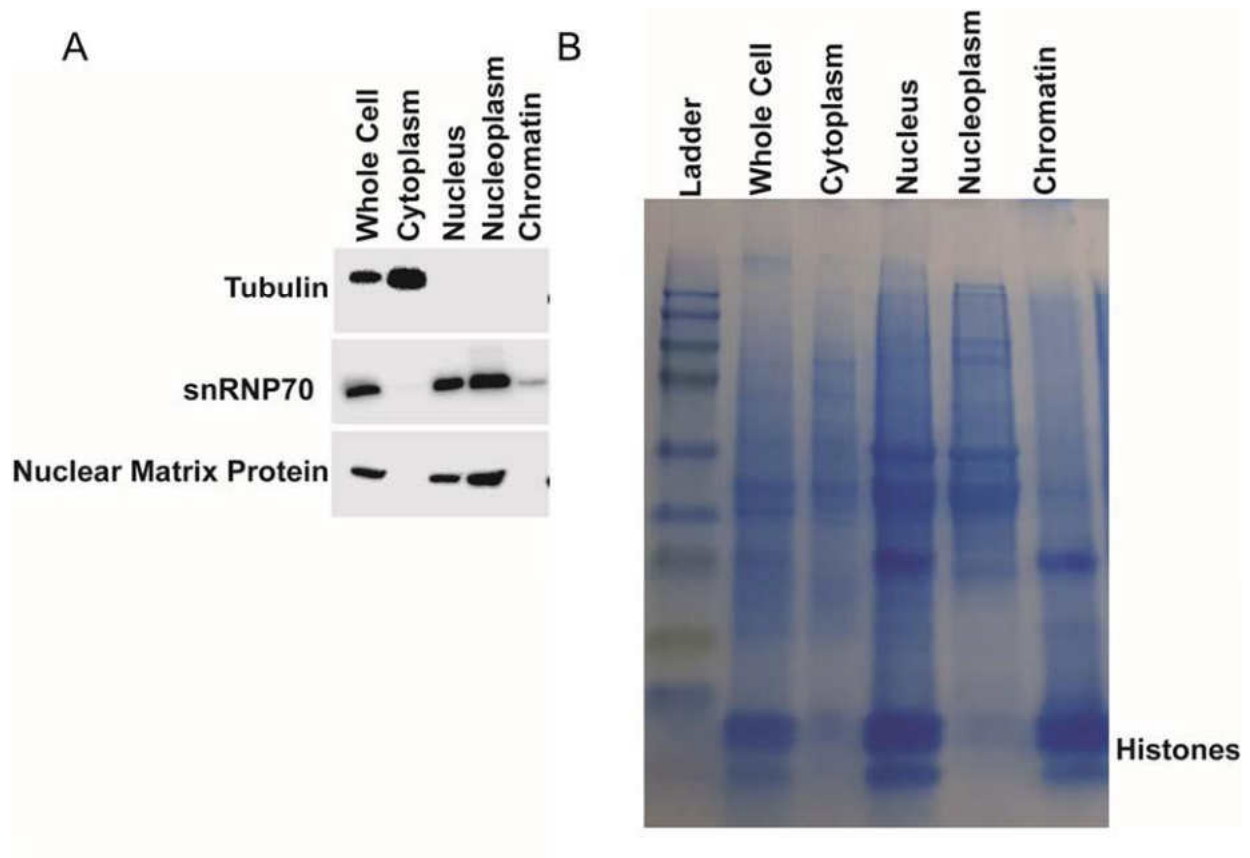


Figure 2-3. A. Western blot validation of cell fractionation in MCF7 cells. A. Subcellular localization of proteins were probed to evaluate the fractionation. The cytoplasmic marker used was Tubulin, nucleoplasm markers were snRNP70 and nuclei matrix protein. B. Coomassie staining of cellular fractions in MCF7 cells. Lane 1: the protein ladder, lane 2: whole cell lysates, lane 3: cytoplasmic lysate, lane 4: nuclear lysate, lane 5: nucleoplasmic lysate, and lane 6: chromatin lysate. Histone proteins served as chromatin markers, which could be seen in whole cell lysates, nuclear, and chromatin lysates.

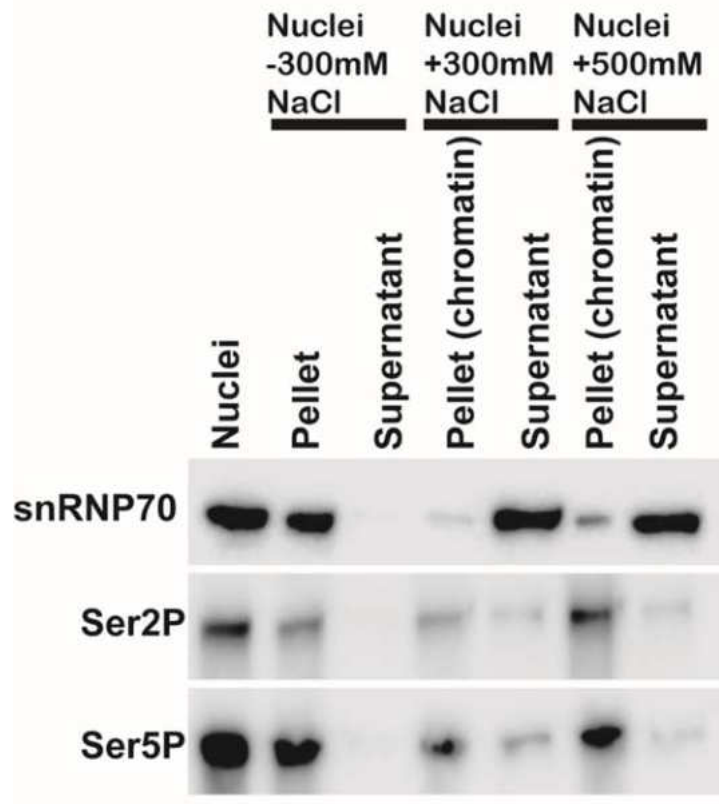


Figure 2-4: Pol II complexes are stable in high salt washes. Subcellular localization of proteins was probed to evaluate the state of Pol II in different salt buffers. Lane 1. Nuclei without any treatment. Lane 2 (Pellet) and lane 3 (supernatant) after nuclei were treated with lysis buffer without 300mM NaCl, Nuclei were fractionated with 300mM NaCl and the Pellet (chromatin) and supernatant (nucleoplasm) were collected (lane 4 and 5 respectively). Nuclei were treated with 500mM fractionation buffer and the supernatant (nucleoplasm) and pellet (chromatin) were collected (lane 6 and 7 respectively). NaCl is important in nuclei fractionation, as shown by enrichment of nucleoplasmic markers in lanes 5 and 7. Pol II phosphorylation (Ser 5P and Ser 2P) states were used to visualize Pol II. Pol II was enriched in nuclear and chromatin fractions (lanes 1,2,4&6).

### **Radioactive Ligation Mediated (LM) PCR detects small RNA**

It is difficult to detect short RNA by conventional qPCR, or polyadenylation qPCR, splinter ligase PCR, and Northern blot analysis (results not shown) because of the abundance detection limits of these approaches and the small size of these RNAs.

Therefore, to detect these RNAs, I used a method based on ligation mediated PCR combined with radioactivity to detect these RNAs in the cytoplasm, nucleoplasm, and chromatin cell fractions. Radioactivity amplifies the signal so that even low abundance RNAs can be detected

Because the RNAs were short in nature and not conducive for standard qPCR, I ligated adapters to the 3' ends of the RNA in a similar manner to small RNA library preparation. This allowed me to design a reverse primer for amplification based on the sequence of the adapter (Figure 2-5). For the forward primer, I selected two genes, SNAI1 and HSPA1B, which were known to exhibit pausing based on previous data (Samarakkody et al., 2015), and designed primers at the promoter region of these genes to amplify these regions. I selected SNORD49A as a control gene. To test whether this approach worked, I selected one primer each for SNAI1 and SNORD49A and two different primers for HSPA1B and performed the reactions in the presence or absence of the enzymes ligase and reverse transcriptase (RT) (Table 2-2). This gave me four conditions: the RNAs with or without an adapter, and with or without the RT reaction to make cDNA.

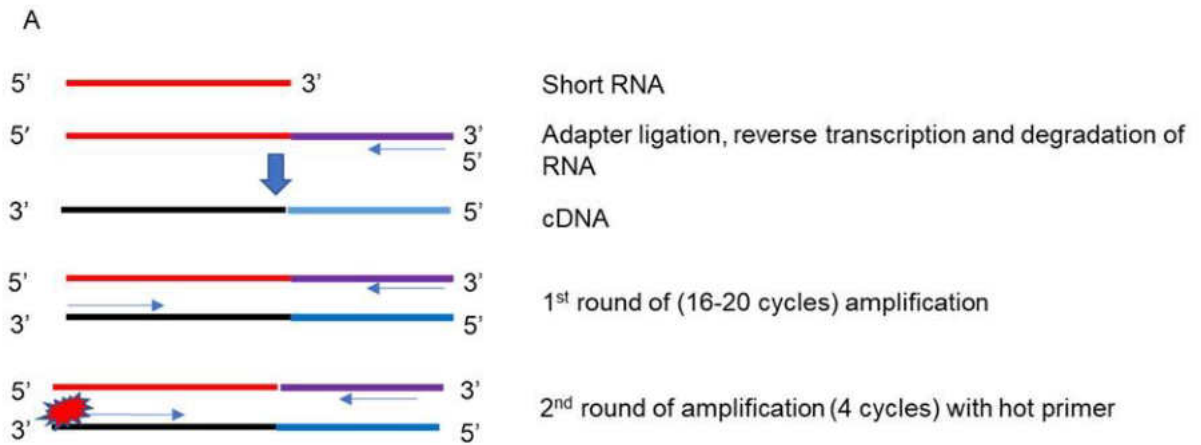


Figure 2-5: Ligation mediated PCR Schematic diagram for Radioactive Ligation Mediated PCR (Radioactive LM PCR). RNAs are ligated at the 3' prime end with adapters. The RNAs were reverse transcribed to generate cDNA. The cDNAs were amplified and subsequently amplified with a radiolabelled primer.



I used the cycle threshold (Ct values), which is the number of cycles for which a fluorescence product can be detected above the background signal, to assess whether the approach worked. A lower Ct signifies more of the product. Here, lower Ct values for the positive control (RNA plus ligase plus the RT) compared to the three negative controls (+ligase -RT, -ligase +RT, -ligase -RT) indicate that the experiment worked. For this experiment, I designed two primers for the HSPA1B gene, and one primer each for SNORD49A and SNAI1. Of these four primers, three resulted in an amplification product. I proceeded to clone the products of the amplification (Figure 2-6A) to further validate the approach. The sequenced products (Figure 2-6B) were BLATed on the UCSC genome browser, and an example (Figure 2-6B,C) shows the sequenced clone mapping to the expected region on the genome.

I tested this approach in different cell fractions, in the presence and absence of reverse transcriptase (Table 2-3), and I was able to detect RNAs in all of the fractions. I also tested this approach on two different batches of *in vitro* transcribed RNA (Set A and Set B) containing RNA (2S, 4S, 5S, 7S) of different lengths using the four aforementioned conditions. These RNAs were generated in the lab using DNA sequences designed from the mouse genome and *in vitro* transcribed using MEGAscript™ T7 Transcription Kit (Ambion) (Table 2-4). The approach worked because the Ct values for the positive controls were again lower than those of the negative controls. 2S and 4S RNAs showed low Ct values compared to 5S and 7S, suggested that the *in vitro* made RNA were of high quality.

I next set out to assess the limits of detection by using different concentrations of these RNAs. I tested the approach on diluted RNAs in whole cell and chromatin fractions

using plus or minus RT conditions. I expected and observed a 3-cycle difference in the 10-fold dilution series, confirming that there was indeed a 10-fold dilution. This affirmed the sensitivity of the approach (Table 2-5).

Table 2-2: CT values of small RNA adapter ligation qPCR tested on specific genes under four conditions.

<b>Conditions</b>	<b>SNORD49A</b>	<b>HSPA1A+5F</b>	<b>HSPA1B+2F</b>	<b>hSNAI1 4T_20</b>
+ligase +RT	24.03*	32.70	26.17*	27.38*
+ligase -RT	Not detected	37.85	36.66	37.51
-ligase +RT	38.01	36.01	34.32	37.57
-ligase -RT	Not detected	38.95	37.20	38.24

Table 2-3: CT values of small RNA adapter ligation qPCR tested on specific genes with and without reverse transcriptase in various cell fractions.

<b>+RT</b>	<b>Whole Cell</b>	<b>Cytoplasm</b>	<b>Nucleus</b>	<b>Chromatin</b>	<b>Nucleoplasm</b>
Snord49AF	23.96	29.75	22.65	20.03	22.32
hspA1B+2F	26.16	25.88	25.39	24.01	25.10
hSNAI14t_20	27.34	29.96	26.27	24.56	27.34
Spike 2sf	20.81	20.53	21.12	20.47	19.70

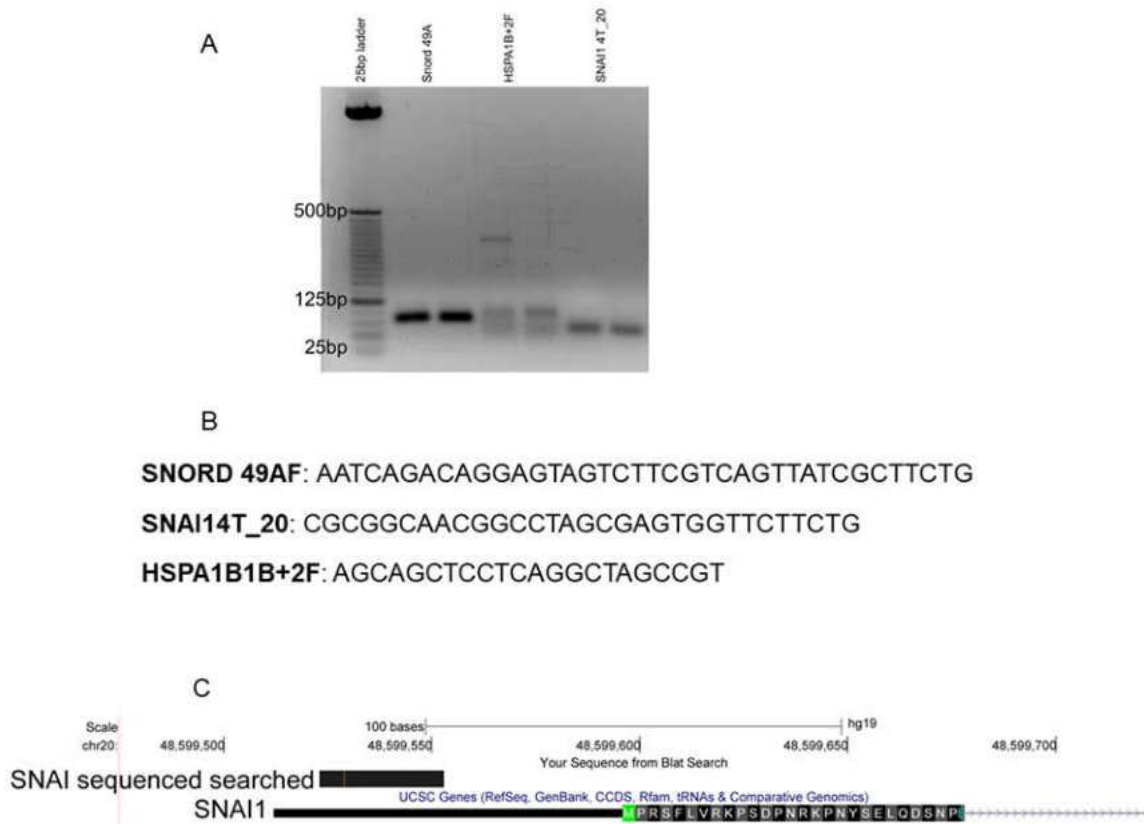


Figure 2-6: Validation of Ligation Mediated PCR A. Agarose gel of qPCR products of ligation mediated PCR. Expected sizes of products. B. Sequencing results SNORD49A, SNAI1, and HSPA1B+2F qPCR product. C. SNAI1 Validated on UCSC genome browser.

Table 2-4: CT values of small RNA adapter ligation qPCR tested on *in vitro* made RNAs with and without reverse transcriptase.

	SetA +ligase +Rt	SetA +ligase -Rt	SetA -ligase +Rt	SetA -ligase -Rt
2Sf	4.37	32.60	14.19	32.92
4Sf	3.58	19.65	13.49	25.60
5Sf	34.15	36.51	38.80	38.96
7Sf	20.91	37.41	35.76	37.63
	SetB +ligase +Rt	SetB +ligase -Rt	SetB -ligase +Rt	SetB -ligase -Rt
2Sf	6.86	28.05	16.47	29.00
4Sf	2.84	16.44	15.04	23.24
5Sf	28.48	36.85	35.27	37.80
7Sf	24.30	38.61	36.03	37.89

Table 2-5: Determining the amount of starting RNA sufficient for adapter ligation qPCR -extracted RNA with and without reverse transcriptase.

+RT	WC	WC 1:10	WC 1:100	Water + Tru3P
SNORD49AF	22.48	25.62	28.69	35.47
SNAI1 4t_20	28.91	31.64	35.25	37.54
HspA1B2F	27.15	30.13	32.98	35.09
+RT	Chromatin	Chromatin 1:10	Chromatin 1:100	Water
SNORD49AF	18.57	21.50	25.79	0.00
Snai 4t_20	25.37	28.51	32.10	0.00
HspA1B2F	23.14	26.45	30.25	37.93

-RT	WC	WC 1:10	WC 1:100	Water + Tru3P
SNORD49AF	0.00	36.15	0.00	0.00
Snai 4t_20	37.13	36.42	36.29	36.74
HspA1B2F	35.03	35.57	35.84	36.24
-RT	Chromatin	Chromatin 1:10	Chromatin 1:100	Water
SNORD49AF	0.00	0.00	0.00	0.00
Snai 4t_20	35.65	37.03	37.24	0.00
HspA1B2F	34.07	35.32	35.48	36.43

I further modified this approach to incorporate a radiolabeled primer. First, I determined the minimum number of amplification cycles that would be required before adding the radiolabeled primer. I found that 16 cycles were sufficient to amplify the RNA before I proceeded to further amplify with 4 cycles of the radiolabeled primer. The product was first run on a polyacrylamide gel to determine the number of amplification cycles needed for detection (Figure 2-7). Subsequent experiments were resolved on a sequencing gel.

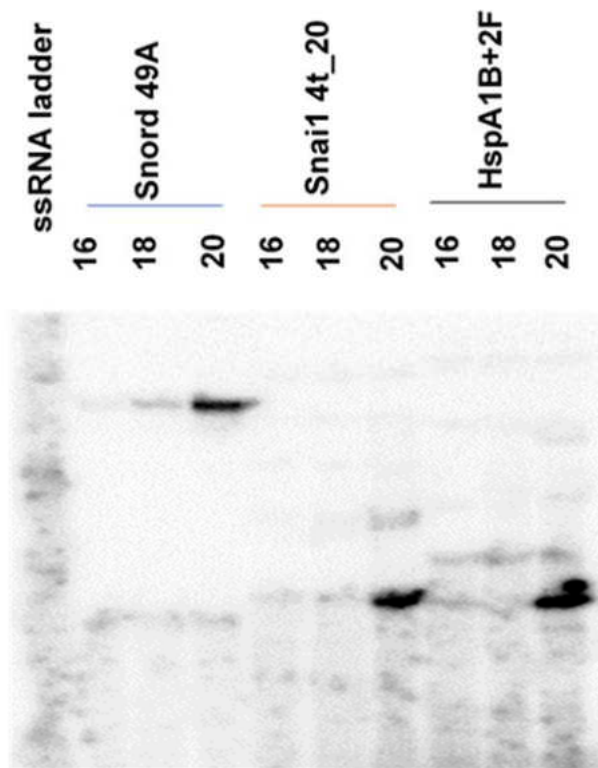


Figure 2-7: Optimization of Amplification cycle for LM radioactive PCR. HspA1B, Snai1 4T\_20, and SNORD49A. PCR products were run on a 15% PAGE gel to determine the optimum amplification cycles for the amplification of short RNA. The expected sizes were SNORD49A: 78. SNAI14T\_20: 51-62. HSPA1B: 50-70

## Testing enzymes to map the 5' End Status of the Observed LM PCR RNA Products

I devised an approach to investigate the modifications associated with the 5' ends of these RNAs. First, I tested the RNA 5' Pyrophospho-hydrolase RppH (NEB), Alkaline Phosphatase, Calf Intestinal (CIP) (NEB), Terminator 5'-Phosphate Dependent Exonuclease (Epicenter) (Terminator), and RNA 5' Polyphosphatase (Epicenter) enzymes (Figure 2-8A) on phosphorylated RNA oligos using enzymatic reactions to check whether the enzymes were working. CIP catalyzes the dephosphorylation of 5' and 3' ends of DNA and RNA. XRN-1 (NEB) is a processive 5'→3' exoribonuclease, requiring 5' monophosphate ends. Terminator degrades RNAs with a 5' monophosphate end. RppH is a decapping enzyme which removes pyrophosphate from the 5' end of triphosphorylated RNA to leave a monophosphate RNA.

I set up different experimental conditions to check the efficacy of these enzymes. I observed that Terminator completely degraded RNA while XRN-1 did not. Next, I tested RNA 5' Polyphosphatase and RppH. I set up different experimental conditions to determine the efficacy of RppH and Polyphosphatase enzyme. I observed that both enzymes could be used for investigating the 5' end modifications of RNA (Figure 2-8B).

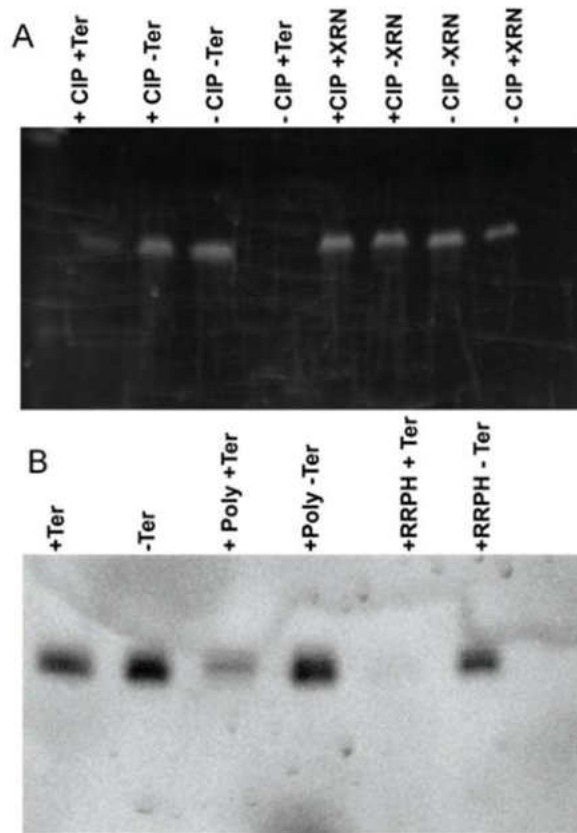


Figure 2-8: Polyacrylamide gel images of enzyme activity. A. Terminator (Ter) enzyme degraded monophosphate RNAs better than XRN1. B. Polyphosphatase enzyme and RPPH enzyme degrade triphosphorylated RNAs as advertised.



## **RNA Preparation**

### **Total RNA isolation**

Total RNAs were isolated from whole cell fractions using the Qiagen miRNeasy kit. The cell pellets were resuspended in 700ul of QIAzol and incubated on the benchtop at room temperature for 5 minutes. The homogenate was mixed with 140ul of chloroform and vigorously mixed for 15 seconds. The homogenate was placed on the benchtop for 2-3 minutes and the RNA isolation was performed according to manufacturer's instruction.

### **qPCR**

Two micrograms of RNA from whole cell aliquots were added to 2.5 uM of Random Hexamer, heated at 70°C for 5 minutes, and reverse transcribed into cDNAs by adding a pre-made mix (0.5x First Strand Synthesis Buffer (Invitrogen), 2.5mM DTT, 0.5mM dNTPS, and SSRT III reverse transcriptase enzyme (Invitrogen)). The reaction was incubated at 50°C for 50 mins and heat inactivated at 85°C. Primers used for qPCR are listed in Table 2-6. GAPDH or Actin primers were used as control housekeeping genes and the analysis.  $\Delta C_t$  and  $2^{-\Delta\Delta C_t}$  are calculated to get the expression fold change.

Table 2-6. List of qPCR primers and their sequences

<b>Primer</b>	<b>Sequence</b>
RT-HSP90AA1 forward	ATG AGC CTG AGG TGA ACT TG
RT-HSP90AA1 reverse	GCT CTT CCT CCG CTC TTT G
RT-HSPH1 forward	GCT TTC TCC CAG GGT TTC TTA
RT-HSPH1 reverse	CAC TCA GAA GGA CAC ACA GAC
hActinF_QiaRT Forward	GCA AGG GAC TTC CTG TAA CAA
hActinR_QiaRT reverse	CAT CCC CCA AAG TTC ACA ATG
hGAPDHF_QiaRT forward	CCC AAT ACG ACC AAA TCC GTT
hGAPDHR_QiaRT reverse	TCT CTG CTC CTC CTG TTC G
hHSP70F_QiaRT forward	GCT GAT GAT GGG GTT ACA CA
hHSP70R_QiaRT reverse	CGA GAA GGA CGA GTT TGA GC
hSnailF_QiaRT forward	AGT GGG GAC AGG AGA AGG G
hSnailR_QiaRT reverse	CAA GAT GCA CAT CCG AAG CC

## Preparation of short-capped RNAs

The major short-coming in the existing short capped RNA detection (Samarakkody et al., 2015) in mammalian systems is the tedious and time consuming phenol chloroform extraction step used for buffer exchanges and extraction of RNA at multiple steps. To overcome this, experiments to test the suitability of columns for extracting small RNA were performed. A small DNA oligo (hCDH1) was labelled by  $\alpha^{32}\text{P}$ -CTP to mimic small RNAs. Three different approaches were used: Zymo Oligo Clean & Concentrator columns kit (Zymo), Qiagen miRNeasy kit columns (Qiagen), and Trizol (Ambion) method of RNA extraction. The  $^{32}\text{P}$  labelled hCDH1 DNA oligo was prepared by adding T4 Polynucleotide Kinase (PNK) (NEB) and ATP (NEB).

