The Regulation of Premature Termination at Divergent Promoters

by

Anthony Chun-yin Chiu

Bachelor of Science
University of Toronto, 2009

Submitted to the Department of Biology
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

SEPTEMBER 2017

Signature of Author: Signature redacted
Department of Biology
August 18, 2017

Certified by: Signature redacted
Phillip A. Sharp
Institute Professor and Professor of Biology
Thesis Supervisor

Certified by: Signature redacted
Amy E. Keating
Professor of Biology
Co-Chair, Biology Graduate Committee

The Regulation of Premature Termination at Divergent Promoters

by

Anthony Chun-yin Chiu

Submitted to the Department of Biology on August 18, 2017
in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

ABSTRACT

Transcription is one of the most fundamental processes in cells, governing the conversion of genetic information to RNA. Numerous regulatory mechanisms function to ensure that desired transcripts are being expressed. Promoters transcribe divergently, producing low-abundant upstream antisense RNAs (uaRNAs) in addition to a stable downstream RNAs. Thus, a central question is what mechanisms are sense RNAs more stable compared to most transcription events. It is proposed that an asymmetric distribution of U1 snRNP binding sites and polyadenylation site (PAS) motifs known as the U1-PAS axis regulates early termination of RNA Polymerase II.

Here, we generated a conditional knockout of the essential RNA exosome subunit, Exosc3, in mouse embryonic stem cells. Removal of Exosc3 resulted in stabilization of polyadenylated uaRNAs, enhancer RNAs and long noncoding RNAs. In addition, promoter proximal pausing increased modestly upon Exosc3 removal. Interestingly, a large class of polyadenylated short transcripts in the sense direction terminate within the first intron, similar to premature termination observed upon U1 inhibition.

Further investigation of these prematurely termination sites revealed they are found at the edges of stable nucleosome free regions demarcated by CpG islands and are suppressed by U1 snRNP. Interestingly, promoter-proximal Pol II pausing consists of two processes: TSS-proximal and +1 stable nucleosome pausing. Genes associated with premature termination have increased +1 stable nucleosome pausing, association of chromatin remodelers and are more sensitive to inhibition by flavopiridol or a Myc inhibitor.

Additionally, the nuclear poly(A) binding protein, Pabpn1, promotes degradation of polyadenylated uaRNAs. Most Pabpn1 sensitive uaRNAs are also Exosc3 substrates, and sensitivity to Pabpn1 inhibition inversely correlates with the proximity of the termination site to the TSS. Interestingly, at uaRNAs and sense RNAs, Pabpn1-sensitive PAS termination events also occur near the first stable nucleosome, similar to Exosc3-sensitive PAS termination events, suggesting that Pabpn1 collaborates with Exosc3 to regulate stability of polyadenylated transcripts.

Hence, this supports a model whereby U1 snRNP, +1 stable nucleosomes and the degradation machinery converge to create a transcription elongation checkpoint downstream of promoter-proximal pausing.

Thesis Supervisor: Phillip A. Sharp
Title: Institute Professor, Professor of Biology

Acknowledgements

First, I must thank Phil for both being my mentor and an important role model over the past few years. I have truly appreciated his insights when things were confusing, as well as his constructive comments that helped me develop professionally. Tolkien once wrote "Not all who wander are lost." He must not have been referring to me, because I went on hundreds of failed wanderings, but Phil was gracious enough to let me go on them and then steering me back to the main question. I will always cherish the time spent in the Sharp Lab.

To my committee members, Tyler Jacks and Rick Young, I have truly appreciated their suggestions and support over the years. Both have been important role models, as I learned from the way they think and communicate about scientific problems, as well as the way they engage with the broader scientific community.

To my colleagues: you have been incredibly patient with my terrible jokes, non-sequiturs and less well-thought out questions. To bay mates Jeremy and Tim: thank you for teaching me every day and making our corner interesting. To Hiroshi: thank you for being an invaluable, late night collaborator, and for our scientific discussions that took this project to a new direction. To my collaborators, Xuebing, Andrea and Jay for their insights. And of course, the rest of the Sharpies for making this journey fun, in particular Margarita, Courtney, Albert, Mohini, Paige, Jay and Amanda, for their companionship.

There are many non-Sharpies that I need to thank. To the class of 2010, especially Courtney, Brian, Lynne, Greg, Daniel, Peter, Kalen and Monica. To everyone else on the 4th floor, especially those in the Jacks lab. To members of the Graduate Christian Fellowship, thanks for helping me grow spiritually. To other cohorts, such as the Sloan MBA community, the Tango community, thank you for giving me a break from the scientific life.

Finally, I am grateful to my parents for their unwavering support, patience and unfiltered suggestions throughout the years, which has made me into who I am. As well, my two siblings Jonathan and Veronica, who walked with me throughout my life.

Anthony

Table of Contents

Men love to wonder, and that is the seed of science.

-Ralph Waldo Emerson

Chapter 1

Introduction: Mechanisms of Transcription

This chapter provides a background to transcription and divergent transcription.

Transcription is one of the most fundamental processes in cells, involving selective amplification of DNA into transitory RNAs. A decade ago, our knowledge about transcription was based on a few principles. RNAs were selectively produced from a small fraction of the genome known as genes. RNA polymerase bound to a promoter, transcribed through the gene and released a transcript that was exported to the cytoplasm to be translated. The major mode of differential transcription involved the activity of gene specific transcription factors that recruit RNA polymerase to the promoter. Most of these principles were challenged with the development of a wide variety of genome-wide sequencing techniques (Table S1), where previous curiosities at individual genes were found to occur at far greater frequencies genome-wide.

Firstly, the depth of high-throughput sequencing revealed that transcription is *pervasive*, whereby greater than 70% of the human genome is transcribed, although most events result in low abundant transcripts (Consortium et al., 2007; Djebali et al., 2012). Secondly, most transcription events produce *noncoding RNAs (ncRNAs)*, including long intergenic noncoding RNAs (lincRNAs), upstream antisense RNAs (uaRNAs) and enhancer RNAs (eRNAs) (Almada et al., 2013; Consortium et al., 2007; Guttman et al., 2009; Kim et al., 2010b; Preker et al., 2008; Seila et al., 2008). Some non-coding RNAs have physiological functions; lincRNA-p21 modulates p53-dependent expression of p21 (Dimitrova et al., 2014). Thirdly, *post-initiation regulation* is a common mode of regulation. Engaged RNA polymerases are paused downstream of the transcription start site (TSS) in metazoans (Muse et al., 2007; Zeitlinger et al., 2007). Regulated splicing of detained introns prevents export of transcripts from the nucleus (Boutz et al., 2015). The RNA exosome has critical roles in regulating steady state RNA levels of

numerous noncoding transcripts (Almada et al., 2013; Gudipati et al., 2012; Preker et al., 2008; Schneider et al., 2012).

Lastly, most mammalian transcription is *divergent*. Work from our lab as well as others found most expressed mammalian promoters and active enhancers transcribe divergently (Core et al., 2008; Djebali et al., 2012; Kim et al., 2010b; Preker et al., 2008; Seila et al., 2008). For instance, divergent promoters generate a low-abundant upstream antisense RNA (uaRNA[1]) and a higher-abundant sense mRNA transcript. In addition, enhancers produce low-abundant divergent enhancer RNAs (Kim et al., 2010b), raising the interesting question about what mechanisms differentiate promoters from enhancers as both possess transcription activity.

My thesis began by aiming to understand what made uaRNAs different from sense mRNAs. An understanding of this process will reveal fundamental insights into the processes cells utilize to determine whether or not to produce substantial RNAs. Moreover, understanding transcriptional regulation has important implications in disease pathology, as a surprising number of mutations linked with tumorigenesis involve mutations in the transcription machinery, often by increasing transcription activity. We serendipitously found evidence of a novel transcriptional checkpoint associated with divergent transcription, nucleosomes and polymerase pausing. In the introduction, I will discuss general properties of transcription, initial discoveries of divergent transcription and outline the various steps of the transcription cycle, focusing on the relationship to divergent transcription. I will conclude by recapping the key questions I sought to address.

[1] uaRNAs are called promoter proximal transcripts (PROMPTs) in humans. I will mostly use uaRNA except if referring to human experiments.

1.1 Transcription Fundamentals

RNA Polymerase II

RNA polymerase is the enzymatic complex that catalyzes the polymerization of ribonucleotides to form the nascent RNA transcript. There are 3 different eukaryotic RNA polymerases, which transcribe different types of genes (Roeder and Rutter, 1969, 1970). Of the three, RNA polymerase II (Pol II[2]) transcribes mRNA genes as well as various noncoding RNAs including lncRNAs and snRNAs, whereas Pol I transcribes the 45S rRNA precursor and Pol III transcribes tRNAs, U6 snRNA and 5.8S rRNAs. Given divergent transcription is primarily associated with Pol II, we will focus primarily on Pol II transcription, though Pol I and Pol III transcription are critical in cells.

Pol II is comprised of 12 subunits (Rpb1-Rpb12[3]) (Kolodziej et al., 1990; Sayre et al., 1992). The core subunits are comprised of Rpb1, Rpb2 and Rpb3, and are evolutionarily conserved all the way to the *E coli* RNA polymerase (Young, 1991). During transcription, DNA and RNA form a DNA:RNA hybrid within the core, and free nucleotides reach the active site through a pore in Pol II (Gnatt et al., 2001). After polymerization, RNA polymerase translocates so that an accessible 3' OH is in the active site. The stability of the hybrid in the elongation complex provides the energetics to drive translocation reaction rather than ATP hydrolysis (Nudler, 2012), so if Pol II transcription slows down due to a pause, the 3' end of the RNA transcript can be extruded from a secondary channel. In this state, Pol II is stalled, since it is incapable of transcribing yet it remains bound to DNA. This situation is resolved by cleavage of

[2] There are three naming conventions for the abbreviation of RNA polymerase II in the scientific literature: RNAPII, Pol2 and Pol II. In this thesis, I will be using Pol II.

[3] In this thesis, the convention for proteins in general will be normal font, first letter capitalized, except when talking about human proteins only, where it will be capitalized. The yeast literature alters between SGD convention (Mtr4p) or Mtr4, so for consistency with mammalian proteins, I will use Mtr4.

the extruded RNA through the activity of TFIIS, allowing Pol II to continue elongating (Izban and Luse, 1992; Reinberg and Roeder, 1987).

Early studies on purified mammalian Pol II found that the Rpb1 ran as multiple bands on a gel (Sklar et al., 1975). Subsequent studies revealed that the upper bands arise from extensive phosphorylation of the Rpb1 subunit at the disordered C-terminal repeat domain (CTD), made of 26-52 repeats of the heptad YSPTSPS (Young, 1991). Each of these 7 positions are post-translationally modified during transcription, and together forms a CTD code that couples RNA processing events to various steps of RNA transcription (Buratowski, 2003). For instance, Ser5P recruits the enzyme linked with 5' capping of RNA (Cho et al., 1997; McCracken et al., 1997a) whereas Ser2P recruits termination factors such as Pcf11 (Ahn et al., 2004; Barilla et al., 2001; McCracken et al., 1997b). Recent genome-wide studies have profiled the spatial association of most of these modifications to specific steps of the transcription cycle (Bataille et al., 2012; Kim et al., 2010a; Mayer et al., 2010; Schlackow et al., 2017). In particular, Ser5P and Ser7P are found at the 5' end of genes whereas Ser2P builds up over transcription and peaks at the 3' end of genes.

The mammalian transcription cycle occurs in 5 steps (**Fig. 1**). In *transcription initiation*, Pol II is recruited to accessible promoters, the double-stranded DNA is unwound, Pol II is loaded onto the DNA forming a preinitiation complex, and Pol II is released from the promoter by TFIIH. Next, *promoter proximal pausing* occurs when Pol II arrests 30-60 nucleotides downstream of transcription start sites (TSS). Release from the pause is induced by the activity of P-TEFb. Subsequently, Pol II transcribes until it stalls at the *+1 stable nucleosome*. During *productive elongation*, Pol II produces the majority of the pre-mRNA transcript while transcribing through nucleosomes. Lastly, *transcription termination* occurs after Pol II

encounters a termination signal, causing cleavage and polyadenylation of the pre-mRNA, and mRNA export. The exposed 5' end of the nascent RNA allows a 5'-to-3' exoribonuclease to displace Pol II from the DNA.

Chromatin

DNA does not exist naked in the nucleus, but is rather wrapped into nucleosomes to form chromatin, enabling a long, linear molecule to be packaged into a small nucleus. The nucleosome is usually comprised of a histone octamer, made of a tetramer core of two H3/H4 heterodimers, and two H2A/H2Bdimers (Luger et al., 1997). The centre of the histone octamer core is called the dyad axis, around which wraps 147 bp of DNA. There are also reports of subnucleosomes or nucleosomal hexamers, which may occur during transcription when Pol II promotes eviction of H2A/H2B dimers.

Nucleosomes are a physical block that impacts all nuclear processes involving the DNA. In vitro transcription experiments found transcription on a chromatinized template was significantly less efficient than transcribing naked DNA, suggesting that nucleosomes create a barrier to Pol II elongation (Izban and Luse, 1991, 1992). Subsequent work would demonstrate that this barrier occurs at a pause site as Pol II enters +40 to +50 bp into the nucleosome due to an interaction with the H3/H4 tetramer core (Bondarenko et al., 2006; Kireeva et al., 2005), which has been confirmed by genome-wide studies of Pol II stalling (Weber et al., 2014).

The nucleosome barriers are regulated in three ways. First, adjusting the composition of the nucleosome particle changes the nucleosome barrier. Processes that introduce the histone variants such as H3.3 and H2A.Z near the promoter create highly unstable nucleosomes (Jin and Felsenfeld, 2007). Secondly, various external enzymes can reduce the barrier (Selth et al., 2010).

Chromatin remodelers use ATP to slide nucleosomes along DNA; mammalian CHD1 has been linked with promoting nucleosome exchange near the promoter (Skene et al., 2014). Alternatively, *histone chaperones* promote the eviction and reassembly of nucleosomes. FACT is an elongation factor that promotes the removal and reassembly of H2A/H2B dimers during transcription elongation (Belotserkovskaya et al., 2003). Lastly, histones possess N-terminal tails that are posttranslationally modified during transcription. These modifications can directly regulate transcription by altering compaction of nucleosomes or alternatively by creating platforms to recruit other proteins that regulate chromatin. Histones are cotranscriptionally modified in the transcription cycle; genome-wide profiling of many histone marks have identified roles for these variants: H3K4me3 and H3K9/14Ac are initiation marks[4] and are found at the 5' ends of genes, H3K36me3 and H3K79me2 are elongation marks of active genes whereas H3K27me3 and H3K9me2 are repressive marks (Barski et al., 2007; Bernstein et al., 2005; Guenther et al., 2007).

Two terms are frequently used to describe the spatial orientation of nucleosomes. *Nucleosome positioning* refers to the precise location of a nucleosome on the DNA compared to the average. If the nucleosome is highly positioned or if deviation is low, in most cells, there will be a nucleosome positioned at roughly the same position. In contrast, if the nucleosome is not highly positioned or deviation is high, then nucleosomes do not necessarily bind to the same series of nucleotides; regions with low nucleosome positioning are said to have *fuzzy nucleosomes*. In contrast, *nucleosome occupancy* refers to the propensity of nucleosomes to associate with a specific base, and indicated by the signal intensity at a specific base. There are

[4] Technically, H3K4me3 is not a mark of initiation in mammals, because it is found at all CpG island promoters. Rather, it is a mark of the 5' end of genes.

many reasons why there would be low occupancy: they could be *unstable nucleosomes* or *subnucleosomes*.

The binding of nucleosomes to DNA is influenced by both intrinsic (sequence-specific) and extrinsic (chromatin remodelers) factors. Intrinsically, nucleosomes disfavor binding to regions with long poly(dA-dT) tracts, due to their rigidity (Kaplan et al., 2009). In addition, nucleosomes disfavor binding to DNA sequences with many CpG dinucleotides (Ramirez-Carrozzi et al., 2009). Lastly, nucleosomes prefer to bind to regions with phased AA/TT/TA dinucleotide sequences every 10 bps, due to differential flexibility of these respective sequences (Satchwell et al., 1986; Segal et al., 2006). Extrinsically, proteins can sterically block nucleosome assembly. Pol II binding near the TSS creates a strongly positioned nucleosome immediately downstream of the TSS (Schones et al., 2008).

1.2 Divergent Transcription

Divergent transcription in mammals was discovered independently in three labs, each using different genome-wide methodologies (Core et al., 2008; Preker et al., 2008; Seila et al., 2008). While analyzing the genome-wide production of small RNAs in mESCs, the Sharp Lab serendipitously discovered that many small RNAs were produced around the transcriptions start site of genes (Seila et al., 2008). Metagene alignments of these transcription start-site associated RNAs (TSSa-RNAs) revealed a sharp peak 50 nts downstream of the TSS and a broader spread 200 nts upstream of the TSS. Importantly, 67% of transcribed genes were found to produce TSSa-RNAs in both directions, suggesting divergent transcription is a common feature of mammalian transcription. In parallel, the Lis lab was developing techniques to study promoter-proximal pausing in a human cell line, IMR90 (Core et al., 2008). To measure nascent

transcription, they combined the nuclear run on assay with high-throughput sequencing, developing Global Run-On sequencing (GRO-seq). Similarly, Pol II was engaged and actively transcribed in both directions, peaking +50 and -250 from the TSS. This assay showed that divergent transcription was common, whereby 77% of active genes or 55% of all promoters produced divergent RNAs. Lastly, the Jensen lab was trying to determine the roles of the RNA exosome in humans. They depleted the core exosome subunits (hRRP40 or hRRP44) using siRNAs in HeLa cells and assayed gene expression using tiling arrays on poly(A)-selected RNA. Surprisingly, many RNAs were upregulated 0.5-2.5 kb upstream from the TSS, so they called this class of RNAs promoter upstream transcripts[5] (PROMPTs). Unlike the other two studies, these transcription events occurred in both directions over the PROMPT region. In hindsight, this may be from contamination with eRNAs and from higher background of tiling arrays.

All three studies found CpG islands promoters were highly correlated with divergent transcription. Moreover, H3K4me3 and Pol II were present in two peaks of similar amounts around divergent TSSs (Core et al., 2008; Seila et al., 2008) and studies on individual uaRNAs found that they were capped (Flynn et al., 2011; Preker et al., 2011), arguing that uaRNAs had undergone transcription initiation similar to mRNAs. In contrast, histone marks of elongation (H3K79me2 and H3K36me3) were depleted in the upstream antisense direction when compared to the sense direction, arguing that the upstream antisense Pol II was not undergoing productive elongation.

Since then, studies in other labs demonstrate that yeast also have widespread divergent transcription (Neil et al., 2009; Xu et al., 2009). In contrast, GRO-seq in Drosophila has been unable to detect widespread divergent transcription, possibly because they have different

[5] For this thesis, these upstream antisense transcripts will be called uaRNAs if either I am describing mouse results or more generally, upstream antisense transcripts in mammals. Occasionally, PROMPTs will be used when the results are human specific. In contrast, the sense transcript will always be called mRNAs.

promoter structures than mammals. Studies into divergent transcription gained prominence when a landmark study in 2010 found that activated neurons produce low copy, bidirectional enhancer RNAs (eRNAs) from enhancers (Kim et al., 2010b). Similar to uaRNAs, eRNAs are regulated by the RNA exosome (Andersson et al., 2014). Despite being low abundant, there are hints in the literature that some eRNAs and uaRNAs may be functional (Schaukowitch et al., 2014). Additionally, most lncRNAs originate from divergent transcription and are initiate similarly with the sense transcript (Sigova et al., 2013). Altogether, uaRNAs and eRNAs have similar functional properties and the biogenesis of both classes of RNAs may be linked. Thus, a key avenue of research focuses on what makes eRNAs and uaRNAs unstable, or alternatively, why are sense mRNAs stable if the majority of transcription events in mammals unstable.

To address, we will discuss the steps of transcription, while focusing on the differences between sense mRNA transcription and uaRNA transcription.

1.3 Transcription Initiation

When divergent transcription was first discovered, all three studies noted that divergent transcription occurred mostly at CpG islands. In contrast, Drosophila do not have CpG islands and do not appear to have divergent transcription. It is unlikely to be due to low sequencing depth as the same technique, GRO-seq, was used to discover divergent transcription in mammals so it is important to discuss the process by which Pol II initiates, with a focus on CpG islands.