Equal aliquots were extracted either by passing the mixture through a Zymo or a Qiagen column and eluted twice. In the case of Trizol extraction, the labelled oligonucleotide was extracted once. The unbound fraction during the process was collected. The same approach was used for  $\alpha^{32}\text{P}$ -CTP alone. Both eluted and unbound fractions were collected, and signals were measured using a scintillation counter. Zymo columns showed more efficient binding when compared to the Trizol extraction and the Qiagen columns (Figure 2-9). Moreover, using Zymo to elute the oligos took 5 minutes, compared to several hours for the phenol chloroform extraction method. Therefore, I adopted the Zymo columns for extracting and concentrating short capped and uncapped RNAs from fractions.

Cdh1 labelled primer with different extraction methods

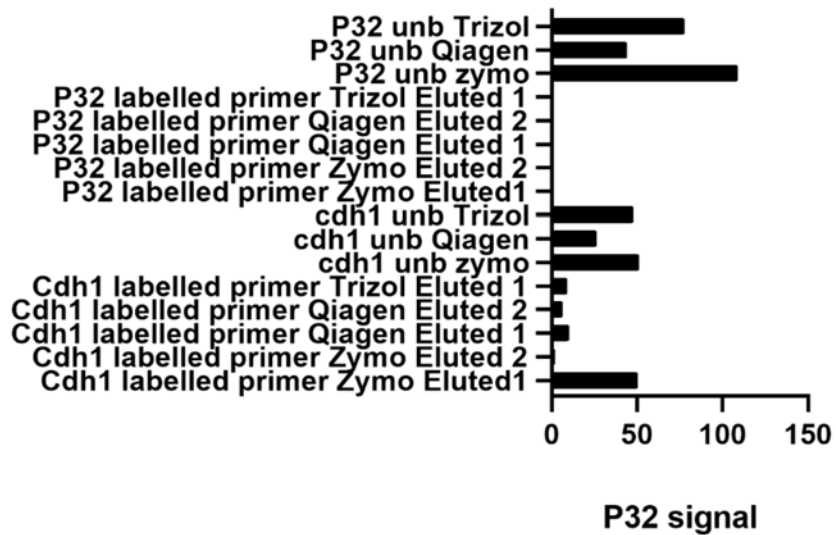


Figure 2-9: Testing different RNA columns from different vendors for extraction of small RNA using a radiolabelled DNA oligonucleotide ( $^{32}\text{P}$  labelled hCDH1). Equal aliquots of the  $^{32}\text{P}$  labelled hCDH1 were extracted using Qiagen columns, Zymo columns and Trizol method of extraction. Two elution steps were performed for the Qiagen and Zymo columns and one extraction step was performed for the Trizol step. The unbound fractions were collected at each step. Signals were measured with a scintillation counter. Zymo columns extracted small oligonucleotides more efficiently than Qiagen columns or Trizol.

## Short Capped RNA library preparation

Short capped RNA libraries were made from RNA purified from chromatin fractions (2 replicates) instead of from nuclei fractions, as published previously (Samarakkody et al., 2015). To address my questions, the extra step of purifying short capped RNAs was necessary to compare these RNAs directly with short uncapped RNAs from chromatin. To generate libraries, chromatin fractions were first treated with 5' polyphosphate dependent polyphosphatase (Epicenter) to reduce the 5' ends of RNAs with tertiary and secondary phosphates to monophosphate RNAs. These were further degraded using the 5' exonuclease, Terminator, to leave only capped RNA in the fraction. The capped RNA was then uncapped using the decapping enzyme RppH. The libraries were made using the NEBNext small RNA kit. I first generated libraries and sequenced on the Miseq to validate the Zymo inclusion in the protocol. Libraries were sequenced to a range of depth of 40 to 70 million reads (Table 2-7).

Table 2-7: Alignment percentages for short capped RNA sequencing

Short capped	Rep 1	Rep 2
Number of raw reads	69304879	44147638
UNIQUE READS:		
Uniquely mapped reads number	15615680	16593788
Uniquely mapped reads %	22.53%	37.59%

## Short uncapped RNA Sequencing

I next developed an approach to characterize short uncapped processed RNA which have 5' monophosphate ends. This enabled me to examine genome-wide RNA processing within the promoter region of genes (Project, 2009; Taft et al., 2009; Valen et al., 2011), by sequencing small 5'-monophosphorylated RNAs prepared from the chromatin, nucleoplasmic, and cytoplasmic fractions of MCF-7 cells.

Total short RNAs (< 200 nucleotides) were extracted using Qiagen miRNeasy kit from the chromatin, nucleoplasmic and cytoplasmic fractions. I included an additional step to concentrate RNAs using the Zymo DNA Clean & Concentrator. These RNAs were then run on a Bioanalyzer to check the concentration, and libraries were made using a NEBNext small RNA kit. To select for uncapped RNAs, I ligated adapters to the 5' and 3' ends of my total short RNA (capped and uncapped RNA). Since these RNAs have 5' monophosphate ends and 3' hydroxyl ends, there was no need for any enzymatic reactions to deplete other RNAs with different 5' ends present in my samples.

The adapters were diluted 1:1 with water and the final PCR was performed for 16 cycles. The libraries were purified from a 6% polyacrylamide gel and then validated on a Bioanalyzer 2100 and quantified by the Biorad QX200 Droplet Digital PCR (ddPCR) system prior to single end sequencing on the Illumina HiSeq 4000 (Novogene) (Figure 2-10). Libraries were sequenced to a range of depth of 40 to 80 million reads (Table 2-8).

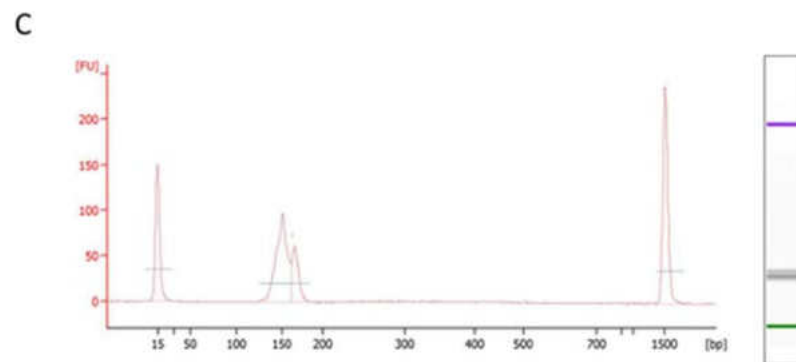
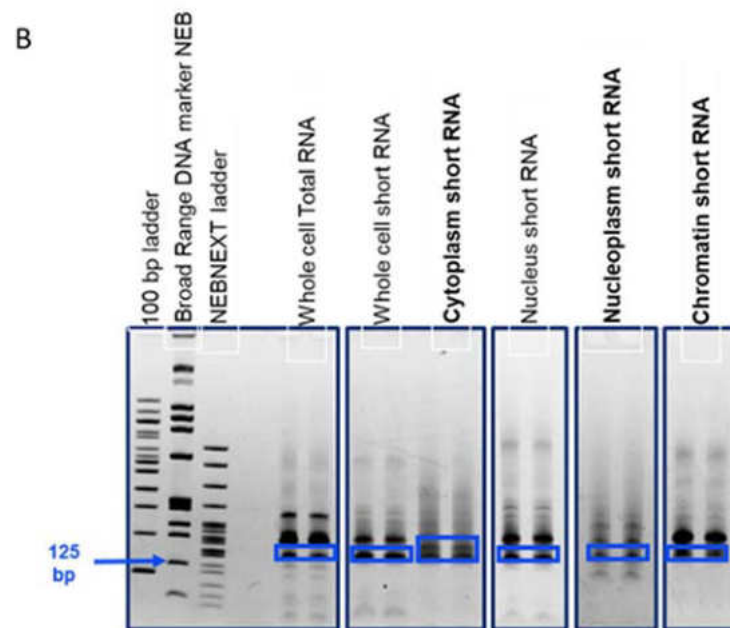
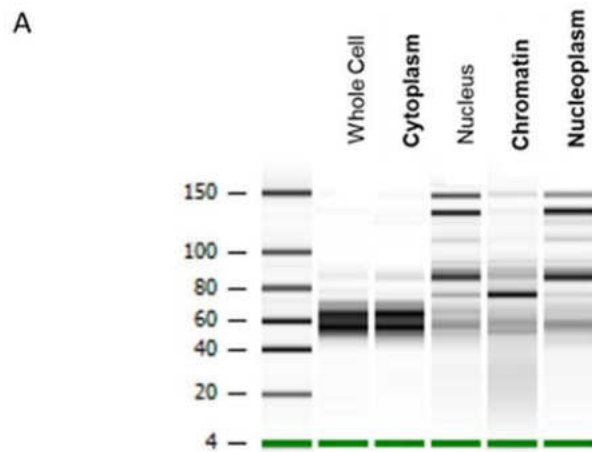


Figure 2-10: Systematic approach for studying short RNA distributions. A. Bioanalyzer validation of short RNA <200 nt from the various cell fraction. The bands show size distribution of the different RNAs in each fraction. Short RNAs in cytoplasmic, nucleoplasmic and chromatin fractions isolated look different. B 6% Polyacrylamide gel of short RNA libraries of different fractions. Bands show different library products amplified. Adaptor to adaptor products are found at about 125bp C. An example of a bioanalyzer image of chromatin library excised from the gel (B) and purified. The library size ranges from about 140nt to 180 nucleotides.



Table 2-8. Alignment statistics for short uncapped RNA

Small uncapped processed	Rep 1	Rep2	Rep 3	Rep 4
Number of raw reads	83749227	64404688	57480821	59538825
UNIQUE READS:				
Uniquely mapped reads number	41191547	31267317	20936275	15136387
Uniquely mapped reads %	49.18%	48.55%	36.42%	25.42%

### Precision Run On sequencing (PRO-Seq)

Nuclei were prepared from one 15cm dish and resuspended in 100 ul of Freezing buffer (50mM Tris-Cl pH 8.3, 40% glycerol, 5mM MgCl<sub>2</sub>, 0.1mM EDTA). The nuclei were treated with biotin-11-NTPs (Perkin Elmer) (Mahat et al., 2016). All four biotinylated NTPs were used in each run-on reaction. Run-on reactions and library preparations was done exactly according to (Mahat et al., 2016). I modified the approach to select for specific 5' end modifications of RNA (see Table 2-9). Libraries were validated on the Bioanalyzer 2100 and quantified by the Biorad QX200 Droplet Digital PCR (ddPCR) system prior to sequencing.

For naturally occurring processed RNAs with 5' monophosphate ends (termed processed PRO-Seq), I carried out the run-on reaction without any enzymatic reaction or fragmentation. To select for RNAs with capped and triphosphate 5' modifications (termed PRO-Start RNAs in this study), I treated the RNA with Terminator to degrade RNA with monophosphate ends and RNA 5' pyrophosphohydrolase to reduce both the capped RNA and triphosphate RNA to monophosphate RNA before proceeding to library

preparation. I also looked at all possible uncapped conditions (triphosphorylated, monophosphorylated, and degraded) RNA by using CIP to remove the phosphate group at the 5' end. I re-phosphorylated by using T4 PNK (Table 2-9). I selected RNAs with only triphosphate ends (PPP PRO-Seq RNAs) as a control. To select for PPP PRO-Seq RNAs, I degraded RNAs with monophosphate ends by using Terminator I further reduced the triphosphate ends with RNA 5' polyphosphatase before proceeding to library preparation.

Table 2-9. Modified PRO seq Approach

	Step 1	Step 2	Captures	Fragmentation	Replicates
PO4 (processed PRO-Seq)			Monophosphate RNA (naturally occurring)	No	2
PRO-Start	Terminator	RppH	7meGpppRNA and triphosphate RNAs	No	2
CIP_PRO-seq	CIP	PNK	Triphosphate RNAs, monophosphate RNAs RNA, degraded RNAs	No	1
PPP PRO-Seq (Triphosphate)	Terminator		Triphosphate RNAs	No	2
PRO-Seq	RppH	PNK	fragmented monophosphate RNAs	Yes	2

### Sequencing and Bioinformatics

Sequencing was done either on a MiSeq instrument (UND) or Novogene. Paired end (50 bp) runs were done on the MiSeq and single end (50 bp) runs were done on the HiSeq (Novogene) using a small RNA sequencing option (on a HiSeq instrument).

## Bioinformatics

Data were analysed using the following software and command lines.

- 1) Quality control analysis was done using an open source software tool FastQC using the following command line:

```
fastqc <filename.gz>
```

where *filename.gz* is the name of the data file.

- 2) Adapters were removed using the Cutadapt tool with the following command lines:

```
cutadapt -a <adapter sequence> -o output.fastq input.fastq.
```

where *adapter sequence* is the adapter sequence to be removed

*output.fastq* is the name of the output file

*input.fastq* is the name of the input file.

- 3) The data was aligned to the reference genome using the STAR aligner (Dobin et al., 2013) to provide a map of transcription across the genome using the following command lines:

```
STAR --runThreadN 10 --genomeDir/ --outFileNamePrefix File.Name. --
```

```
outFilterMismatchNoverLmax 0.05 --outFilterMatchNmin 10 --
```

```
outFilterScoreMinOverLread 0 --outFilterMatchNminOverLread 0 --alignIntronMax
```

```
1 --readFilesCommand zcat --readFilesIn File.Name.fq.gz;
```

The following options were used:

--runThreadN: the number of cores

--genomeDir: path to the folder containing STAR index

--outFileNamePrefix: File name

--outFilterMismatchNoverLmax: number of mismatches  $\leq 5\%$  of mapped length (for example, 0 mismatch is allowed for 16-19b, 1 mismatch for 20-39b, etc.)

--outFilterMatchNmin 10: reads  $\geq 10$ b matched to the genome

--outFilterScoreMinOverLread 0 --outFilterMatchNminOverLread 0 --alignIntronMax 1: switching off splicing.

--readFilesCommand zcat --readFilesIn File.fq.gz: to use gz compressed files as input

The STAR aligner trim reads at the 5' ends for quality control purposes; however, that removes important information for my experiment. Hence, I filtered out reads that STAR trimmed. To do so, the following awk command was run on the output SAM files:

```
awk '{S=0; split($6,C,/[[0-9]]*/); n=split($6,L,/[[NMSID]]/); if (and($2,0x10)>0 && C[n]=="S") {S=L[n-1]} else if (and($2,0x10)==0 && C[2]=="S") {S=L[1]}; if (S<=1) print }' Aligned.out.sam > Aligned.filtered.sam
```

I filtered out reads  $>48$  bp using samtools as a quality measure since the sequencing length was 50 bp. The bam files were intersected with a repeat list to remove repeats using bedtools.

The data was analysed using bedtools and samtools to generate 5' ends, 3' ends, and the length distribution. I used an in-house gene curated list as reference point to

obtain the 5' ends, 3' ends, and lengths of reads located between 100 nucleotides upstream and downstream of the TSS.

### **Normalization**

For gene expression, I obtained reads that were located between 200 nucleotides downstream of the TSS and 1000 nucleotides upstream of the TES. I divided this number by the length of the gene body for each gene.

The figures were drawn using PRISM.

### **Gene List and Curation**

Gene definitions were downloaded from UCSC. Genes with duplicate names with the same TSS were consolidated by selecting the longest isoform. Gene isoforms with different TSSs (beyond 100 nt difference) were resolved by considering the TSS with the highest Pol II ChIP signal in MCF-7 cells. Using a list of protein coding genes downloaded from HGNC, only protein coding genes were selected. The TSS were reannotated using 5' Start-seq RNA positions. The nucleotide with the highest number of “sense” reads within +/- 500 bps of the TSS was identified as the reannotated TSS. Genes with fewer than 5 Start-seq reads were discarded. The resulting list contained 10,114 entries.

## CHAPTER 3

### RESULTS: DEVELOPMENT OF PROTOCOL TO INVESTIGATE PREMATURE TERMINATION DURING POL II PAUSING

#### Introduction

Premature termination has been proposed to occur within the promoters of genes for decades. How this occurs is not known. In my initial studies, I hypothesized that the full-length transcripts from the paused Pol II (Figure 3-1A) is terminated (Figure 3-1B) into the nucleoplasm. Most approaches used to study premature termination have focused on the suggested proteins involved in regulating the termination (Brannan et al., 2012; Nojima et al., 2015). A few studies have focused on the properties of the RNAs (Project, 2009; Taft et al., 2009; Valen et al., 2011) within the Pol II complex during Pol II pausing and their functional relevance by focusing on RNAs in nuclear or whole cell lysates.

I was particularly interested in characterizing the various RNAs associated with Pol II at the promoter region in different fractions (chromatin, nucleoplasm, and cytoplasm) in order to delineate the premature termination pathway. In order to select for these RNAs, I had to perform several fractionation steps to first isolate nuclei and then further separate the nuclei into the chromatin and nucleoplasm fractions. Cell fractionation has been previously used to study gene regulation and splicing (A. & D.L., 2009; Bhatt et al., 2012). In this instance, the method allowed me to not only focus on the chromatin associated RNAs, but also to identify and further study the various species of RNAs within each specific fraction (see Methods).

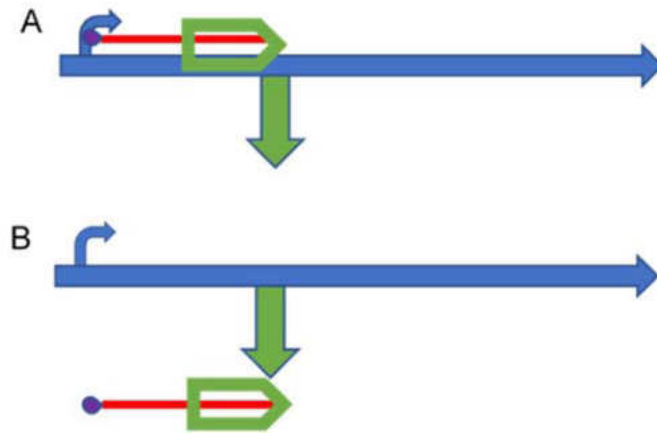


Figure 3-1. Model of premature termination around the promoter region of genes. A. Paused RNA on a gene is generated when Pol II pauses within 100 nucleotides downstream of the TSS. B. During Pol II pausing, paused Pol II can undergo premature termination to release paused RNA. In this model, the released RNA transcript length corresponds to the distance between the TSS and the paused Pol II.

### **LM PCR detects short RNA Products at the promoter region**

To test my hypothesis that during Pol II pausing, Pol II complexes are dissociated at the promoter region to release a short RNA, I applied the radioactive LM PCR method I developed to RNA extracted from whole cell, cytoplasm, nucleoplasm, and chromatin fractions. To determine the size of the products of amplification, I made ladders that could be detected via radioactivity. I end labelled DNA primers of known sizes using PNK (NEB). These markers were run alongside the LM PCR products.

My control gene (SNORD49A) showed an approximate length of 67 nucleotides (Figure 3-2). This is based on the product length obtained after cloning and sequencing (Figure 2-6B). The difference between the position of amplification product to the ladder could be attributed to the properties of the DNA being compared. The ladder was single stranded while the amplified products were double stranded. Single stranded DNA travels faster through the gel than double stranded DNA. Despite the molecular weight mismatch, the results indicated in principle that this approach could be used to study low abundance and short RNAs (Figure 3-2).

SNAI1 and HSPA1B showed products that were not easily interpretable due to the distribution of fragment sizes (Figure 3-3 and 3-4, respectively) in the various fractions. The SNAI1 products showed enrichment in the chromatin fractions for RNAs which were 36 nucleotides in length. This is similar to the distance of SNAI1-Pol II complexes from the TSS, as seen in permanganate footprinting studies, as well as the size of short-capped RNA, as seen in short-capped RNA sequencing (Samarakkody et al., 2015). However, there were smaller sizes also observed in the nucleoplasmic fractions. Similarly, for the HSPA1B gene, products 28 nucleotides in length were observed in the



whole cell, nuclei, cytoplasm, and chromatin fractions, but smaller products at 21 and 23 nucleotides in length appeared to be enriched in the nucleoplasm. Based on these observations, I envisaged that the shorter products may have been generated through further processing of RNA associated with paused Pol II and leading to premature termination. For some of these shorter RNAs, there may be different 5' end modifications depending on the mechanism involved (e.g. capping versus 5' monophosphate).

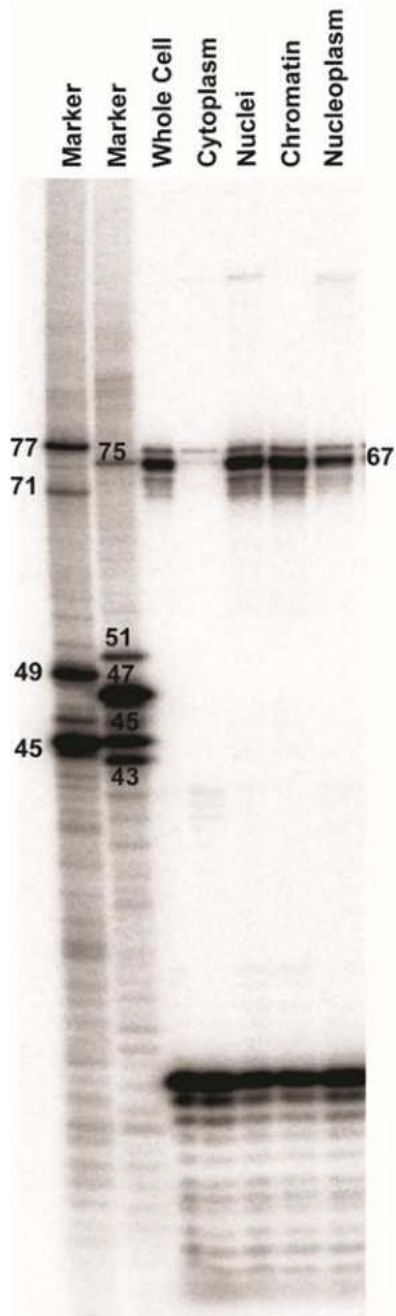


Figure 3-2. Radioactive LM PCR of SNORD49A in whole cell, cytoplasm, nuclei, chromatin and nucleoplasm fractions. The fractions were ligated at the 3' adapter, amplified, and subjected to further amplification with a labelled primer. This length was enriched in whole cell, nuclei, chromatin, and nucleoplasm fractions. The length of the product was calculated using the formula  $(a-b)=c$ , a= The length of RNA from start site to paused site, b=The start site of the primer, c=The length of the adapter (e.g. SNORD49A:  $(71-14)+21=78$ ).

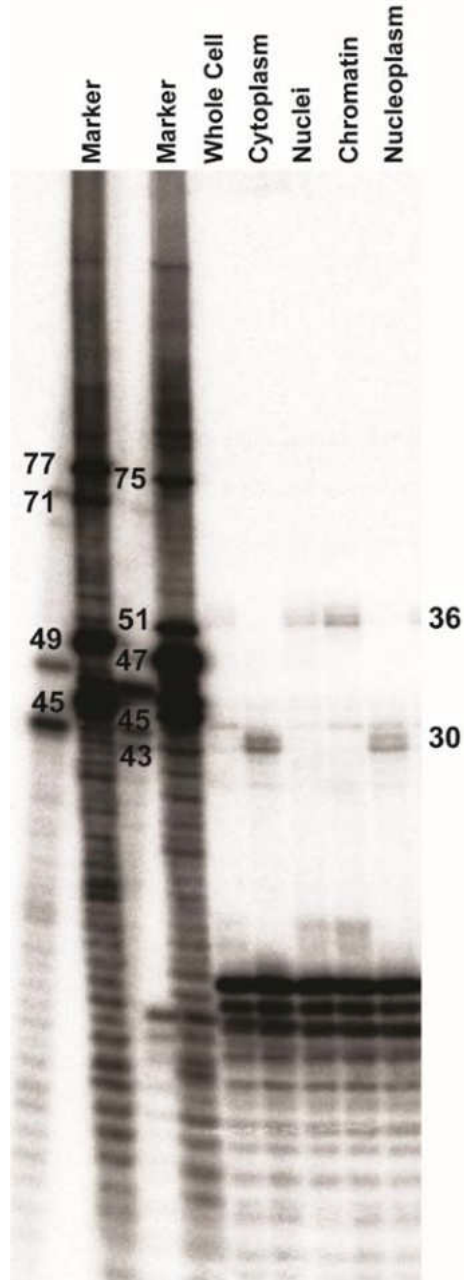


Figure 3-3. Radioactive LM radioactive PCR of SNAI1 in whole cell, cytoplasm, nuclear, chromatin, and nucleoplasm fractions. The fractions were ligated at the 3' adapter, amplified, and subjected to a second amplification with a labelled primer. The fragments which were 36 nucleotides were observed in the whole cell, nuclear, and chromatin fractions. The cytoplasm and nucleoplasm were enriched in products 30 nucleotides in length. The length of the product was calculated using the formula  $(a-b)=c$ , a=The length of RNA from start site to paused site, b=The start site of the primer, c=The length of the adapter (e.g. SNAI1 4T\_20:  $((39 \text{ to } 50) - 9) + 21 = (51 \text{ to } 62)$ ).

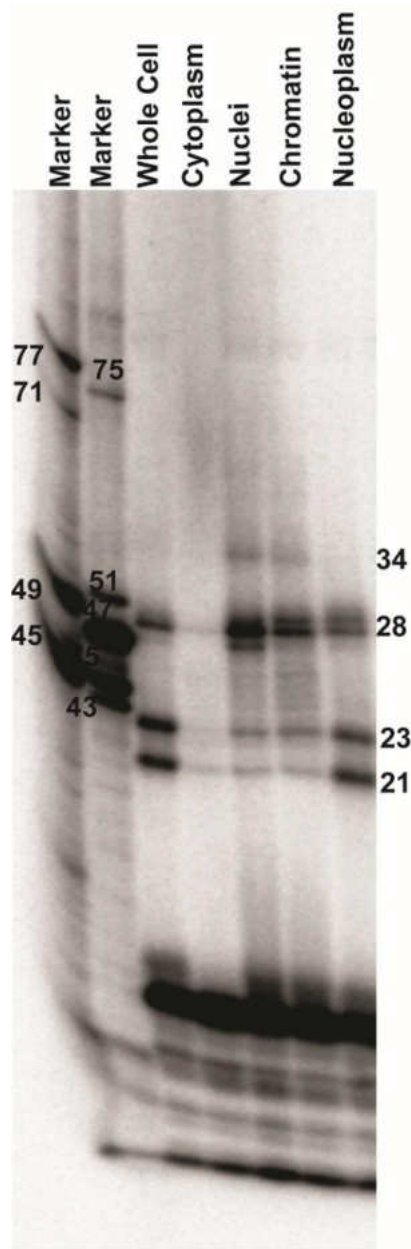


Figure 3-4. LM PCR of HSPA1B in whole cell, cytoplasm, nuclei, chromatin, and nucleoplasm fractions. The fractions were ligated at the 3' adapter, amplified, and subjected to a second amplification with a labelled primer. The observed size amplified were 21, 23, 28, and 34 nucleotides. The length of 34 nucleotides was enriched in chromatin and nuclei fractions. The lengths of 21, 23, and 28 were observed in all fractions tested except for the cytoplasm. The length of the product was calculated using the formula  $(a-b)=c$ ,  $a$ =The length of RNA from start site to paused site,  $b$ =The start site of the primer,  $c$ = The length of the adapter (e.g. HSPA1B 70:  $((31 \text{ to } 51)-2)+21= (50 \text{ to } 70)$ ).

### Mapping the 5' End Status of the Observed LM PCR RNA Products

To differentiate the various 5' end modifications (capped, triphosphate, diphosphate, and monophosphate), I devised an approach by using enzymatic reactions to select for RNAs with capped, triphosphorylated, or monophosphorylated ends (Table 3-1).

Table 3-1: Approach used to delineate 5' end modification using radioactive LM PCR

RppH	-	-	-	+
RNA 5' Polyphosphatase	-	-	+	-
Terminator	-	+	+	+
RNAs degraded		5'P	5P 5'PP 5'PPP	5P 5'PP 5'PPP Capped products

I observed that when RNAs from the nuclei were probed for SNORD49A, the signal was decreased in Terminator treatment and decapping+ Terminator treatment. However, there was no change in polyphosphatase and Terminator exonuclease treatment, which confounded the results. This is because I observed a decrease in exonuclease alone; hence, I expected to also see a decrease in the polyphosphatase and exonuclease condition, irrespective of the activity of the polyphosphatase enzyme in the presence of triphosphate RNAs (Figure 3-5). However, the signal intensity of the chromatin associated RNAs in Figure 3-6 do not decrease in the exonuclease reaction sample, which suggested that monophosphate RNAs may be present in the nucleoplasm.

This may also explain the decrease in signal for SNORD49A when probed in the nucleus. Polyphosphatase and decapping conditions showed a decrease in signal, suggesting that these SNORD49A RNAs possess capped or triphosphate ends on chromatin, as expected (Figure 3-6).

I observed several distinct species of HSP RNA (Figure 3-5 B, C, D) in the nuclei. Observations suggested that most of these RNAs were capped (Figure 3-5 B, C) because these RNAs were susceptible to the decapping enzyme. However, the digestion pattern gave an incomplete picture about the nature of the other types of RNA. For example, although the small RNA species lane 7 was expected to be susceptible to both decapping and polyphosphatase enzyme treatment, it was sensitive to the polyphosphatase enzyme treatment only. Tests using two different preparations *in vitro* prepared RNA - one that was dephosphorylated RNA and a second that was capped – confirmed that the enzyme treatments degraded RNA in the expected manner. As seen in Figure 3-7, the unphosphorylated RNA showed no changes in size or signal intensity when treated with the various enzymes. This was expected because the Terminator requires a monophosphate end for activity. The capped RNA showed a change in signal with decapping enzyme and exonuclease as expected because these *in vitro* transcribed RNAs were capped *in vitro*.

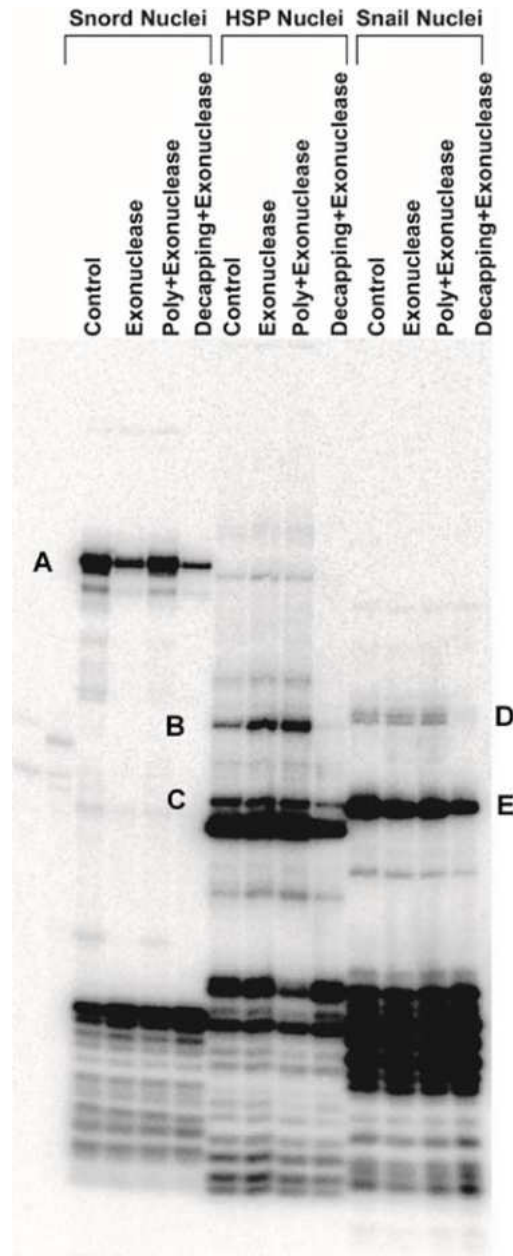


Figure 3-5: Mapping RNA ends in radioactive LM PCR reactions. SNORD49A, HSPA1B, and SNAI1 genes were tested in nuclei fractions. SNORD49A in nuclei showed susceptibility to decapping enzymes, which suggested they were capped, as well as susceptibility to exonuclease activity, which suggested they had monophosphate ends (A). HSPA1B showed some species of RNA that were capped (B&C) and SNAI1 showed some species that were capped (D).

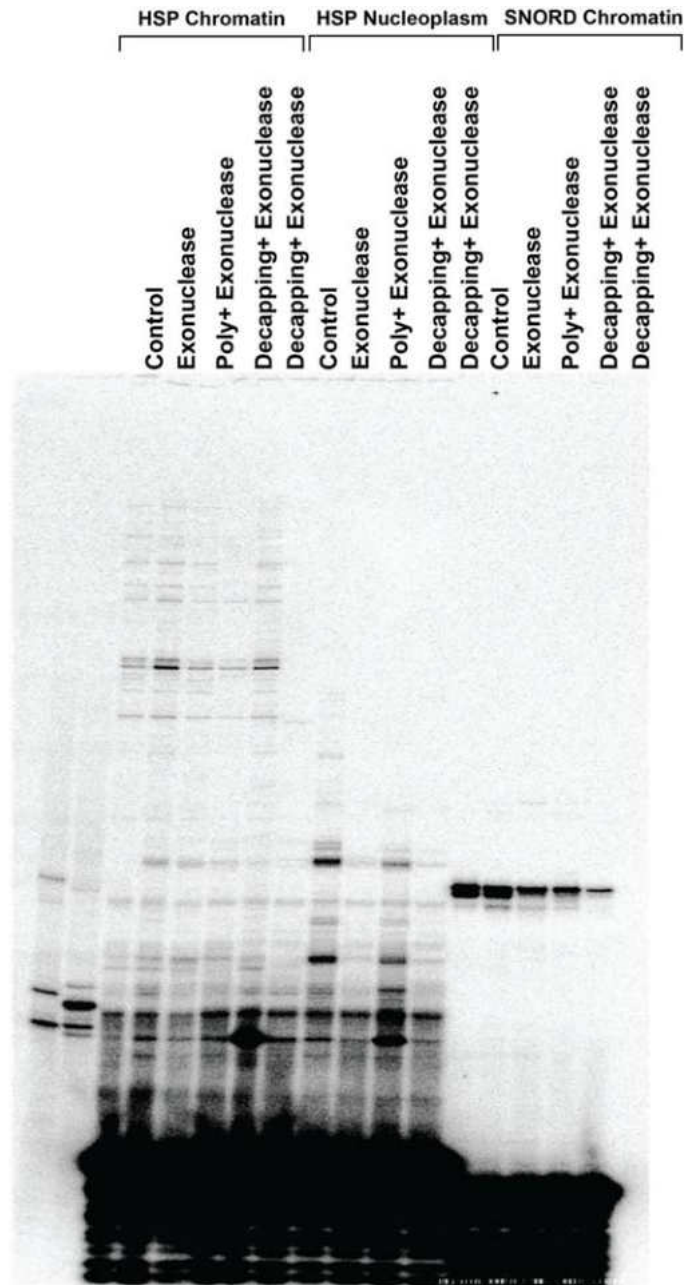


Figure 3-6: Mapping RNA ends in radioactive LM PCR reactions. SNORD49A in chromatin fractions, HSPA1B in chromatin and nucleoplasm fractions were tested. SNORD49A showed susceptibility to polyphosphatase and decapping enzymes on chromatin. Initial transcribed RNA has a triphosphate 5' end and capped structure.



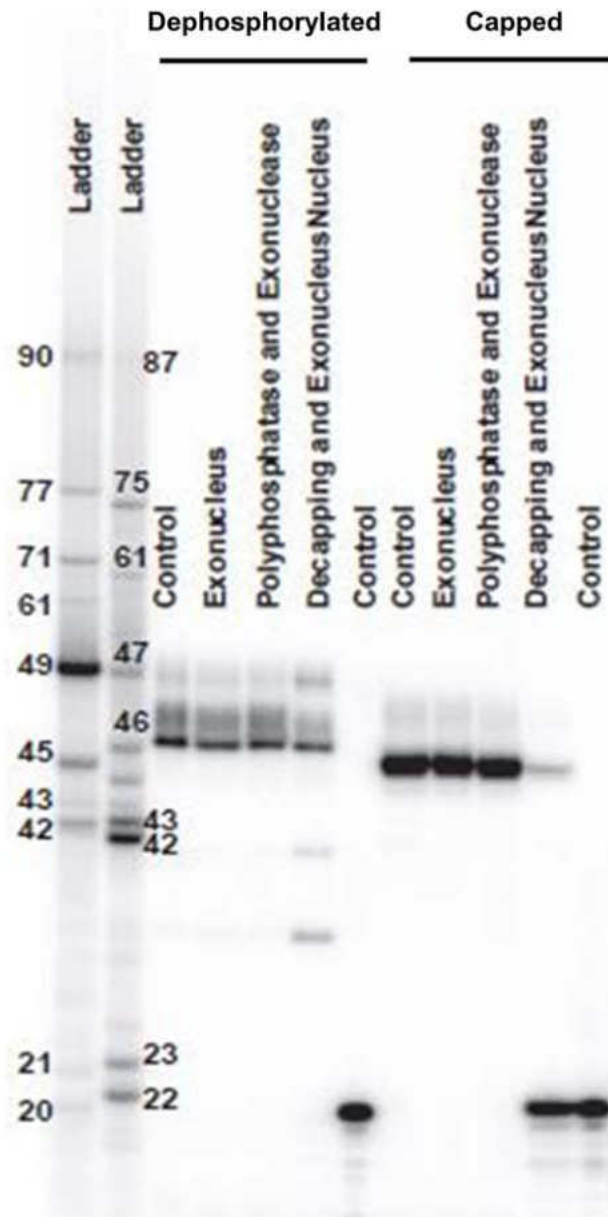


Figure 3-7. Radioactive LM PCR detection of dephosphorylated and capped RNA. Dephosphorylated RNA was not susceptible to enzymatic treatments because exonuclease required a monophosphate end for activity, however Capped RNA was susceptible to decapping enzyme.

Based on the above results, I concluded that Radioactive LM PCR may need further modification to distinguish the 5' end modification. Additionally, RNAs associated with the promoter region may not only be RNAs generated from paused Pol II, but also may be RNAs generated from other processes such as:

1. Early transcribing Pol II prior to reaching paused region (Figure 3-8 A)
2. RNAs that have been generated as a result of processing from the 5' end of the RNA (Valen et al., 2011) (Figure 3-8 B and C)
3. RNAs that have been internally cleaved between the transcription start site and the paused region (Wagschal et al., 2012). (Figure 3-8 D)

These different permutations are not easily distinguishable using the LM PCR approach. Thus, there was a need to design alternative experiments to study these phenomena. I decided to approach the problem on a genome-wide level by sequencing these short RNAs and then using PRO-Seq with special modifications to detect the different 5' end modifications associated with small RNAs at the promoter region.

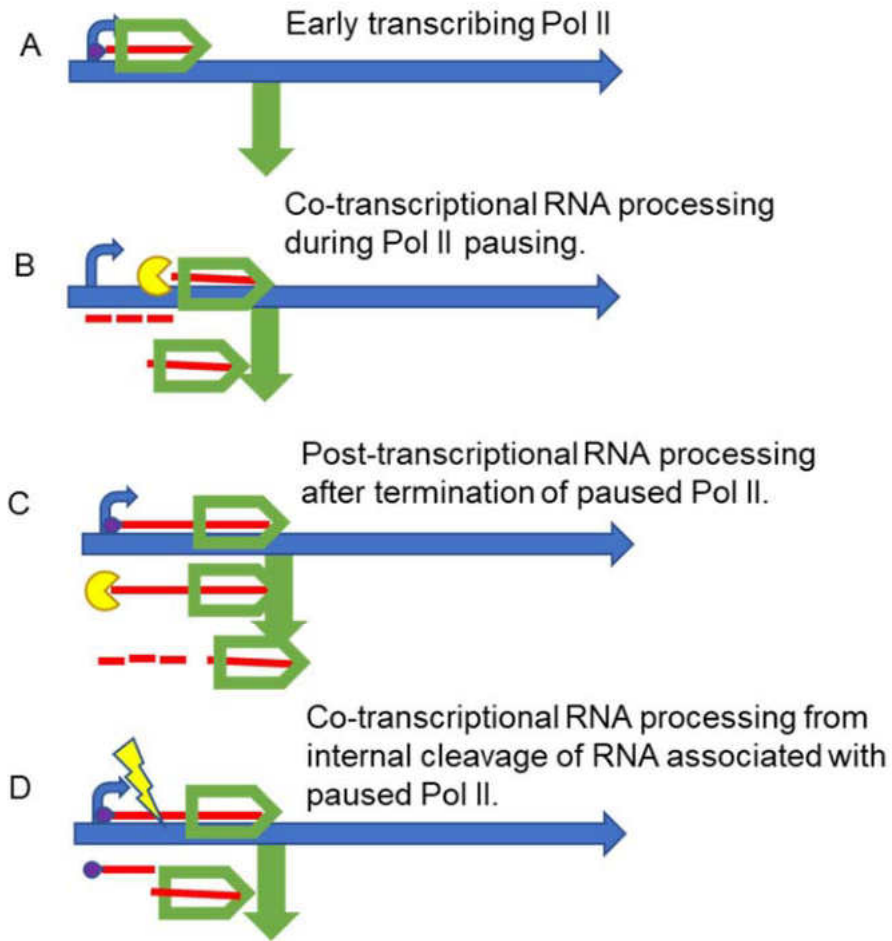


Figure 3-8 Different scenarios that can lead to shorter RNA fragments around the TSS of genes. A. Early transcription complex generates short RNA. B. Co-transcriptional processing of RNA from the 5' to 3' end during Pol II pausing. C. Post transcriptional processing of 5' ends after premature termination of paused Pol II. D. Internal cleavage during Pol II pausing to generate two products. RNA associated with Pol II can be dissociated from chromatin into the nucleoplasm.

## CHAPTER 4

### **RESULTS: PROCESSED RNAS ARE THE INTERMEDIATE PRODUCT BETWEEN PAUSED RNAS AND PREMATURE TERMINATION PRODUCTS**

#### **Genome-wide approach to characterizing RNAs at the 5' ends of genes**

I recognized that the observed RNAs in the radioactive LM PCR include more than just full-length transcripts from Pol II paused complexes. There were shorter fragments which could have been generated from many different processes. RNAs at the promoter were either capped or uncapped (Figure 4-1). Uncapped RNAs (Figure 4-1), which I subsequently refer to as processed RNA, have been observed at the promoter region, and were proposed to have monophosphate ends and to be generated from paused Pol II (Taft et al., 2010; Valen et al., 2011).

RNAs with monophosphate ends have also been proposed to be generated through 5' to 3' exonuclease activity or internal cleavage (Figure 3-8B,D). I found that these uncapped processed RNAs were not degraded products because degradation by RNAses and chemical agents leads to 5' OH formation rather than monophosphate ends, and these degraded RNAs cannot be ligated with adapters for library preparation unless the 5' and 3' ends are repaired. Whether these RNAs are generated co-transcriptionally or post-transcriptionally, however, has yet to be determined.

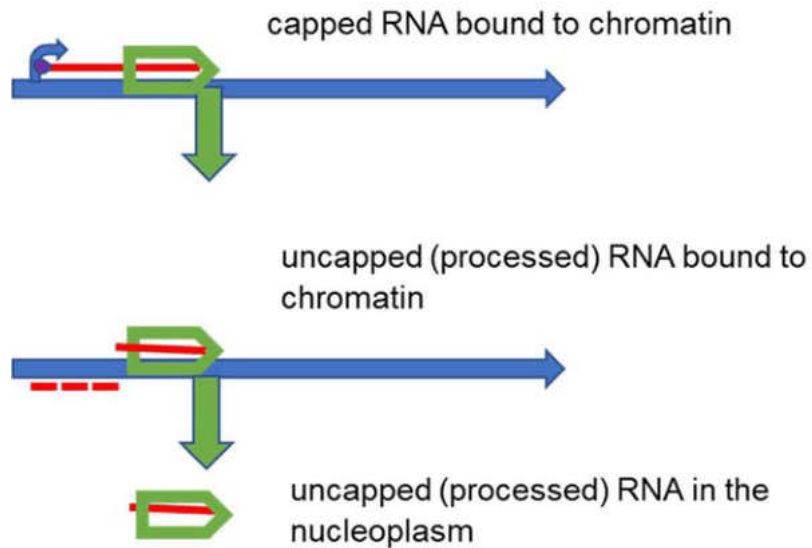


Figure 4-1: Different types of RNAs (capped and uncapped) around the promoter region of genes. Short capped RNAs are generated during Pol II pausing and may form a substrate for processing. Uncapped processed RNAs can be generated co-transcriptionally (processing occurs on chromatin). Alternatively, uncapped processed RNAs can be generated post transcriptionally in the nucleoplasm. Both scenarios may involve decapping and 5' to 3' exonuclease activity to generate RNAs with 5' monophosphate ends or may result from internal cleavage of the paused capped RNAs.

To address the biogenesis of these uncapped processed RNAs, I adopted an approach in which small capped and uncapped processed RNAs linked to the promoter were extracted from different cell fractions (Figure 1-5) followed by high resolution sequencing of 5' and 3' ends of these RNAs in MCF7 cells. This approach enabled me to investigate the relationship between pausing and processing by analyzing paused capped RNAs and small uncapped RNAs linked to the promoter on chromatin. It also allowed me to determine whether the biogenesis of these uncapped processed RNA was co-transcriptional or post-transcriptional. The rationale was that if these RNAs linked to the promoter were generated from paused Pol II, the 3' ends of both paused capped RNAs and small uncapped RNAs would be the same. After comparing 5'- and 3'- end positions of processed RNAs with those that are “extendable” using global run-on assay, I demonstrated that RNA processing is co-transcriptional in MCF7 cells.

### **Characteristics of short capped RNA**

To visualize the location of the 5' and 3' ends of the short capped RNAs, I aligned the sequenced reads using the STAR aligner and looked at the distribution of short RNAs sequenced in regions between 100 nucleotides upstream and downstream of the transcription start site (TSS) using my in-house gene curated list (refer to Methods). The 5' ends signify the start sites, and the 3' ends signify the location of the paused Pol II. The length of the of RNAs range from 14 nucleotides to 47 nucleotides. The most abundant length was 31 nucleotides (Figure 4-2).

I also compared the short capped RNA sequenced from the chromatin to short capped RNA sequenced from the nuclear fraction, which had been previously published

(Samarakkody et al., 2015), and re-analyzed it using my parameters. The two data sets showed similar length distribution (Figure 4-3) and were strongly correlated with a Spearman correlation of 0.92 (Figure 4-4). The 5' ends start at the TSS (0) and the 3' ends of the RNA range from about 20 to 50 nucleotides (Figure 4-5A,B, Figure 4-6). The SNAIL gene, an example of a gene in which pausing occurs (as established by Samarakkody et al., (2015)), showed similarities between short capped RNA from the chromatin and nucleus with regards to the 5' and 3' ends on the UCSC browser shot (Figure 4-7). These findings are noteworthy because most of the 5' ends occurred at the TSS, which confirmed that the isolation of capped RNA was successful.

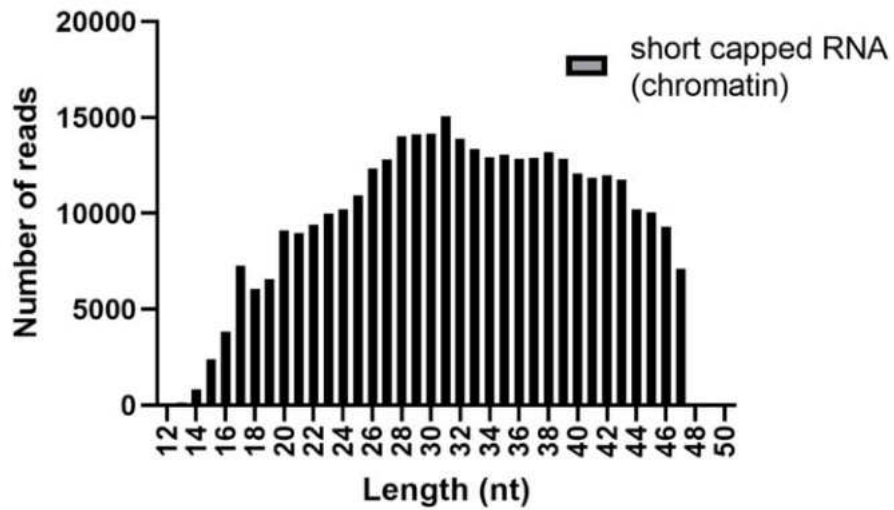


Figure 4-2: Short capped RNA size distribution of reads (1 of 2 replicates). The reads were obtained from the region 100 nucleotides upstream and downstream of the TSS. The RNAs range from 14 nucleotides to 47 nucleotides in length. The most abundant RNA length was 31 nucleotides.



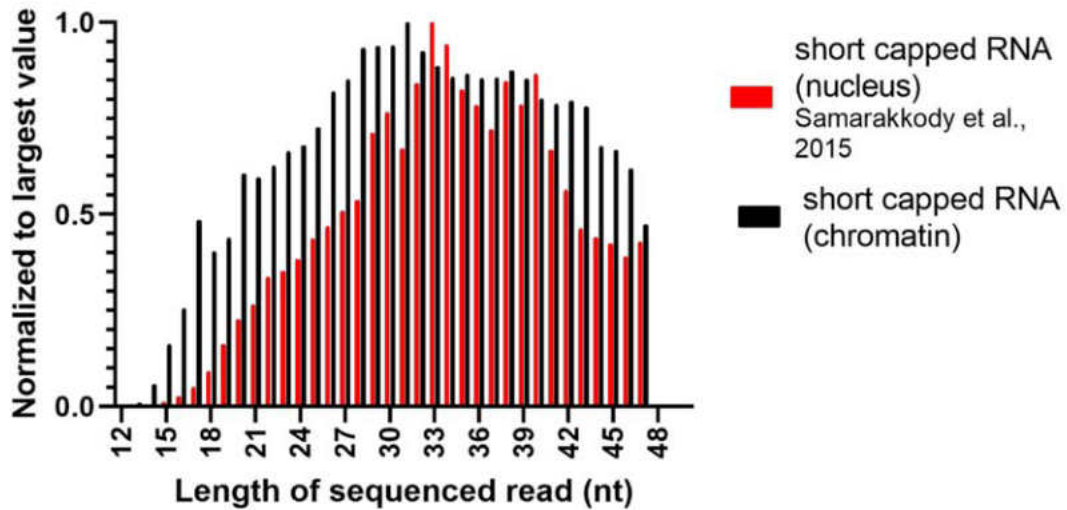


Figure 4-3: Comparison in length distribution between short capped RNAs from chromatin and short capped RNAs published from nuclei (Samarakkody et al., 2015) (1 of 2 replicates each). The fragments fall within 100 nucleotides upstream and downstream of the TSS of genes. The RNAs range from size 14 nucleotides to size 47 nucleotides. The lengths were normalized internally to the most abundant length. The most abundant length for short capped RNAs from chromatin was 31 nucleotides and for short capped RNAs from the nuclei was 33 nucleotides. Each data point was normalized to its largest value.

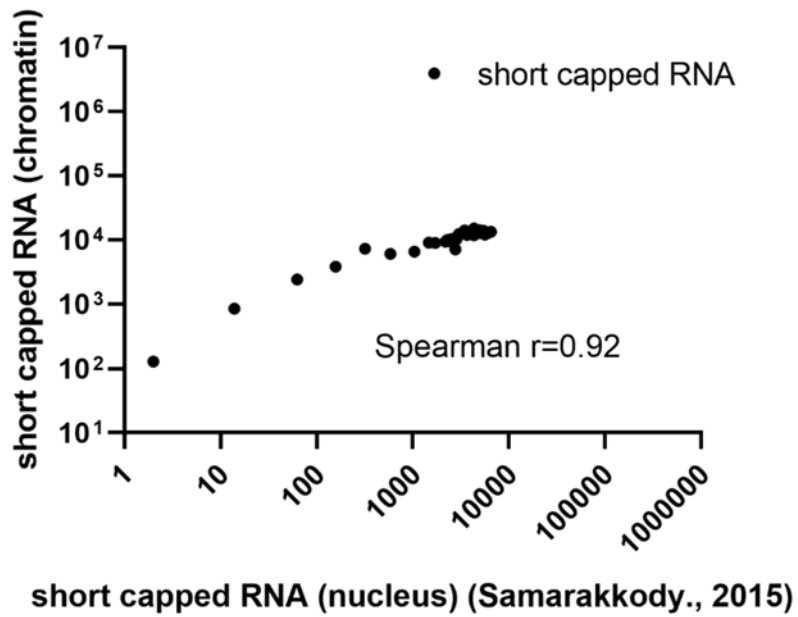


Figure 4-4: Correlation plot for promoter-proximal counts between short capped RNAs from chromatin and short capped RNAs from nuclei in 10,114 genes. Short capped RNAs from chromatin and short capped RNAs from nucleus showed a strong correlation (Spearman) of 0.92.