Two Classes of Promoters

The promoter refers to the region of DNA from where transcription initiates, and includes not only the transcription start site (TSS) but also gene regulatory elements to which basal

transcription factors[6] and gene-specific transcription factors (TFs) bind (Kadonaga, 2012). Historically, the core promoter element is thought to comprise 3 components: an upstream motif known as the TFIIB recognition element (BRE), a TATA-box, and an Initiator element. These motifs collaborate to recruit the basal transcription factors (TFIIA, TFIIB, TFIID, TFIIE, TFIIE, TFIIH), Pol II and Mediator. TFIID is the main DNA-binding complex, which bind directionally to TATA boxes through its TBP subunit, whereas the BRE binds TFIIB. In addition to the core promoter elements, distal elements such as enhancers increase the likelihood that the basal transcription machinery is recruited to the core promoter through DNA looping and interactions with the Mediator and TFIID. The precise assembly pathway in vivo is unclear, but recent studies suggest the sequential assembly of basal transcription factors and Pol II to promoters is gene specific. Nevertheless, the association of all basal transcription factors (or their functional orthologs) to promoters is required to initiate transcription.

For some time, it was known that CpG dinucleotides are selected against due to their higher mutagenic properties, except at the 5' ends of genes where they cluster at CpG islands (CGI). An important study by Carninici linked CpG islands to different types of transcription initiation (Carninci et al., 2006). Genome-wide mapping of the 5' ends of capped RNAs suggested that most TSSs in mammals were distributed in two major types of promoters: promoters with narrow initiation around a single site (focal promoters) and promoters with dispersed initiation (up to 100 nts). While both promoter classes have Initiator elements, focal promoters uniquely possess TATA-boxes whereas almost all dispersed promoters were associated with CpG islands.

[6] In earlier literature, basal transcription factors are called general transcription factors. However, not all promoters have these 'general' transcription factors. For example, there are functional alternatives to TFIID. Consequently, Kadonaga advocates the term basal transcription factor to reflect this new understanding.

Mammalian promoters now are classified into TATA-containing and CGI promoters, whereby the majority of promoters are CpG island promoters (Saxonov et al., 2006). CGI promoters are older and associated with house-keeping genes, whereas TATA-containing genes are associated with tissue-specific genes. The same basal transcription factors that bind to TATA promoters also bind to CpG islands promoters, but it is currently unclear what sequence elements aside from the Initiator are important for transcription initiation in CpG island promoters.

Nucleosome Structure at Promoters

Genome-wide studies demonstrate that the average gene has a depletion of nucleosomes around the TSS, forming a Nucleosome Free Region[7] (NFR) (Kaplan et al., 2009; Schones et al., 2008; Yuan et al., 2005). The +1 nucleosome refers to the first nucleosome downstream of the TSS, whereas the first nucleosome upstream is the -1 nucleosome. There is substantial evidence that initiating Pol II binds immediately upstream of the +1 nucleosome, both from precise mapping of the basal transcription factors using ChIP-exo (Rhee and Pugh, 2012) and from GRO-cap of nascent transcripts (Core et al., 2014). Thus, the open nature of the NFR permits Pol II to access sequence elements necessary for initiating transcription and is critical for positioning the pre-initiation complex (PIC).

Coactivators collaborate with basal transcription factors and Pol II to promote initiation. Many coactivators regulate access to nucleosome-bound promoters, by functioning as histone modifying enzymes (ex. PRMT5, GCN5 or p300/CBP) or as chromatin remodelers (ex. SWI/SNF). In mammals, analysis of the activation of the IFN-β gene revealed a sequential recruitment of factors, culminating in the eviction of the +1 nucleosome at the TATA promoter

[7] Alternatively called Nucleosome Depleted Region (NDR). In this thesis, I will use NFR and later introduce the term Stable Nucleosome Free Region (SNFR) for the nucleosome depletion over CpG islands.

(Agalioti et al., 2000). Initially GCN5 acetylates the +1 nucleosome, which recruits the Pol II holoenzyme and p300/CBP. Subsequently, SWI/SNF is recruited to the promoter to move the +1 nucleosome, uncovering the TATA box and allowing TBP to bind. Similarly SWI/SNF has been shown to function in regulating other TATA promoters by a similar mechanism (Ramirez-Carrozzi et al., 2009).

Unlike TATA-promoters where nucleosomes must be removed, mammalian CpG islands generally have lower nucleosome occupancy and are SWI/SNF independent. Work analyzing LPS-induced genes in human macrophages found a depletion of nucleosomes across the CpG island (Ramirez-Carrozzi et al., 2009). CpG rich sequences were found to disfavor the assembly of nucleosomes in an *in vitro* nucleosome assembly assay (Ramirez-Carrozzi et al., 2009). Another study found a strong correlation between the length of CpG islands and the NFR (Fenouil et al., 2012). As the GC content of the CGI increased, the position of the peak nucleosome signal moved to the +2, +3 then +4 nucleosome position, though there remains a weakly associating +1 nucleosome immediately positioned after the TSS. Two independent methods confirmed that the reduction in nucleosome occupancy was robust, rather than an artifact of MNase. Hence, this depletion likely explains why CGI promoters do not have a requirement for the SWI/SNF complex to evict nucleosomes, since it is already nucleosome depleted to promote efficient transcription (Ramirez-Carrozzi et al., 2009). It should be noted that despite there being a depletion for nucleosomes across the entire CpG island, there is still a stronger NFR around 110 bp between the +1 and -1 nucleosome within the CpG, which can be detected using DNase or lower concentrations of MNase (Core et al., 2014; de Dieuleveult et al., 2016).

Irrespective of gene activity, CGIs are enriched for H3K4me3 and acetylated histones, and depleted for H3K36me2 (Guenther et al., 2007; Mikkelsen et al., 2007). Many of these marks are deposited by proteins which recognized unmethylated CpGs. Deposition of H3K4me3 at CGIs is promoted through an interaction of CpG dinucleotides with Cfp1, a subunit of the H3K4me3 histone methyltransferase complex (Thomson et al., 2010), whereas active demethylation of H3K36 at CGIs occurs due to the binding of the H3K36 demethylase KDM2A (Blackledge et al., 2010). At promoters, H3K4me3 can recruit the H4 histone acetyltransferase HBO1, the chromatin remodeling complex CHD1 and the basal transcription factor TFIID to CpG islands (Vermeulen et al., 2007). In contrast, H3K36 methylation suppresses transcription initiation by recruiting a histone deacetyltransferase (Carrozza et al., 2005; Keogh et al., 2005). Thus, CGIs promote initiation by being refractory to stable nucleosome assembly, by promoting histone marks that promote nucleosome remodeling, and by inhibiting pathways that suppress initiation.

Thus, TATA promoters are regulated by eviction of the +1 nucleosome whereas CpG promoters already have lower nucleosome occupancy. This is consistent with observations that CGI promoters are found mostly at housekeeping genes that are constantly active, whereas TATA promoters function in regulated transcription. Divergent transcription in mammals is strongly biased to CGI island promoters, likely because it is easier to initiate transcription divergently if the promoter already possesses unstable nucleosomes (Core et al., 2008; Preker et al., 2008; Seila et al., 2008). In contrast, nucleosomes are typically bound to TATA promoters and must be evicted for transcription to initiate.

The +1 Nucleosome Barrier: Roles of H2A.Z, H3K56 and Chd1

Studies into precise positioning of Pol II revealed that Pol II frequently backtracks at nucleosomes. The largest nucleosomal barrier occurs at the +1 nucleosome (Weber et al., 2014). There are several approaches for reducing the +1 nucleosome barrier, many of which have been linked with divergent transcription.

H2A.Z is most frequently found at the +1 nucleosome and -1 nucleosome at both active and inactive genes (Barski et al., 2007; Mavrich et al., 2008; Raisner et al., 2005). Incorporation of the histone variant H2A.Z reduces the +1 nucleosome barrier in Drosophila (Weber et al., 2014). Weber found a strong anti-correlation between H2A.Z signal and Pol II stalling at the +1 nucleosome, and that reduction of H2A.Z by siRNA was associated with increased Pol II stalling. Studies of nucleosome core particles revealed nucleosomes containing H2A.Z or another histone variant, H3.3, are highly unstable (Jin and Felsenfeld, 2007), supporting the idea that H2A.Z reduces the nucleosome barrier. The distribution of H2A.Z around the TSS is due to the activities of the SWR complex in yeast (SRCAP complex and Ep400 in mammals), which binds to the edges of the NFR (Ranjan et al., 2013; Yen et al., 2013). Consistent with this, ChIP-seq data showing that Ep400 associates with the edges of the NFR in mESCs (de Dieuleveult et al., 2016).

The roles of H2A.Z and H3K56ac were more apparent in two recent papers suggesting that chromatin-based mechanisms regulate promoter directionality. In a yeast genome-wide screen, Marquardt found mutations that modulated the expression of divergent noncoding RNAs were enriched for components of the H3K56ac-nucleosome assembly pathway (Marquardt et al., 2014). In this pathway, H3K56 is acetylated by Rtt109 in *S cerevisiae* or p300/CBP in mammals and is incorporated into chromatin by CAF-I (Das et al., 2009; Li et al., 2008). In addition, the -1

nucleosome was also regulated by the SWI/SNF complex (Marquardt et al., 2014), perhaps because H3K56ac regulates the activity of SWI/SNF (Xu et al., 2005).

In another paper, mutations of Rtt109 reduced transcription overall, yet effects on steady state RNAs were only observed in the absence of the RNA exosome (Rege et al., 2015). This suggested that a subset of transcripts are degraded by the RNA exosome when there is high Pol II density, perhaps due to increased collision frequency and greater Pol II stalling, which results in exosome-mediated degradation (Lemay et al., 2014). This study also found that removing both H2A.Z and the RNA exosome resulted in similar phenotypes as loss of H3K56ac and the RNA exosome. Importantly, the upregulation of uaRNAs upon removal of the RNA exosome was suppressed in the absence of H2A.Z, suggesting that divergent transcription requires a permissive initiation environment at the -1 nucleosome at mammalian promoters.

Other chromatin remodelers are linked with regulating the +1 nucleosome. In one study, knockdown of the chromatin remodeler Smarca4 (also called esBAF) in mESCs resulted in delocalized nucleosomes in the gene body and increased transcription at both sense mRNAs and uaRNAs from NFRs genome-wide (Hainer et al., 2015). In mouse, studies of a dominant negative mutant of CHD1 revealed that the chromatin remodeler regulates the +1 nucleosome stall and is responsible for the majority of Pol II-directed nucleosome turnover around promoters (Skene et al., 2014). The impact of CHD1 on the +1 nucleosome has not been observed in yeast, likely because mammalian Chd1 has a chromodomain that binds to promoter-proximal H3K4me3 signals (Flanagan et al., 2005).

Thus, the regulation of the +1/-1 nucleosome barrier through nucleosome dynamics is critical for modulating transcription. Reducing the amount of -1 nucleosomes by reducing its incorporation or promoting its disassembly results in increased divergent transcription.

Steps in Transcription Initiation

After the DNA sequence is made accessible, transcription factors synergistically recruit the general transcription apparatus, usually through interactions with TFIID and Mediator. Mediator is the central chaperone of this machinery, recruiting basal transcription factors as well as Pol II through its CTD (Kim et al., 1994; Thompson et al., 1993). Mediator also recruits the key initiator of transcription, TFIIH. TFIIH uses its helicase activity to unwind DNA, creating the open initiation complex. Pol II attempts to transcribe but is usually unsuccessful, because the initial RNA-DNA hybrid is too short to be stably bound (Luse, 2013). These abortive transcripts are up to 10 nts in length and occur for several cycles before creating a stable hybrid. In addition, TFIIH phosphorylates Ser5 of the CTD of Pol II, which allows promoter escape by weakening the interaction between Mediator and Pol II.

A 5' methylguanosine cap is added shortly after transcription initiation to most RNAs to prevent degradation by 5'-to-3' exonucleases. Recruitment of the capping enzyme occurs by a direct interaction with Ser5P modification on the Pol II CTD (Cho et al., 1997; Komarnitsky et al., 2000; McCracken et al., 1997a). Studies of individual divergent transcripts demonstrate that uaRNAs were capped (Flynn et al., 2011), suggesting that divergent transcripts have undergone productive initiation and their instability is not from being uncapped.

A central component of splicing, U1 snRNA, also regulates initiation. Work identifying noncoding RNAs that interact with TFIIH found that U1 snRNA promotes the initial catalytic steps during the abortive initiation phase (Kwek et al., 2002). Moreover, after promoter escape, portions of the preinitiation complex remains associated with the promoter for rapid re-initiation (Yudkovsky et al., 2000), a process enhanced by U1 snRNA (Damgaard et al., 2008; Kwek et al., 2002).

1.4 Promoter Proximal Pausing

Discovery

The Lis lab was interested in understanding the heat shock response in *Drosophila.* Using the Hsp70 model gene, heat shocking cells resulted in increased Pol II association throughout the gene body (Gilmour and Lis, 1986). Surprisingly, Pol II associated with the 5' end of the Hsp70 gene, even when the gene was 'off.' Furthermore, these polymerases were transcription competent (Rougvie and Lis, 1988), suggesting that Pol II is paused immediately after transcription initiation and poised for regulated activation..

High-throughput sequencing techniques reveal the majority of Pol II is paused near the TSS rather than spread out throughout the gene (Guenther et al., 2007; Muse et al., 2007; Zeitlinger et al., 2007), and are elongation competent (Core et al., 2008; Kwak et al., 2013). Functionally, genes with higher pausing have a Pol II ready for rapid activation. In *Drosophila*, genes with higher pausing respond to environmental or developmental stimuli (Muse et al., 2007; Zeitlinger et al., 2007), whereas in mESCs, genes with higher pausing tend to be components of ESC signaling pathways (Williams et al., 2015). Nevertheless, pausing occurs at all Pol II bound genes in mouse embryonic stem cells, since inhibitors of pause release universally blocked transcription elongation (Jonkers et al., 2014).

Mechanisms of Pausing

Investigations into the transcriptional inhibitor DRB revealed that two factors were necessary for the promoter-proximal pause: negative elongation factor (NELF) and DRB-sensitivity inducing factor (DSIF) (Wada et al., 1998a; Yamaguchi et al., 1999). NELF is specific to higher metazoans, whereas DSIF is made up of Spt4/Spt5. Spt5 binds to nascent RNAs as well as other elongation factors when it emerges from the Pol II core (Missra and

Gilmour, 2010), and recruits NELF to establish pausing. Pausing is released through the activity of a kinase called positive transcription elongation factor-b (P-TEFb), which phosphorylates DSIF, NELF and Serine 2 of the Pol II CTD to promote pause release (Cheng and Price, 2007; Kim and Sharp, 2001; Marshall and Price, 1995; Wada et al., 1998b). Phosphorylation of NELF promotes its dissociation from the transcriptional complex (Fujinaga et al., 2004), whereas phosphorylation of Spt5 converts DSIF into a positively acting elongation factor which travels with Pol II throughout the gene (Yamada et al., 2006). The role of P-TEFb at stimulating elongation is also conserved in yeast. While yeast do not have promoter proximal pausing due to a lack of NELF, the yeast homolog Bur1/Bur2 also phosphorylates Ser2 of the CTD and Spt5 to promote transcription elongation (Liu et al., 2009; Zhou et al., 2009).

Due to the importance of P-TEFb at promoting pause release, various mechanisms exist to regulate P-TEFb activity. Many gene-specific transcription factors previously thought to function in initiation also function in pause release. For instance, cMyc and NF-kappaB both interact and recruit P-TEFb to promoters (Barboric et al., 2001; Rahl et al., 2010). Additionally, more than half the P-TEFb molecules are inactive, bound to the 7SK snRNP complex (Nguyen et al., 2001; Yang et al., 2001; Yik et al., 2003). Acetylated histones recruit Brd4 near the promoter, releasing P-TEFb from the 7SK complex and activating its kinase activity (Jang et al., 2005; Loven et al., 2013; Yang et al., 2005). Independent of Brd4, P-TEFb is found as a part of the Super Elongation Complex (SEC) (He et al., 2010; Lin et al., 2010; Sobhian et al., 2010). The complex itself is highly heterogeneous, but generally consists of P-TEFb, EAF1/2, AFF1/4, AF9/ENL and ELL1/2/3. SEC is recruited to genes to promote pause release, likely through interactions with the Mediator complex (Takahashi et al., 2011).

Mutations in pathways associated with promoter proximal pausing are frequently associated with cancer, enabling greater production of RNAs necessary for cellular growth and division. *MYC* is the most frequently amplified oncogene in cancer (Beroukhim et al., 2010), and amplification of the gene product results in a global increase in RNA production due to enhanced recruitment of P-TEFb and pause release (Lin et al., 2012). Additionally, SEC components such as AF9, AF10, ENL, AFF1 and ELL1 are frequently translocated in leukemias associated with translocations of MLL (Tenney and Shilatifard, 2005), resulting in constant recruitment of P-TEFb and pause release at MLL target genes such as *HOXA9* and *HOXA10* in leukemia (Lin et al., 2010). Thus, regulation of promoter-proximal pausing plays a critical role in human disease.

One hypothesis for the lower abundance at uaRNAs is that Pol II were not undergoing pause release in the antisense direction. Two lines of evidence argues against this model. First, uaRNA production was suppressed upon inhibition of P-TEFb with flavopiridol (Flynn et al., 2011). Moreover, the product of P-TEFb activity, Ser2P, was enriched over uaRNA regions (Preker et al., 2011). Together, this argues that the decision point on whether or not to create a stable transcript occurs subsequent to promoter-proximal pausing.

1.4 Transcription Elongation I (Pausing to Nucleosome Barrier)

Currently, Pol II has received the "go" signal, marked by phosphorylation of DSIF. A plethora of proteins operate with clockwork precision to promote transcription. These factors are commonly called 'elongation factors,' and comprise of histone modifiers, histone chaperones, nucleosome remodelers and adaptor proteins. Three critical things happen in this first stage of elongation: H3K4me3/H3K79me2 deposition, the first instance of RNA splicing, and +1 nucleosome pausing.

The H3K4me3/H3K79me2 pathway

H3K4 methylation is one of the most commonly described histone marks in the literature and associated with active genes. Analysis of divergent transcription around promoters revealed that H3K4me3, a mark commonly associated with initiation, is enriched in both directions, whereas H3K79me2 and H3K36me2, marks associated with transcription elongation, were enriched solely at the sense direction (Core et al., 2008; Seila et al., 2008). To get a better understanding of this difference, an understanding of how these marks are deposited is necessary.

The switch to productive elongation depends on the phosphorylation of the Spt5 subunit of the elongation factor Spt4/Spt5 by Bur1/Bur2 (P-TEFb in mammals), which activates it and creates a docking site for a key adaptor of transcription elongation, the PAF complex (Liu et al., 2009; Zhou et al., 2009). The PAF complex was initially identified in screens for Pol II interacting proteins (Wade et al., 1996) and functions in numerous pathways including H3K4 methylation, H3K36 methylation, Ser2P levels, transcription through chromatin and proper 3' end processing. The PAF complex promotes both H3K4me3 and H3K79me2 by stimulating the activity of the H2B ubiquitylating enzymes, Bre1/Rad6 (Dover et al., 2002; Sun and Allis, 2002; Wood et al., 2003b). H2Bub1 promotes H3K4 dimethylation and trimethylation (Schneider et al., 2005) or the deposition of H3K79me2 (Krogan et al., 2003a; Ng et al., 2002; Wood et al., 2003a). H3K4 methylation is facilitated by the activity of the Set1 (or KMT2) within a larger complex called COMPASS (Briggs et al., 2001; Miller et al., 2001), whereas H3K79me2 is deposited by the Dot1 enzyme (Lacoste et al., 2002).

This pathway is conserved in mammals but with somewhat more complications. After phosphorylation by P-TEFb, DSIF recruits the hPAF complex to transcribed genes and is

necessary for H2B ubiquitylation by Bre1 homologs, RNF20 and RNF40 (Kim et al., 2009; Zhu et al., 2005). Overexpressing RNF20/40 promotes H3K4me3 levels, whereas knocking it down decreases H3K4me3, arguing that the yeast histone pathway functions in mammals (Zhu et al., 2005). The mammalian COMPASS complex was initially identified from attempts to purify interaction partners of the tumor suppressor MEN1 (Hughes et al., 2004). While there are 6 Set1 homologs in humans, the SET1A/SET1B complexes deposit the majority of H3K4me3 (Wu et al., 2008). Similar to yeast, H2B ubiquitylation is also important for hDOT1L methylation of H3K120 (McGinty et al., 2008; Mohan et al., 2010). DOT1L is found in a complex which includes interaction partners of MLL-fusion proteins, such as ENL and AF9 (Biswas et al., 2011; Mohan et al., 2010). One characteristic of MLL-AFF4 fusions in MLL-fusion leukemias is ectopic H3K79me2 signal and aberrant expression of HOX genes, though how it is unclear how this histone mark influences transcription.