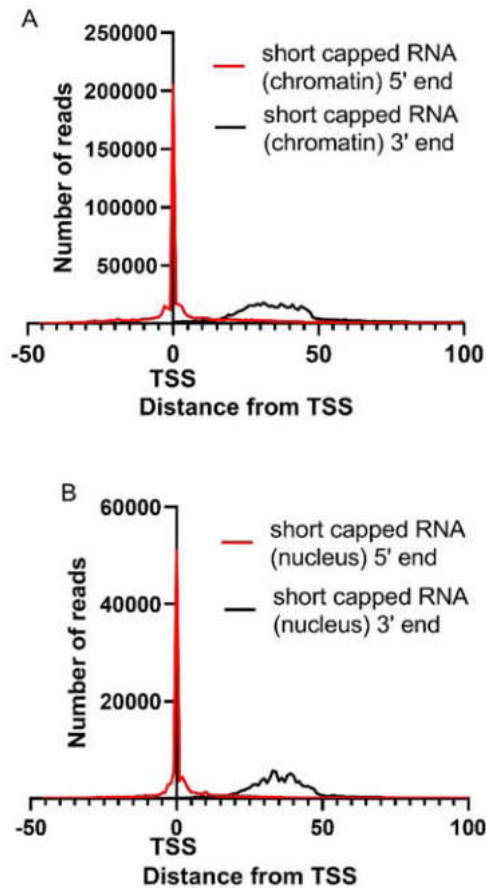


Figure 4-5: Short capped RNAs 5' and 3' positions within 50 nucleotides upstream and 100 nucleotides downstream from the TSS (1 of 2 replicates). A. Short capped RNAs from chromatin fractions. B. Short capped RNAs from nuclei fractions (Samarakkody et al., 2015). Both short capped RNA from chromatin and nuclei show 5' ends at 0 (TSS) and 3' ends range from 20 to 50 nucleotides. The 3' ends signify the paused region.

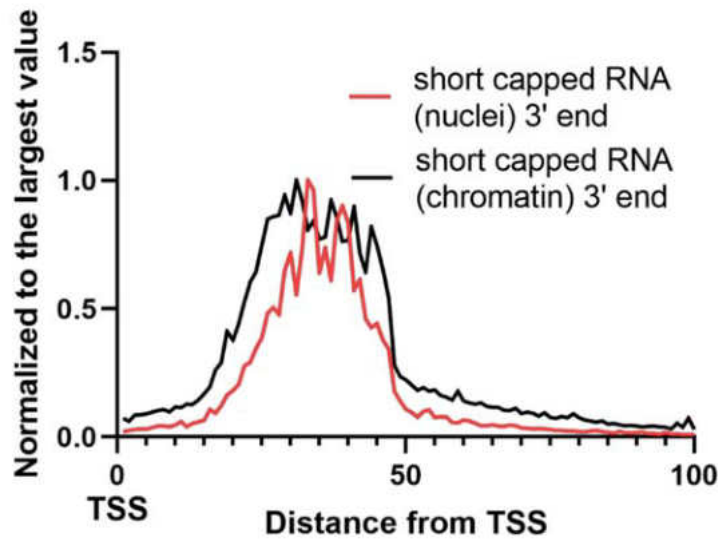


Figure 4-6: Short capped RNAs 3' positions within the region 50 nucleotides upstream and 100 nucleotides downstream of the TSS (1 of 2 replicates shown). Short capped RNAs (chromatin) 3' ends are situated in the same location as short capped RNAs (nuclei) 3' ends (Samarakkody et al., 2015). Each profile was normalized to its largest value. This shows similarity between the two data sets.

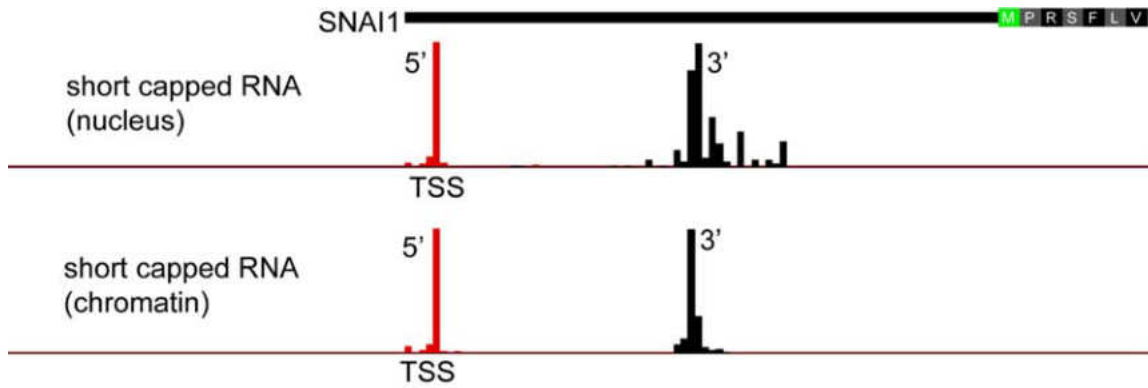


Figure 4-7: A UCSC genome browser shot of the SNAI1 gene illustrating the 5' and 3' end positions of short capped RNAs from chromatin and published short capped RNA from nuclei (Samarakkody et al., 2015). The start site (5' end) and paused site (3' end) relative to the RefSeq (NCBI Reference Sequence Database) annotated start site is shown. Both data sets exhibit similar location patterns of 5' and 3' ends.

### **Characteristics of short uncapped RNAs from chromatin fractions (Processed RNA)**

To test whether short uncapped RNAs were being dissociated from promoter complexes by virtue of their association with RNA processing machinery such as Ago2 (Zamudio et al., 2014), I sequenced short uncapped (processed) RNAs in the nucleoplasmic and cytoplasmic (2 replicates) and chromatin (4 replicates) fractions. My goal was to determine if short uncapped RNAs associated with the chromatin fraction at the 5' end of genes. To do so, I identified the following properties of the RNAs from the cytoplasmic, nucleoplasmic and chromatin fractions: their sizes, 5' end locations, and 3' end locations.

Analysis of uncapped (processed) RNAs in these subcellular fractions showed that they were enriched in the chromatin fractions (1 of 4 replicates shown) (Figure 4-8). Genes with short uncapped (processed) RNA profiles showed a peak density at 15 nucleotides from the TSS at the 5' ends and about 30 to 50 nucleotides downstream the TSS for the 3' ends for chromatin fractions (Figure 4-8C). For the cytoplasmic and nucleoplasmic fractions, the 5' ends were distributed broadly between the TSS and the location 20 nucleotides downstream of the TSS. The 3' ends were also broadly distributed and were located between 20-60 nucleotides downstream of the TSS. The predominant size observed in the chromatin fraction was 18 nucleotides, which matched previous reports (Taft et al., 2009) (Figure 4-9). The difference in the location of the 5' ends amongst the chromatin, nucleoplasmic, and cytoplasmic fractions suggested that processing of RNA is co-transcriptional because it occurred on chromatin.

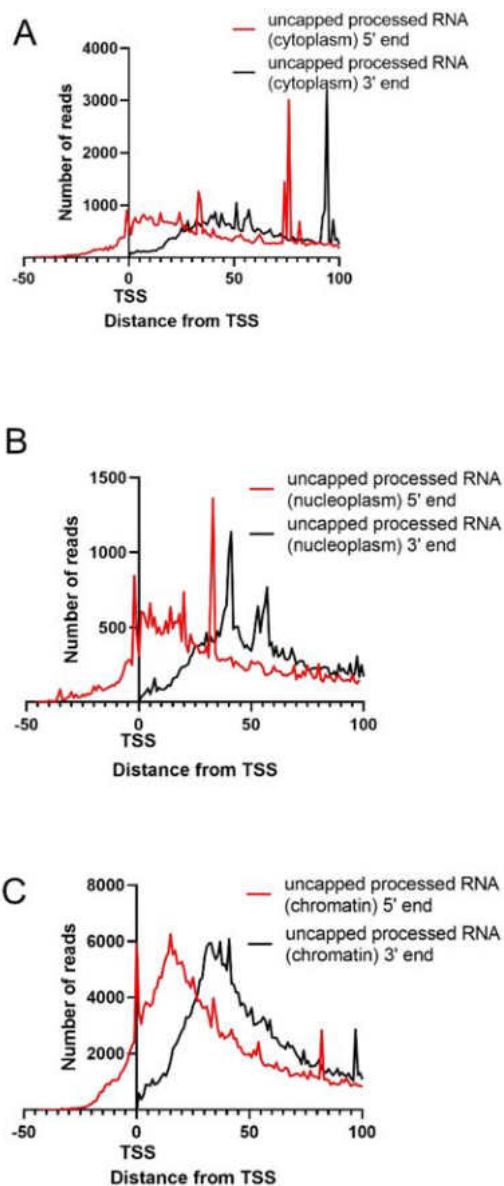


Figure 4-8: Metagenome profile of short uncapped processed RNA 50 nucleotides upstream and 100 nucleotides downstream of the TSS in A. cytoplasmic, B. nucleoplasmic, and C. chromatin fractions. 5' ends have a distinct peak 15 nucleotides downstream of the TSS for chromatin but a broad peak from 0-20 nucleotides for the cytoplasm and nucleoplasm. The 3' ends have a distinct shape between 20-50 nucleotides downstream of the TSS for chromatin compared to the cytoplasm and nucleoplasm that exhibit a broader, flatter shape between 20-60 nucleotides. Short uncapped processed RNAs are enriched in chromatin fractions suggesting co-transcriptional processing is occurring as opposed to post-transcriptional processing. The cytoplasm and nucleoplasm represent 1 of 2 replicates, and the chromatin is 1 of 4 replicates.

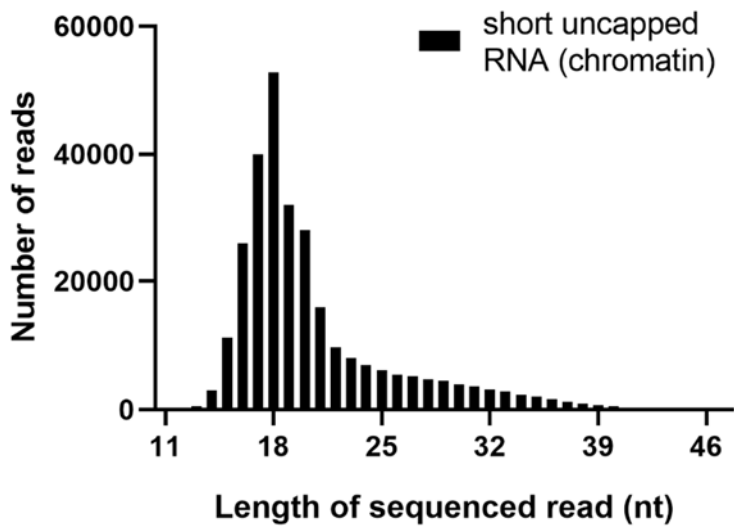


Figure 4-9: Short uncapped processed RNAs size distribution of reads (1 of 4 replicates each). The reads were obtained from regions 100 nucleotides upstream and downstream of the TSS. The RNAs range from 14 nucleotides to 40 nucleotides in length. The most abundant RNA was 18 nucleotides.



### **Short uncapped processed RNAs which are found at the 5' ends of genes are generated from paused complexes**

In order to determine if uncapped processed RNAs are generated from paused Pol II and whether the 3' ends undergo processing via backtracking, I compared the location of the 5' and 3' ends of short capped RNAs and uncapped processed RNAs.

In comparison to the 5' ends of short capped RNAs (which map to the start site), the 5' ends of uncapped (processed) RNAs differed because the 5' ends were downstream of the TSS (Figure 4-10A), confirming 5' end processing. From the metagene plots, the 3' ends of short capped RNAs (which map to the paused site) and uncapped (processed) RNAs were positioned in the same location (Figure 4-10B and Figure 4-12), which implied that the processed RNAs are generated from paused RNAs.

To further characterize uncapped processed RNA biogenesis, I investigated the locations of both capped and uncapped (processed) RNAs by subtracting the averages of the locations of both 5' and 3' ends of these two types of RNAs. As expected, the maximum difference in position between the start sites of short capped RNAs and processed RNAs ranged from 9-15 nucleotides at the 5' ends (Figure 4-11), confirming that there is a difference in the location of 5' ends between capped RNAs and uncapped (processed) RNAs. This reinforced the idea that there is a possibility of 5' to 3' exoribonuclease activity or an internal cleavage of the RNA.

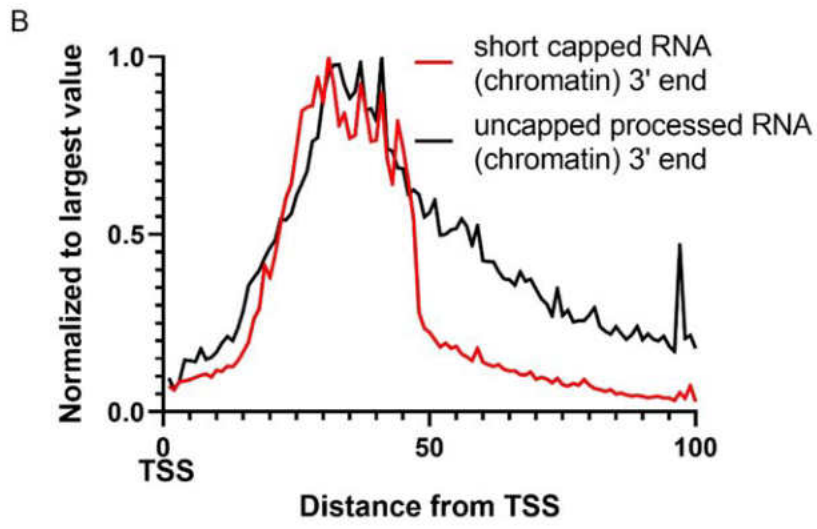
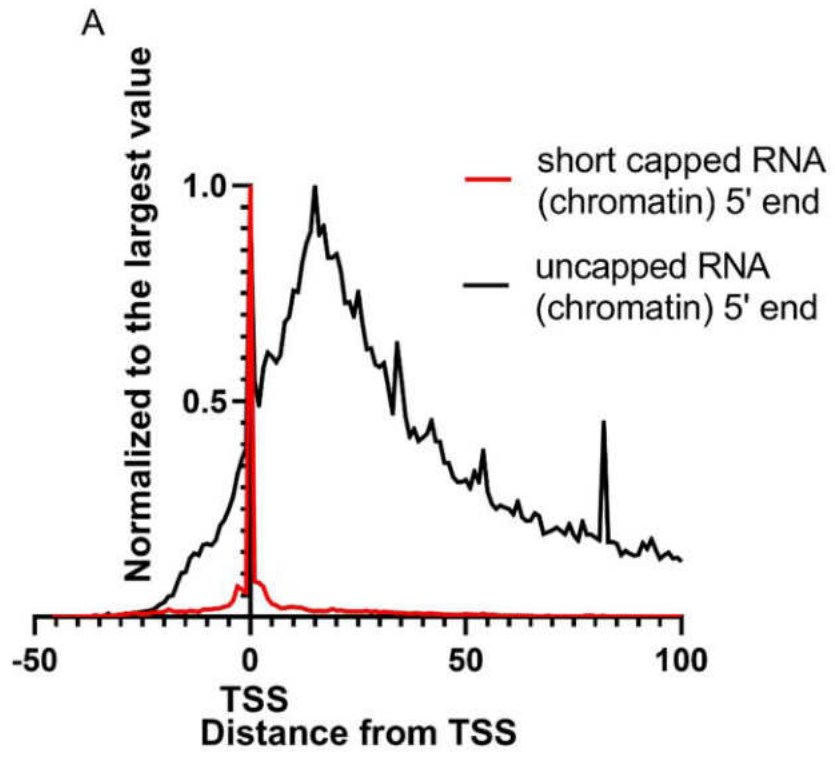


Figure 4-10: Short capped and small uncapped processed RNA 5' and 3' positions 50 nucleotides upstream and 100 nucleotides downstream of TSS. A. The 5' location of short capped RNA is different from the 5' location of short uncapped RNA. Short capped RNA is located at the TSS whereas uncapped processed RNA has a peak density 15 nucleotides downstream of the start site. B. The 3' location of short uncapped RNA and the 3' location of short capped RNA are found 20-50 nucleotides downstream the TSS. This observation suggested an absence of backtracking in the generation of uncapped processed RNA.

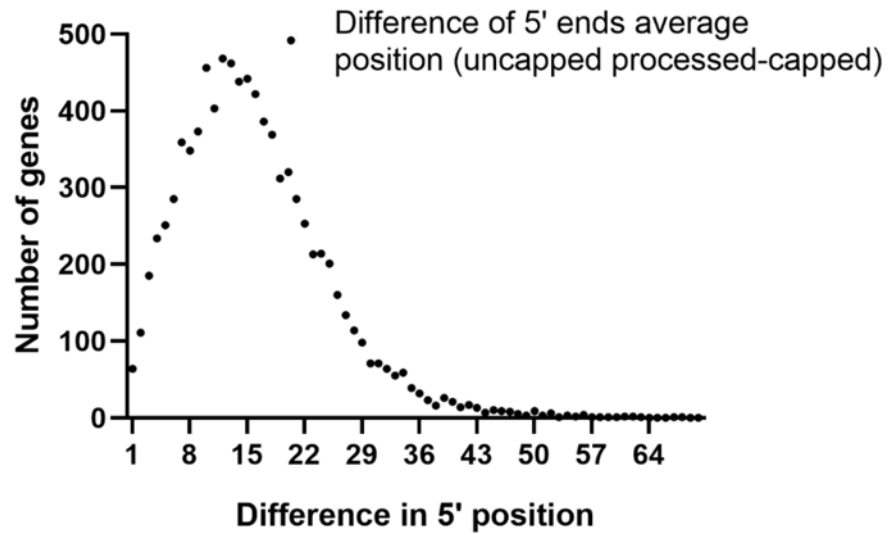


Figure 4-11: Distribution of average difference of 5' end positions of short capped and short uncapped RNAs. The 5' position of short capped RNAs was subtracted from the average 5' position of uncapped processed RNAs. The predominant difference is between 13-15 nucleotides, confirming 5' end processing co-transcriptionally.

To determine if backtracking may be involved in the generation of these short RNAs, I calculated the average difference in locations of the 3' ends of genes in both capped and uncapped RNAs. I subtracted the average location of the capped 3' ends of genes from the average location of the 3' ends of the uncapped processed genes. An average distance of 0 signified the same location of 3' ends of short capped RNAs and processed RNAs. Most genes showed a difference of -3 to 3 nucleotides between short capped RNAs and uncapped (processed) RNAs. This suggested that the 3' ends of both RNAs reside in the same location (Figure 4-12). A negative value indicated that the 3' ends of short capped RNAs are upstream of the 3' end of processed RNAs. Genes with a negative value  $>-3$  may undergo backtracking. For example, SNAIL1 shows a difference in 5' end positions between short capped and processed RNAs but shows similar 3' end positions for the two types of RNA (Figure 4-13 and Figure 4-14). Similarly, scatter plots which show the 5' locations, 3' locations, and the length of the RNAs for either the short capped RNAs or the processed RNAs reveal that short-capped RNA 5' ends are localized at the 0 position (the TSS), while the 5' ends of the processed RNAs are located away from the TSS. For both RNAs, the 3' ends are located between 20-50 nucleotides downstream of the TSS, indicating the paused site.

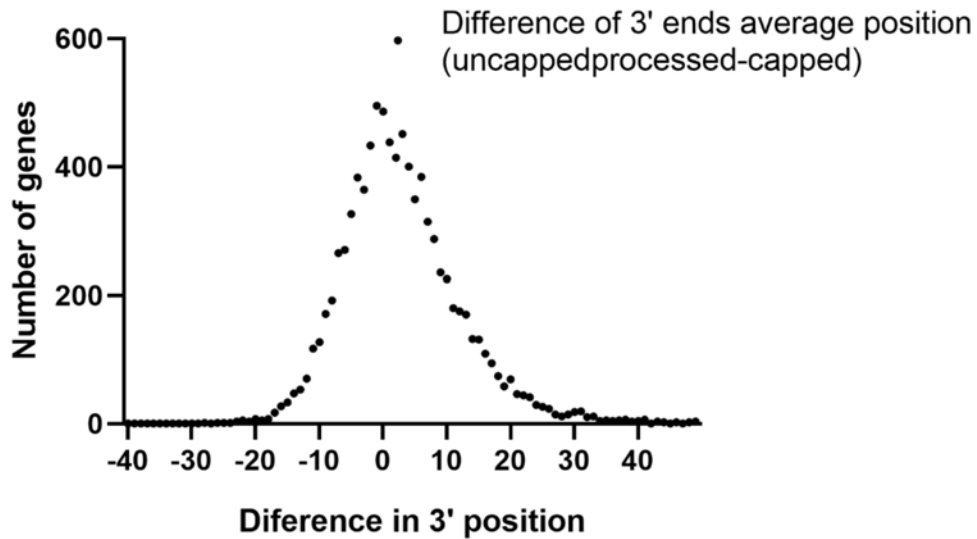


Figure 4-12: Distribution of average difference of 3' end position of short capped and short uncapped RNAs. The average 3' position of short capped RNAs was subtracted from the average 3' position of uncapped processed RNAs. An average distance of 0 signified the same location of 3' ends of short capped RNAs and processed RNAs. A positive value indicated that the 3' end of processed RNAs is downstream the 3' end of short capped RNAs. A negative value indicated that the 3' ends of short capped RNAs are upstream of the 3' ends of processed RNAs. Most genes were centered between -3 to 3 nucleotides, making backtracking an unlikely event for these genes. Genes with a negative value  $>-3$  may undergo backtracking.

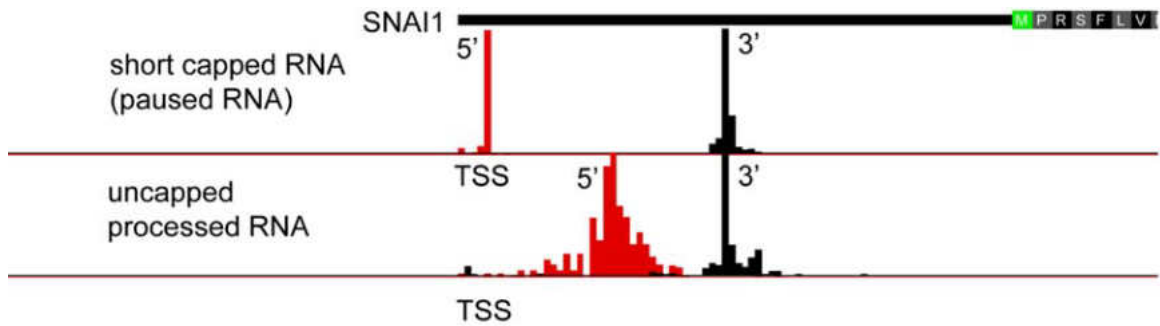


Figure 4-13. UCSC genome browser shot of SNAI1 gene showing short capped RNAs (paused RNA) and uncapped processed RNAs on chromatin. The 3' ends of short capped RNAs are in the same location as the 3' ends of uncapped processed RNAs. The 5' end of uncapped processed RNAs is downstream of the 5' ends of short capped (paused) RNAs. This suggests co-transcriptional processing of paused RNA to generate processed RNA.

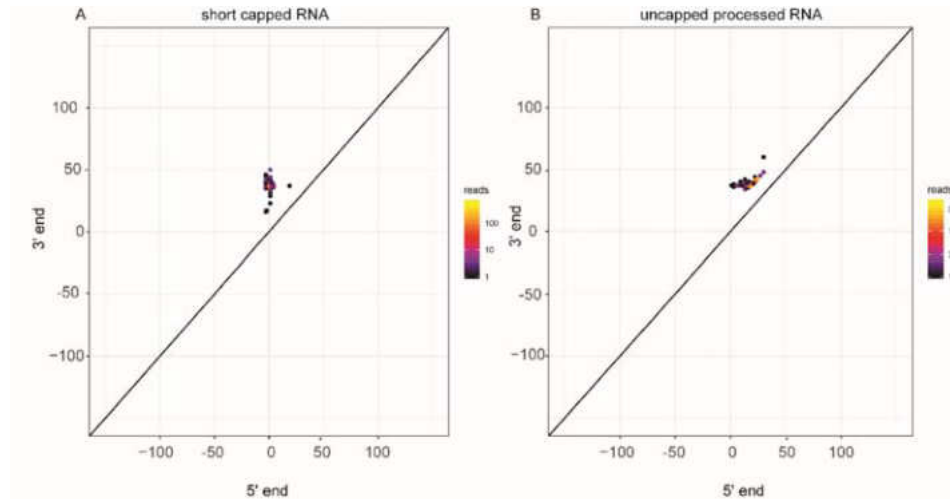


Figure 4-14. Scatter plots show the 5' start site, 3' position, and distance transcribed by Pol II in short capped RNAs and uncapped processed RNAs on the SNAI1 gene. A. SNAI1 gene short capped data shows most of the 5' ends of the transcribed RNAs at 0 (TSS). The 3' ends are located between 20 and 50 nucleotides downstream of the start site. B. SNAI1 uncapped processed RNAs show start sites that are away from the TSS and have 3' locations about 20 to 50 nucleotides downstream of the TSS. SNAI shows an example of a gene that undergoes co-transcriptional processing from paused RNAs.



### **Processed RNAs which are enriched in highly paused genes**

Next, I sought to determine whether processed RNAs are enriched in highly paused genes. I visualized this on a heatmap (Figure 4-15A) by sorting the reads according to highly paused genes. Highly paused genes were determined by the frequency of short capped RNA reads and sorted by decreasing intensity on heatmaps. I observed that there was a relationship between decreasing paused genes and decreasing processed genes, as indicated by the decreasing intensity of the color code on the heatmap. I next divided the short capped RNA read frequencies into deciles of decreasing order and determined which group was enriched in processed RNAs (Figure 4-15B). I observed that processed RNAs are indeed enriched in highly paused genes. I further determined the correlation relationship between short capped RNAs and processed RNAs. The Spearman correlation coefficient was 0.73, which indicated a moderate positive correlation between these RNAs. I noted that paused genes and genes involved in processing at the 5' ends of genes correlated weakly with gene expression. (Figure 4-16).

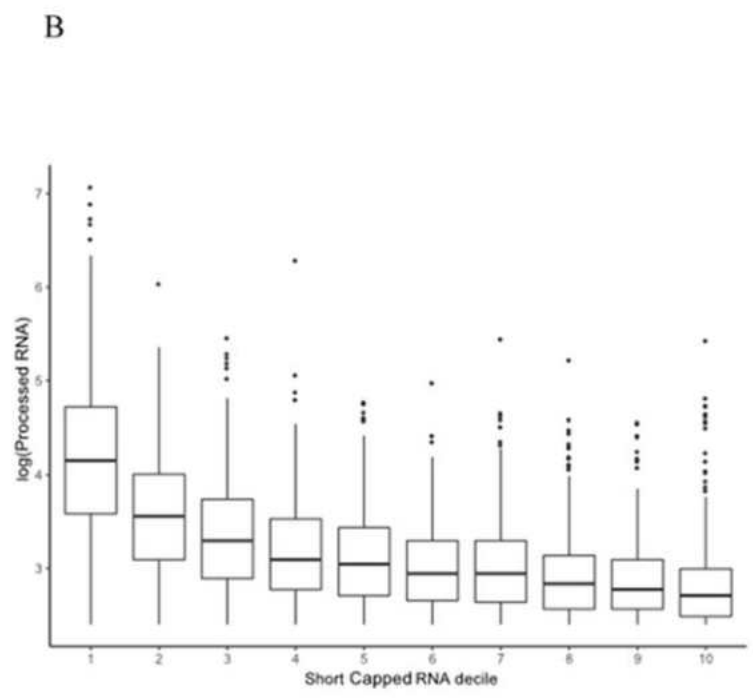
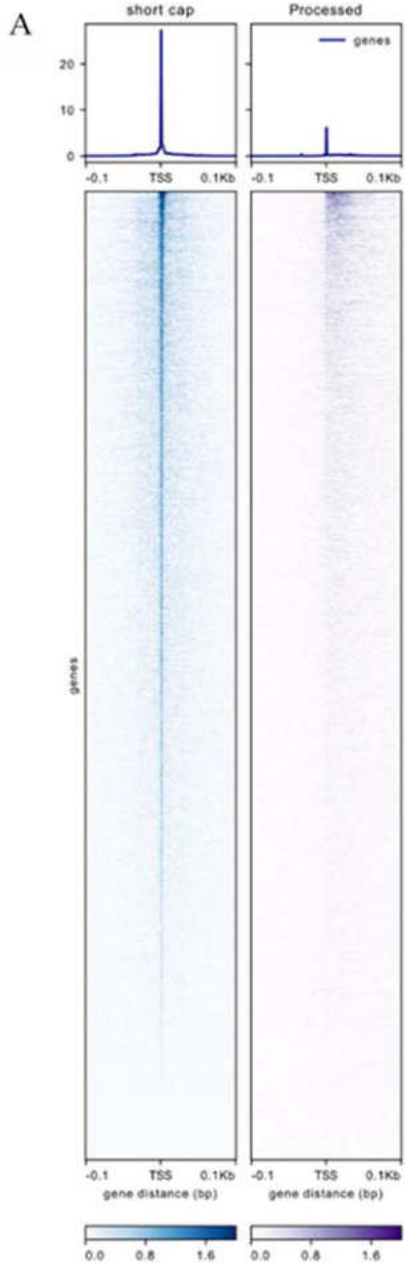


Figure 4-15. Processed RNAs are enriched in highly paused genes (short capped RNA). A. Heatmap of short capped and processed RNAs. The genes are sorted in decreasing order of frequency of short capped RNAs. B. Boxplot of short capped RNAs categorized into deciles with 1 representing genes with high pausing and 10 with least pausing. Genes with processed RNAs are enriched in highly paused genes.

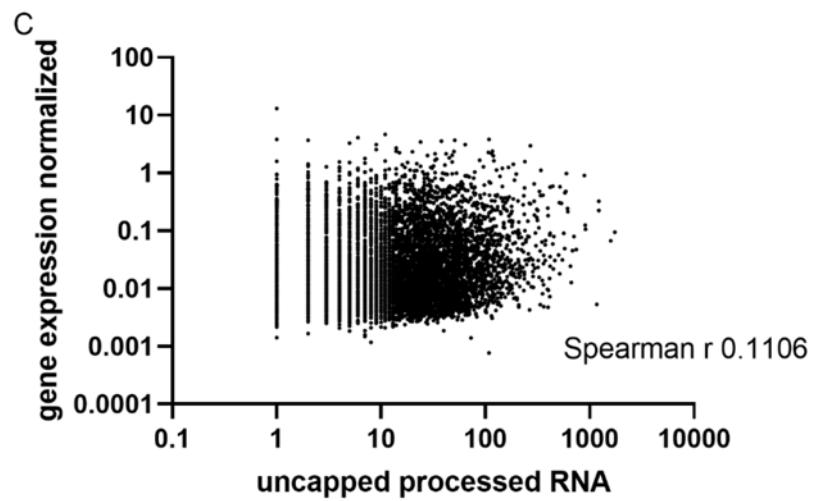
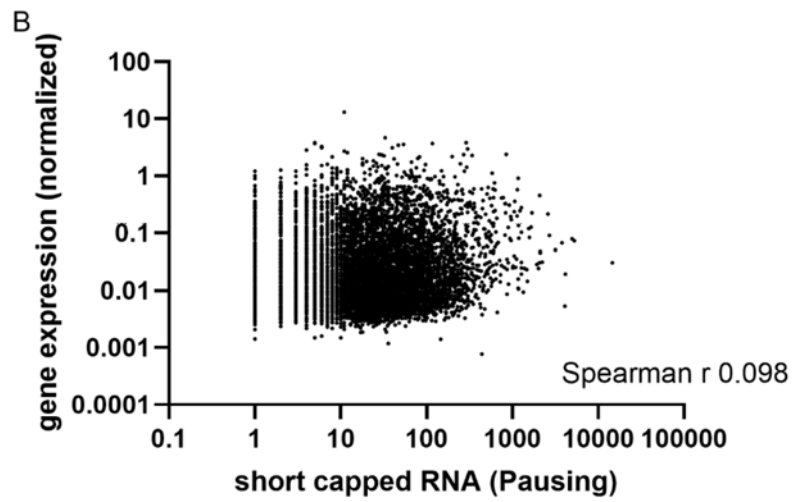
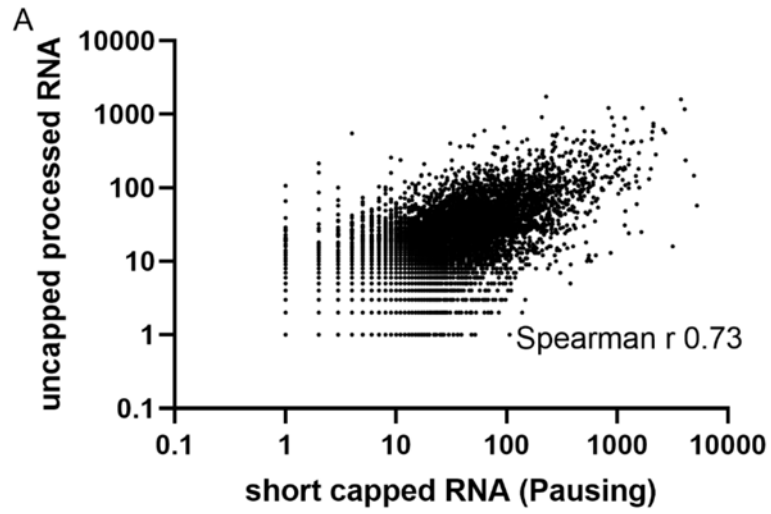


Figure 4-16: Processed RNAs correlate with pausing but not with gene expression within the region between 100 nucleotides upstream and downstream of the TSS. A. Genes with processed RNAs moderately correlated with pausing ( $r = 0.73$ ). B. Paused genes correlated weakly with gene expression (0.98) C. Genes with uncapped processed RNAs correlated weakly with gene expression (0.011). Expression was determined by Pol II occupancy 200 nucleotides downstream of the TSS and 1000 nucleotides upstream of the transcription end site. The reads were normalized by the length of the gene. This suggests that processing may be an obligatory process within the cell.

## **Role of XRN2 in generation of processed RNA**

XRN2, a 5' to 3' exoribonuclease which degrades monophosphorylated RNAs, has been implicated in premature termination (Brannan et al., 2012; Nojima et al., 2015; Valen et al., 2011). To determine the role of XRN2 in the generation of uncapped processed RNA, I analyzed the effect of depleting XRN2 (Figure 4-17) on uncapped RNA levels and corresponding sequence characteristics. The predominant RNA length observed in cells depleted of XRN2 was 18 nucleotides (Figure 4-18). Similar to control cells, XRN2 knockdown cells have uncapped (processed) RNA 5' ends that differ from the 5' ends of short capped RNAs. I observed that the 3' ends of processed RNAs and XRN2-depleted RNAs reside in the same location (Figure 4-19)

I expected that XRN2 depletion would halt processing and a length distribution profile that is shifted to the right with an enrichment of RNAs whose 5' ends would be closer to or at the TSS. However, as illustrated in Figure 4-20, read lengths showed an increase in the level of RNAs in the range of 18 to 22 nucleotides in XRN2-depleted cells compared to uncapped processed RNAs when the data were normalized by the number of reads. I compared the 5' ends of XRN2 depleted uncapped processed RNAs to the control processed RNAs by normalizing the reads to total reads sequenced and also to reads on the gene body.

The ratio between the maximum peak positions of 5' processed RNAs and 5' processed RNAs in XRN2 knockdown was 1.76 in the case of normalization to the total reads (Figure 4-21). Unlike Valen et al., 2011's findings, the 3' ends also exhibited a similar pattern of enrichments. Compared to wild type, they noted a reduction in occupancy of the 3' ends of genes in XRN2 depleted cells. Heat maps depicted an

association between genes with paused genes, and between genes that undergo processing and genes that had undergone XRN2 depletion.

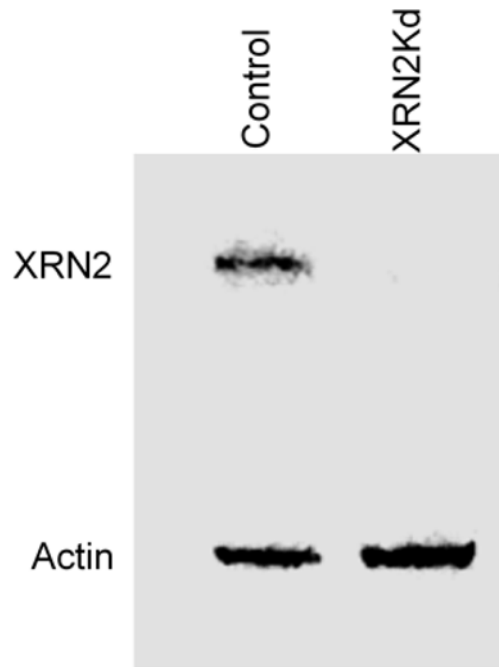


Figure 4-17: Western blot analysis of XRN2 siRNA treated cells. Equal amounts of cell lysates were separated by SDS and transferred to PVDF membranes. The membranes were probed with anti-XRN2. Actin served as the loading control.



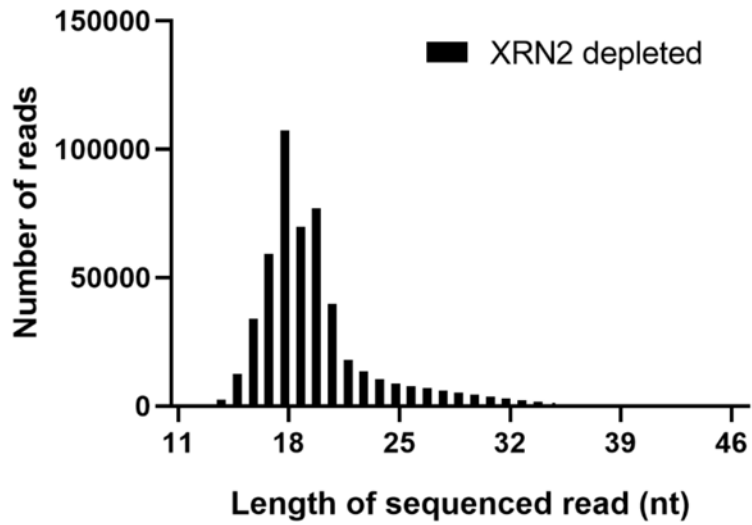


Figure 4-18: Size distribution of short uncapped processed RNAs of XRN2 depleted cells (1 of 2 replicates each). The reads were obtained from the region 100 nucleotides upstream and downstream of the TSS. The RNAs ranged from 14 nucleotides to 34 nucleotides in length. The most abundant RNAs were 18 nucleotides in length.

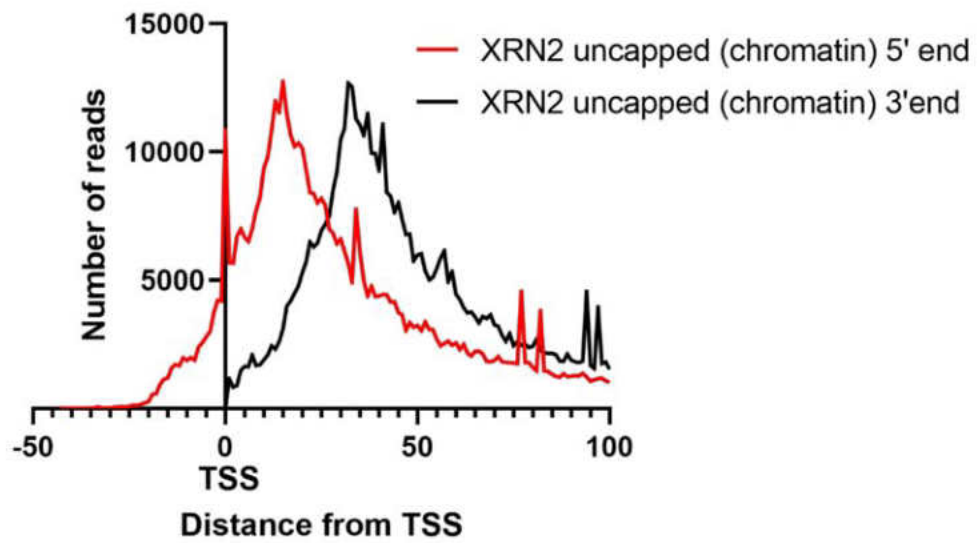


Figure 4-19: Metagene plots of 5' and 3' positions of XRN2 processed RNAs within the first 100 nucleotides of genes (1 of 2 replicates). The 5' ends differ from the 3' ends.

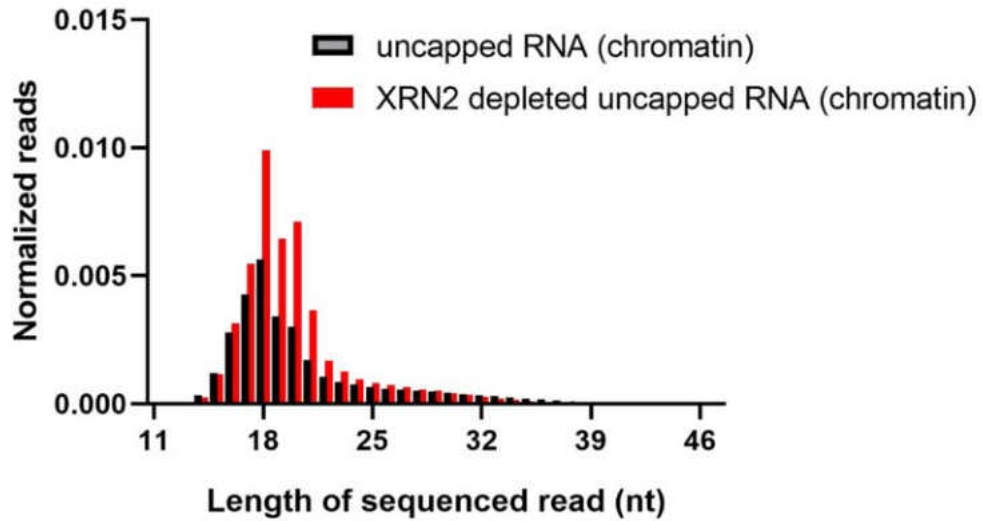


Figure 4-20: Normalized processed RNA length reads distribution for the region between 100 nucleotides upstream and downstream from TSS between processed RNAs and XRN2-KD processed RNAs. Black denotes the length distribution of processed RNA and red denotes the length distribution of XRN2 knockdown processed RNA. Reads were normalized to the total sequenced read. RNAs 18-22 nucleotides are more abundant in XRN2 depleted cells compared to the mock.

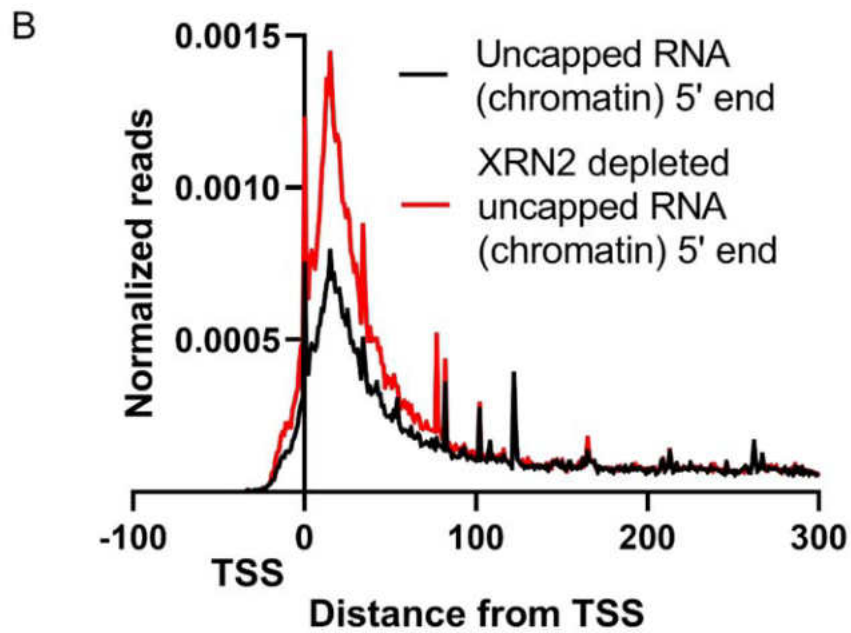
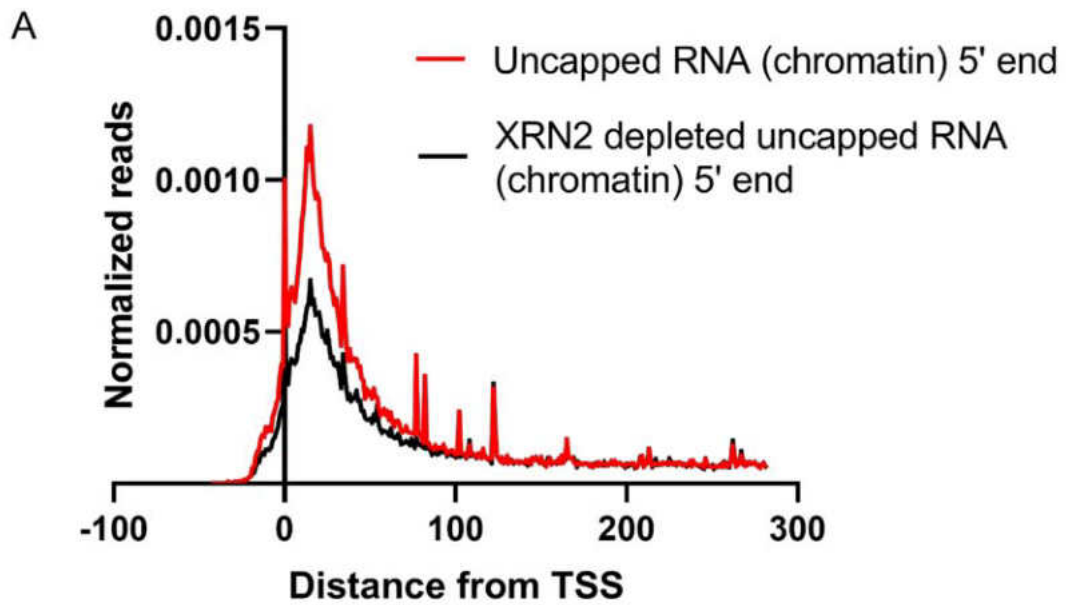


Figure 4-21: Normalized 5' ends of processed RNAs and 5' ends of processed XRN2 depleted distance from TSS. A. 5' end location of processed RNA and 5' end location of XRN2 depleted processed RNA normalized by the number of sequenced reads. B. 5' end location of processed RNA and 5' end location of XRN2 depleted processed RNA normalized by reads in the gene body. XRN2 depleted cells have enrichment of processed RNA compared to the mock cells.

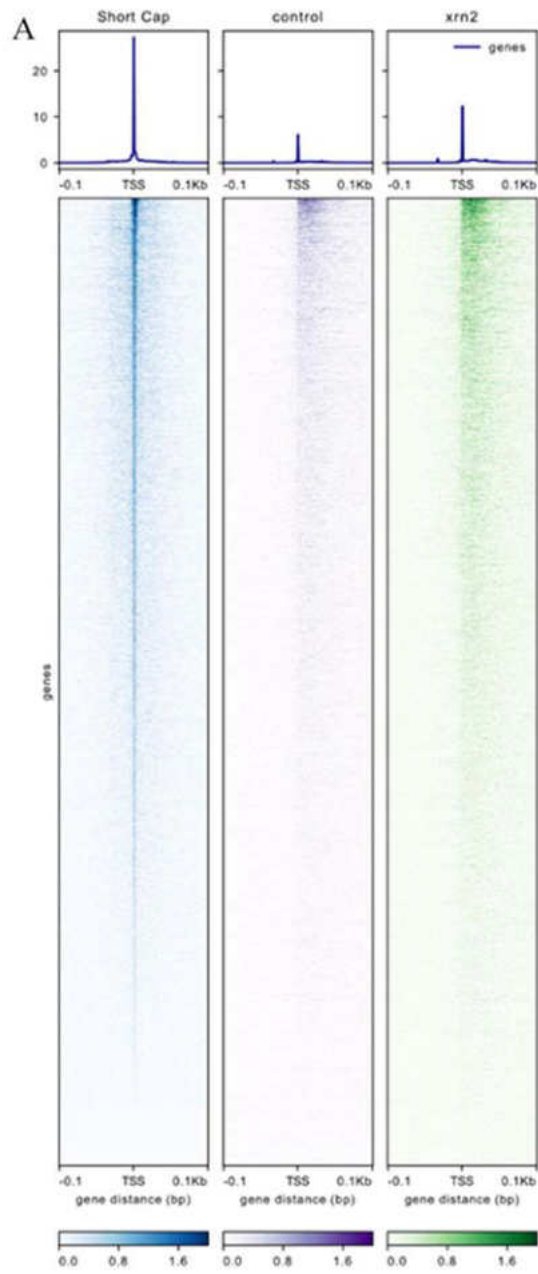


Figure 4-22. XRN2 depleted processed RNAs are enriched in highly paused genes (short capped RNA). A. Heatmap of XRN2 compared to short capped and processed RNAs and sorted by short capped RNAs. B. Heatmap of XRN2 depleted RNAs compared to short capped and processed RNAs and sorted by processed RNA signal.

## CHAPTER 5

### RESULTS: MODIFIED PRO-SEQ REVEALS PROCESSING OF RNAs AT THE 5' ENDS OF GENES FROM PAUSED COMPLEXES

#### Introduction

In the third part of my study, I adapted the PRO-Seq technique to study promoter dynamics by focusing on the 5' end status of RNAs during early transcription. This was expected to be an efficient way of studying the different 5' end status of these RNAs (monophosphorylated, triphosphorylated, and capped) since the approach generates more reads and only RNAs bound to Pol II are extended, making it more specific to elongating Pol II which contains RNA. This approach also avoids contamination from abundant miRNAs and other small noncoding RNAs, a major limitation in the preparation of uncapped (processed) small RNA libraries.

To emulate my previous study by selecting for RNAs whose 5' ends determine the start site of genes, I used enzymatic reactions to select RNAs with capped and triphosphate 5' modifications. I called this technique PRO-Start. In this study, PRO-Start selects for both capped RNAs and RNAs with triphosphorylated ends, unlike other studies called PRO-cap, which select for only capped RNAs (Kwak et al., 2013; Mahat et al., 2016).

I also selected for RNAs with monophosphate ends, which I subsequently called processed-PRO-Seq to differentiate from the processed RNAs in the earlier chapter. My main goal was to test if these processed RNAs at the 5' ends of genes could be extended using run-on. The approach I developed differs slightly from a recent publication (Tome et al., 2018). The main difference was my development of a modified PRO-Seq approach



to select for uncapped RNAs with a monophosphorylated end, while in that study they selected for all uncapped RNAs (triphosphorylated 5' ends, monophosphorylated ends, and degraded RNAs). Additionally, they determined the start site by selecting for RNAs with caps.

### **Characteristics of PRO-Start and processed PRO-Seq**

To determine the applicability of this new approach, I subjected the data to the parameters used in the previous section. I determined the length distribution, 5' end positions, and 3' end positions of the RNA. PRO-Start length distribution profiles (Figure 5-1) ranged from 14 to 47 nucleotides. As expected, the 5' ends were located at the TSS (0), and the 3' prime ends distributed from about 20 - 50 nucleotides (Figure 5-2).

The length distribution of processed-PRO-Seq reads around the TSS showed a maximum length of 19 nucleotides (Figure 5-3). However, processed RNA reads were 18 nucleotides in length (Figure 4-6). The one nucleotide difference is due to the addition of a labelled nucleotide during the run-on reaction in the PRO-Seq experiments.

For processed PRO-Seq profiles, the 5' ends peak at about 15 nucleotides from the TSS, which confirms processing occurs around the TSS (Figure 5-4). When the 5' ends and 3' ends are compared to the PRO-Start data, I observe that the 5' ends of monophosphate PRO-Seq differ from the 5' end of PRO-Start (Figure 5-5). The 3' ends of processed PRO-Seq are distributed in the same location as the 3' ends of PRO-Start reads (Figure 5-5). The 3' distribution is slightly skewed to the right of the edge of the PRO-Start 3' ends. This difference may be due to the fact that PRO-Start is composed of both capped RNAs and triphosphorylated RNAs. The triphosphorylated RNAs are RNAs in the process of being capped. Their 3' ends are likely to occupy the medial portion of

PRO-Start 3' ends. This may suggest that for processing to occur, the RNAs must have been capped before undergoing uncapping to generate monophosphorylates ends.