Since H2Bub1 promotes H3K4me3 and H3K79me2, why does both sides of divergent promoters have H3K4me3 but H3K79me2 is restricted to sense transcription? Unlike yeast, most mammalian promoters have H3K4me3, irrespective of whether Pol II is elongating (Bernstein et al., 2006; Guenther et al., 2007). One component of the mammalian COMPASS complex, Cfp1, binds to nonmethylated CpGs, promoting H3K4 methylation over CpG islands (Thomson et al., 2010). In this study, inserting an artificial promoter-less CpG-rich DNA into the genome created novel H3K4me3 enriched regions. Hence, while H3K4me3 has been called a transcription initiation mark in mammals due to its proximity to the TSS, it is more precisely a mark of CpG islands. The levels of H3K4me3 observed at a gene are a combination of basal levels of H3K4me3 deposited by CpG islands as well as additional H3K4me3 deposited during transcription elongation through the H2B ubiquitylation pathway. This explains the histone

marks associated with divergent transcription. Despite both sites having H3K4me3 signal, the sense gene has more HK4me3 signal (Core et al., 2008; Seila et al., 2008). Additionally, this relationship between H3K4me3 and CpG islands provides a molecular explanation for the historical use of H3K27ac and H3K4me1 as marks of active enhancers: enhancers lack CpG islands so have lower levels of H3K4me3, but are acetylated to promote transcription.

RNA Splicing

In the mid-1970s, scientists were trying to understand the relationship between heterogeneous nuclear RNAs[8] and cytoplasmic RNAs. EM microscopy of R-loops between transcribed RNA and the genomic DNA revealed that intervening sequences in DNA were spliced out to make mature mRNA (Berget et al., 1977). Since those discoveries, many years of analyses has identified a plethora of proteins to form the spliceosome, including the U snRNAs, RNA helicases and numerous RNA-binding proteins. The splicing reaction is a RNA-catalyzed process, in which the 5' splice site (5'SS) is joined to the 3' splice site (3'SS), while excising the intron as a lariat (Wahl et al., 2009). Initially, the 5'SS is recognized by U1 snRNP, whereas the branchpoint sequence upstream of the 3'SS is recognized by U2 snRNP. Subsequently, U4/U5/U6 join as a preformed tri-snRNP, where U6 snRNA substitutes for U1 snRNA at binding to the 5'SS, and U4 snRNA is evicted. This activated complex catalyzes two SN2-like reactions: the branchpoint adenosine attacks the 5'SS to form a lariat intermediate, and then the 3'OH of the 5' exon attacks the 3'SS to join the exons together. After splicing, the spliceosome is disassembled and the exon junction complex is deposited at the splice-junction to promote nuclear export.

[8] Heteronuclear RNAs (hnRNAs) are now called pre-mRNAs.

Splicing has roles beyond producing a spliced transcript; numerous studies showed that splicing is important for gene expression. Initial attempts to express cDNAs in mammalian systems found that splicing was essential to express proteins at high levels (Brinster et al., 1988; Palmiter et al., 1991). Later genome-wide analyses would reveal that almost all genes in mammals are spliced. One reason is that splicing promotes mRNA export through an interaction between the exon-junction complex and mRNA export factors (Le Hir et al., 2001). However, the 5' SS can also enhance gene expression independent of the splicing reaction. While U2, U4,U5 and U6 are generally stoichiometric, U1 snRNA is found at significantly higher levels, suggesting the U1 snRNA may have roles beyond splicing. Mutations of a promoter-proximal 5'SS reduced transcription in nuclear-run on experiments, and compensatory mutations in U1 snRNA rescued it (Furger et al., 2002). Another study demonstrated that this may be due to the activity of U1 snRNA at stimulating formation of the first phosphodiester bond as well as TFIIH-dependent transcription reinitiation (Kwek et al., 2002). A 5'SS promotes the recruitment of basal transcription factors TFIID, TFIIH and TFIIB (Damgaard et al., 2008). The U1 snRNA can also regulate promoter-proximal termination globally through an interaction with poly(A) signals (Berg et al., 2012; Kaida et al., 2010), which will be discussed later. Altogether, these results suggest that splicing signals have roles in regulating transcription. 5' splice sites are preferentially enriched near the TSS for mRNAs, but not for uaRNAs (Almada et al., 2013; Ntini et al., 2013), suggesting that binding sites for U1 snRNP may have important roles in regulating transcription. In addition, we found the majority of uaRNAs are not spliced, further suggesting splicing may be a key regulator of uaRNA stability.

Transcribing through the CpG-Island

Nucleosomes are a general barrier to transcription (Churchman and Weissman, 2011; Kwak et al., 2013; Weber et al., 2014). Recent work using techniques that maps the precise 3' end of transcribing Pol II suggests that the +1 nucleosome is the site of the most stalling/backtracked events (Weber et al., 2014). However, the properties of the +1 nucleosome differs between *Drosophila* and humans due to the tendency for mammalian promoters to have CpG islands (Saxonov et al., 2006). Due to the propensity of CGIs to deter nucleosome assembly, the initial steps in the mammalian transcription cycle occur over unstable nucleosomes[9], which facilitates Pol II transcription through the CpG island (de Dieuleveult et al., 2016). Eventually Pol II reaches a point where nucleosomes associate more strongly with DNA, creating a transcription barrier. The first nucleosome with a strong association is the +1 stable nucleosome.

One intriguing question is what modulates this stable nucleosome barrier. While studies have not focused on this specific barrier, the +1 nucleosome near the TSS can be modulated by H2A.Z and Chd1 (Weber et al., 2014). Some of these factors may also regulate the +1 stable nucleosome pause. Genome-wide mapping of chromatin remodelers at stable nucleosomes found the +1 stable nucleosome is most associated with Chd1 and Chd8, whereas the -1 nucleosome is mostly associated with Ep400 and Chd4 (de Dieuleveult et al., 2016). Human CHD1 is recruited to the genome by a chromodomain that recognizes H3K4me3 (Flanagan et al., 2005), so the CpG island can directly regulate its own nucleosome stability.

There are also numerous additional pauses that occur cotranscriptionally. During transcription, Pol II pauses at 3'SS of introns, likely to ensure proper spliceosome assembly

[9] Unstable nucleosomes are also called fragile nucleosomes, due to their sensitivity to MNase and are best detected at low MNase-concentrations

(Alexander et al., 2010). Additionally, inhibition of CDK9 using two inhibitors (KM05283 and DRB) revealed examples of additional P-TEFb mediated pauses shortly after the promoter proximal pause and also at the 3' end of genes (Laitem et al., 2015). Analysis of data from this paper further suggests that there may be pausing at the edge of CpG islands.

1.5 Transcription Elongation II: Transcribing through Nucleosomes

Several studies on divergent transcription suggests that the stage subsequent to +1 nucleosome pausing is linked with regulating divergent transcription. In this step, H3K36me3 is deposited and nucleosomes must be traversed.

SETD2 and H3K36 methylation

Analyses of H3K36me3 have revealed potential links with divergent transcription. Initial studies found that sense transcription was enriched for H3K36me3 in comparison to antisense transcription (Core et al., 2014). Moreover, one pathway dependent on H3K36me3 has been implicated in modulating a subset of divergent transcription events in yeast (Churchman and Weissman, 2011). Hence, we need to take a step back and discuss how H3K36me3 is deposited. In yeast, two different kinases phosphorylate Ser2 of the CTD. The first, Bur1, primarily functions near the promoter and phosphorylates Spt5. The second, Ctk1, functions later in the transcription cycle and is associated with H3K36me3. Ser2P creates a platform for the binding of Set2, the H3K36 methyltransferase (Krogan et al., 2003b; Li et al., 2002). It is currently unclear how Ctk1 promotes the recruitment of Set2 and not Bur1, though it is suggested that Ctk1 promotes the deposition of the majority of Ser2P in contrast with Bur1.

In mammals, H3K36me3 is associated with actively transcribed genes. Just like Set1, there are multiple versions of Set2 in mammals, though SETD2 mediates the majority of H3K36 trimethylation (Edmunds et al., 2008). Various studies suggest that mRNA splicing plays an important role in promoting H3K36me3. First, experiments using the splicing inhibitor spliceostatin A revealed that pre-mRNA splicing is critical for H3K36me3 (Kim et al., 2011). In another paper, H3K36me3 was enriched in genes with introns compared to genes without introns, and treatment with a different splicing inhibitor meamycin reduced H3K36me3 levels and SETD2 association at genes (de Almeida et al., 2011). Moreover, genome-wide studies reveal a reciprocal link between H3K79me2 and H3K36me3, whereby H3K79me2 and H2Bub1 is present up until the first 3'SS, after which H3K36me3 starts accumulating (Huff et al., 2010). In summary, these studies suggest splicing promotes the deposition of H3K36me3. The reduced H3K36me3 signals at uaRNAs in contrast with sense mRNAs may be a byproduct of the lack of splicing signals and/or splicing events at uaRNAs.

Transcribing through Chromatin

Transcription through chromatin requires a balancing act. First, the nucleosome barriers must be reduced so that Pol II can transcribe through it; this usually involves histone exchange by histone chaperones. Secondly, after Pol II passes through chromatin, stable nucleosomes must be deposited back on the DNA; otherwise additional Pol II molecules can spuriously associate with nucleosome free regions generated in the wake of actively elongating Pol II and initiate undesired transcription events.

Several histone chaperones are recruited during transcription to promote transcription elongation. The histone chaperone FACT was identified as a factor that promoted transcription

through chromatinized templates (Orphanides et al., 1998). FACT functions by promoting the exchange of H2A/H2B dimers on nucleosomes, resulting in temporary histone hexamers, before re-depositing the dimer to regenerate the histone octamer (Belotserkovskaya et al., 2003; Xin et al., 2009). Spt6 is a different histone chaperone that primarily interacts with H3 (Bortvin and Winston, 1996). In vivo experiments in S2 cells found that knockdown of Spt6 resulted in a slower Pol II elongation (Ardehali et al., 2009). Both FACT and Spt6 are recruited cotranscriptionally during elongation, the former due to an association with the PAF complex (Simic et al., 2003) and the latter through an interaction with a CTD phosphorylated on both Ser5 and Ser2 (Sun et al., 2010).

In addition, the H3K36me2/3 marks deposited by Set2 suppress cryptic initiation from within gene bodies. As Pol II transcribes, it brings along histone acetylatransferases to acetylate histones and alters the transcribed chromatin. Since acetylated histones decompact chromatin and promote transcription initiation, the chromatin needs to be reset to the non-acetylated state prior to Pol II passage; otherwise unwanted transcription would initiate from within the gene body and interfere with properly transcribing polymerases. H3K36me3 marks recruit the Rpd3S complex, which functions as the H4 deacetylase to reset chromatin (Carrozza et al., 2005; Keogh et al., 2005). Mutations in this complex or any of the upstream factors result in cryptic initiation within genes. Curiously, deletions of the Rco1 subunit of Rpd3S in *S cerevisiae* resulted in increased divergent transcription from genes that were arranged head-to-tail (Churchman and Weissman, 2011), probably due to the role of Rpd3S at suppressing cryptic initiation because yeast genes are packed very close to one another. This pathway also suppress cryptic initiation in mESCs, where the chromodomain MRG15 binds to H3K36me3 and recruits several of Rpd3S complex subunits including HDAC1 and Sin3A, as well as the H3K4me3-demethylase KDM5B (JARID1B) (Xie

et al., 2011). Importantly, knockdown of either MRG15 or KDM5B result in increased cryptic transcription from within the gene body.

In addition, H3K36me3 associates with the chromatin remodeler Isw1b (Smolle et al., 2012). Similar to the H3K36me2-Rpd3S pathway, deletions of Isw1 and another chromatin remodeler Chd1 resulted in increased histone exchange over ORFs and increased cryptic initiation from within the gene body. In yeast, Chd1 is recruited through an interaction with various elongation factors, including PAF, Spt4/5 and FACT (Simic et al., 2003). In mammals, CHD1 promotes nucleosome turnover at the promoter but also suppresses nucleosome turnover within the gene body (Skene et al., 2014).

1.6 Transcription Termination and the U1-PAS Axis

A major focus of divergent transcription studies has been on how these RNAs terminate. Elongation appears to be different between uaRNAs and mRNAs, exemplified by the differences in H3K36me3 and H3K79me2 marks. Moreover, uaRNAs are frequently shorter than mRNAs, suggesting that one source of uaRNA instability may be premature termination, which prevents the transcribing polymerase from maturing into a productive transcription elongation complex. Supporting this view, several studies found a key protein, the RNA exosome, actively degrades uaRNAs (Flynn et al., 2011; Preker et al., 2011; Preker et al., 2008). Complicating these studies were different reports about the nature of the 3' end of uaRNAs. Some studies suggested uaRNAs were polyadenylated (Almada et al., 2013; Ntini et al., 2013; Preker et al., 2008), whereas other studies suggested that uaRNAs were nonpolyadenylated (Flynn et al., 2011; Preker et al., 2011). It is likely that there are two redundant pathways that function to degrade

noncoding RNAs, as cells prefer not producing RNAs when they are not needed. Importantly, both pathways center on a key complex called the RNA exosome.

The RNA Exosome

RNA exosome is 10 subunit complex that acts as the major 3'-to-5' exoribonuclease. The RNA exosome comes in multiple flavors: the nuclear exosome has an additional Rrp6 subunit. Rrp40 (mammalian EXOSC3) is thought to be a critical subunit of the RNA exosome as mutations in this subunit abrogate the activity of the RNA exosome. Structurally, the yeast RNA exosome is a barrel surrounding a channel, through which single-stranded RNA of at least 25 nucleotides is threaded through to be degraded at the other end by Rrp44 (Dis3) (Bonneau et al., 2009; Makino et al., 2013), similar to the mechanism of protein degradation by the proteasome. The activity of the RNA exosome is regulated by adapter proteins that recruit RNAs to the exosome and by cofactors that either stimulate exosome activity or prepare the substrates through processes like adenylation.

The RNA exosome has numerous cytoplasmic and nuclear processes. In the cytoplasm, most mRNAs are degraded by deadenylation followed by decapping and 5'-to-3' decay. However, a complementary pathway exists whereby deadenylation is followed by exosome decay, through the involvement of the SKI complex (Ski2/Ski3/Ski7) (Anderson and Parker, 1998). Moreover, translation quality control frequently involves the RNA exosome, such as nonsense mediated decay where a premature stop codon signals the mRNA for degradation (Lejeune et al., 2003; Mitchell and Tollervey, 2003).

Most studies of the RNA exosome have focused on their nuclear roles. It was initially characterized to function in 3' end processing of 5.8S rRNA, snRNA and snoRNA precursors

(Allmang et al., 1999; van Hoof et al., 2000). Additionally, the RNA exosome functions in numerous nuclear quality control pathways. For instance, the RNA exosome performs transcriptional surveillance to suppress the expression of noncoding RNAs including cryptic unstable transcripts (Wyers et al., 2005), uaRNAs (Flynn et al., 2011; Preker et al., 2008) and eRNAs (Andersson et al., 2014). The RNA exosome has also been proposed to resolve backtracked polymerases in collaboration with TFIIS (Lemay et al., 2014). More recently, the RNA exosome has been suggested to function in class switching of B cells, regulating R-loops (DNA:RNA hybrids) and suppressing genome instability (Mischo et al., 2011; Pefanis et al., 2014; Pefanis et al., 2015). Amplification of RNA exosome subunits, especially EXOSC4, is frequently found in cancers (cBioPortal). While the mechanism is unclear, it likely protects rapidly dividing tumor cells from genomic instability that arises from R-loops during amplified transcription.

Non-Polyadenylated Pathway

One hypothesis is that uaRNAs resemble yeast cryptic unstable transcripts (CUTs) due to their short lengths, oligo(A) ends and sensitivity to the RNA exosome. CUTs are a class of short, unstable noncoding RNA transcripts initially identified in mutations of various RNA exosome subunits (Wyers et al., 2005). These substrates were found to be oligoadenylated[10] by the TRAMP complex. The trimeric TRAMP complex is made up of the RNA helicase Mtr4, RNA-binding proteins Air2 and a non-canonical, distributive poly(A) polymerase Trf4, which adds 3-4 adenosines to promote exosome decay (LaCava et al., 2005). It is speculated that short A tail

[10] The literature is unclear when it comes to the term polyadenylation. In this thesis, oligoadenylation will refer to short adenosine tails of between 3-10 adenosines. In contrast, polyadenylation will refer to longer adenosine tails of around 80-150 adenosines. Hyperadenylation refers to the longest adenosine tails of over 200 adenosines.

helps provide a foothold for the RNA exosome to bind to, similar to how short A tails promote degradation by RNases in bacteria and the mitochondria.

Knowing this, the Jensen lab used SILAC/MS to identify mammalian homologs of the TRAMP complex. hMTR4 is found in two complexes that occupy different sub-compartments of the nucleus (Lubas et al., 2011). The nucleolar complex is the hTRAMP complex, made of hMTR4, TRF4-2 and ZCCHC7, whereas the nucleoplasmic Nuclear Exosome Targeting (NEXT) complex is made of hMTR4, ZCCHC8 and RBM7. While hTRAMP regulates the processing of rRNAs, the NEXT complex promotes degradation of uaRNAs and eRNAs (Lubas et al., 2011).

The NEXT complex is recruited to the 5' ends of RNAs through a series of interactions involving the adapter protein ZC3H18 and the cap-binding complex CBCA (CBP20-CBP80-AR2) (Andersen et al., 2013; Hallais et al., 2013). CBCA promotes degradation of uaRNAs as double knockdowns of CBCA and the RNA exosome synergistically stabilizes uaRNAs (Andersen et al., 2013). In addition, at reporter genes, ARS2 promotes the use of cap-proximal PAS motifs, due to the association of ARS2 with the CFIIm complex of the CPA machinery. This role of CBCA at recruiting TRAMP to short transcripts is also conserved in *S pombe*.

In *S cerevisiae*, CUTs are also targeted for exosome decay through the Nrd1-Nab3-Sen1 pathway (Arigo et al., 2006; Thiebaut et al., 2006; Vasiljeva and Buratowski, 2006), whereby Nrd1 and Nab3 binds to RNA motifs to recruit the RNA exosome in addition to the Sen1 RNA helicase (Carroll et al., 2007; Steinmetz and Brow, 1998; Steinmetz et al., 2006). It is unclear whether the NNS pathway functions in mammals; while there are homologous proteins, there is currently no evidence suggesting that they function at uaRNAs. However, Nrd1 regulates divergent transcription in yeast, as deletion of Nrd1 results in stabilization of NUTs (Nrd1-unstable transcripts) genome-wide (Schulz et al., 2013). Interestingly, sequence analysis found

yeast mRNAs are depleted for Nrd1-binding motifs, arguing natural selection selected for stabilization of mRNAs.

Additionally there may be a role for the Integrator in regulating non-polyadenylated uaRNAs. The Integrator is a 14-subunit complex in mammals initially identified for its ability to bind to the CTD of Pol II and function in 3' end processing of U1 and U2 snRNAs (Baillat et al., 2005), but more recently linked with histone processing and promoter proximal termination (Skaar et al., 2015). Recently the Integrator has also been found to be involved in the biogenesis of most non-polyadenylated eRNAs in HeLa cells by promoting earlier termination (Lai et al., 2015). It is possible that the integrator is the machine that stimulates release of nonpolyadenylated RNAs from Pol II for targeting by NEXT and the RNA exosome. RNA-seq of knockdowns of Integrator subunits found a modest upregulation of uaRNAs, supporting this idea (Stadelmayer et al., 2014). Curiously, two integrator subunits share homology to the CPSF complex linked with cleavage and polyadenylation: Ints11 is the main catalytic subunit homologous to CPSF73, whereas Ints9 is homologous to CPSF100 (Baillat et al., 2005).

Polyadenylation Pathway

While one series of studies focused on investigating uaRNAs as unstable, nonpolyadenylated transcripts, another series of studies focused on revelations that some uaRNAs were polyadenylated, arguing there must be a pathway to degrade polyadenylated RNAs. These observations would become the basis for the discovery of the U1-PAS axis and subsequent work done in this thesis.

Polyadenylation refers to the sequential addition of adenosine residues to create a tail at the 3' end of RNAs. Polyadenylation is a critical step in transcription because the binding of

cytoplasmic poly(A) binding proteins (PABP) protects mRNAs from RNA decay (Ford et al., 1997), analogous to how the 5' cap protects mRNAs from 5'-to-3' decay. Polyadenylation occurs in three steps: the RNA is cleaved 20-30 nucleotides downstream of a polyadenylation signal (PAS) motif, the 5' product is polyadenylated and the 3' product is removed from the chromatin to terminate transcription.

In mammals, the PAS motif is made up of 3 elements: a core hexamer (AAUAAA), the poly(A) site and downstream GU-rich elements (Chan et al., 2011). The core hexamer was initially identified at the 3' ends of many individually sequenced mRNAs (Proudfoot and Brownlee, 1976), and since then has been found at the 3' ends of most mRNAs. Other PAS variants have also been identified, but the top two most frequently used PAS motifs A[A/U]UAAA make up the majority of sequenced transcripts. The distance between the PAS motif and the cleavage site is highly conserved (between 20-30 bases), because increasing the distance between the PAS motif and the cleavage site causes transcripts to be destabilized (Wu and Bartel, 2017). In contrast, the downstream elements are far less conserved than the core hexamer and provide additional binding sites for the cleavage and poly(A) (CPA) machinery. Interestingly, poly(A) uaRNAs frequently have canonical PAS motifs, arguing that uaRNAs are polyadenylated through the CPA machinery (Almada et al., 2013; Ntini et al., 2013).

The CPA machinery is comprised of 6 major proteins: 4 cleavage factors (CPSF, CstF, CFIm, CFIIm), the poly(A) polymerase (PAP) and a nuclear poly(A)-binding protein (PABPN1). CPSF, CstF and CFIm associate with actively elongating Pol II (Hirose and Manley, 2000), the latter two through binding to the CTD of Pol II. Pausing of Pol II frequently occurs at the 3' ends of RNAs over G-rich sequences, allowing time for the CPA machinery to recognize the PAS motif (Yonaha and Proudfoot, 1999). Interestingly, CpG islands also have G-rich sequences,

which may contribute to the +1 stable nucleosome pause at the boundaries of CpG islands. Transcription termination begins with recognition of PAS motifs by CPSF, CstF and CFIm, followed by recruitment of CFIIm and PAP. The mature termination complex then cleaves the RNA via the CPSF73 subunit (Mandel et al., 2006). Next, CPSF anchors PAP to the 5' cleavage product to add adenosines distributively. PABPN1 binds after the first 10 As, and promotes processive polyadenylation by PAP, creating a long poly(A) tail, before the mature mRNA is released and exported to the cytoplasm. The 3' cleavage product remains attached to a transcribing Pol II and must be released. The 5' uncapped end is a substrate for the 5'->3' exoribonuclease Xrn2, which torpedoes down the RNA to both degrade it and promote Pol II release from the chromatin (Kim et al., 2004; West et al., 2004). While this process functions at the 3' ends of genes, most of this core termination machinery has also been found in human CLIP-seq studies to bind uaRNAs as well as promoter-proximal regions of genes (Almada et al., 2013; Nojima et al., 2015), indicating that CPA can also functions near the promoter.

The U1-PAS Axis

Two studies focused on examining polyadenylation at uaRNAs identified a U1-PAS axis that functions to regulate divergent transcription in mammals. Analysis of sequence motifs found that PAS motifs (A[A/T]TAAA) are selectively depleted within mRNA genes, whereas U1 splicing signals (AGGURAGU) are selectively enriched near the TSS (Almada et al., 2013; Ntini et al., 2013). Various lines of evidence support that these PAS sequences were used and terminated by the CPA machinery. First, a large fraction of 3' ends of polyadenylated uaRNAs possess canonical PAS motifs within 20-30 nucleotides upstream of the cleavage site (Almada et al., 2013; Ntini et al., 2013). Secondly, mutations at the predicted PAS motifs of 2 PROMPTs

abrogated the use of those sites (Ntini et al., 2013). Lastly, analysis of CLIP-seq datasets revealed that CPA subunits were recruited to PAS sites at PROMPTs in humans (Almada et al., 2013). Moreover, there is a strong link between U1 snRNA and PAS-mediated termination. Initially U1 was found to suppress use of late stage poly(A) sites in the life cycle of the bovine papillomavirus (Furth et al., 1994) through multiple direct interactions between U1 snRNA and the CPA machinery (Gunderson et al., 1998; Lutz et al., 1996). Since then, the inhibitory effect of U1 on PAS termination has been observed to regulate PAS usage globally (Almada et al., 2013; Berg et al., 2012; Kaida et al., 2010).

Together, the U1-PAS axis model postulates that the decision of where to terminate is the key driver for exosome sensitivity. Early termination by PAS motifs have been linked with RNA instability; decreasing the distance between the TSS and a canonical PAS motif at individual genes or PROMPTs results in transcript instability and exosome sensitivity (Andersen et al., 2012; Ntini et al., 2013). Older genes have stronger U1/PAS biases than younger genes, suggesting that natural selection preferred moving termination sites later (Almada et al., 2013). This selective depletion for PAS motifs at mRNAs is similar to the selective depletion of Nrd1-binding sites in *S cerevisiae*. Altogether, this suggesting that the default pathway in all cells is to degrade transcripts and stable transcripts have been specifically selected for over time.

How are polyadenylated RNAs being degraded? One candidate is the nuclear poly(A) binding protein (PABPN1), which normally functions to promote polyadenylation at 3' ends of mRNAs. One study found PABPN1 and the poly(A) tail promoted degradation of a subset of lncRNAs, including snoRNA host genes and some divergent lncRNAs (Beaulieu et al., 2012). The RNA exosome and hMTR4 was necessary to degrade these RNAs, but not the poly(A) polymerase found in the hTRAMP complex. Supporting the connection between PABPN1 and

the RNA exosome, PABPN1 promoted degradation of polyadenylated viral transcripts (PANΔENE) as well as improperly spliced RNAs (Bresson and Conrad, 2013). Interestingly, PABPN1 autoregulates its own production. Binding of PABPN1 to genomically encoded poly(A) tracts in the terminal intron blocks splicing and promotes exosome decay of the PABPN1 transcript (Bergeron et al., 2015).

Several recent studies including work from this thesis suggest that PABPN1 promotes degradation of a subset of uaRNAs (Bresson et al., 2015; Li et al., 2015; Meola et al., 2016). Identification of additional interaction partners of hMTR4 revealed that ZFC3H1 bridges an interaction between hMTR4 and PABPN1, analogous to how ZCCHC8 bridges an interaction between hMTR4 and RBM7 in the NEXT complex (Meola et al., 2016). This Poly(A) Exosome Targeting (PAXT) connection preferentially targeted longer substrates with poly(A) tails including snoRNA host genes, whereas NEXT preferentially targeted PROMPTs and eRNAs, though a subset of PROMPTs are targeted by PABPN1 (Meola et al., 2016).

This polyadenylation-associated degradation pathway is highly conserved in *S pombe*. The nuclear poly(A) binding protein Pab2 along with Rmn1 recruits a TRAMP-like complex called MTREC (Mtl1-Red1) to poly(A) tails (Lee et al., 2013; Yamanaka et al., 2010). MTREC then subsequently recruits the RNA exosome to target transcripts (Zhou et al., 2015). Each of these proteins are orthologs of the PAXT connection: Mlt1 is an ortholog of hMTR4, Red1 is ZFC3H1 and Pab2 is Pabpn1. Interestingly, Rmn1 is a homolog of RBM27, which is a candidate interaction partner of hMTR4 in SILAC experiments, suggesting it likely functions in the PAXT connection.

Altogether, this argues that a nuclear polyadenylation degradation pathway is a major mode of regulating gene expression. Moreover, there are likely two different pathways that

function to degrade uaRNAs. A subset are non-polyadenylated and degraded with the NEXT complex and another portion are polyadenylated and degraded through the PAXT connection.

1.8 Summary

In the previous pages, I briefly summarized our existing understanding of transcription, with a focus on divergent transcription and what differs between downstream sense and upstream antisense transcription. While it may seem that there is a linear transcription pathway, in reality there is frequent variation across genes, where one step may be bypassed rendering earlier checkpoints less critical. This is especially frequent in mutations in the transcription apparatus that promote tumor progression.

In summary, here is what we now know about uaRNAs. uaRNAs initiate primarily from CpG island promoters, potentially due to the nucleosome free structure. H2A.Z is also necessary to promote production of uaRNAs, likely reflective of the necessity for initiating at accessible nucleosomes. uaRNAs are capped and escape promoter proximal pausing. Chromatin around uaRNAs possess regions of high H3K4me3, which is likely due to an underlying basal level of H3K4me3 that is promoted by Cfp1 at CpG islands. Antisense transcription also lacks H3K36me3 or H3K79me2 when compared to sense transcription, which may be due to a lack of splicing signals at uaRNAs. uaRNAs are primarily nuclear, which may be due to the lack of splicing as exon-junction complexes promote RNA export. The ends of uaRNAs are heterogeneous and are either polyadenylated or nonpolyadenylated. The former is involved in the U1-PAS axis, where the lack of 5' splice sites and the higher frequency of PAS motifs promotes early termination of uaRNAs. Through an unknown mechanism (perhaps Pabpn1 and the PAXT connection), polyadenylated uaRNAs are degraded by the RNA exosome. Alternatively, there is

a backup pathway for non-polyadenylated uaRNAs, involving the NEXT complex and possibly the Integrator. Both pathways involve the association of the RNA with an exosome that is bound to the 5' cap through physical protein-protein interactions.

Given this, my thesis sought to answer three questions:

First, is terminating through the PAS pathway a frequent event? Moreover, is the RNA exosome involved in degrading premature PAS transcripts and how is it linked with U1 activity? In Chapter 2, using a combination of RNA-seq and 2P-seq, we show PAS termination is detected in approximately 40% of uaRNAs and 30% of eRNAs, and the exosome degrades polyadenylated uaRNAs. Moreover, U1 and Exosc3 collaborate at degrading premature termination products within the first intron of the sense transcript.

Where is termination happening and what factors regulate selection of those sites? In Chapter 3, PAS termination happens at the edge of the -1 and +1 stable nucleosome beyond the Stable Nucleosome Free Region (SNFR) at the Stable Nucleosome Termination Area (SNTA). Moreover, genes with premature termination frequently exhibit increased +1 stable nucleosome pausing, which is regulated by the activity of P-TEFb and cMyc.

Since polyadenylation tends to be associated with stable transcripts, how are these polyadenylated transcripts being degraded? In Chapter 4, Pabpn1 targets a subset of uaRNAs for degradation; namely the ones that are polyadenylated. Pabpn1 sensitivity is conferred by terminating close to the promoter by PAS signals. Notably, Pabpn1 also functions at promoting termination at the -1 and +1 stable nucleosome, suggesting Pabpn1 functions in the U1-PAS axis.

Thus, we begin in the next section by generating a system to define uaRNAs.

1.9 Figures

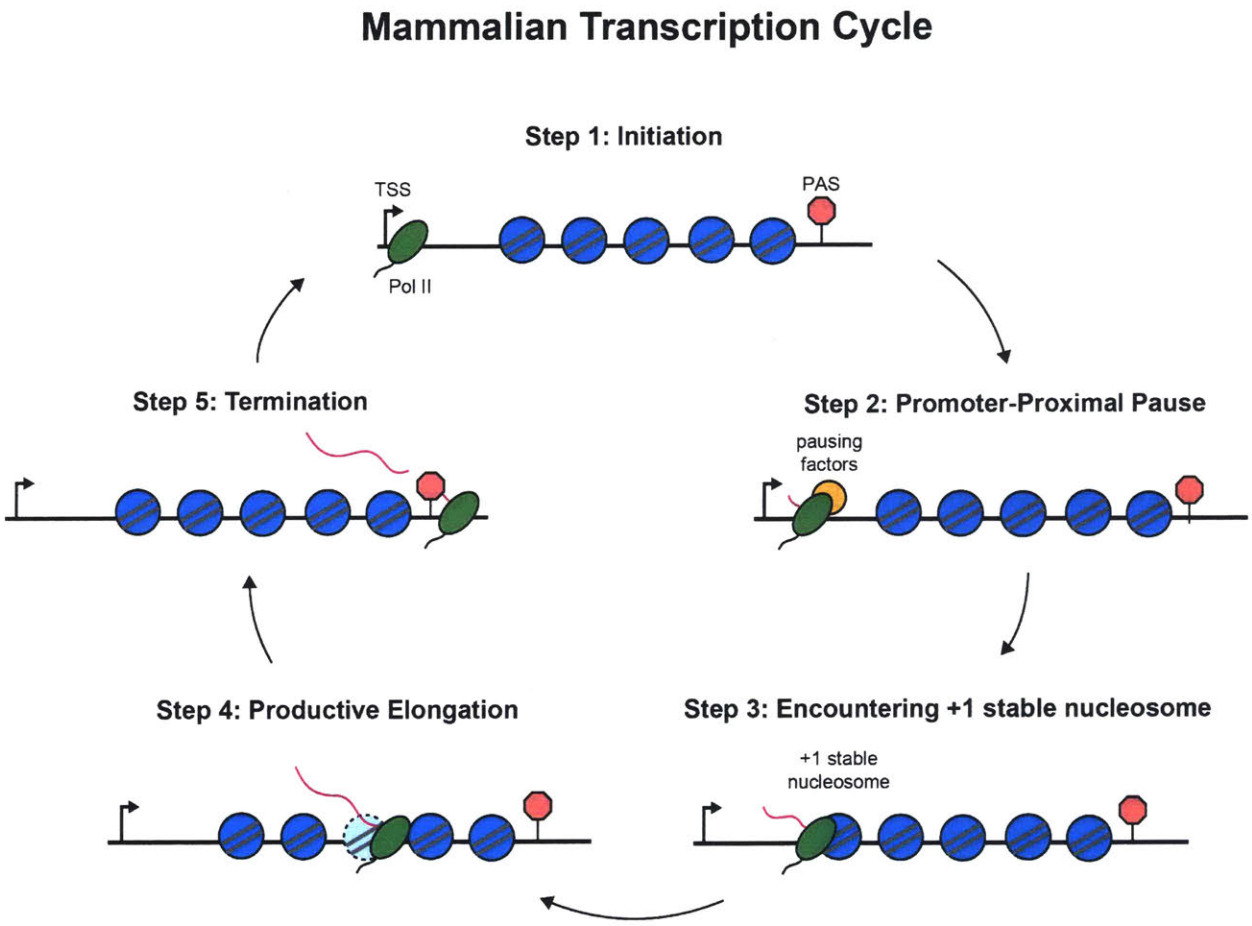

Figure 1. The 5 Steps of the Mammalian Transcription Cycle

Step 1: Transcription Initiation
Step 2: Promoter-Proximal Pause
Step 3: Contact with the +1 Stable Nucleosome
Step 4: Productive Elongation
Step 5: Transcription Termination

1.10 Supplemental Materials

Table S1. Sequencing Technologies for Studying Transcription Pathways

Technique	Key Question	Summary of Assay
RNA-seq	Steady state RNA levels	Poly(A) selected or rRNA-depleted RNA is converted to cDNA.
4sU-seq	Nascent RNA	Metabolically label cells with 4-sU. Biotinylate, isolate with streptavidin, convert to cDNA.
ChIP-seq (chromatin immunoprecipitation sequencing)	Association of DNA with protein	Formaldehyde crosslink, immunoprecipitate protein of interest and purify associated DNA.
ChIP-exo	Association of DNA with protein (precise)	Formaldehyde crosslink, immunoprecipitate protein of interest. Exonuclease treat, purify associated DNA
GRO-seq (global run on sequencing)	Nascent Transcription, elongation competent[11]	Isolate nuclei, run-on reaction in the presence of BrdUTP. RNA is hydrolyzed and isolated with anti-BrdU antibody, prior to cDNA synthesis.
PRO-seq (precise run on sequencing)	Nascent Transcription, elongation competent (precise)	Isolate nuclei, run-on reaction in the presence of biotinylated-UTP. RNA is hydrolyzed and purified by streptavidin, prior to cDNA synthesis.
NET-seq (nascent RNA)	Nascent Transcription (precise)	Immunoprecipitate elongating Pol II complex. Extract RNA and convert to cDNA.
3'NT	Nascent Transcription (precise)	Isolate transcription elongation complexes. Extract RNA and convert to cDNA.
RIP-seq (RNA immunoprecipitation)	Association of RNA with protein	Immunoprecipitate protein of interest from whole cell lysate. Associated RNA is extracted and converted to cDNA.
PAR-clip	Association of RNA with protein	Metabolically label cells with 4sU. UV crosslink, immunoprecipitate protein of interest, extract RNA and convert to

[11] Elongation competent refers to the fact that this techniques requires a 3' OH in the active site. If Pol II is backtracked, this technique would not work.

		cDNA.
Technique	**Key Question**	**Summary of Assay**
CLIP-seq (crosslinking and immunoprecipitation)	Association of RNA with protein	UV crosslink, immunoprecipitate protein of interest, extract RNA and convert to cDNA.
CHIRP-seq (chromatin isolation by RNA purification), RAP-RNA (RNA antisense purification), CHART (capture hybridization analysis of RNA targets)	Position where noncoding RNAs bind DNA	Crosslink, use biotinylated oligos to pull out RNA of interest and associated DNAs, extract DNA.
MNase-seq	Open chromatin.	MNase digest chromatin, purify mononucleosomes, and extract DNA.
FAIRE-seq (formaldehyde assisted isolation of regulatory elements)	Open chromatin	Formaldehyde crosslink, sonicate, phenol/chloroform extract, isolate aqueous layer and extract DNA.
ATAC-seq (assay for transposase-accessible chromatin)	Open chromatin	Incubate DNA with Tn5 transposon, fragment.
DNase-seq	Regulatory elements	Chromatin is treated with DNase, extracted and sequenced.
CAGE (cap-analysis gene expression)	5' ends	Cap trap and add a 5' linker, cleave with MmeI, ligate 3' adapter, PCR amplify and sequence.
GRO-cap	5' ends	Isolate nuclei, run-on reaction in the presence of Brd-UTP. Isolate RNA in the using streptavidin beads. Ligate 3' adapter. Degrade RNAs with no or mono-phosphate. Remove 5' cap. Ligate to 5' adapter, PCR amplify and sequence.
PAL-seq (poly(A)-tail length)	Length of poly(A) tail	Ligate using biotinylated splint oligo, fragment, streptavidin select, create cDNA. On sequencer, extend with dTTP and biotin-dUTP, sequence, and flow in fluorescent streptavidin.

Technique	Key Question	Summary of Assay
TAIL-seq	Length of poly(A) tail	Ligate adapter, fragment, sequence from 5' and 3' end, computationally measure poly(A) length.
2P-seq (Poly(A)-primed sequencing)	Position of poly(A) tail	Poly(A) select RNA, fragment RNA, reverse-transcribe with oligo-dT
3P-seq (Poly(A)-position profiling by sequencing)	Position of poly(A)tail	Poly(A) select RNA, ligate splint biotinylated oligos, fragment RNA, select with streptavidin and convert to cDNA.

1.11 REFERENCES

Agalioti, T., Lomvardas, S., Parekh, B., Yie, J., Maniatis, T., and Thanos, D. (2000). Ordered recruitment of chromatin modifying and general transcription factors to the IFN-beta promoter. Cell *103*, 667-678.

Ahn, S.H., Kim, M., and Buratowski, S. (2004). Phosphorylation of serine 2 within the RNA polymerase II C-terminal domain couples transcription and 3' end processing. Mol Cell *13*, 67-76.

Alexander, R.D., Innocente, S.A., Barrass, J.D., and Beggs, J.D. (2010). Splicing-dependent RNA polymerase pausing in yeast. Mol Cell *40*, 582-593.

Allmang, C., Kufel, J., Chanfreau, G., Mitchell, P., Petfalski, E., and Tollervey, D. (1999). Functions of the exosome in rRNA, snoRNA and snRNA synthesis. EMBO J *18*, 5399-5410.

Almada, A.E., Wu, X., Kriz, A.J., Burge, C.B., and Sharp, P.A. (2013). Promoter directionality is controlled by U1 snRNP and polyadenylation signals. Nature *499*, 360-363.

Andersen, P.K., Lykke-Andersen, S., and Jensen, T.H. (2012). Promoter-proximal polyadenylation sites reduce transcription activity. Genes Dev *26*, 2169-2179.