The scatter plot of the SNAI1 gene (Figure 5-6) is an example of processing in a gene. The scatterplot confirms that all the 5' ends are at the TSS (0) and the 3' ends are located 20-50 nucleotides downstream of the TSS for PRO-Start RNAs. However, the processed PRO-Seq data showed that the 5' ends of these RNAs range from the TSS to about 50 nucleotides downstream, and are within the same range as the PRO-Start RNAs. This suggested that processing of the RNAs begins at the start site and involves a decapping or triphosphate removal mechanism to generate RNAs with a monophosphate end.

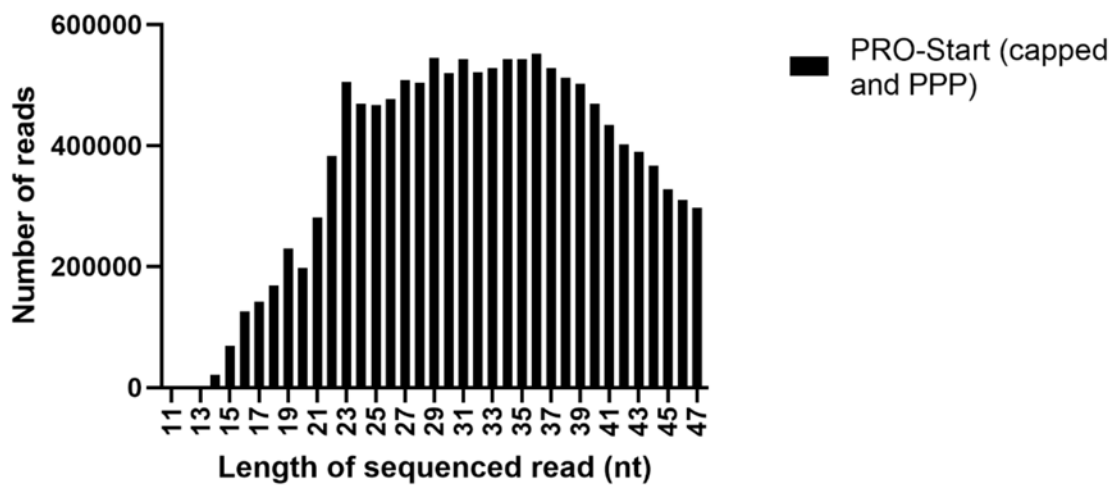


Figure 5-1: PRO-Start RNA size distribution of reads within the first 100 nucleotides of genes. (1 of 2 replicates each). The RNAs range from 14 nucleotides to 47 nucleotides in length.

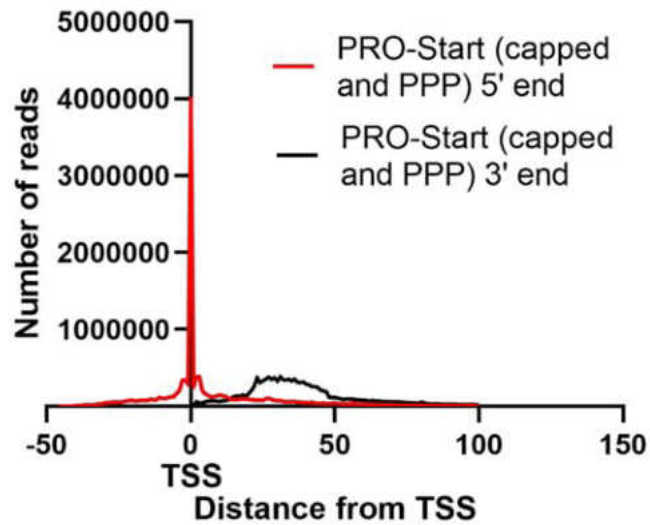


Figure 5-2: PRO-cap RNA 5' and 3' positions within the region between 100 nucleotides upstream and downstream from the TSS of genes. (1 of 2 replicates shown for short capped RNAs from chromatin). 5' end positions are centered around the TSS and 3' end positions range between 20 and 50 nt downstream of the TSS.

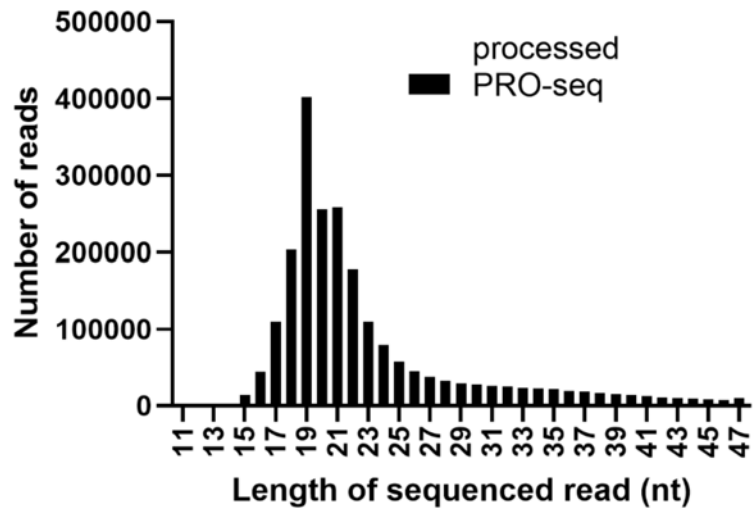


Figure 5-3: Size distribution of short uncapped processed PRO-Seq RNAs reads within the region between 100 nucleotides upstream and downstream of the TSS of genes (1 of 4 replicates shown). 19 nucleotides was the most abundant read length observed.

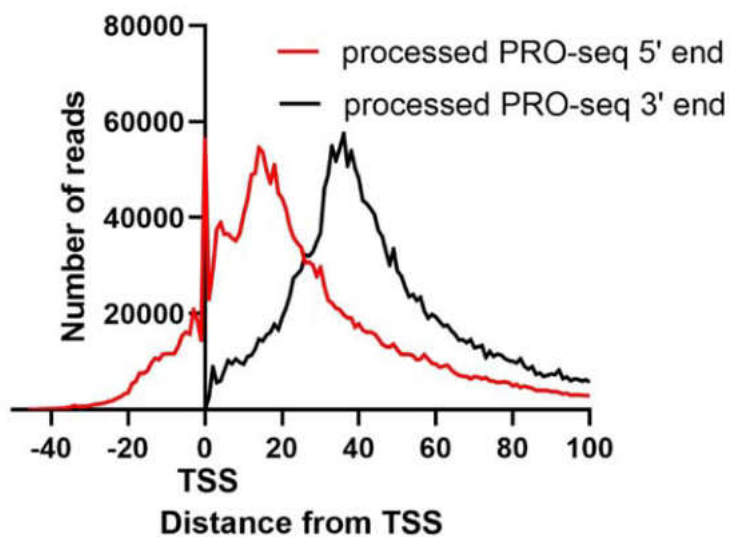


Figure 5-4: Metagene plots of 5' ends and 3' ends of processed PRO-Seq within the region between 100 nucleotides upstream and downstream of the TSS. The 5' end positions differ from the 3' end positions. The 5' ends have a peak location at 15 nucleotides downstream the TSS, and the 3' ends peak at about 38 nucleotides downstream from the TSS.

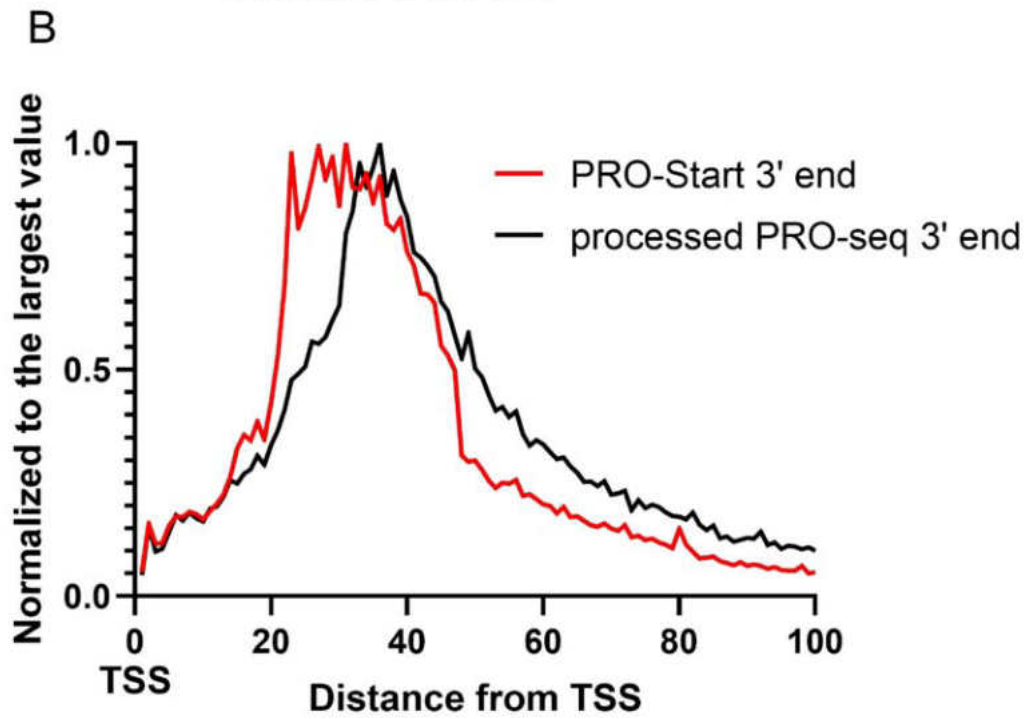
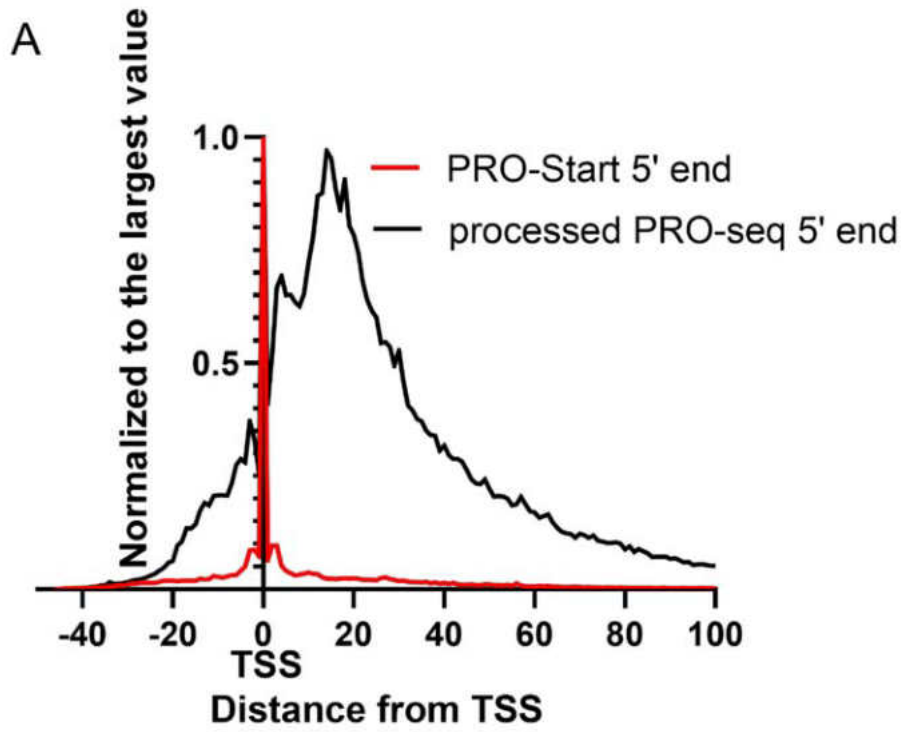


Figure 5-5: Figure 4-10: PRO-Start and small uncapped processed RNAs 5' and 3' positions within the region between 100 nucleotides upstream and downstream of the TSS. A. The 5' location of PRO-Start RNAs is different from the 5' location of processed PRO-Seq RNAs. B. The distribution of PRO-Start and processed PRO-Seq 3' end locations does not fully overlap, in contrast to Fig 4-10B. This difference can be attributed to triphosphorylated RNAs that were isolated together with capped RNAs in PRO-Start. These triphosphorylated RNAs are in the early stages of transcription and their peak location for 3' ends may be 20-30 nucleotides downstream of the TSS.



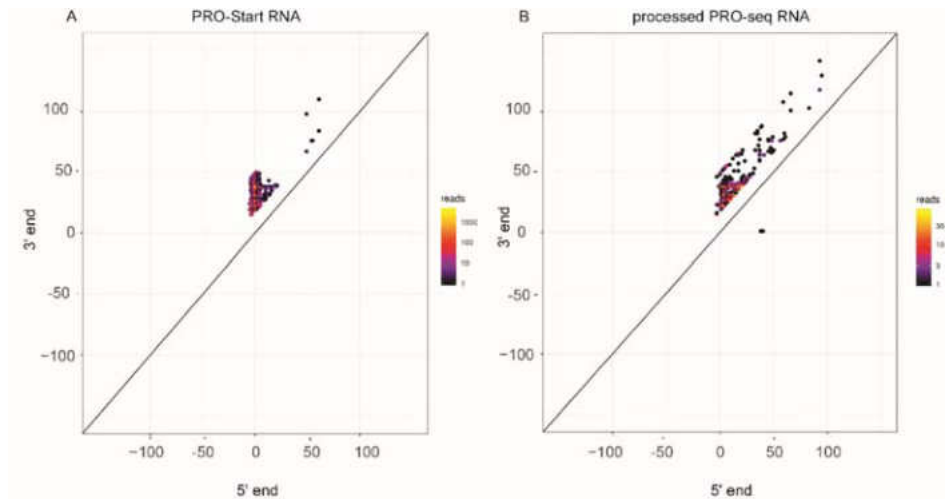


Figure 5-6: Scatter plots show the 5' start site, 3' position, and distance transcribed by Pol II in PRO-Start RNAs and processed PRO-Seq RNAs on the SNAI1 gene. A. SNAI1 gene PRO-Start data shows most of the 5' ends of the transcribed RNAs at 0 (TSS). The 3' ends are located between 20 and 50 nucleotides downstream of the start site. B. SNAI1 processed PRO-Seq RNAs show start sites that are away from the TSS and have 3' locations about 20 to 50 nucleotides downstream of the TSS. SNAI shows an example of a gene that undergoes co-transcriptional processing from paused RNAs.

### **Using modified PRO-Seq, most genes do not show backtracking at the 3' ends of genes**

I next investigated whether a high depth of sequencing would help me to confirm my earlier findings regarding the phenomenon of Pol II backtracking. I observed that the average difference between PRO-Start and processed PRO-Seq at the 5' end was around 10-15 nucleotides, which confirmed that processing occurred (Figure 5-7A). Most genes showed a difference in the 3' ends at -3 to +3 nucleotides, which suggests that the 3' ends of PRO-Start and processed PRO-Seq are in close proximity. This does not support the idea of backtracking (Figure 5-7B).

### **Modified PRO-Seq shows that processed RNAs are strongly correlated to pausing, but weakly correlated to gene expression.**

I determined the relationship between processed RNA and Pol II pausing and their relationship with gene expression. I determined pausing by two ways:

1. The number of reads in the region between 100 nucleotides upstream and downstream of the TSS from PRO-Start RNAs.
2. The 3' reads of RNAs from PRO-Seq data within 100 nucleotides upstream and downstream of the TSS.

Gene expression was obtained for the region 200 nucleotides downstream of the TSS and 1000 nucleotides upstream of the transcription end site. I observed that both PRO-start and PRO-Seq, which show pausing, strongly correlated with processing with a Spearman correlation of 0.92 and 0.72, respectively (Figure 5-8 A and B). However, pausing and processing weakly correlated with gene expression, with a Spearman correlation of 0.084 and 0.15, respectively. This suggested that processing may be an obligatory process associated with genes.

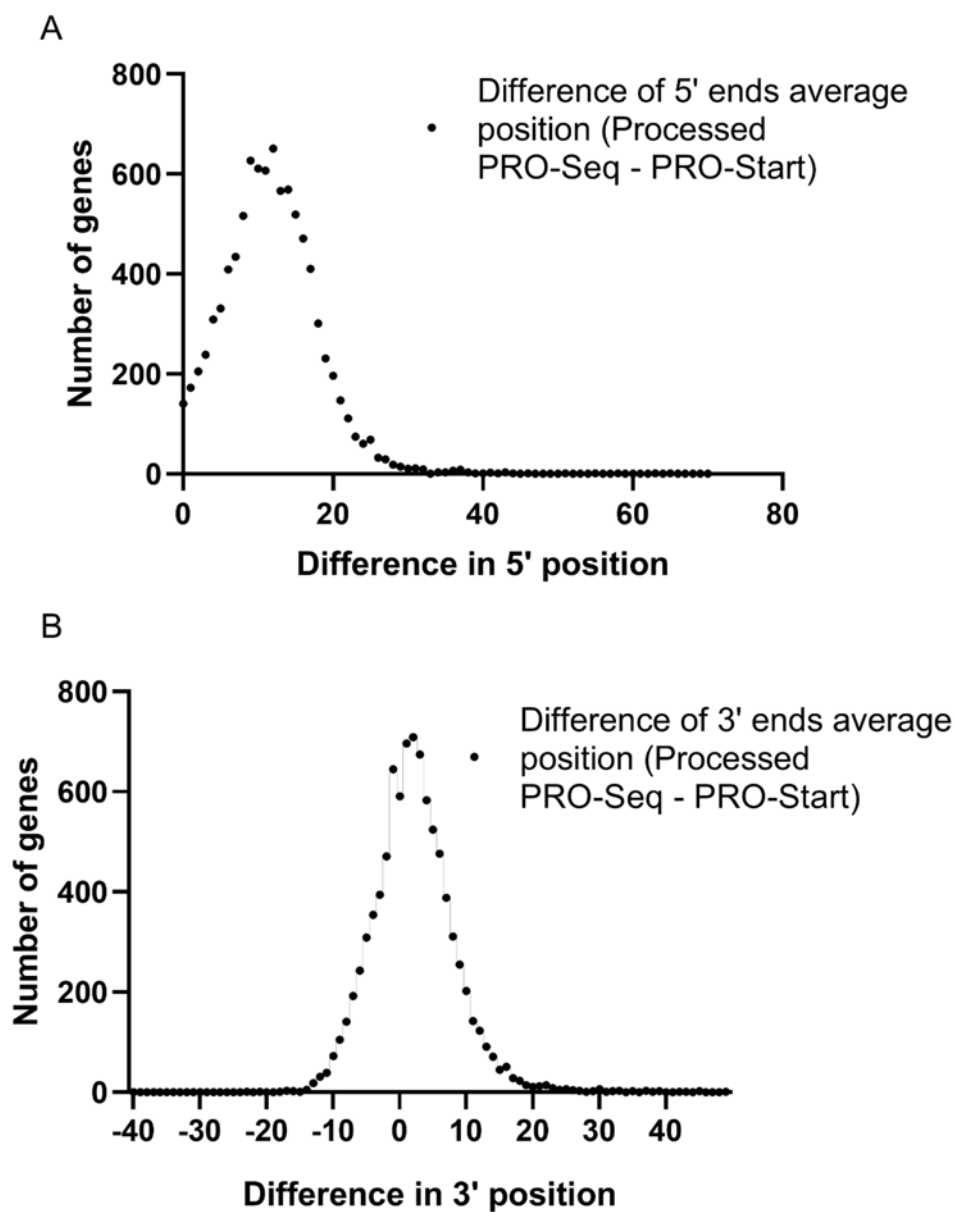


Figure 5-7: Distribution of average difference of 5' and 3' ends in modified PRO-Seq data. A. 5' ends of PRO-Start RNAs and short uncapped processed RNAs. Most reads had a difference of 13-15 nucleotides indicating processing on genes. B. 3' ends of PRO-Start RNAs and short uncapped processed RNAs. Most reads are centered around 0, making backtracking an unlikely event for most genes.

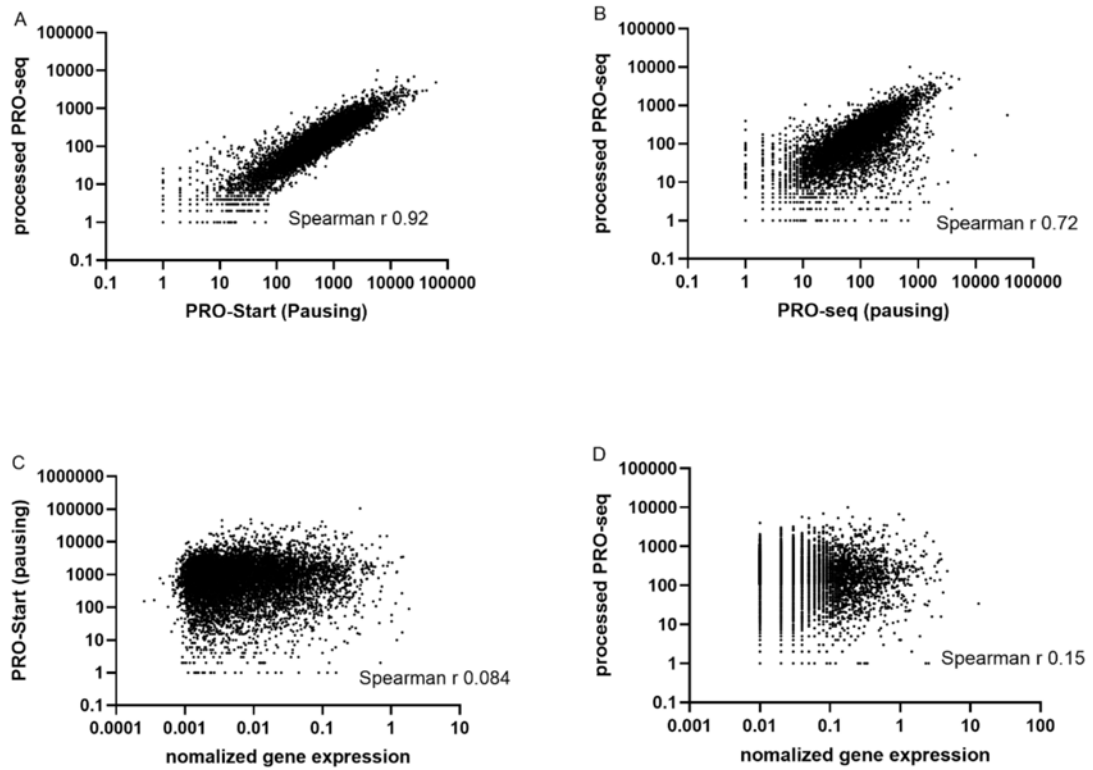


Figure 5-8: The relationship between processing of paused RNAs to pausing and gene expression. A. PRO-Start within the promoter region correlated strongly with processed PRO-Seq with a Spearman coefficient of 0.92. B. The 3' ends of PRO-Seq data which also signified pausing within the promoter was used independently to assess the relationship between pausing and processing. This showed a positive correlation of 0.72. C. Pausing weakly correlated with gene expression. D. Processed RNAs generated by modified PRO-Seq weakly correlated with gene expression.

**Modified PRO-Seq is comparable to the small RNA sequencing approach to studying generation of processed RNA on chromatin.**

To compare my results with my previous study in chapter 4, I determined the lengths of the sequenced RNA reads which were located in the region between 100 nucleotides upstream and downstream of the TSS, the 5' end locations, and 3' end locations of these RNAs.

I observed similarities between short capped RNAs and processed RNAs in comparison to PRO-Start and processed PRO-Seq profiles respectively. A UCSC genome browser snapshot of the CALM1 gene showed similarities between the short capped RNA and its equivalent experiment PRO-Start, as well as processed RNAs and PRO-Seq processed RNA, by focusing on the location of their 5' end locations and 3' end locations (Figure 5-9). The 5' ends and 3' ends of both data were in the same location, although the processed PRO-Seq 5' ends show a broader distribution, which is attributed to the greater depth of sequencing for the modified PRO-Seq experiment.

I next compared the length of RNAs sequenced between short capped RNAs and PRO-Start RNAs and observed similar pattern of distribution (Figure 5-10). The RNAs ranged from 14 nucleotides to 47 nucleotides. The 3' end locations between PRO-Start RNAs and short capped RNAs were in the same location, 20-50 nucleotides from the TSS (Figure 5-11). The 5' ends of processed RNAs and processed PRO-Seq showed the same pattern of deviation from the TSS at about 15 nucleotides downstream (Figure 5-12), and their 3' ends were distributed in a similar fashion (Figure 5-13).

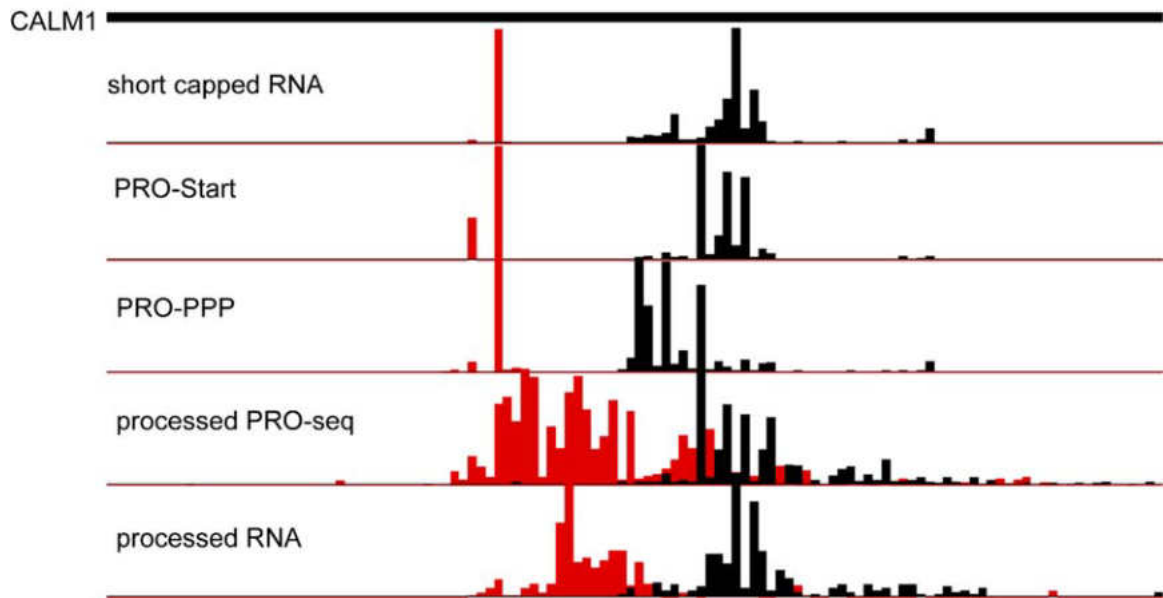


Figure 5-9: UCSC browser shots comparing the small RNA sequencing to Modified PRO-Seq. Short capped and processed RNAs are comparable to PRO-Start (capped and triphosphorylated RNAs), and processed RNA PRO-Seq. The profiles are similar, although the 5' ends of the processed PRO-Seq track exhibit a broader distribution than the processed RNA track due to the greater sequencing depth for this experiment.