Andersen, P.R., Domanski, M., Kristiansen, M.S., Storvall, H., Ntini, E., Verheggen, C., Schein, A., Bunkenborg, J., Poser, I., Hallais, M., *et al.* (2013). The human cap-binding complex is functionally connected to the nuclear RNA exosome. Nat Struct Mol Biol *20*, 1367-1376.

Anderson, J.S., and Parker, R.P. (1998). The 3' to 5' degradation of yeast mRNAs is a general mechanism for mRNA turnover that requires the SKI2 DEVH box protein and 3' to 5' exonucleases of the exosome complex. EMBO J *17*, 1497-1506.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., *et al.* (2014). An atlas of active enhancers across human cell types and tissues. Nature *507*, 455-461.

Ardehali, M.B., Yao, J., Adelman, K., Fuda, N.J., Petesch, S.J., Webb, W.W., and Lis, J.T. (2009). Spt6 enhances the elongation rate of RNA polymerase II in vivo. EMBO J *28*, 1067-1077.

Arigo, J.T., Eyler, D.E., Carroll, K.L., and Corden, J.L. (2006). Termination of cryptic unstable transcripts is directed by yeast RNA-binding proteins Nrd1 and Nab3. Mol Cell *23*, 841-851.

Baillat, D., Hakimi, M.A., Naar, A.M., Shilatifard, A., Cooch, N., and Shiekhattar, R. (2005). Integrator, a multiprotein mediator of small nuclear RNA processing, associates with the C-terminal repeat of RNA polymerase II. Cell *123*, 265-276.

Barboric, M., Nissen, R.M., Kanazawa, S., Jabrane-Ferrat, N., and Peterlin, B.M. (2001). NF-kappaB binds P-TEFb to stimulate transcriptional elongation by RNA polymerase II. Mol Cell *8*, 327-337.

Barilla, D., Lee, B.A., and Proudfoot, N.J. (2001). Cleavage/polyadenylation factor IA associates with the carboxyl-terminal domain of RNA polymerase II in Saccharomyces cerevisiae. Proc Natl Acad Sci U S A *98*, 445-450.

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

Bataille, A.R., Jeronimo, C., Jacques, P.E., Laramee, L., Fortin, M.E., Forest, A., Bergeron, M., Hanes, S.D., and Robert, F. (2012). A universal RNA polymerase II CTD cycle is orchestrated by complex interplays between kinase, phosphatase, and isomerase enzymes along genes. Mol Cell *45*, 158-170.

Beaulieu, Y.B., Kleinman, C.L., Landry-Voyer, A.M., Majewski, J., and Bachand, F. (2012). Polyadenylation-dependent control of long noncoding RNA expression by the poly(A)-binding protein nuclear 1. PLoS Genet *8*, e1003078.

Belotserkovskaya, R., Oh, S., Bondarenko, V.A., Orphanides, G., Studitsky, V.M., and Reinberg, D. (2003). FACT facilitates transcription-dependent nucleosome alteration. Science *301*, 1090-1093.

Berg, M.G., Singh, L.N., Younis, I., Liu, Q., Pinto, A.M., Kaida, D., Zhang, Z., Cho, S., Sherrill-Mix, S., Wan, L., *et al.* (2012). U1 snRNP determines mRNA length and regulates isoform expression. Cell *150*, 53-64.

Bergeron, D., Pal, G., Beaulieu, Y.B., Chabot, B., and Bachand, F. (2015). Regulated Intron Retention and Nuclear Pre-mRNA Decay Contribute to PABPN1 Autoregulation. Mol Cell Biol *35*, 2503-2517.

Berget, S.M., Moore, C., and Sharp, P.A. (1977). Spliced segments at the 5' terminus of adenovirus 2 late mRNA. Proc Natl Acad Sci U S A *74*, 3171-3175.

Bernstein, B.E., Kamal, M., Lindblad-Toh, K., Bekiranov, S., Bailey, D.K., Huebert, D.J., McMahon, S., Karlsson, E.K., Kulbokas, E.J., 3rd, Gingeras, T.R., *et al.* (2005). Genomic maps and comparative analysis of histone modifications in human and mouse. Cell *120*, 169-181.

Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., *et al.* (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell *125*, 315-326.

Beroukhim, R., Mermel, C.H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., Barretina, J., Boehm, J.S., Dobson, J., Urashima, M., *et al.* (2010). The landscape of somatic copy-number alteration across human cancers. Nature *463*, 899-905.

Biswas, D., Milne, T.A., Basrur, V., Kim, J., Elenitoba-Johnson, K.S., Allis, C.D., and Roeder, R.G. (2011). Function of leukemogenic mixed lineage leukemia 1 (MLL) fusion proteins through distinct partner protein complexes. Proc Natl Acad Sci U S A *108*, 15751-15756.

Blackledge, N.P., Zhou, J.C., Tolstorukov, M.Y., Farcas, A.M., Park, P.J., and Klose, R.J. (2010). CpG islands recruit a histone H3 lysine 36 demethylase. Mol Cell *38*, 179-190.

Bondarenko, V.A., Steele, L.M., Ujvari, A., Gaykalova, D.A., Kulaeva, O.I., Polikanov, Y.S., Luse, D.S., and Studitsky, V.M. (2006). Nucleosomes can form a polar barrier to transcript elongation by RNA polymerase II. Mol Cell *24*, 469-479.

Bonneau, F., Basquin, J., Ebert, J., Lorentzen, E., and Conti, E. (2009). The yeast exosome functions as a macromolecular cage to channel RNA substrates for degradation. Cell *139*, 547-559.

Bortvin, A., and Winston, F. (1996). Evidence that Spt6p controls chromatin structure by a direct interaction with histones. Science *272*, 1473-1476.

Boutz, P.L., Bhutkar, A., and Sharp, P.A. (2015). Detained introns are a novel, widespread class of post-transcriptionally spliced introns. Genes Dev *29*, 63-80.

Bresson, S.M., and Conrad, N.K. (2013). The human nuclear poly(a)-binding protein promotes RNA hyperadenylation and decay. PLoS Genet *9*, e1003893.

Bresson, S.M., Hunter, O.V., Hunter, A.C., and Conrad, N.K. (2015). Canonical Poly(A) Polymerase Activity Promotes the Decay of a Wide Variety of Mammalian Nuclear RNAs. PLoS Genet *11*, e1005610.

Briggs, S.D., Bryk, M., Strahl, B.D., Cheung, W.L., Davie, J.K., Dent, S.Y., Winston, F., and Allis, C.D. (2001). Histone H3 lysine 4 methylation is mediated by Set1 and required for cell growth and rDNA silencing in Saccharomyces cerevisiae. Genes Dev *15*, 3286-3295.

Brinster, R.L., Allen, J.M., Behringer, R.R., Gelinas, R.E., and Palmiter, R.D. (1988). Introns increase transcriptional efficiency in transgenic mice. Proc Natl Acad Sci U S A *85*, 836-840.

Buratowski, S. (2003). The CTD code. Nat Struct Biol *10*, 679-680.

Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C.A., Taylor, M.S., Engstrom, P.G., Frith, M.C., *et al.* (2006). Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet *38*, 626-635.

Carroll, K.L., Ghirlando, R., Ames, J.M., and Corden, J.L. (2007). Interaction of yeast RNA-binding proteins Nrd1 and Nab3 with RNA polymerase II terminator elements. RNA *13*, 361-373.

Carrozza, M.J., Li, B., Florens, L., Suganuma, T., Swanson, S.K., Lee, K.K., Shia, W.J., Anderson, S., Yates, J., Washburn, M.P., *et al.* (2005). Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription. Cell *123*, 581-592.

Chan, S., Choi, E.A., and Shi, Y. (2011). Pre-mRNA 3'-end processing complex assembly and function. Wiley Interdiscip Rev RNA *2*, 321-335.

Cheng, B., and Price, D.H. (2007). Properties of RNA polymerase II elongation complexes before and after the P-TEFb-mediated transition into productive elongation. J Biol Chem *282*, 21901-21912.

Cho, E.J., Takagi, T., Moore, C.R., and Buratowski, S. (1997). mRNA capping enzyme is recruited to the transcription complex by phosphorylation of the RNA polymerase II carboxy-terminal domain. Genes Dev *11*, 3319-3326.

Churchman, L.S., and Weissman, J.S. (2011). Nascent transcript sequencing visualizes transcription at nucleotide resolution. Nature *469*, 368-373.

Consortium, E.P., Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigo, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., *et al.* (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature *447*, 799-816.

Core, L.J., Martins, A.L., Danko, C.G., Waters, C.T., Siepel, A., and Lis, J.T. (2014). Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. Nat Genet *46*, 1311-1320.

Core, L.J., Waterfall, J.J., and Lis, J.T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science *322*, 1845-1848.

Damgaard, C.K., Kahns, S., Lykke-Andersen, S., Nielsen, A.L., Jensen, T.H., and Kjems, J. (2008). A 5' splice site enhances the recruitment of basal transcription initiation factors in vivo. Mol Cell *29*, 271-278.

Das, C., Lucia, M.S., Hansen, K.C., and Tyler, J.K. (2009). CBP/p300-mediated acetylation of histone H3 on lysine 56. Nature *459*, 113-117.

de Almeida, S.F., Grosso, A.R., Koch, F., Fenouil, R., Carvalho, S., Andrade, J., Levezinho, H., Gut, M., Eick, D., Gut, I., *et al.* (2011). Splicing enhances recruitment of methyltransferase HYPB/Setd2 and methylation of histone H3 Lys36. Nat Struct Mol Biol *18*, 977-983.

de Dieuleveult, M., Yen, K., Hmitou, I., Depaux, A., Boussouar, F., Bou Dargham, D., Jounier, S., Humbertclaude, H., Ribierre, F., Baulard, C., *et al.* (2016). Genome-wide nucleosome specificity and function of chromatin remodellers in ES cells. Nature *530*, 113-116.

Dimitrova, N., Zamudio, J.R., Jong, R.M., Soukup, D., Resnick, R., Sarma, K., Ward, A.J., Raj, A., Lee, J.T., Sharp, P.A., *et al.* (2014). LincRNA-p21 activates p21 in cis to promote Polycomb target gene expression and to enforce the G1/S checkpoint. Mol Cell *54*, 777-790.

Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., *et al.* (2012). Landscape of transcription in human cells. Nature *489*, 101-108.

Dover, J., Schneider, J., Tawiah-Boateng, M.A., Wood, A., Dean, K., Johnston, M., and Shilatifard, A. (2002). Methylation of histone H3 by COMPASS requires ubiquitination of histone H2B by Rad6. J Biol Chem *277*, 28368-28371.

Edmunds, J.W., Mahadevan, L.C., and Clayton, A.L. (2008). Dynamic histone H3 methylation during gene induction: HYPB/Setd2 mediates all H3K36 trimethylation. EMBO J *27*, 406-420.

Fenouil, R., Cauchy, P., Koch, F., Descostes, N., Cabeza, J.Z., Innocenti, C., Ferrier, P., Spicuglia, S., Gut, M., Gut, I., *et al.* (2012). CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters. Genome Res *22*, 2399-2408.

Flanagan, J.F., Mi, L.Z., Chruszcz, M., Cymborowski, M., Clines, K.L., Kim, Y., Minor, W., Rastinejad, F., and Khorasanizadeh, S. (2005). Double chromodomains cooperate to recognize the methylated histone H3 tail. Nature *438*, 1181-1185.

Flynn, R.A., Almada, A.E., Zamudio, J.R., and Sharp, P.A. (2011). Antisense RNA polymerase II divergent transcripts are P-TEFb dependent and substrates for the RNA exosome. Proc Natl Acad Sci U S A *108*, 10460-10465.

Ford, L.P., Bagga, P.S., and Wilusz, J. (1997). The poly(A) tail inhibits the assembly of a 3'-to-5' exonuclease in an in vitro RNA stability system. Mol Cell Biol *17*, 398-406.

Fujinaga, K., Irwin, D., Huang, Y., Taube, R., Kurosu, T., and Peterlin, B.M. (2004). Dynamics of human immunodeficiency virus transcription: P-TEFb phosphorylates RD and dissociates negative effectors from the transactivation response element. Mol Cell Biol *24*, 787-795.

Furger, A., O'Sullivan, J.M., Binnie, A., Lee, B.A., and Proudfoot, N.J. (2002). Promoter proximal splice sites enhance transcription. Genes Dev *16*, 2792-2799.

Furth, P.A., Choe, W.T., Rex, J.H., Byrne, J.C., and Baker, C.C. (1994). Sequences homologous to 5' splice sites are required for the inhibitory activity of papillomavirus late 3' untranslated regions. Mol Cell Biol *14*, 5278-5289.

Gilmour, D.S., and Lis, J.T. (1986). RNA polymerase II interacts with the promoter region of the noninduced hsp70 gene in Drosophila melanogaster cells. Mol Cell Biol *6*, 3984-3989.

Gnatt, A.L., Cramer, P., Fu, J., Bushnell, D.A., and Kornberg, R.D. (2001). Structural basis of transcription: an RNA polymerase II elongation complex at 3.3 A resolution. Science *292*, 1876-1882.

Gudipati, R.K., Xu, Z., Lebreton, A., Seraphin, B., Steinmetz, L.M., Jacquier, A., and Libri, D. (2012). Extensive degradation of RNA precursors by the exosome in wild-type cells. Mol Cell *48*, 409-421.

Guenther, M.G., Levine, S.S., Boyer, L.A., Jaenisch, R., and Young, R.A. (2007). A chromatin landmark and transcription initiation at most promoters in human cells. Cell *130*, 77-88.

Gunderson, S.I., Polycarpou-Schwarz, M., and Mattaj, I.W. (1998). U1 snRNP inhibits pre-mRNA polyadenylation through a direct interaction between U1 70K and poly(A) polymerase. Mol Cell *1*, 255-264.

Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P., *et al.* (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature *458*, 223-227.

Hainer, S.J., Gu, W., Carone, B.R., Landry, B.D., Rando, O.J., Mello, C.C., and Fazzio, T.G. (2015). Suppression of pervasive noncoding transcription in embryonic stem cells by esBAF. Genes Dev *29*, 362-378.

Hallais, M., Pontvianne, F., Andersen, P.R., Clerici, M., Lener, D., Benbahouche Nel, H., Gostan, T., Vandermoere, F., Robert, M.C., Cusack, S., *et al.* (2013). CBC-ARS2 stimulates 3'-end maturation of multiple RNA families and favors cap-proximal processing. Nat Struct Mol Biol *20*, 1358-1366.

He, N., Liu, M., Hsu, J., Xue, Y., Chou, S., Burlingame, A., Krogan, N.J., Alber, T., and Zhou, Q. (2010). HIV-1 Tat and host AFF4 recruit two transcription elongation factors into a bifunctional complex for coordinated activation of HIV-1 transcription. Mol Cell *38*, 428-438.

Hirose, Y., and Manley, J.L. (2000). RNA polymerase II and the integration of nuclear events. Genes Dev *14*, 1415-1429.

Huff, J.T., Plocik, A.M., Guthrie, C., and Yamamoto, K.R. (2010). Reciprocal intronic and exonic histone modification regions in humans. Nat Struct Mol Biol *17*, 1495-1499.

Hughes, C.M., Rozenblatt-Rosen, O., Milne, T.A., Copeland, T.D., Levine, S.S., Lee, J.C., Hayes, D.N., Shanmugam, K.S., Bhattacharjee, A., Biondi, C.A., *et al.* (2004). Menin associates with a trithorax family histone methyltransferase complex and with the hoxc8 locus. Mol Cell *13*, 587-597.

Izban, M.G., and Luse, D.S. (1991). Transcription on nucleosomal templates by RNA polymerase II in vitro: inhibition of elongation with enhancement of sequence-specific pausing. Genes Dev *5*, 683-696.

Izban, M.G., and Luse, D.S. (1992). The RNA polymerase II ternary complex cleaves the nascent transcript in a 3'----5' direction in the presence of elongation factor SII. Genes Dev *6*, 1342-1356.

Jang, M.K., Mochizuki, K., Zhou, M., Jeong, H.S., Brady, J.N., and Ozato, K. (2005). The bromodomain protein Brd4 is a positive regulatory component of P-TEFb and stimulates RNA polymerase II-dependent transcription. Mol Cell *19*, 523-534.

Jin, C., and Felsenfeld, G. (2007). Nucleosome stability mediated by histone variants H3.3 and H2A.Z. Genes Dev *21*, 1519-1529.

Jonkers, I., Kwak, H., and Lis, J.T. (2014). Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. Elife *3*, e02407.

Kadonaga, J.T. (2012). Perspectives on the RNA polymerase II core promoter. Wiley Interdiscip Rev Dev Biol *1*, 40-51.

Kaida, D., Berg, M.G., Younis, I., Kasim, M., Singh, L.N., Wan, L., and Dreyfuss, G. (2010). U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. Nature *468*, 664-668.

Kaplan, N., Moore, I.K., Fondufe-Mittendorf, Y., Gossett, A.J., Tillo, D., Field, Y., LeProust, E.M., Hughes, T.R., Lieb, J.D., Widom, J., *et al.* (2009). The DNA-encoded nucleosome organization of a eukaryotic genome. Nature *458*, 362-366.

Keogh, M.C., Kurdistani, S.K., Morris, S.A., Ahn, S.H., Podolny, V., Collins, S.R., Schuldiner, M., Chin, K., Punna, T., Thompson, N.J., *et al.* (2005). Cotranscriptional set2 methylation of histone H3 lysine 36 recruits a repressive Rpd3 complex. Cell *123*, 593-605.

Kim, H., Erickson, B., Luo, W., Seward, D., Graber, J.H., Pollock, D.D., Megee, P.C., and Bentley, D.L. (2010a). Gene-specific RNA polymerase II phosphorylation and the CTD code. Nat Struct Mol Biol *17*, 1279-1286.

Kim, J., Guermah, M., McGinty, R.K., Lee, J.S., Tang, Z., Milne, T.A., Shilatifard, A., Muir, T.W., and Roeder, R.G. (2009). RAD6-Mediated transcription-coupled H2B ubiquitylation directly stimulates H3K4 methylation in human cells. Cell *137*, 459-471.

Kim, J.B., and Sharp, P.A. (2001). Positive transcription elongation factor B phosphorylates hSPT5 and RNA polymerase II carboxyl-terminal domain independently of cyclin-dependent kinase-activating kinase. J Biol Chem *276*, 12317-12323.

Kim, M., Krogan, N.J., Vasiljeva, L., Rando, O.J., Nedea, E., Greenblatt, J.F., and Buratowski, S. (2004). The yeast Rat1 exonuclease promotes transcription termination by RNA polymerase II. Nature *432*, 517-522.

Kim, S., Kim, H., Fong, N., Erickson, B., and Bentley, D.L. (2011). Pre-mRNA splicing is a determinant of histone H3K36 methylation. Proc Natl Acad Sci U S A *108*, 13564-13569.

Kim, T.K., Hemberg, M., Gray, J.M., Costa, A.M., Bear, D.M., Wu, J., Harmin, D.A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., *et al.* (2010b). Widespread transcription at neuronal activity-regulated enhancers. Nature *465*, 182-187.

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

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

Kolodziej, P.A., Woychik, N., Liao, S.M., and Young, R.A. (1990). RNA polymerase II subunit composition, stoichiometry, and phosphorylation. Mol Cell Biol *10*, 1915-1920.

Komarnitsky, P., Cho, E.J., and Buratowski, S. (2000). Different phosphorylated forms of RNA polymerase II and associated mRNA processing factors during transcription. Genes Dev *14*, 2452-2460.

Krogan, N.J., Dover, J., Wood, A., Schneider, J., Heidt, J., Boateng, M.A., Dean, K., Ryan, O.W., Golshani, A., Johnston, M., *et al.* (2003a). The Paf1 complex is required for histone H3 methylation by COMPASS and Dot1p: linking transcriptional elongation to histone methylation. Mol Cell *11*, 721-729.

Krogan, N.J., Kim, M., Tong, A., Golshani, A., Cagney, G., Canadien, V., Richards, D.P., Beattie, B.K., Emili, A., Boone, C., *et al.* (2003b). Methylation of histone H3 by Set2 in Saccharomyces cerevisiae is linked to transcriptional elongation by RNA polymerase II. Mol Cell Biol *23*, 4207-4218.

Kwak, H., Fuda, N.J., Core, L.J., and Lis, J.T. (2013). Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. Science *339*, 950-953.

Kwek, K.Y., Murphy, S., Furger, A., Thomas, B., O'Gorman, W., Kimura, H., Proudfoot, N.J., and Akoulitchev, A. (2002). U1 snRNA associates with TFIIH and regulates transcriptional initiation. Nat Struct Biol *9*, 800-805.