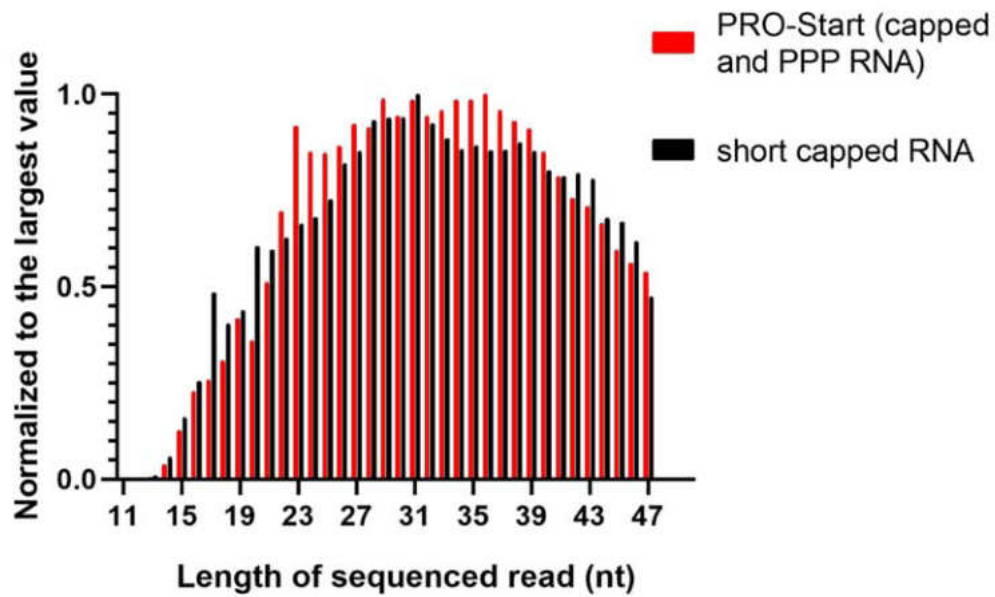


Figure 5-10: Figure 4-3: Comparison in length distribution between PRO-Start RNA from short capped RNAs from chromatin within the first 100 nucleotides of genes. (1 of 2 replicates each). The RNAs range from 14 nucleotides to 47 nucleotides in length. The data was normalized to the largest value in each data set.

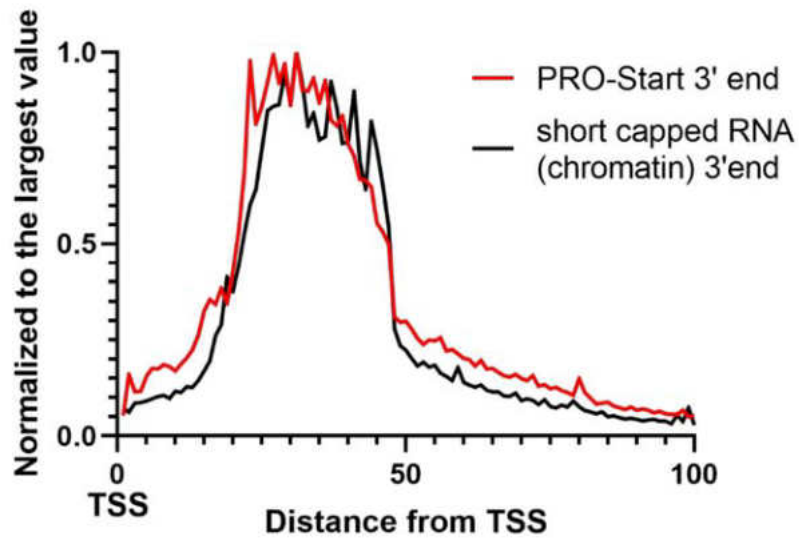


Figure 5-11 Comparison between processed PRO-Start and short capped RNAs (chromatin) at 3' positions within the region between 100 nucleotides upstream and downstream of the TSS. A. The 3' locations of processed PRO-Start and short capped RNAs are distributed between 20 to 50 nucleotides downstream of the TSS. The data was normalized to the largest value in each data set.



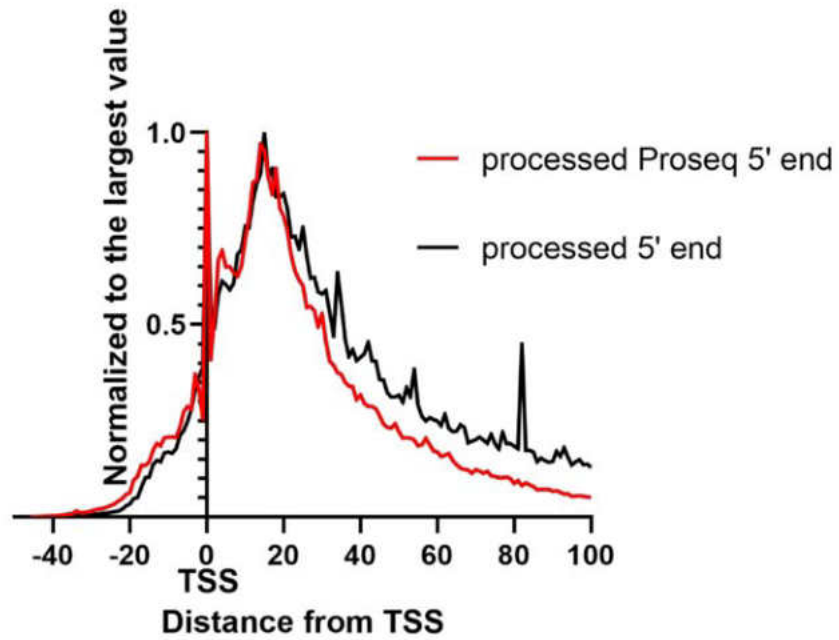


Figure 5-12. Comparison between processed PRO-Seq and processed RNAs (chromatin) at 5' positions within the region between 100 nucleotides upstream of and downstream of the TSS. A. The 5' location of processed PRO-Seq and processed RNAs peak about 15 nucleotides downstream of the TSS. The data was normalized to the largest value in each data set.

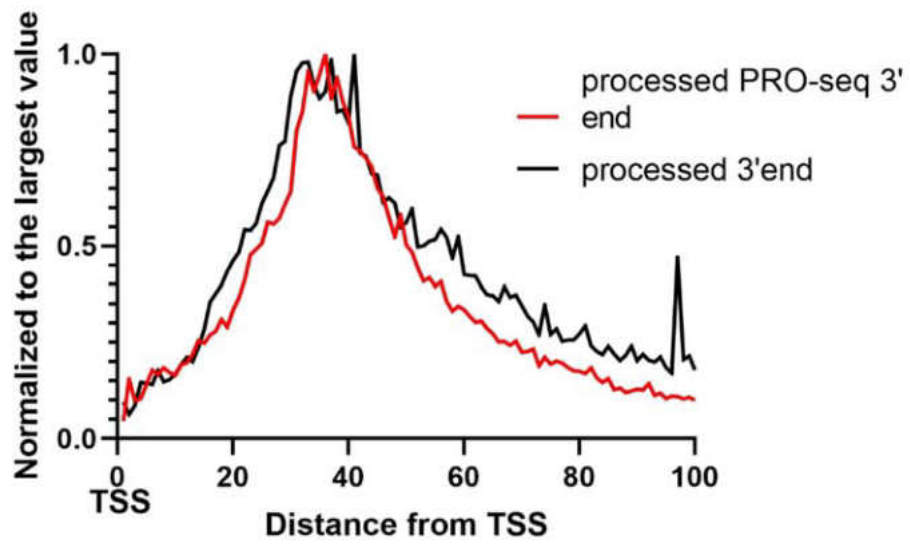


Figure 5-13: Comparison between processed PRO-Seq and processed RNAs (chromatin) at 3' positions within the region 100 nucleotides upstream and downstream of the TSS. A. The 3' locations of processed PRO-Seq and processed RNAs within the same location. The data was normalized to the largest value in each data set.

## CHAPTER 6

### DISCUSSION AND CONCLUSION

#### Discussion

Genes undergo a phenomenon called Pol II pausing, which has been implicated in regulation. Evidence of Pol II pausing has been shown using a myriad of approaches. These approaches have focused on either the transcription bubble associated with transcribing Pol II through permanganate footprinting (Gilmour et al., 2009), Pol II-DNA interactions through ChIP (Johnson et al., 2007), or the RNAs associated with the paused Pol II through short capped RNA sequencing (Nechaev et al., 2010), GRO-seq (Core et al., 2008), and PRO-Seq (Kwak et al., 2013). Across different organisms such as *Drosophila* (Nechaev et al., 2010), mice (Scruggs et al., 2015), rats (Scheidegger et al., 2019), and humans (Samarakkody et al., 2015), Pol II pausing location is conserved. At the paused site, Pol II can either continue into elongation or terminate in a process called premature termination to generate a short transcript. The question of whether premature termination occurs has been a conundrum in the field since Pol II pausing was first discovered. In addition, it is unknown whether all transcribing polymerases that pause within the first 100 nucleotides continue to elongation.

I took a different approach to this problem by focusing on the RNAs associated with paused Pol II to determine whether RNA processing occurs co-transcriptionally or post-transcriptionally. This question was intriguing because RNA processing during Pol II pausing may serve as an intermediary step that leads to premature termination. As I discussed previously, I envisaged that short RNAs in the paused state would be

associated with chromatin, while short RNAs generated by premature termination would be contained in either the nucleoplasm or cytoplasm fractions (Fig 1-5).

### **Radioactive LM PCR**

In the first part of the study, we initially hypothesized that during the premature termination event, the full-length transcript from the start site to the paused site would be terminated. We first developed an approach to visualize these RNAs on a gene by gene basis using ligation mediated radioactive PCR. The technique worked, in principle, but it had limitations. Firstly, I had to design individual gene-specific primers to detect RNAs associated with paused Pol II. These primers had to amplify very short RNAs, making designing forward and reverse primers difficult. Because of this, adapters had to be ligated to the primers. Secondly, GC rich content in some genes of the mammalian genome provided a challenge in designing primers from the start site of the gene.

I observed from the LM PCR that in addition to the expected full-length transcripts, there were shorter transcripts. I observed that there were different species of RNAs associated with each promoter for the *SNAI1* and *HSPA1B* gene, but because the detection of these RNAs was dependent on the primer design, I could only detect the RNAs if I designed primers for those specific promoter regions. For this experiment, it was better to design primers a few nucleotides downstream of the start site to ensure that I could capture these RNAs. For example, the *SNAI1* primer design was over a region 10 nucleotides downstream of its annotated TSS. This meant that I was likely to detect either RNAs that were paused or, as subsequent studies show, RNAs that were uncapped and processed.

HSPA1B's primer was designed such that it was two nucleotides downstream of the TSS. This primer theoretically would help to amplify the RNA in the region between two nucleotides downstream of the TSS to the paused Pol II on the HSPA1B gene. However, while these primers allowed me to amplify RNAs from this region that may have been associated with different paused 3' locations, the design of the primers caused the amplified region to be upstream of some uncapped processed RNAs, which usually peak 15 nucleotides downstream of the TSS. Thus, the LM PCR method did not allow me to capture all of the processed RNAs generated by pausing.

Also, sequencing gels for both SNAIL and HSPA1B showed longer fragments of RNA in the nuclei and chromatin fractions which were not present in the nucleoplasm or cytoplasm. This suggests that premature termination may not involve the full-length transcript from the start site to the paused site and supports the observation by Henriques et al., (2013) that the full length transcript is not terminated. However, their observation that premature termination might be infrequent may have arisen due to the limitations of their approach, which only looked at short capped RNAs. I considered that premature termination may involve processing at the 5' ends of genes to trigger termination, and thus may occur more frequently than has previously been thought. In fact, two recent publications have presented findings demonstrating genome wide high turnover of paused Pol II at the promoter region (Erickson et al., 2018; Krebs et al., 2017; Steurer et al., 2018), supporting this notion.

Unfortunately, it was difficult to differentiate the 5' status of these observed RNAs with this particular experimental setup because primer location in this setup can pick up both capped and uncapped RNAs. Subsequent genome wide studies showed that

uncapped RNAs are present at the TSS. To circumvent this problem, one could design gene specific primers and ligate adaptors to the 5' ends instead of the 3' ends. However, ligation would work on only RNAs that are uncapped. Thus, because this approach would not allow me to ligate adapters to the 5' caps of RNAs, I would be able to study only uncapped processed RNAs. To sum up, while this approach is valuable because the radiolabelled primers allow for visible amplification of low signal, it will require further optimizations in order to overcome its remaining limitations.

### **Capped and uncapped RNA sequencing**

Studies that have focused on premature termination usually investigate proteins that could play roles in this critically understudied phenomenon. However, other studies have discovered an array of short RNAs less than 200 nucleotides, such as transcription initiation RNAs (tiRNAs) (Taft et al., 2010) and transcription start site RNAs (TSSa RNAs) (Valen et al., 2011), which have been associated with the promoters of genes that feature Pol II pausing. In my study, we called these RNAs uncapped processed RNAs. To rule out the argument that these short RNAs are not linked to Pol II pausing or premature termination, but are rather byproducts of degradation, Valen et al., (2011) isolated short RNAs from both nuclei and whole cell samples and compared the two sets. If degradation had produced these RNAs, they would be found in the cytoplasm. However, while they found that these short RNAs were present in the nuclear fraction, the biogenesis of these RNAs remained unclear.

Due to the conclusions from my previous LM PCR study - namely, that shorter length RNA transcripts were observed in the region between the TSS of genes and the paused Pol II location - I revised my hypothesis that during termination, the full length

transcript and paused Pol II dissociate from the DNA. The tiRNAs and TSSa RNAs observed within the promoter region, proposed to be generated from paused Pol II, suggested that premature termination may involve processing of the RNAs associated with the paused complex.

I thus designed an experiment to sequence RNAs from the chromatin, nucleoplasm, and chromatin fractions. I sequenced both uncapped and capped RNAs in the case of the chromatin fractions and sequenced uncapped RNA for the cytoplasm and nucleoplasm fractions. I observed an enrichment of uncapped processed RNAs which were 17-20 nucleotides in length in the chromatin fraction, as previously seen (Taft et al., 2009; Valen et al., 2011). By comparing the location of the 5' ends of uncapped processed RNAs to the 5' ends of capped RNAs, I confirmed that these RNAs were processed at the 5' ends.

Previous studies had ruled out the generation of uncapped processed RNAs in the cytoplasm (Taft et al., 2010; Valen et al., 2011). However, it is difficult to delineate whether the processing is co-transcriptional or post-transcriptional. If Pol II terminates and is still bound to the paused full-length transcript, this transcript will get processed in the nucleoplasm (Figure 3-8). Since Pol II can still be bound to RNA, the fragment protected by Pol II would still be 18 nucleotides in length. This is an example of post-transcriptional processing of RNA.

Alternatively, processing can occur on chromatin, as proposed by Valen et al, (2011). This second kind of processing is an example of co-transcriptional processing. To differentiate between these two scenarios, it was again necessary to fractionate cells into chromatin, nucleoplasm, and cytoplasm. Fractionation also provided me with a platform

to estimate the amount of short RNAs in these fractions. In the future, quantifying these RNAs will aid in comparisons between fractions and can potentially provide a measure of pausing turnover. For example, if I were to observe that a particular gene exhibits a higher enrichment of these short RNAs in the nucleoplasm relative to chromatin, I could argue that there is a high rate of premature termination for that gene.

### **Backtracking**

Taft et al., (2010) proposed that uncapped processed RNA was generated from backtracking. In the backtracking scenario, Pol II backtracks towards the TSS, leaving a protruding RNA at the 3' end. The exposed end is cleaved by TFIIS. This feature has been shown *in vitro* (Cheung & Cramer, 2011; Izban & Luse, 1993) and *in vivo* (Nechaev et al., 2010). However, Valen et al., (2011) disputed this hypothesis, since backtracking alone cannot generate these RNAs. Backtracking generates RNA products that are 7-14 nucleotides in length (Izban & Luse, 1993) while uncapped processed RNAs are greater than 18 nucleotides in length. I cannot exclude the possibility that backtracking resulted in the formation of some of the uncapped processed RNAs I studied because the approach used to define the 3' ends was of low resolution.

Additionally, the *in vivo* detection of backtracking in *Drosophila* suggested that generation of these uncapped processed RNAs could be dependent on backtracking. By comparing their data to Pol II ChIP, Taft et al., (2010) and Valen et al., (2011) proposed that these RNAs are generated from paused Pol II. The limitation in their study is that ChIP has low resolution because it cannot map protein-chromatin interactions to nucleotide resolution. This meant that Pol II ChIP could not delineate either the 3' boundary of their small RNA data or the exact location of Pol II accurately.



To rule out backtracking, I generated two different sets of libraries from the chromatin fraction. Sequencing short capped RNA libraries allowed me to determine the start site of a gene and the location of paused Pol II. On the other hand, sequencing of the uncapped processed libraries allowed me to determine the location of the 5' ends and 3' ends of uncapped processed RNAs. By comparing the capped RNAs to the uncapped RNAs generated in the same location, I could tell if these RNAs were generated from paused complexes, processed at their 5' ends, and whether they were products of backtracked Pol II.

At the 3' ends, we observed that a majority of genes do not undergo backtracking to generate uncapped processed RNAs. This was because the average position of their 3' ends were in the same location. This suggests that the biogenesis of these RNAs does not predominantly involve the process of backtracking (Valen et al., 2011), as suggested by Taft et al., (2009). It further suggests that uncapped RNAs are likely to be produced by Pol II pausing.

The stability of these RNAs can be investigated by labelling cells with a labelled nucleotide, chasing the labelled nucleotide by washing the media and adding unlabelled nucleotide, and collecting these cells at different time points (Imamachi et al., 2014; Paulsen et al., 2014). This approach would enable one to determine both whether these RNAs are stable and whether they are quickly degraded.

Additionally, the role of these RNAs that are generated can be investigated by transfecting cells with these RNA oligos of known sequences to determine their potential functions within the cell (Project, 2009).

## **PRO-Seq**

Our study was limited by contamination from abundant stable non-coding RNAs (miRNAs, tRNAs, rRNAs, snRNAs, and snoRNAs), which possess monophosphorylated ends. To circumvent this, we modified the PRO-Seq technique to select for 5' modifications of RNA at the promoter such as capped RNAs, triphosphorylated RNAs and monophosphorylated RNAs. The depth of sequencing in the modified PRO-Seq is higher than that of the traditional small RNA sequencing. In theory, this increased depth would give me better coverage over the promoter-proximal loci such that I would be able to better map the 5' ends and 3' ends of short RNAs as well as the paused Pol II locations.

Most techniques used to study transcription initiation do not identify the 5' end status of the RNA, since Pol II is the focus. A recent study (Nojima et al., 2015) suggests that three states of Pol II phosphorylation exist within the promoter region (unphosphorylated, serine 5 phosphorylation, and serine 2 phosphorylation) in mammalian cells. The link between the phosphorylation states and the 5' end status of the RNAs is unknown. Serine-2 phosphorylation is known to be enriched at the 3' ends of genes where termination occurs to produce a full-length transcript. The presence of Serine-2 at the promoter may serve as a signal to induce premature termination.

The uncapped processed PRO-Seq experiment hinged on the ability of Pol II containing processed RNA to resume transcription. The success of this experiment suggested that uncapped processed RNAs bound to Pol II are not stalled, but rather have the ability to proceed into elongation. However, these Pol II complexes associated with uncapped processed RNAs may be terminated by XRN2 through the torpedo mechanism

(Fong et al., 2015). We compared the uncapped processed PRO-Seq to PRO-Start (capped and triphosphorylated RNAs). Similar to the observations between capped RNA and uncapped processed RNA sequencing, we observed that the 5' ends of the uncapped processed RNA in both methods were downstream from the start site, indicating the occurrence of processing. Further, the location of their 3' ends suggested that these RNAs were generated from paused Pol II complexes.

We noticed that between the start sites of the PRO-Start and its 3' ends were signals that were not always present in the short capped RNA sequencing. We noted that these were usually the 3' ends of the triphosphorylated RNA. We observed the same patterns between the processed RNA and start site RNA at an equally higher resolution and depth. The two independent techniques of PRO-Start and short capped RNA sequencing allowed us to visualize uncapped processed RNAs at the promoter. Our observations indicate that major approaches which study Pol II-DNA interaction at the promoter must take into account the various RNAs associated with Pol II.

### **Model of premature termination**

Currently, there are two main models of premature termination in the transcriptional regulation field. The decapping model suggests the involvement of decapping proteins which leave RNAs with 5' monophosphate ends. These form substrates for XRN2, a 5' to 3' exoribonuclease which degrades the RNA and prematurely terminates Pol II in a torpedo fashion, akin to termination which occurs at the transcription end site to produce a matured mRNA (Brannan et al., 2012).

The second model of premature termination has been described on HIV-1 promoters. It also involves XRN2, but in this instance, the TAR RNA associated with the

paused complex is cleaved by the microprocessor complex Drosha and DGCR8. The remaining RNA bound to Pol II is further processed by XRN2, and Pol II is prematurely terminated. Interestingly, the cleaved RNA medial to the TSS is processed by Rrp6, a 3' to 5' exoribonuclease, into a smaller RNA which has been implicated in gene repression. (Wagschal et al., 2012). This leads to a possible function of premature termination to produce RNAs that can modulate the gene's expression, possibly by repression. This phenomenon has been observed in human cells as well (Project, 2009).

Valen et al., 2011 proposed that these RNAs are processed at the 5' of genes. They tested this idea by performing double knockdown of XRN1 and XRN2. Though inconclusive, when both XRN1 and 2 were depleted in the study, Valen et al., 2011 observed an increase in the length of these uncapped processed RNAs. This suggested that XRN1 and 2 played a role in post transcriptional processing of these small RNAs, which are linked to the promoter. However, XRN1 is known to be a cytoplasmic exonuclease. Because Pol II pausing happens exclusively in the nucleus, XRN1 may not have contributed much to the observed increase in the length of these RNAs. Unfortunately, Valen's study did not differentiate between the role of XRN1 versus XRN2.

I investigated the role of 5' to 3' exoribonucleases in the generation of these RNAs. Two such exoribonucleases XRN1 and XRN2 which are cytoplasmic and nucleoplasmic proteins were knockdown and their impact on the generation of these RNAs were investigated. I investigated the link between the generation of short RNAs presumably from paused RNA, whether the biogenesis was co-transcriptional or post-transcriptional, and whether the biogenesis was independent of backtracking.

Valen et al., (2011) also proposed that these 18-nucleotide long RNAs were remnants of RNA 5' to 3' processing, and that these RNAs were protected by stalled Pol II, because Pol II crystal structure can contain 17-20 nucleotides of RNA (Brueckner 2009). However, there is also a possibility that an internal cleavage mechanism could result in an RNA fragment which is 18 nucleotides in length. If this is the case, then , then DIS3, a protein which is both a 3'-5' exonuclease and endonuclease (Szczepińska et al., 2015), and CPSF73, a subunit of the 3' end processing endonuclease CPSF, may be implicated in this process (Nojima et al., 2015).

I next moved to test whether XRN2, the nuclear 5' to 3' exoribonuclease, was alone involved in the biogenesis of uncapped processed RNAs. Brannan et al., (2012), using ChIP, showed that profiles of Pol II were altered when XRN2 was depleted. Consistent with studies by Valen et al., (2011), we observed an increase in the length of the transcripts. However, with the proposed function of XRN2, it is expected that metagene profiles should reveal enrichment of 5' ends of processed RNA at the start site or close to the start site.

Rather, I observed an abundance of these short RNAs away from the TSS with XRN2 depletion with a moderate increase in length, which supports the idea that there may be internal cleavage of paused RNA before XRN2 exonuclease activity. Nojima et al., (2015) used XRN2 depletion to show that there is an abundance of RNAs within the promoter region and these RNAs are associated with serine 2 phosphorylation. Their model implicates the protein complex Cleavage and polyadenylation apparatus (CPA), especially the CPSF73 subunit, in an internal cleavage of RNA. Wagschal et al., (2012) showed that on the HIV promoter, there is internal cleavage involving the microprocessor

complex (Drosha and Dgcr8), XRN2, and Rrp6. Alternatively, XRN2 may work with other partners to process capped RNAs.

Some studies (reviewed in Mugridge & Gross, 2013) suggest that malformed capped RNAs are targeted for degradation and these RNAs may be the byproducts of this processing. This pathway can involve decapping as suggested by Brannan et al., (2012), or there are also proteins capable of removing pyrophosphates from triphosphorylated RNAs during the capping process. DXO, a decapping and exoribonuclease protein, may play a role in the generation of these processed RNA.

Studies show that DXO has decapping, pyrophosphatase, and exonuclease activities (Chang et al., 2012; Jiao et al., 2013; Picard-Jean et al., 2018). DXO has been implicated in the quality assurance of mRNA capping. Aberrant capped RNAs are shown to affect mRNA splicing. Capping is important to protect the 5' ends of RNAs from 5' to 3' exoribonucleases. The products of premature terminated Pol II may be the remnants of irregularly capped RNA.

This suggests that DXO may be localized around the TSS and decaps RNAs which have malformed cap structures. XRN2 then takes over the processing to destabilize Pol II and prematurely terminate the RNA as described in a proposed model (Figure 6-1). I hypothesize that depletion of DXO and XRN2 would abolish processing at the 5' ends of genes. My findings suggest that biogenesis of these RNAs may involve other partners as well.