LaCava, J., Houseley, J., Saveanu, C., Petfalski, E., Thompson, E., Jacquier, A., and Tollervey, D. (2005). RNA degradation by the exosome is promoted by a nuclear polyadenylation complex. Cell *121*, 713-724.

Lacoste, N., Utley, R.T., Hunter, J.M., Poirier, G.G., and Cote, J. (2002). Disruptor of telomeric silencing-1 is a chromatin-specific histone H3 methyltransferase. J Biol Chem *277*, 30421-30424.

Lai, F., Gardini, A., Zhang, A., and Shiekhattar, R. (2015). Integrator mediates the biogenesis of enhancer RNAs. Nature *525*, 399-403.

Laitem, C., Zaborowska, J., Isa, N.F., Kufs, J., Dienstbier, M., and Murphy, S. (2015). CDK9 inhibitors define elongation checkpoints at both ends of RNA polymerase II-transcribed genes. Nat Struct Mol Biol *22*, 396-403.

Le Hir, H., Gatfield, D., Izaurralde, E., and Moore, M.J. (2001). The exon-exon junction complex provides a binding platform for factors involved in mRNA export and nonsense-mediated mRNA decay. EMBO J *20*, 4987-4997.

Lee, N.N., Chalamcharla, V.R., Reyes-Turcu, F., Mehta, S., Zofall, M., Balachandran, V., Dhakshnamoorthy, J., Taneja, N., Yamanaka, S., Zhou, M., *et al.* (2013). Mtr4-like protein coordinates nuclear RNA processing for heterochromatin assembly and for telomere maintenance. Cell *155*, 1061-1074.

Lejeune, F., Li, X., and Maquat, L.E. (2003). Nonsense-mediated mRNA decay in mammalian cells involves decapping, deadenylating, and exonucleolytic activities. Mol Cell *12*, 675-687.

Lemay, J.F., Larochelle, M., Marguerat, S., Atkinson, S., Bahler, J., and Bachand, F. (2014). The RNA exosome promotes transcription termination of backtracked RNA polymerase II. Nat Struct Mol Biol *21*, 919-926.

Li, J., Moazed, D., and Gygi, S.P. (2002). Association of the histone methyltransferase Set2 with RNA polymerase II plays a role in transcription elongation. J Biol Chem *277*, 49383-49388.

Li, Q., Zhou, H., Wurtele, H., Davies, B., Horazdovsky, B., Verreault, A., and Zhang, Z. (2008). Acetylation of histone H3 lysine 56 regulates replication-coupled nucleosome assembly. Cell *134*, 244-255.

Li, W., You, B., Hoque, M., Zheng, D., Luo, W., Ji, Z., Park, J.Y., Gunderson, S.I., Kalsotra, A., Manley, J.L., *et al.* (2015). Systematic profiling of poly(A)+ transcripts modulated by core 3' end processing and splicing factors reveals regulatory rules of alternative cleavage and polyadenylation. PLoS Genet *11*, e1005166.

Lin, C., Smith, E.R., Takahashi, H., Lai, K.C., Martin-Brown, S., Florens, L., Washburn, M.P., Conaway, J.W., Conaway, R.C., and Shilatifard, A. (2010). AFF4, a component of the ELL/P-TEFb elongation complex and a shared subunit of MLL chimeras, can link transcription elongation to leukemia. Mol Cell *37*, 429-437.

Lin, C.Y., Loven, J., Rahl, P.B., Paranal, R.M., Burge, C.B., Bradner, J.E., Lee, T.I., and Young, R.A. (2012). Transcriptional amplification in tumor cells with elevated c-Myc. Cell *151*, 56-67.

Liu, Y., Warfield, L., Zhang, C., Luo, J., Allen, J., Lang, W.H., Ranish, J., Shokat, K.M., and Hahn, S. (2009). Phosphorylation of the transcription elongation factor Spt5 by yeast Bur1 kinase stimulates recruitment of the PAF complex. Mol Cell Biol *29*, 4852-4863.

Loven, J., Hoke, H.A., Lin, C.Y., Lau, A., Orlando, D.A., Vakoc, C.R., Bradner, J.E., Lee, T.I., and Young, R.A. (2013). Selective inhibition of tumor oncogenes by disruption of super-enhancers. Cell *153*, 320-334.

Lubas, M., Christensen, M.S., Kristiansen, M.S., Domanski, M., Falkenby, L.G., Lykke-Andersen, S., Andersen, J.S., Dziembowski, A., and Jensen, T.H. (2011). Interaction profiling identifies the human nuclear exosome targeting complex. Mol Cell *43*, 624-637.

Luger, K., Mader, A.W., Richmond, R.K., Sargent, D.F., and Richmond, T.J. (1997). Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature *389*, 251-260.

Luse, D.S. (2013). Promoter clearance by RNA polymerase II. Biochim Biophys Acta *1829*, 63-68.

Lutz, C.S., Murthy, K.G., Schek, N., O'Connor, J.P., Manley, J.L., and Alwine, J.C. (1996). Interaction between the U1 snRNP-A protein and the 160-kD subunit of cleavage-

polyadenylation specificity factor increases polyadenylation efficiency in vitro. Genes Dev *10*, 325-337.

Makino, D.L., Baumgartner, M., and Conti, E. (2013). Crystal structure of an RNA-bound 11-subunit eukaryotic exosome complex. Nature *495*, 70-75.

Mandel, C.R., Kaneko, S., Zhang, H., Gebauer, D., Vethantham, V., Manley, J.L., and Tong, L. (2006). Polyadenylation factor CPSF-73 is the pre-mRNA 3'-end-processing endonuclease. Nature *444*, 953-956.

Marquardt, S., Escalante-Chong, R., Pho, N., Wang, J., Churchman, L.S., Springer, M., and Buratowski, S. (2014). A chromatin-based mechanism for limiting divergent noncoding transcription. Cell *157*, 1712-1723.

Marshall, N.F., and Price, D.H. (1995). Purification of P-TEFb, a transcription factor required for the transition into productive elongation. J Biol Chem *270*, 12335-12338.

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

Mayer, A., Lidschreiber, M., Siebert, M., Leike, K., Soding, J., and Cramer, P. (2010). Uniform transitions of the general RNA polymerase II transcription complex. Nat Struct Mol Biol *17*, 1272-1278.

McCracken, S., Fong, N., Rosonina, E., Yankulov, K., Brothers, G., Siderovski, D., Hessel, A., Foster, S., Shuman, S., and Bentley, D.L. (1997a). 5'-Capping enzymes are targeted to pre-mRNA by binding to the phosphorylated carboxy-terminal domain of RNA polymerase II. Genes Dev *11*, 3306-3318.

McCracken, S., Fong, N., Yankulov, K., Ballantyne, S., Pan, G., Greenblatt, J., Patterson, S.D., Wickens, M., and Bentley, D.L. (1997b). The C-terminal domain of RNA polymerase II couples mRNA processing to transcription. Nature *385*, 357-361.

McGinty, R.K., Kim, J., Chatterjee, C., Roeder, R.G., and Muir, T.W. (2008). Chemically ubiquitylated histone H2B stimulates hDot1L-mediated intranucleosomal methylation. Nature *453*, 812-816.

Meola, N., Domanski, M., Karadoulama, E., Chen, Y., Gentil, C., Pultz, D., Vitting-Seerup, K., Lykke-Andersen, S., Andersen, J.S., Sandelin, A., *et al.* (2016). Identification of a Nuclear Exosome Decay Pathway for Processed Transcripts. Mol Cell *64*, 520-533.

Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.K., Koche, R.P., *et al.* (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature *448*, 553-560.

Miller, T., Krogan, N.J., Dover, J., Erdjument-Bromage, H., Tempst, P., Johnston, M., Greenblatt, J.F., and Shilatifard, A. (2001). COMPASS: a complex of proteins associated with a trithorax-related SET domain protein. Proc Natl Acad Sci U S A *98*, 12902-12907.

Mischo, H.E., Gomez-Gonzalez, B., Grzechnik, P., Rondon, A.G., Wei, W., Steinmetz, L., Aguilera, A., and Proudfoot, N.J. (2011). Yeast Sen1 helicase protects the genome from transcription-associated instability. Mol Cell *41*, 21-32.

Missra, A., and Gilmour, D.S. (2010). Interactions between DSIF (DRB sensitivity inducing factor), NELF (negative elongation factor), and the Drosophila RNA polymerase II transcription elongation complex. Proc Natl Acad Sci U S A *107*, 11301-11306.

Mitchell, P., and Tollervey, D. (2003). An NMD pathway in yeast involving accelerated deadenylation and exosome-mediated 3'-->5' degradation. Mol Cell *11*, 1405-1413.

Mohan, M., Herz, H.M., Takahashi, Y.H., Lin, C., Lai, K.C., Zhang, Y., Washburn, M.P., Florens, L., and Shilatifard, A. (2010). Linking H3K79 trimethylation to Wnt signaling through a novel Dot1-containing complex (DotCom). Genes Dev *24*, 574-589.

Muse, G.W., Gilchrist, D.A., Nechaev, S., Shah, R., Parker, J.S., Grissom, S.F., Zeitlinger, J., and Adelman, K. (2007). RNA polymerase is poised for activation across the genome. Nat Genet *39*, 1507-1511.

Neil, H., Malabat, C., d'Aubenton-Carafa, Y., Xu, Z., Steinmetz, L.M., and Jacquier, A. (2009). Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. Nature *457*, 1038-1042.

Ng, H.H., Xu, R.M., Zhang, Y., and Struhl, K. (2002). Ubiquitination of histone H2B by Rad6 is required for efficient Dot1-mediated methylation of histone H3 lysine 79. J Biol Chem *277*, 34655-34657.

Nguyen, V.T., Kiss, T., Michels, A.A., and Bensaude, O. (2001). 7SK small nuclear RNA binds to and inhibits the activity of CDK9/cyclin T complexes. Nature *414*, 322-325.

Nojima, T., Gomes, T., Grosso, A.R., Kimura, H., Dye, M.J., Dhir, S., Carmo-Fonseca, M., and Proudfoot, N.J. (2015). Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing. Cell *161*, 526-540.

Ntini, E., Jarvelin, A.I., Bornholdt, J., Chen, Y., Boyd, M., Jorgensen, M., Andersson, R., Hoof, I., Schein, A., Andersen, P.R., *et al.* (2013). Polyadenylation site-induced decay of upstream transcripts enforces promoter directionality. Nat Struct Mol Biol *20*, 923-928.

Nudler, E. (2012). RNA polymerase backtracking in gene regulation and genome instability. Cell *149*, 1438-1445.

Orphanides, G., LeRoy, G., Chang, C.H., Luse, D.S., and Reinberg, D. (1998). FACT, a factor that facilitates transcript elongation through nucleosomes. Cell *92*, 105-116.

Palmiter, R.D., Sandgren, E.P., Avarbock, M.R., Allen, D.D., and Brinster, R.L. (1991). Heterologous introns can enhance expression of transgenes in mice. Proc Natl Acad Sci U S A *88*, 478-482.

Pefanis, E., Wang, J., Rothschild, G., Lim, J., Chao, J., Rabadan, R., Economides, A.N., and Basu, U. (2014). Noncoding RNA transcription targets AID to divergently transcribed loci in B cells. Nature *514*, 389-393.

Pefanis, E., Wang, J., Rothschild, G., Lim, J., Kazadi, D., Sun, J., Federation, A., Chao, J., Elliott, O., Liu, Z.P., *et al.* (2015). RNA exosome-regulated long non-coding RNA transcription controls super-enhancer activity. Cell *161*, 774-789.

Preker, P., Almvig, K., Christensen, M.S., Valen, E., Mapendano, C.K., Sandelin, A., and Jensen, T.H. (2011). PROMoter uPstream Transcripts share characteristics with mRNAs and are produced upstream of all three major types of mammalian promoters. Nucleic Acids Res *39*, 7179-7193.

Preker, P., Nielsen, J., Kammler, S., Lykke-Andersen, S., Christensen, M.S., Mapendano, C.K., Schierup, M.H., and Jensen, T.H. (2008). RNA exosome depletion reveals transcription upstream of active human promoters. Science *322*, 1851-1854.

Proudfoot, N.J., and Brownlee, G.G. (1976). 3' non-coding region sequences in eukaryotic messenger RNA. Nature *263*, 211-214.

Rahl, P.B., Lin, C.Y., Seila, A.C., Flynn, R.A., McCuine, S., Burge, C.B., Sharp, P.A., and Young, R.A. (2010). c-Myc regulates transcriptional pause release. Cell *141*, 432-445.

Raisner, R.M., Hartley, P.D., Meneghini, M.D., Bao, M.Z., Liu, C.L., Schreiber, S.L., Rando, O.J., and Madhani, H.D. (2005). Histone variant H2A.Z marks the 5' ends of both active and inactive genes in euchromatin. Cell *123*, 233-248.

Ramirez-Carrozzi, V.R., Braas, D., Bhatt, D.M., Cheng, C.S., Hong, C., Doty, K.R., Black, J.C., Hoffmann, A., Carey, M., and Smale, S.T. (2009). A unifying model for the selective regulation of inducible transcription by CpG islands and nucleosome remodeling. Cell *138*, 114-128.

Ranjan, A., Mizuguchi, G., FitzGerald, P.C., Wei, D., Wang, F., Huang, Y., Luk, E., Woodcock, C.L., and Wu, C. (2013). Nucleosome-free region dominates histone acetylation in targeting SWR1 to promoters for H2A.Z replacement. Cell *154*, 1232-1245.

Rege, M., Subramanian, V., Zhu, C., Hsieh, T.H., Weiner, A., Friedman, N., Clauder-Munster, S., Steinmetz, L.M., Rando, O.J., Boyer, L.A., *et al.* (2015). Chromatin Dynamics and the RNA Exosome Function in Concert to Regulate Transcriptional Homeostasis. Cell Rep *13*, 1610-1622.

Reinberg, D., and Roeder, R.G. (1987). Factors involved in specific transcription by mammalian RNA polymerase II. Transcription factor IIS stimulates elongation of RNA chains. J Biol Chem *262*, 3331-3337.

Rhee, H.S., and Pugh, B.F. (2012). Genome-wide structure and organization of eukaryotic pre-initiation complexes. Nature *483*, 295-301.

Roeder, R.G., and Rutter, W.J. (1969). Multiple forms of DNA-dependent RNA polymerase in eukaryotic organisms. Nature *224*, 234-237.

Roeder, R.G., and Rutter, W.J. (1970). Specific nucleolar and nucleoplasmic RNA polymerases. Proc Natl Acad Sci U S A *65*, 675-682.

Rougvie, A.E., and 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*, 795-804.

Satchwell, S.C., Drew, H.R., and Travers, A.A. (1986). Sequence periodicities in chicken nucleosome core DNA. J Mol Biol *191*, 659-675.

Saxonov, S., Berg, P., and Brutlag, D.L. (2006). A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc Natl Acad Sci U S A *103*, 1412-1417.

Sayre, M.H., Tschochner, H., and Kornberg, R.D. (1992). Reconstitution of transcription with five purified initiation factors and RNA polymerase II from Saccharomyces cerevisiae. J Biol Chem *267*, 23376-23382.

Schaukowitch, K., Joo, J.Y., Liu, X., Watts, J.K., Martinez, C., and Kim, T.K. (2014). Enhancer RNA facilitates NELF release from immediate early genes. Mol Cell *56*, 29-42.

Schlackow, M., Nojima, T., Gomes, T., Dhir, A., Carmo-Fonseca, M., and Proudfoot, N.J. (2017). Distinctive Patterns of Transcription and RNA Processing for Human lincRNAs. Mol Cell *65*, 25-38.

Schneider, C., Kudla, G., Wlotzka, W., Tuck, A., and Tollervey, D. (2012). Transcriptome-wide analysis of exosome targets. Mol Cell *48*, 422-433.

Schneider, J., Wood, A., Lee, J.S., Schuster, R., Dueker, J., Maguire, C., Swanson, S.K., Florens, L., Washburn, M.P., and Shilatifard, A. (2005). Molecular regulation of histone H3 trimethylation by COMPASS and the regulation of gene expression. Mol Cell *19*, 849-856.

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

Schulz, D., Schwalb, B., Kiesel, A., Baejen, C., Torkler, P., Gagneur, J., Soeding, J., and Cramer, P. (2013). Transcriptome surveillance by selective termination of noncoding RNA synthesis. Cell *155*, 1075-1087.

Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thastrom, A., Field, Y., Moore, I.K., Wang, J.P., and Widom, J. (2006). A genomic code for nucleosome positioning. Nature *442*, 772-778.

Seila, A.C., Calabrese, J.M., Levine, S.S., Yeo, G.W., Rahl, P.B., Flynn, R.A., Young, R.A., and Sharp, P.A. (2008). Divergent transcription from active promoters. Science *322*, 1849-1851.

Selth, L.A., Sigurdsson, S., and Svejstrup, J.Q. (2010). Transcript Elongation by RNA Polymerase II. Annu Rev Biochem *79*, 271-293.

Sigova, A.A., Mullen, A.C., Molinie, B., Gupta, S., Orlando, D.A., Guenther, M.G., Almada, A.E., Lin, C., Sharp, P.A., Giallourakis, C.C., *et al.* (2013). Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells. Proc Natl Acad Sci U S A *110*, 2876-2881.

Simic, R., Lindstrom, D.L., Tran, H.G., Roinick, K.L., Costa, P.J., Johnson, A.D., Hartzog, G.A., and Arndt, K.M. (2003). Chromatin remodeling protein Chd1 interacts with transcription elongation factors and localizes to transcribed genes. EMBO J *22*, 1846-1856.

Skaar, J.R., Ferris, A.L., Wu, X., Saraf, A., Khanna, K.K., Florens, L., Washburn, M.P., Hughes, S.H., and Pagano, M. (2015). The Integrator complex controls the termination of transcription at diverse classes of gene targets. Cell Res *25*, 288-305.

Skene, P.J., Hernandez, A.E., Groudine, M., and Henikoff, S. (2014). The nucleosomal barrier to promoter escape by RNA polymerase II is overcome by the chromatin remodeler Chd1. Elife *3*, e02042.

Sklar, V.E., Schwartz, L.B., and Roeder, R.G. (1975). Distinct molecular structures of nuclear class I, II, and III DNA-dependent RNA polymerases. Proc Natl Acad Sci U S A *72*, 348-352.

Smolle, M., Venkatesh, S., Gogol, M.M., Li, H., Zhang, Y., Florens, L., Washburn, M.P., and Workman, J.L. (2012). Chromatin remodelers Isw1 and Chd1 maintain chromatin structure during transcription by preventing histone exchange. Nat Struct Mol Biol *19*, 884-892.

Sobhian, B., Laguette, N., Yatim, A., Nakamura, M., Levy, Y., Kiernan, R., and Benkirane, M. (2010). HIV-1 Tat assembles a multifunctional transcription elongation complex and stably associates with the 7SK snRNP. Mol Cell *38*, 439-451.

Stadelmayer, B., Micas, G., Gamot, A., Martin, P., Malirat, N., Koval, S., Raffel, R., Sobhian, B., Severac, D., Rialle, S., *et al.* (2014). Integrator complex regulates NELF-mediated RNA polymerase II pause/release and processivity at coding genes. Nat Commun *5*, 5531.

Steinmetz, E.J., and Brow, D.A. (1998). Control of pre-mRNA accumulation by the essential yeast protein Nrd1 requires high-affinity transcript binding and a domain implicated in RNA polymerase II association. Proc Natl Acad Sci U S A *95*, 6699-6704.

Steinmetz, E.J., Warren, C.L., Kuehner, J.N., Panbehi, B., Ansari, A.Z., and Brow, D.A. (2006). Genome-wide distribution of yeast RNA polymerase II and its control by Sen1 helicase. Mol Cell *24*, 735-746.

Sun, M., Lariviere, L., Dengl, S., Mayer, A., and Cramer, P. (2010). A tandem SH2 domain in transcription elongation factor Spt6 binds the phosphorylated RNA polymerase II C-terminal repeat domain (CTD). J Biol Chem *285*, 41597-41603.

Sun, Z.W., and Allis, C.D. (2002). Ubiquitination of histone H2B regulates H3 methylation and gene silencing in yeast. Nature *418*, 104-108.

Takahashi, H., Parmely, T.J., Sato, S., Tomomori-Sato, C., Banks, C.A., Kong, S.E., Szutorisz, H., Swanson, S.K., Martin-Brown, S., Washburn, M.P., *et al.* (2011). Human mediator subunit MED26 functions as a docking site for transcription elongation factors. Cell *146*, 92-104.

Tenney, K., and Shilatifard, A. (2005). A COMPASS in the voyage of defining the role of trithorax/MLL-containing complexes: linking leukemogensis to covalent modifications of chromatin. J Cell Biochem *95*, 429-436.

Thiebaut, M., Kisseleva-Romanova, E., Rougemaille, M., Boulay, J., and Libri, D. (2006). Transcription termination and nuclear degradation of cryptic unstable transcripts: a role for the nrd1-nab3 pathway in genome surveillance. Mol Cell *23*, 853-864.

Thompson, C.M., Koleske, A.J., Chao, D.M., and Young, R.A. (1993). A multisubunit complex associated with the RNA polymerase II CTD and TATA-binding protein in yeast. Cell *73*, 1361-1375.

Thomson, J.P., Skene, P.J., Selfridge, J., Clouaire, T., Guy, J., Webb, S., Kerr, A.R., Deaton, A., Andrews, R., James, K.D., *et al.* (2010). CpG islands influence chromatin structure via the CpG-binding protein Cfp1. Nature *464*, 1082-1086.

van Hoof, A., Lennertz, P., and Parker, R. (2000). Yeast exosome mutants accumulate 3'-extended polyadenylated forms of U4 small nuclear RNA and small nucleolar RNAs. Mol Cell Biol *20*, 441-452.

Vasiljeva, L., and Buratowski, S. (2006). Nrd1 interacts with the nuclear exosome for 3' processing of RNA polymerase II transcripts. Mol Cell *21*, 239-248.

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

Wada, T., Takagi, T., Yamaguchi, Y., Ferdous, A., Imai, T., Hirose, S., Sugimoto, S., Yano, K., Hartzog, G.A., Winston, F., *et al.* (1998a). DSIF, a novel transcription elongation factor that regulates RNA polymerase II processivity, is composed of human Spt4 and Spt5 homologs. Genes Dev *12*, 343-356.

Wada, T., Takagi, T., Yamaguchi, Y., Watanabe, D., and Handa, H. (1998b). Evidence that P-TEFb alleviates the negative effect of DSIF on RNA polymerase II-dependent transcription in vitro. EMBO J *17*, 7395-7403.

Wade, P.A., Werel, W., Fentzke, R.C., Thompson, N.E., Leykam, J.F., Burgess, R.R., Jaehning, J.A., and Burton, Z.F. (1996). A novel collection of accessory factors associated with yeast RNA polymerase II. Protein Expr Purif *8*, 85-90.

Wahl, M.C., Will, C.L., and Luhrmann, R. (2009). The spliceosome: design principles of a dynamic RNP machine. Cell *136*, 701-718.

Weber, C.M., Ramachandran, S., and Henikoff, S. (2014). Nucleosomes are context-specific, H2A.Z-modulated barriers to RNA polymerase. Mol Cell *53*, 819-830.

West, S., Gromak, N., and Proudfoot, N.J. (2004). Human 5' --> 3' exonuclease Xrn2 promotes transcription termination at co-transcriptional cleavage sites. Nature *432*, 522-525.

Williams, L.H., Fromm, G., Gokey, N.G., Henriques, T., Muse, G.W., Burkholder, A., Fargo, D.C., Hu, G., and Adelman, K. (2015). Pausing of RNA polymerase II regulates mammalian developmental potential through control of signaling networks. Mol Cell *58*, 311-322.

Wood, A., Krogan, N.J., Dover, J., Schneider, J., Heidt, J., Boateng, M.A., Dean, K., Golshani, A., Zhang, Y., Greenblatt, J.F., *et al.* (2003a). Bre1, an E3 ubiquitin ligase required for recruitment and substrate selection of Rad6 at a promoter. Mol Cell *11*, 267-274.

Wood, A., Schneider, J., Dover, J., Johnston, M., and Shilatifard, A. (2003b). The Paf1 complex is essential for histone monoubiquitination by the Rad6-Bre1 complex, which signals for histone methylation by COMPASS and Dot1p. J Biol Chem *278*, 34739-34742.

Wu, M., Wang, P.F., Lee, J.S., Martin-Brown, S., Florens, L., Washburn, M., and Shilatifard, A. (2008). Molecular regulation of H3K4 trimethylation by Wdr82, a component of human Set1/COMPASS. Mol Cell Biol *28*, 7337-7344.

Wu, X., and Bartel, D.P. (2017). Widespread Influence of 3'-End Structures on Mammalian mRNA Processing and Stability. Cell *169*, 905-917 e911.

Wyers, F., Rougemaille, M., Badis, G., Rousselle, J.C., Dufour, M.E., Boulay, J., Regnault, B., Devaux, F., Namane, A., Seraphin, B., *et al.* (2005). Cryptic pol II transcripts are degraded by a nuclear quality control pathway involving a new poly(A) polymerase. Cell *121*, 725-737.

Xie, L., Pelz, C., Wang, W., Bashar, A., Varlamova, O., Shadle, S., and Impey, S. (2011). KDM5B regulates embryonic stem cell self-renewal and represses cryptic intragenic transcription. EMBO J *30*, 1473-1484.

Xin, H., Takahata, S., Blanksma, M., McCullough, L., Stillman, D.J., and Formosa, T. (2009). yFACT induces global accessibility of nucleosomal DNA without H2A-H2B displacement. Mol Cell *35*, 365-376.

Xu, F., Zhang, K., and Grunstein, M. (2005). Acetylation in histone H3 globular domain regulates gene expression in yeast. Cell *121*, 375-385.

Xu, Z., Wei, W., Gagneur, J., Perocchi, F., Clauder-Munster, S., Camblong, J., Guffanti, E., Stutz, F., Huber, W., and Steinmetz, L.M. (2009). Bidirectional promoters generate pervasive transcription in yeast. Nature *457*, 1033-1037.

Yamada, T., Yamaguchi, Y., Inukai, N., Okamoto, S., Mura, T., and Handa, H. (2006). P-TEFb-mediated phosphorylation of hSpt5 C-terminal repeats is critical for processive transcription elongation. Mol Cell *21*, 227-237.

Yamaguchi, Y., Takagi, T., Wada, T., Yano, K., Furuya, A., Sugimoto, S., Hasegawa, J., and Handa, H. (1999). NELF, a multisubunit complex containing RD, cooperates with DSIF to repress RNA polymerase II elongation. Cell *97*, 41-51.

Yamanaka, S., Yamashita, A., Harigaya, Y., Iwata, R., and Yamamoto, M. (2010). Importance of polyadenylation in the selective elimination of meiotic mRNAs in growing S. pombe cells. EMBO J *29*, 2173-2181.

Yang, Z., Yik, J.H., Chen, R., He, N., Jang, M.K., Ozato, K., and Zhou, Q. (2005). Recruitment of P-TEFb for stimulation of transcriptional elongation by the bromodomain protein Brd4. Mol Cell *19*, 535-545.

Yang, Z., Zhu, Q., Luo, K., and Zhou, Q. (2001). The 7SK small nuclear RNA inhibits the CDK9/cyclin T1 kinase to control transcription. Nature *414*, 317-322.

Yen, K., Vinayachandran, V., and Pugh, B.F. (2013). SWR-C and INO80 chromatin remodelers recognize nucleosome-free regions near +1 nucleosomes. Cell *154*, 1246-1256.

Yik, J.H., Chen, R., Nishimura, R., Jennings, J.L., Link, A.J., and Zhou, Q. (2003). Inhibition of P-TEFb (CDK9/Cyclin T) kinase and RNA polymerase II transcription by the coordinated actions of HEXIM1 and 7SK snRNA. Mol Cell *12*, 971-982.

Yonaha, M., and Proudfoot, N.J. (1999). Specific transcriptional pausing activates polyadenylation in a coupled in vitro system. Mol Cell *3*, 593-600.

Young, R.A. (1991). RNA polymerase II. Annu Rev Biochem *60*, 689-715.

Yuan, G.C., Liu, Y.J., Dion, M.F., Slack, M.D., Wu, L.F., Altschuler, S.J., and Rando, O.J. (2005). Genome-scale identification of nucleosome positions in S. cerevisiae. Science *309*, 626-630.

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

Zeitlinger, J., Stark, A., Kellis, M., Hong, J.W., Nechaev, S., Adelman, K., Levine, M., and Young, R.A. (2007). RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo. Nat Genet *39*, 1512-1516.

Zhou, K., Kuo, W.H., Fillingham, J., and Greenblatt, J.F. (2009). Control of transcriptional elongation and cotranscriptional histone modification by the yeast BUR kinase substrate Spt5. Proc Natl Acad Sci U S A *106*, 6956-6961.

Zhou, Y., Zhu, J., Schermann, G., Ohle, C., Bendrin, K., Sugioka-Sugiyama, R., Sugiyama, T., and Fischer, T. (2015). The fission yeast MTREC complex targets CUTs and unspliced pre-mRNAs to the nuclear exosome. Nat Commun *6*, 7050.

Zhu, B., Zheng, Y., Pham, A.D., Mandal, S.S., Erdjument-Bromage, H., Tempst, P., and Reinberg, D. (2005). Monoubiquitination of human histone H2B: the factors involved and their roles in HOX gene regulation. Mol Cell *20*, 601-611.

Chapter 2

The RNA Exosome Promotes Premature Termination in the First Intron

This chapter is adapted from the first half of the following manuscript as well as unpublished data:

Anthony C. Chiu, Hiroshi I. Suzuki, Xuebing Wu, Dig B. Mahat, Andrea J. Kriz, and Phillip A Sharp. U1 snRNP Suppresses Premature Polyadenylation at Transcription Pause Sites Associated with Stable Nucleosomes.

Contributions:

AC and HS designed the experiments, performed U1 inhibition and performed computational analyses. AC and XW generated 2P-seq data. AK generated the Exosc3 CKO cell line. AC generated RNA-seq libraries. AC and DM generated PRO-seq libraries.

2.1 ABSTRACT

Divergent transcription is observed at the promoters of most active mammalian protein-coding genes, but productive transcription elongation is primarily limited to the mRNA or sense direction. Previously, we defined a U1-PAS axis, in which enrichment of polyadenylation signal (PAS) motifs and depletion of U1 snRNP binding sites promote early termination at upstream antisense RNAs. By generating a conditional *Exosc3* deletion mouse embryonic stem cell (mESC) system, we find the majority of low-abundant noncoding RNAs including long noncoding RNAs (lncRNAs), enhancer RNAs (eRNAs) and uaRNAs were upregulated upon Exosc3 depletion. Approximately 40% of uaRNAs have detectable poly(A) ends and are substrates of the RNA exosome. Surprisingly, Exosc3 depletion and U1 inhibition both result in an increase in detectable premature termination events in the first intron, suggesting they may function together in a similar pathway. Additionally, Exosc3 loss results in increased promoter-proximal pausing. Our results further expand the roles of the RNA exosome to include regulating the stability of prematurely terminated transcripts and regulation of promoter proximal pausing.

2.2 INTRODUCTION

High-throughput sequencing of the mammalian transcriptome revealed two major phenomena. First, transcription is primarily divergent; at most mammalian promoters, RNA Polymerase II (Pol II) transcribes divergently from the transcription start sites (TSS), forming a stable mRNA and an unstable transcript called an upstream antisense RNA (uaRNA) in mice or PROMPT in humans (Core et al., 2008; Preker et al., 2008; Seila et al., 2008), whereas active enhancers are also known to produce divergent, unstable enhancer RNAs (eRNAs) (Kim et al., 2010). Secondly, transcription is pervasive but selective; a significant fraction the genome is transcribed (Birney et al 2007, Djebali et al 2012), but the majority of these transcripts (including uaRNAs and eRNAs) are found at low copies per cell except for situations where a stable, protein-coding transcript is produced.

The link between these two phenomena centers on the activity of the RNA exosome, a 3'-to-5' exoribonuclease that functions in many pathways including degrading improperly spliced transcripts (Bousquet-Antonelli et al., 2000), resolution of backtracked Pol II (Lemay et al., 2014), ribosomal RNA processing (Mitchell et al., 1997) and nonsense mediated decay (Mitchell and Tollervey, 2003). Importantly, the RNA exosome also regulates the production of various low-abundant noncoding RNAs including cryptic unstable transcripts in *S cerevisiae* (Wyers et al., 2005), uaRNAs (Flynn et al., 2011; Preker et al., 2008) and eRNAs (Andersson et al., 2014).

The relative frequency of U1 splicing signals and polyadenylation signal (PAS) motifs (U1-PAS axis) plays a critical role in suppressing antisense transcription (Almada et al, 2013; Ntini et al, 2013; Core et al 2014). High levels of PAS motifs throughout the genome promote early transcription termination wherever there is an initiation event, a process suggested to destabilize transcripts (Andersen et al., 2012). However, mRNA genes have evolved a low level of PAS motifs across the transcription unit and an enrichment for 5' splice sites proximal to the transcription start

site (TSS). Recognition of the 5' splice site by U1 snRNP suppresses the use of nearby PAS motifs by the 3'-end processing machinery, promoting elongation and synthesis of mature RNA (Berg et al., 2012; Kaida et al., 2010).

It remains unclear how premature termination by the RNA exosome interacts with the polyadenylated RNA substrates from the U1-PAS axis. Here, we investigate the link between the U1-PAS axis and the RNA exosome by creating a conditional *Exosc3* knockout cell line. Transcriptome profiling suggests that many noncoding RNAs are upregulated upon Exosc3 removal. Of those, about 40% of uaRNAs and 30% of eRNAs had detectable poly(A) ends, many of which were upregulated upon Exosc3 knockout suggesting a role for the RNA exosome at degrading polyadenylated transcripts. Surprisingly, Exosc3 removal also stabilizes premature termination events within the first intron, similar to U1 inhibition. Moreover, promoter proximal pausing is modestly increased upon Exosc3 loss, but it does not appear to be related to CpG islands. These results suggest that the RNA exosome broadly functions in regulating sense transcription, in addition to its more well-known role at regulating noncoding RNAs.

2.3 RESULTS

Generation of Conditional Exosc3-deletion System

To further identify exosome-targeted transient RNA species, we generated a doxycycline (dox)-inducible Exosc3 conditional knockout (CKO) mESC line where the core RNA exosome subunit, Exosc3, was conditionally depleted. The CRISPR-Cas9 system was used to delete the entire endogenous Exosc3 gene in a mESC line expressing dox-inducible C-terminus FLAG-HA-tagged Exosc3 (Exosc3-FH) (**Fig. 1A**). Since we initially obtained a clone with a heterozygote deletion (data not shown), this cell line was transfected with additional sgRNAs to inactivate the

other allele, generating a dox-inducible *Exosc3* conditional knockout (CKO) cell line (**Fig. S1A**). Deletion of the two alleles was validated by Sanger sequencing, revealing different deletions at the two alleles (**Fig. S1B**). After 3 days of dox withdrawal, quantitative real time PCR (qRT-PCR) demonstrated high efficiency of conditional Exosc3 depletion in our system (**Fig. S1C**). An increase in cell death was observed at this time point, suggesting Exosc3 is an essential gene.

We performed RNA-Seq on rRNA-depleted RNA from Exosc3 CKO mESCs after 3 days of dox removal and confirmed specific loss of Exosc3 transcripts (**Fig. 1B**). Metaplots of RNA-seq reads around the transcription start site (TSS) of UCSC canonical transcripts revealed a 6-fold stabilization of uaRNAs upon Exosc3 depletion, but little change in overall RNA reads in the sense direction (**Fig. 1C**). Similarly, alignment of RNA-seq reads around intergenic enhancers defined by Oct4, Sox2 and Nanog (OSN) chromatin immunoprecipitation (ChIP)-seq peaks revealed that eRNAs are stabilized by 8-fold upon loss of Exosc3 (**Fig. 1D**), confirming the role of the RNA exosome in suppressing eRNA transcripts.

De Novo Transcriptome Assembly

Next, to obtain profiles of normally suppressed transcripts in the presence of RNA exosomes with accurate transcript architectures, the RNA transcriptome was assembled de novo using the Stringtie algorithm (Pertea et al., 2015) after pooling RNA-Seq libraries, followed by various filtering steps to categorize transcript classes (**Fig. 2A**). For instance, uaRNAs were defined as divergent transcripts with a 5' end within 1 kb upstream and antisense of the closest gene TSS (**Fig. S2A**), whereas convergent transcripts were antisense RNAs that overlapped the gene TSS (**Fig. S2B**) (Mayer et al., 2015). Enhancer eRNAs (eRNAs) were defined as transcripts overlapping a 1 kb window of an OSN enhancer peak (**Fig. S2C**).

We identified 3,336 high-confidence uaRNAs, with a median distance of 135 bp to the closest corresponding gene TSS (**Fig. S2D**). Approximately 29% of surveyed expressed genes (FPKM > 0.5) were producing uaRNAs, less than the roughly 2/3 of expressed promoters that prior studies have found to be divergent (Core et al., 2008; Seila et al., 2008). This is likely due to three reasons. First, prior approaches did not focus strictly on antisense transcripts that initiated upstream of genes, but rather called transcription events upstream and antisense of genes as divergent. Including convergent transcripts as well as bidirectional genes to capture all upstream antisense events results in 48% of expressed genes being divergent, a much closer number than based on uaRNAs alone. We attribute the remaining differences due to detection limits of RNA-seq and more stringent thresholds. GRO-seq (Core et al., 2008) and small RNA sequencing (Seila et al., 2008) are both of capable of capturing shorter RNAs, whereas RNA-seq is limited by a size-selection that removes adapter dimers. Moreover, the previous threshold for defining a divergent gene required at least one antisense read and one sense read within 1.5 kb of the TSS (Seila et al., 2008). Using the same thresholds, 68% of expressed genes from our RNA-seq are divergent, similar to the 67% from the initial divergent transcription paper (**Fig. S2E**).

Many Noncoding RNAs are Upregulated Following Removal of Exosc3

Based on *de novo* transcript assembly, we found that most uaRNAs, super-enhancer-associated enhancer RNAs (seRNA), and typical enhancer-associated enhancer RNAs (teRNA) were significantly upregulated upon loss of the RNA exosome (minimum 2-fold change, FDR<0.1) (**Fig. 2C-D**). Previously identified long noncoding RNAs (lncRNAs) changed more modestly, 1.5 fold, upon Exosc3 depletion. Novel lncRNAs identified in this study were more significantly upregulated than previously identified lncRNAs (data not shown), but this class may

be contaminated with eRNAs originating from enhancers other than Oct4/Sox2/Nanog enhancers. Nevertheless, this suggests genome-wide studies identifying lncRNAs may have missed lncRNAs that are normally degraded by the RNA exosome.

A substantial fraction (28%) of mRNAs encoding protein changed upon Exosc3 depletion (FDR<0.1, minimum 2-fold change). Gene set enrichment analysis (GSEA) revealed that p53-target genes were significantly changing (**Fig. 3A-B**), consistent with observations that Exosc3 CKO cells detached the longer doxycycline was withdrawn. Accordingly, p53 protein levels increased upon Exosc3 depletion, peaking 2 days after dox removal, though the increase was lower than a doxorubicin-treated control (**Fig. 3C**). A gradual increase in cleaved caspase signal was also observed, indicating apoptosis (**Fig. 3D**). We further observed an increase in γ-H2AX upon Exosc3 removal, consistent with other reports suggesting that Exosc3 removal leads to genomic instability (**Fig. 3C**) (Pefanis et al., 2015). In addition, changes in genes linked with differentiation of mESCs were also detected by GSEA, consistent with studies showing that p53 activation in mESCs promotes differentiation (Li et al., 2012; Lin et al., 2005). While Oct4 and Sox2 expression levels did not change, levels of Klf4, Nanog and Esrrb decreased (**Fig. 3E**), suggesting a possible conversion from a naïve stem cell state to a primed stem cell state marked by lower expression of Nanog, Esrrb and Klf4 (Hackett and Surani, 2014).

Polyadenylated uaRNAs and eRNAs are Substrates of the RNA Exosome

Using poly(A)-primed sequencing (2P-seq) (Spies et al., 2013), we generated a genome-wide dataset of cleavage sites from polyadenylated transcripts from the Exosc3 CKO mESCs, whereby the cleavage site is defined as the last nucleotide before the addition of a poly(A) tail. The putative cleavage sites were further filtered to remove sequencing artifacts from priming to

internal A-stretches, and subdivided into those containing one of 36 PAS motif variants within an upstream 80 bp window and those without (Almada et al., 2013). Cleavage sites with nearby PAS motifs comprise the majority of all used cleavage sites (**Fig. 4A,S3A**). At mRNA ends, termination with a PAS motif accounts for greater than 95% of all poly(A) reads. In contrast, at uaRNAs and eRNAs, termination with PAS motif is less frequent, accounting for roughly three quarters of poly(A) reads. About 40% of individual unique cleavages sites do not have an associated PAS hexamer variant (**Fig. S3A**) and are used less frequently than the two canonical PAS hexamer motifs (A[A/T]TAAA) (**Fig. S3B**). These may be degradation intermediates with short oligo(A) tails captured at low frequency in our poly(A) selection. Hence, we focused on cleavage sites with nearby PAS motifs.