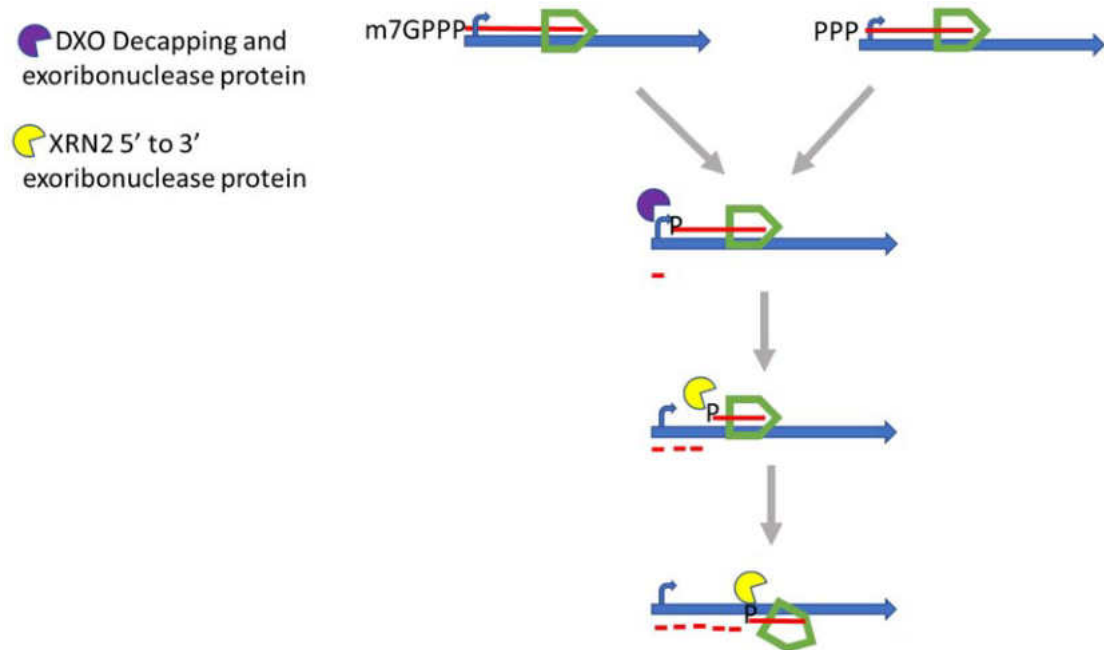


Figure 6-1. Proposed model of biogenesis of processed RNA by DXO, decapping and exoribonuclease protein and XRN2, 5' to 3' exoribonuclease protein. Adapted from (Mugridge & Gross, 2013). DXO can decap RNA (m7GPPP) or remove pyrophosphates from triphosphate RNA (PPP). These processes leave RNAs with monophosphate ends (P) and DXO degrades RNA till XRN2 takes over and prematurely terminates paused Pol II.

## **Conclusion**

From this study, I found that: (1) processed RNA around the promoter region are generated from paused Pol II on chromatin. (2) Biogenesis of processed RNAs does not involve backtracking. (3) Processed RNAs are remnants protected by Paused Pol II. (4) Ligated mediated PCR can be modified to study small and low abundant RNA on a gene specific basis. (5) Enzymatic modification of PRO-Seq to study various 5' end modification of RNAs can be used to study transcription initiation and its relation to gene expression. (6) XRN2 plays a role in the biogenesis of processed RNA.



## References

- A., P.-J., & D.L., B. (2009). Co-transcriptional splicing of constitutive and alternative exons. *Rna*, *15*(10), 1896–1908. <https://doi.org/10.1261/rna.1714509>
- Adelman, K., & Lis, J. T. (2012). Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nature Reviews. Genetics*, *13*(10), 720–731. <https://doi.org/10.1038/nrg3293>
- Austenaa, L. M. I., Barozzi, I., Simonatto, M., Masella, S., Della Chiara, G., Ghisletti, S., ... Natoli, G. (2015). Transcription of Mammalian cis-Regulatory Elements Is Restrained by Actively Enforced Early Termination. *Molecular Cell*, *60*(3), 460–474. <https://doi.org/10.1016/j.molcel.2015.09.018>
- Bhatt, D. M., Pandya-Jones, A., Tong, A. J., Barozzi, I., Lissner, M. M., Natoli, G., ... Smale, S. T. (2012). Transcript dynamics of proinflammatory genes revealed by sequence analysis of subcellular RNA fractions. *Cell*, *150*(2), 279–290. <https://doi.org/10.1016/j.cell.2012.05.043>
- Brannan, K., Kim, H., Erickson, B., Glover-Cutter, K., Kim, S., Fong, N., ... Bentley, D. L. (2012). MRNA Decapping Factors and the Exonuclease Xrn2 Function in Widespread Premature Termination of RNA Polymerase II Transcription. *Molecular Cell*, *46*(3), 311–324. <https://doi.org/10.1016/j.molcel.2012.03.006>
- Buckley, M. S., Kwak, H., Zipfel, W. R., & Lis, J. T. (2014). Kinetics of promoter Pol II on Hsp70 reveal stable pausing and key insights into its regulation. *Genes and Development*, *28*(1), 14–19. <https://doi.org/10.1101/gad.231886.113>
- Buratowski, S. (2009). Progression through the RNA Polymerase II CTD Cycle. *Molecular Cell*, *36*(4), 541–546. <https://doi.org/10.1016/j.molcel.2009.10.019>
- Chaitankar, V., Karakülah, G., Ratnapriya, R., Giuste, F. O., Brooks, M. J., & Swaroop, A. (2016). Next generation sequencing technology and genome-wide data analysis: Perspectives for retinal research. *Progress in Retinal and Eye Research*, *55*, 1–31. <https://doi.org/10.1016/J.PRETEYERES.2016.06.001>
- Chang, J. H., Jiao, X., Chiba, K., Oh, C., Martin, C. E., Kiledjian, M., & Tong, L. (2012). Dxo1 is a new type of eukaryotic enzyme with both decapping and 5'-3' exoribonuclease activity. *Nature Structural & Molecular Biology*, *19*(10), 1011–1017. <https://doi.org/10.1038/nsmb.2381>
- Cheng, B., & Price, D. H. (2007). Properties of RNA polymerase II elongation complexes before and after the P-TEFb-mediated transition into productive elongation. *Journal of Biological Chemistry*, *282*(30), 21901–21912. <https://doi.org/10.1074/jbc.M702936200>

- Cheung, A. C. M., & Cramer, P. (2011). Structural basis of RNA polymerase II backtracking, arrest and reactivation. *Nature*, *471*(7337), 249–253. <https://doi.org/10.1038/nature09785>
- Churchman, L. S., & Weissman, J. S. (2011). Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature*, *469*(7330), 368–373. <https://doi.org/10.1038/nature09652>
- Clark, M. B., Choudhary, A., Smith, M. A., Taft, R. J., & Mattick, J. S. (2013). The dark matter rises: the expanding world of regulatory RNAs. *Essays in Biochemistry*, *54*, 2013. <https://doi.org/10.1042/BSE0540001>
- Corden, J. L. (1990). Tails of RNA polymerase II. *Trends in Biochemical Sciences*, *15*(10), 383–387. [https://doi.org/10.1016/0968-0004\(90\)90236-5](https://doi.org/10.1016/0968-0004(90)90236-5)
- Core, L. J., & Lis, J. T. (2008). Transcription regulation through promoter-proximal pausing of RNA polymerase II. *Science*, *319*(5871), 1791–1792. <https://doi.org/10.1126/science.1150843>
- Core, L. J., Waterfall, J. J., & Lis, J. T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, *322*(5909), 1845–1848. <https://doi.org/10.1126/science.1162228>
- Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., ... Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, *29*(1), 15–21. <https://doi.org/10.1093/bioinformatics/bts635>
- Erickson, B., Sheridan, R. M., Cortazar, M., & Bentley, D. L. (2018). Dynamic turnover of paused Pol II complexes at human promoters. *Genes & Development*, *32*(17–18), 1215–1225. <https://doi.org/10.1101/gad.316810.118>
- Fong, N., Brannan, K., Erickson, B., Nguyen, T., Karp, S., Correspondence, D. L. B., ... Bentley, D. L. (2015). Effects of Transcription Elongation Rate and Xrn2 Exonuclease Activity on RNA Polymerase II Termination Suggest Widespread Kinetic Competition. *Molecular Cell*, *60*(2), 256–267. <https://doi.org/10.1016/j.molcel.2015.09.026>
- Fujita, T., & Schlegel, W. (2010). Promoter-proximal pausing of RNA polymerase II: An opportunity to regulate gene transcription. *Journal of Receptors and Signal Transduction*, *30*(1), 31–42. <https://doi.org/10.3109/10799890903517921>
- Giardina, C., Perez-Riba, M., & Lis, J. T. (1992). Promoter melting and TFIID complexes on *Drosophila* genes in vivo. *Genes and Development*, *6*(11), 2190–2200. <https://doi.org/10.1101/gad.6.11.2190>

- Gilmour, D S, & Lis, J. T. (1986). RNA polymerase II interacts with the promoter region of the noninduced hsp70 gene in *Drosophila melanogaster* cells. *Molecular and Cellular Biology*, 6(11), 3984–3989. <https://doi.org/10.1128/mcb.6.11.3984>
- Gilmour, David S., & Fan, R. (2009). Detecting transcriptionally engaged RNA polymerase in eukaryotic cells with permanganate genomic footprinting. *Methods*, 48(4), 368–374. <https://doi.org/10.1016/J.YMETH.2009.02.020>
- Guenther, M. G., Levine, S. S., Boyer, L. A., Jaenisch, R., & Young, R. A. (2007). A Chromatin Landmark and Transcription Initiation at Most Promoters in Human Cells. *Cell*, 130(1), 77–88. <https://doi.org/10.1016/j.cell.2007.05.042>
- Henriques, T., Gilchrist, D. A., Nechaev, S., Bern, M., Muse, G. W., Burkholder, A., ... Adelman, K. (2013). Stable pausing by rna polymerase II provides an opportunity to target and integrate regulatory signals. *Molecular Cell*, 52(4), 517–528. <https://doi.org/10.1016/j.molcel.2013.10.001>
- Imamachi, N., Tani, H., Mizutani, R., Imamura, K., Irie, T., Suzuki, Y., & Akimitsu, N. (2014). BRIC-seq: A genome-wide approach for determining RNA stability in mammalian cells. *Methods*, 67(1), 55–63. <https://doi.org/10.1016/J.YMETH.2013.07.014>
- Lander, E.S et al. Initial sequencing and analysis of the human genome. (2001). *Nature*, 409(6822), 860–921. <https://doi.org/10.1038/35057062>
- Izban, M. G., & Luse, D. S. (1993). The increment of SII-facilitated transcript cleavage varies dramatically between elongation competent and incompetent RNA polymerase II ternary complexes. *Journal of Biological Chemistry*, 268(17), 12874–12885.
- Jiao, X., Chang, J. H., Kilic, T., Tong, L., & Kiledjian, M. (2013). A Mammalian Pre-mRNA 5' End Capping Quality Control Mechanism and an Unexpected Link of Capping to Pre-mRNA Processing. <https://doi.org/10.1016/j.molcel.2013.02.017>
- Johnson, D. S., Mortazavi, A., Myers, R. M., & Wold, B. (2007). Genome-Wide Mapping of in Vivo Protein-DNA Interactions. *Science*, 316(5830), 1497–1502. <https://doi.org/10.1126/science.1141319>
- Kanhere, A., Brookes, E., Fisher, A. G., Bouwman, R. D., Pereira, C. F., Pombo, A., ... Viiri, K. (2010). Short RNAs Are Transcribed from Repressed Polycomb Target Genes and Interact with Polycomb Repressive Complex-2. *Molecular Cell*, 38(5), 675–688. <https://doi.org/10.1016/j.molcel.2010.03.019>
- Kao, S.-Y., Calman, A. F., Luciw, P. A., & Peterlin, B. M. (1987). Anti-termination of transcription within the long terminal repeat of HIV-1 by tat gene product. *Nature*, 330(6147), 489–493. <https://doi.org/10.1038/330489a0>

- Kapranov, P., Cheng, J., Dike, S., Nix, D. A., Duttagupta, R., Willingham, A. T., ... Gingeras, T. R. (2007). RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science*, *316*(5830), 1484–1488. <https://doi.org/10.1126/science.1138341>
- Kim, T. H., Barrera, L. O., Zheng, M., Qu, C., Singer, M. A., Richmond, T. A., ... Ren, B. (2005). A high-resolution map of active promoters in the human genome. *Nature*, *436*(7052), 876–880. <https://doi.org/10.1038/nature03877>
- Komarnitsky, P., Cho, E. J., & Buratowski, S. (2000). Different phosphorylated forms of RNA polymerase II and associated mRNA processing factors during transcription. *Genes & Development*, *14*(19), 2452–2460. <https://doi.org/10.1101/gad.824700>
- Krebs, A. R., Imanci, D., Hoerner, L., Gaidatzis, D., Burger, L., & Schübeler, D. (2017). Genome-wide Single-Molecule Footprinting Reveals High RNA Polymerase II Turnover at Paused Promoters. *Molecular Cell*, *67*(3). <https://doi.org/10.1016/j.molcel.2017.06.027>
- Krumm, A., Meulia, T., Brunvand, M., & Groudine, M. (1992). The block to transcriptional elongation within the human c-myc gene is determined in the promoter-proximal region. *Genes & Development*, *6*(11), 2201–2213. <https://doi.org/10.1101/gad.6.11.2201>
- Kwak, H., Fuda, N. J., Core, L. J., & Lis, J. T. (2013). Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science*, *339*(6122), 950–953. <https://doi.org/10.1126/science.1229386>
- Law, A., Hirayoshi, K., O'Brien, T., & Lis, J. T. (1998). Direct cloning of DNA that interacts in vivo with a specific protein: application to RNA polymerase II and sites of pausing in *Drosophila*. *Nucleic Acids Research*, *26*(4), 919–924. <https://doi.org/10.1093/nar/26.4.919>
- Lee, T. I., & Young, R. A. (2013). Transcriptional Regulation and Its Misregulation in Disease. *Cell*, *152*(6), 1237–1251. <https://doi.org/10.1016/J.CELL.2013.02.014>
- Mahat, D. B., Kwak, H., Booth, G. T., Jonkers, I. H., Danko, C. G., Patel, R. K., ... Lis, J. T. (2016). Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-Seq). *Nature Protocols*, *11*(8), 1455–1476. <https://doi.org/10.1038/nprot.2016.086>
- Margaritis, T., & Holstege, F. C. P. (2008). Poised RNA Polymerase II Gives Pause for Thought. *Cell*, *133*(4), 581–584. <https://doi.org/10.1016/j.cell.2008.04.027>

- Marshall, N. F., Peng, J., Xie, Z., & Price, D. H. (1996). Control of RNA polymerase II elongation potential by a novel carboxyl-terminal domain kinase. *The Journal of Biological Chemistry*, *271*(43), 27176–27183. <https://doi.org/10.1074/jbc.271.43.27176>
- Mayer, A., Di Iulio, J., Maleri, S., Eser, U., Vierstra, J., Reynolds, A., ... Churchman, L. S. (2015). Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution. *Cell*, *161*(3), 541–554. <https://doi.org/10.1016/j.cell.2015.03.010>
- Mugridge, J. S., & Gross, J. D. (2013). Judge, Jury and Executioner: DXO functions as a decapping enzyme and exoribonuclease in pre-mRNA quality control. *Mol Cell*, *50*(1), 2–4. <https://doi.org/10.1016/j.molcel.2013.03.025>
- Muse, G. W., Gilchrist, D. A., Nechaev, S., Shah, R., Parker, J. S., Grissom, S. F., ... Adelman, K. (2007). RNA polymerase is poised for activation across the genome. *Nature Genetics*, *39*(12), 1507–1511. <https://doi.org/10.1038/ng.2007.21>
- Nechaev, S., & Adelman, K. (2008). Promoter-proximal Pol II: When stalling speeds things up. *Cell Cycle*, *7*(11), 1539–1544. <https://doi.org/10.4161/cc.7.11.6006>
- Nechaev, S., Fargo, D. C., Santos, G. Dos, Liu, L., Gao, Y., & Adelman, K. (2010). Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in *Drosophila*. *Science*, *327*(5963), 335–338. <https://doi.org/10.1126/science.1181421>
- Nojima, T., Gomes, T., Grosso, A. R. F., Kimura, H., Dye, M. J., Dhir, S., ... Proudfoot, N. J. (2015). Mammalian NET-seq reveals genome-wide nascent transcription coupled to RNA processing. *Cell*, *161*(3), 526–540. <https://doi.org/10.1016/j.cell.2015.03.027>
- Orphanides, G., & Reinberg, D. (2002). A unified theory of gene expression. *Cell*, *108*(4), 439–451. [https://doi.org/10.1016/S0092-8674\(02\)00655-4](https://doi.org/10.1016/S0092-8674(02)00655-4)
- Paulsen, M. T., Veloso, A., Prasad, J., Bedi, K., Ljungman, E. A., Magnuson, B., ... Ljungman, M. (2014). Use of Bru-Seq and BruChase-Seq for genome-wide assessment of the synthesis and stability of RNA. *Methods*, *67*(1), 45–54. <https://doi.org/10.1016/j.ymeth.2013.08.015>
- Paulsen, M. T., Veloso, A., Prasad, J., Bedi, K., Ljungman, E. A., Tsan, Y.-C., ... Ljungman, M. (2013). Coordinated regulation of synthesis and stability of RNA during the acute TNF-induced proinflammatory response. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(6), 2240–2245. <https://doi.org/10.1073/pnas.1219192110>

- Pertea, M. (2012). The human transcriptome: an unfinished story. *Genes*, 3(3), 344–360. <https://doi.org/10.3390/genes3030344>
- Phatnani, H. P., & Greenleaf, A. L. (2006). Phosphorylation and functions of the RNA polymerase II CTD. *Genes & Development*, 20(21), 2922–2936. <https://doi.org/10.1101/gad.1477006>
- Picard-Jean, F., Brand, C., Tremblay-Létourneau, M., Allaire, A., Beaudoin, M. C., Boudreault, S., ... Bisailon, M. (2018). 2'-O-methylation of the mRNA cap protects RNAs from decapping and degradation by DXO. *PLOS ONE*, 13(3), e0193804. <https://doi.org/10.1371/journal.pone.0193804>
- Preker, P., Mapendano, C. K., Schierup, M. H., Kammler, S., Nielsen, J., Jensen, T. H., ... Christensen, M. S. (2008). RNA Exosome Depletion Reveals Transcription Upstream of Active Human Promoters. *Science*, 322(5909), 1851–1854. <https://doi.org/10.1126/science.1164096>
- Project, A. S. H. L. E. T. (2009). Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs. *Nature*, 457(7232), 1028–1032. <https://doi.org/10.1038/nature07759>
- Rasmussen, E. B., & Lis, J. T. (1993). In vivo transcriptional pausing and cap formation on three Drosophila heat shock genes (RNA polymerase II/elongation arrest/RNA capping). *Biochemistry*, 90, 7923–7927. Retrieved from <https://www.pnas.org/content/pnas/90/17/7923.full.pdf>
- Reuter, J. A., Spacek, D. V., & Snyder, M. P. (2015). High-Throughput Sequencing Technologies. *Molecular Cell*, 58(4), 586–597. <https://doi.org/10.1016/J.MOLCEL.2015.05.004>
- Rougvie, A. E., & Lis, J. T. (1988). The RNA polymerase II molecule at the 5' end of the uninduced hsp70 gene of *D. melanogaster* is transcriptionally engaged. *Cell*, 54(6), 795–804. [https://doi.org/10.1016/S0092-8674\(88\)91087-2](https://doi.org/10.1016/S0092-8674(88)91087-2)
- Samarakkody, A., Abbas, A., Scheidegger, A., Warns, J., Nnoli, O., Jokinen, B., ... Nechaev, S. (2015). RNA polymerase II pausing can be retained or acquired during activation of genes involved in the epithelial to mesenchymal transition. *Nucleic Acids Research*, 43(8), 3938–3949. <https://doi.org/10.1093/nar/gkv263>
- Scheidegger, A., Dunn, C. J., Samarakkody, A., Koney, N. K.-K., Perley, D., Saha, R. N., & Nechaev, S. (2019). Genome-wide RNA pol II initiation and pausing in neural progenitors of the rat. *BMC Genomics*, 20(1), 477. <https://doi.org/10.1186/s12864-019-5829-4>

- Schroeder, S. C., Schwer, B., Shuman, S., & Bentley, D. (2000). Dynamic association of capping enzymes with transcribing RNA polymerase II. *Genes & Development*, *14*(19), 2435–2440. <https://doi.org/10.1101/gad.836300>
- Scruggs, B. S., Nechaev, S., Gilchrist, D. A., Burkholder, A., Muse, G. W., Adelman, K., & Fargo, D. C. (2015). Bidirectional Transcription Arises from Two Distinct Hubs of Transcription Factor Binding and Active Chromatin. *Molecular Cell*, *58*(6), 1101–1112. <https://doi.org/10.1016/j.molcel.2015.04.006>
- Seila, A. C., Calabrese, J. M., Levine, S. S., Yeo, G. W., Rahl, P. B., Flynn, R. A., ... Sharp, P. A. (2008). Divergent transcription from active promoters. *Science*, *322*(5909), 1849–1851. <https://doi.org/10.1126/science.1162253>
- Sharma, S., Kelly, T. K., & Jones, P. A. (2009, January 1). Epigenetics in cancer. *Carcinogenesis*. Narnia. <https://doi.org/10.1093/carcin/bgp220>
- Steurer, B., Janssens, R. C., Geverts, B., Geijer, M. E., Wienholz, F., Theil, A. F., ... Marteijn, J. A. (2018). Live-cell analysis of endogenous GFP-RPB1 uncovers rapid turnover of initiating and promoter-paused RNA Polymerase II. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(19), E4368–E4376. <https://doi.org/10.1073/pnas.1717920115>
- Strobl, L. J., & Eick, D. (1992). Hold back of RNA polymerase II at the transcription start site mediates down-regulation of c-myc in vivo. *The EMBO Journal*, *11*(9), 3307–3314. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/1505520>
- Szczepińska, T., Kalisiak, K., Tomecki, R., Labno, A., Borowski, L. S., Kulinski, T. M., ... Dziembowski, A. (2015). DIS3 shapes the RNA polymerase II transcriptome in humans by degrading a variety of unwanted transcripts. *Genome Research*, *25*(11), 1622–1633. <https://doi.org/10.1101/gr.189597.115>
- Taft, R. J., Glazov, E. A., Cloonan, N., Simons, C., Stephen, S., Faulkner, G. J., ... Mattick, J. S. (2009). Tiny RNAs associated with transcription start sites in animals. *Nature Genetics*, *41*(5), 572–578. <https://doi.org/10.1038/ng.312>
- Taft, R. J., Simons, C., Nahkuri, S., Oey, H., Korbie, D. J., Mercer, T. R., ... Mattick, J. S. (2010). Nuclear-localized tiny RNAs are associated with transcription initiation and splice sites in metazoans. *Nature Structural & Molecular Biology*, *17*(8), 1030–1034. <https://doi.org/10.1038/nsmb.1841>
- Tome, J. M., Tippens, N. D., & Lis, J. T. (2018). Single-molecule nascent RNA sequencing identifies regulatory domain architecture at promoters and enhancers. *Nature Genetics*, *50*(11), 1533–1541. <https://doi.org/10.1038/s41588-018-0234-5>

- Valen, E., Preker, P., Andersen, P. R., Zhao, X., Chen, Y., Ender, C., ... Jensen, T. H. (2011). Biogenic mechanisms and utilization of small RNAs derived from human protein-coding genes. *Nature Structural and Molecular Biology*, *18*(9), 1075–1082. <https://doi.org/10.1038/nsmb.2091>
- Wada, T., Takagi, T., Yamaguchi, Y., Ferdous, A., Imai, T., Hirose, S., ... Handa, H. (1998). *DSIF, a novel transcription elongation factor that regulates RNA polymerase II processivity, is composed of human Spt4 and Spt5 homologs We report the identification of a transcription elongation factor from HeLa cell nuclear extracts that causes pausing of RNA polymerase II (Pol II) in conjunction with the.* Retrieved from [www.genesdev.org](http://www.genesdev.org)
- Wagschal, A., Rousset, E., Basavarajaiah, P., Contreras, X., Harwig, A., Laurent-Chabalier, S., ... Kiernan, R. (2012). Microprocessor, Setx, Xrn2, and Rrp6 cooperate to induce premature termination of transcription by RNAPII. *Cell*, *150*(6), 1147–1157. <https://doi.org/10.1016/j.cell.2012.08.004>
- Wuarin, J., & Schibler, U. (1994). Physical isolation of nascent RNA chains transcribed by RNA polymerase II: evidence for cotranscriptional splicing. *Molecular and Cellular Biology*, *14*(11), 7219–7225. <https://doi.org/10.1128/mcb.14.11.7219>
- Xie, M., Li, M., Vilborg, A., Lee, N., Shu, M.-D., Yartseva, V., ... Steitz, J. A. A. (2013). Mammalian 5'-Capped MicroRNA Precursors that Generate a Single MicroRNA. *Cell*, *155*(7), 1568–1580. <https://doi.org/10.1016/j.cell.2013.11.027>
- Yamaguchi, Y., Takagi, T., Wada, T., Yano, K., Furuya, A., Sugimoto, S., ... Handa, H. (1999). NELF, a multisubunit complex containing RD, cooperates with DSIF to repress RNA polymerase II elongation. *Cell*, *97*(1), 41–51. [https://doi.org/10.1016/S0092-8674\(00\)80713-8](https://doi.org/10.1016/S0092-8674(00)80713-8)
- Zamudio, J. R., Kelly, T. J., & Sharp, P. A. (2014). Argonaute-bound small RNAs from promoter-proximal RNA polymerase II. *Cell*, *156*(5), 920–934. <https://doi.org/10.1016/j.cell.2014.01.041>
- Zeitlinger, J., Stark, A., Kellis, M., Hong, J. W., Nechaev, S., Adelman, K., ... Young, R. A. (2007). RNA polymerase stalling at developmental control genes in the *Drosophila melanogaster* embryo. *Nature Genetics*, *39*(12), 1512–1516. <https://doi.org/10.1038/ng.2007.26>
- Zhu, Y., Pe'ery, T., Peng, J., Ramanathan, Y., Marshall, N., Marshall, T., ... Price, D. H. (1997). Transcription elongation factor P-TEFb is required for HIV-1 Tat transactivation in vitro. *Genes and Development*, *11*(20), 2622–2632. <https://doi.org/10.1101/gad.11.20.2622>