Alignment of unique cleavage sites with the 36 PAS variants (PAS termination) around the TSS revealed stabilization and thus detection of significantly more PAS-linked cleavage sites in the uaRNA direction upon Exosc3 loss (**Fig. 4B**). These sites peaked around -1 kb from the TSS, mirroring the point where the frequency of predicted PAS motifs reach intergenic levels (**Fig. 4C**). Consistent with the major effect of Exosc3 loss being stabilization of uaRNAs, the half-lives of individual uaRNAs increased by 2-3 fold upon depletion of exosome activity following transcription arrest with flavopiridol (**Fig. 4D, S3C**). 40% of annotated uaRNAs detected by RNA-seq had detectable cleavage sites with a PAS motif in this analysis, suggesting there are also additional PAS-independent pathways that degrade uaRNAs (Meola et al., 2016).

Many enhancers generate bidirectional transcripts (Kim et al., 2010), and a fraction are thought to be polyadenylated (Hsieh et al., 2014). Alignment of unique PAS-mediated cleavage sites around OSN enhancers detected few polyadenylated cleavage sites in the presence of Exosc3, but showed a substantial increase in detectable polyadenylated cleavage sites upon removal of

Exosc3 (**Fig. 4E**), suggesting a subset of enhancers generates exosome sensitive polyadenylated RNAs. There is a depletion for predicted PAS motifs immediately flanking the center of the Oc4/Sox2/Nanog peaks, which matches where termination occurs (**Fig. 4F**). This depletion is not as striking as that near the TSS, so it could be due to enrichment for sequences that bind Oct4, Sox2 or Nanog, or active selection against having PAS motifs. The latter model would indicate a selection towards producing longer eRNAs, suggesting a functional role. Approximately 31% of defined eRNAs or 23% of 2 kb regions flanking OSN peaks generated detectable cleavage sites with PAS motifs, consistent with reports that many eRNAs are not polyadenylated, but rather cleaved through an integrator-dependent mechanism (Kim et al., 2010; Lai et al., 2015).

Premature Termination in the First Intron upon Exosc3 Removal

Unexpectedly, upon Exosc3 depletion, there was a dramatic increase in unique cleavage sites within the gene body peaking at 800 nts downstream of the TSS for approximately 3500 of all genes (**Fig. 4B**). These sites match the position where PAS motif frequency reaches the intragenic background levels (**Fig. 4C**). These results suggest that prematurely-terminated sense transcripts with these PAS sites are additional substrates of the RNA exosome. Since our RNA-seq alignments around the TSS did not reveal Exosc3-dependent stabilization (**Fig. 1C**), we hypothesized that abundant cytoplasmic mRNAs were masking low-abundant reads linked with premature termination. After filtering out exonic reads, we find that Exosc3 depletion causes an increase in intronic RNA-seq reads proximal to the first 5' splice site, and gradually diminishes throughout the first intron (**Fig. 5A**).

One potential explanation for the increase in the first intron signal is a stabilization of lariat intermediates. However, this possibility is unlikely because there is no increase in intronic RNA-

seq reads at the fourth intron upon Exosc3 removal (**Fig. 5B**). This also suggests that the increase in reads in the first intron is not due to a general stimulation of transcription upon exosome depletion. Moreover, cleavage sites stabilized by exosome depletion are found almost exclusively in the first intron and not in the fourth intron, further arguing that Pol II termination is primarily restricted to TSS proximal sequences (**Fig. 5C,D**). This is illustrated by profile of a representative gene, Gnpat, showing increased cleavage site usage proximal to the TSS (**Fig. 5G**). The increase in termination remains specific to the first intron after normalizing for gene expression (**Fig. S3D-E**), further reinforcing the specificity of PAS-dependent termination within the first intron. Intriguingly, some of these premature events have been previously sequenced in cDNA annotations (premature Rad23b is AK163379, premature Pcf11 is BC048838 and premature Psmd14 is AK014293), consistent with these RNAs being contiguous transcripts from the TSS.

Upon Exosc3 removal, a spike in poly(A) reads appears within 5 nucleotides of the annotated 3' splice site specifically at all 3' splice sites (**Fig. S4A**). Since this phenotype is different than the phenotype observed in the first intron, this suggests that fully spliced out lariats are degraded by the RNA exosome in a process involving oligo(A). Supporting this, these termination events are not associated with strong PAS motifs (**Fig. S4B**). Since RNA-seq alignments did not reveal a general stabilization of introns upon Exosc3 removal (**Fig. 5B**), stabilization of spliced introns is rare, and a compensatory pathway such as the previously described Xrn1 pathway (Hilleren and Parker, 2003) is likely the main mode of degrading spliced-out lariats.

Pol II Density is Increased in First Intron upon Exosc3 Removal

The increase in RNAs within the first intron upon depletion of the RNA exosome could be due to either increased production of RNA within the first intron or alternatively to stabilization

of RNAs. Precision run-on sequencing (PRO-seq) was performed to maps the position of actively transcribing Pol II upon Exosc3 removal (Kwak et al., 2013; Mahat et al., 2016). While this library is limited by low read depth[1], Exosc3 removal results in an increase in Pol II occupancy at the first 5' SS, but not to the same extent as the RNA-seq or 2P-seq (**Fig. 5E-F**). This suggests the increase in RNAs in the first intron upon Exosc3 loss is largely due to RNA stabilization, but transcriptional changes may also contribute. Similar to results from RNA-seq and 2P-seq, actively elongating Pol II diminishes throughout the first intron but not the fourth intron, probably reflecting a combination of inefficient Pol II gradually becoming more processive (Jonkers et al., 2014) and early termination within the first intron.

More generally, there is increased promoter proximal pausing in both directions of the TSS upon Exosc3 loss (**Fig. S5A**), quantified using the traveling ratio (promoter reads/body reads) (KS test, $p<2.12x10^{-16}$) (Rahl et al., 2010) (**Fig. S5B**). Mammalian promoters tend to have CpG islands (CGI), which promote formation of R-loops (Ginno et al., 2012). Given the RNA exosome promotes the resolution of R-loops (Pefanis et al., 2015), the increase in pausing upon Exosc3 removal may be due to increased stability of CpG-mediated R-loops, which are known to interfere with transcription elongation. While CGI promoters have more pausing than those without CGIs (**Fig. S5C**), both promoter types exhibit an increase in promoter-proximal pausing upon Exosc3 removal, arguing that increase in pausing upon Exosc3 removal is not due to the presence of CpG islands. There was no difference in nascent transcription around enhancers upon Exosc3 removal (**Fig. S5D**), suggesting the RNA exosome primarily functions to degrade eRNAs instead of regulating its transcription.

[1] The majority of events mapped to rRNA repeats. In the future, we will have to select away rRNAs when doing the modified run-on reaction.

Suppression of Sense Direction PAS Termination by U1 snRNP

It has been previously reported that inhibition of U1 activity promotes the use of early PAS motifs in mammalian cells (Almada et al., 2013; Kaida et al., 2010). Since exosome-regulated promoter-proximal termination of uaRNA and sense RNA within the first intron are linked to PAS motifs, we investigated their dependence on U1 snRNP recognition. Exosc3 CKO cells were either cultured in the presence or absence of doxycycline for 40 hours and then either treated with scrambled (Scr) control antisense morpholino oligonucleotide (AMO) or U1 AMO, antisense to sequences recognizing the 5' splice site, for an additional 8 hours (**Fig. 6A**). We chose to use a 2 day treatment off doxycycline due to technical limitations of nucleofection as well as trying to minimize off-target gene expression from cells undergoing apoptosis. As expected from previous results, in 2P-seq analysis, there was a dramatic increase in detectable PAS-linked cleavage sites in the upstream antisense direction following Exosc3 depletion, but the effects of U1 inhibition were minor (**Fig. 6B**). In contrast, in the sense direction from the TSS, both Exosc3 depletion and inhibition of U1 recognition significantly increased the number of detectable PAS linked cleavage sites. More importantly, the combination of U1 inhibition and Exosc3 depletion resulted in a further increase in PAS cleavage sites. This suggests that U1 recognition suppresses production of PAS-terminated transcripts in the first intron that are rapidly degraded by the exosome. As for eRNAs, the effects of U1 inhibition were almost negligible (**Fig. 6C**). We ascribe the lack of change upon U1 inhibition in uaRNAs and eRNAs to the absence of 5' splice site signal enrichment upstream of uaRNAs (Almada et al., 2013) or flanking the enhancer ChIP sites (**Fig. 6D**). In contrast, a strong 5' splice site signal is commonly found downstream of the TSS in the sense direction, suppressing premature PAS termination.

Similarly, the combinatorial effects of U1 inhibition and Exosc3 depletion were similarly observed in the first intron but not in the fourth intron (**Fig. 6E**). In contrast to a small impact of Exosc3 depletion, U1 inhibition led to about 2 fold increase in PAS-linked unique cleavage sites in the 4th intron, consistent with the idea that U1 suppresses usage of nearby PAS sites throughout the gene (Kaida et al, 2010). 3' RACE analysis and sequencing analysis using gene-specific primers for several genes confirmed that RNA terminated at the predicted site in the 1st intron (**Fig. S6A**).

Because sites of cleavage and polyadenylation can vary locally downstream of a PAS hexamer (**Fig. S6B**), we combined neighboring cleavage sites within 25 nucleotides into cleavage clusters, and focused on 2P clusters with PAS motifs or its variants. Hierarchical clustering of 2P clusters that overlapped the first intron and had at least 10 reads confirmed reproducibility among replicates and showed that over half of the clusters showed significantly higher 2P-seq signals when both Exosc3 and U1 activity are reduced, suggesting they may function in a similar pathway, perhaps by the RNA exosome stabilizing U1-regulated early termination products (**Fig. 6F,S6C**). Although it is possible that there are classes of PAS-terminated RNAs that are differentially regulated by U1 and exosome, it is difficult to confidently assess validity of possible subclasses due to technical limitations. Unlike 2P clusters within the first intron, almost all PAS termination in uaRNAs were primarily Exosc3-responsive (**Fig. 6G**).

2.4 DISCUSSION

Most early studies of divergent transcription in mammals were limited by the use of RNAi to knock down subunits of the RNA exosome (Flynn et al., 2011; Preker et al., 2011; Preker et al., 2008). A previous study as well as ours both argue that the RNA exosome is essential (Pefanis et al., 2014), so early RNAi studies would have selected against strong knockdown. Here, we used CRISPR-mediated approaches to create a doxycycline-regulated Exosc3 knockout cell line for studying RNAs modulated by the RNA exosome. This system allowed us to define specific transcribed regions with higher confidence than prior studies, which simply looked for transcription in an arbitrary window upstream of the TSS. uaRNAs and eRNAs were the most sensitive to RNA exosome depletion, followed by lncRNAs in between and coding mRNAs were the least sensitive, consistent with reports from RNAi-based approaches (Andersson et al., 2014; Preker et al., 2011; Preker et al., 2008). This suggests that the RNA exosome broadly acts a general pathway for degrading unwanted noncoding RNA transcription in mammals, similar to its roles in yeast at regulating cryptic unstable transcripts (CUTs) (Wyers et al., 2005).

We previously proposed that a differential distribution of U1 and PAS sites were regulating premature termination, but there were several unknowns (Almada et al., 2013). Firstly, given PAS termination is usually associated with stable transcripts, were PAS-terminated uaRNAs a rare stable subset or were PAS-terminated uaRNAs unstable? Here, we show that these PAS-terminated uaRNAs are degraded by the RNA exosome and that knockout of the RNA exosome increases their half-lives, arguing that polyadenylated uaRNAs are actively degraded. Secondly, how frequent were these PAS sites being used? We now show that they are found in a subset of uaRNAs (approximately 40%), supporting recent discoveries suggesting that there are numerous pathways degrading ncRNAs (Meola et al., 2016). In *S cerevisiae*, degradation of cryptic unstable transcripts

are dependent on the RNA exosome and the TRAMP complex (Wyers et al., 2005). Analogous to yeast, the nucleoplasmic mammalian TRAMP-homolog called the NEXT complex is important for degrading uaRNAs (or PROMPTs) in humans (Lubas et al., 2015; Lubas et al., 2011). Alternatively, a subset of noncoding RNAs are degraded by the nuclear poly(A)-binding protein PABPN1, though it is unclear whether it functions at uaRNAs (Beaulieu et al., 2012; Bresson and Conrad, 2013; Bresson et al., 2015; Meola et al., 2016).

In addition to its roles at destabilizing uaRNAs, the RNA exosome functions to suppress prematurely PAS terminated transcripts in the sense direction. Previously reports did not observe major change in the sense transcript, likely because the exon signal was swamping out intron signal and siRNAs prevented complete loss of the RNA exosome (Flynn et al., 2011; Preker et al., 2008). Interestingly, several genome-wide assays in this study support the idea that Pol II normally experiences some frequency of early termination, and that loss of the RNA exosome stabilizs those termination events. These results are consistent with a series of protein-protein interaction that bring the RNA exosome in close proximity to the 5'-methyl cap (Andersen et al., 2013; Hallais et al., 2013; Lubas et al., 2015; Lubas et al., 2011). U1 inhibition is also linked with preventing premature termination at PAS motifs in the sense direction (Almada et al., 2013; Berg et al., 2012; Kaida et al., 2010). We now show that while many U1 AMO-responsive clusters are independent of Exosc3-responsive clusters, more than half the clusters show combinatorial outcomes when you inhibit both, suggesting that these processes collaborate to downregulate premature termination. At these combinatorial sites, we hypothesize that early termination after U1 inhibition promotes early usage of PAS motifs, but they are degraded by the RNA exosome due to promoter proximal termination (Andersen et al., 2012). Currently we do not know the precise rules for U1 inhibition nor is our read-depth deep enough to clearly demarcate clusters so future work will need to be

done to identify molecular mechanisms defining responsiveness to U1 inhibition and Exosc3 removal.

The RNA exosome has been found to associate with transcribing Pol II (Andrulis et al., 2002). While this may be to enable co-transcriptional quality control, it also suggests that the RNA exosome may regulate transcription itself. In accordance with that, we show that the depletion of the RNA exosome increases promoter-proximal pausing. Promoter-proximal pausing is a common feature among metazoans, occurring when transcribing Pol II arrests shortly downstream of the TSS due to the activities of NELF and DSIF (Muse et al., 2007; Rahl et al., 2010; Zeitlinger et al., 2007). Since the RNA exosome degrades RNAs, how does degrading RNAS feedback upon promoter-proximal pausing? In *S pombe*, the RNA exosome promotes the resolution of backtracked Pol II in conjunction with TFIIS (Lemay et al., 2014). We speculate a subset of RNA polymerases do not escape from the pause so Exosc3 may be necessary to resolve the backtracked Pol II, which would stimulate recycling of Pol II and reduce promoter proximal pausing. Supporting this idea is the discovery that the Integrator complex promotes pause release (Gardini et al., 2014; Stadelmayer et al., 2014). The Integrator promotes cleavage of eRNAs during transcription (Lai et al., 2015), so it may cleave RNAs bound to stalled Pol II molecules to promote Pol II recycling and release a 3' OH for exosome decay. Additionally, genes with CGIs have a GC skew at the 5' end that promotes the R-loop formation (Ginno et al., 2012). Given R-loops create a barrier for efficient transcription elongation, increases in promoter-proximal pausing upon Exosc3 loss may reflect an increased stability of R-loops (Pefanis et al., 2015), but genes with CGIs showed a similar pausing response to Exosc3 removal as genes without CGIs.

In conclusion, the RNA exosome plays a critical role in many pathways within the cell. From its well-characterized roles in 3' end processing to suppressing noncoding RNA, we have

identified two new activities, namely stabilizing premature-termination events, especially in the first intron, and modulating promoter-proximal pausing.. These two events may not be mutually exclusive, as increased pausing may enable more time for early termination and exosome decay. An understanding of the broad roles of the RNA exosome will have significant impact on both molecular biological processes as well as physiological disease. Recently, the RNA exosome has been linked with tumorigenesis. RNA exosome subunits (in particular *EXOSC4*) are found to be amplified in many sequenced tumors (cBioPort). In contrast, knockdown of *EXOSC4* has been reported to halt tumor progression (Stefanska et al., 2014), whereas the common chemotherapeutic 5-fluorouracil has been shown to phenocopy RNA exosome mutations in causing defects in rRNA-processing (Lum et al., 2004). These may reflect the role of the RNA exosome in promoting ribosome maturation, in suppressing genomic instability, or alternatively, at suppressing premature termination to enable the production of more full-length transcripts. A further understanding of these premature transcription events will provide interesting insights into both transcription biology and disease phenotypes.

2.5 FIGURES

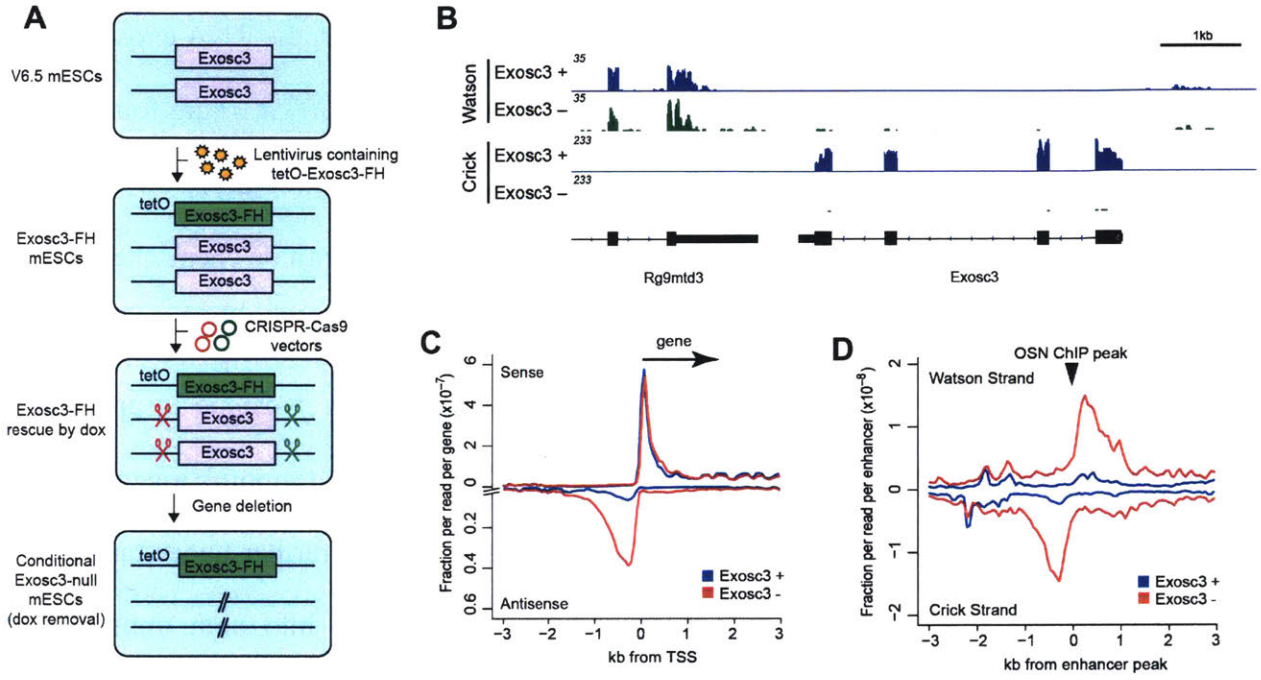

Figure 1. Generation of Exosc3 CKO mESC Cell Line

(A) Schematic showing strategy to knockout the endogenous Exosc3 gene.
(B) Genome browser shot of Exosc3 after doxycycline withdrawal for 3 days.
(C) Metaplot of RNA-seq reads around a 3 kb window flanking TSS of non-overlapping UCSC canonical genes for Control (blue) and Exosc3 KO (3 days off dox, red).
(D) Metaplot of RNA-seq reads around a 3 kb window flanking centers of Oct4, Sox2 and Nanog ChIP-seq peaks, filtering out non-overlapping enhancers.

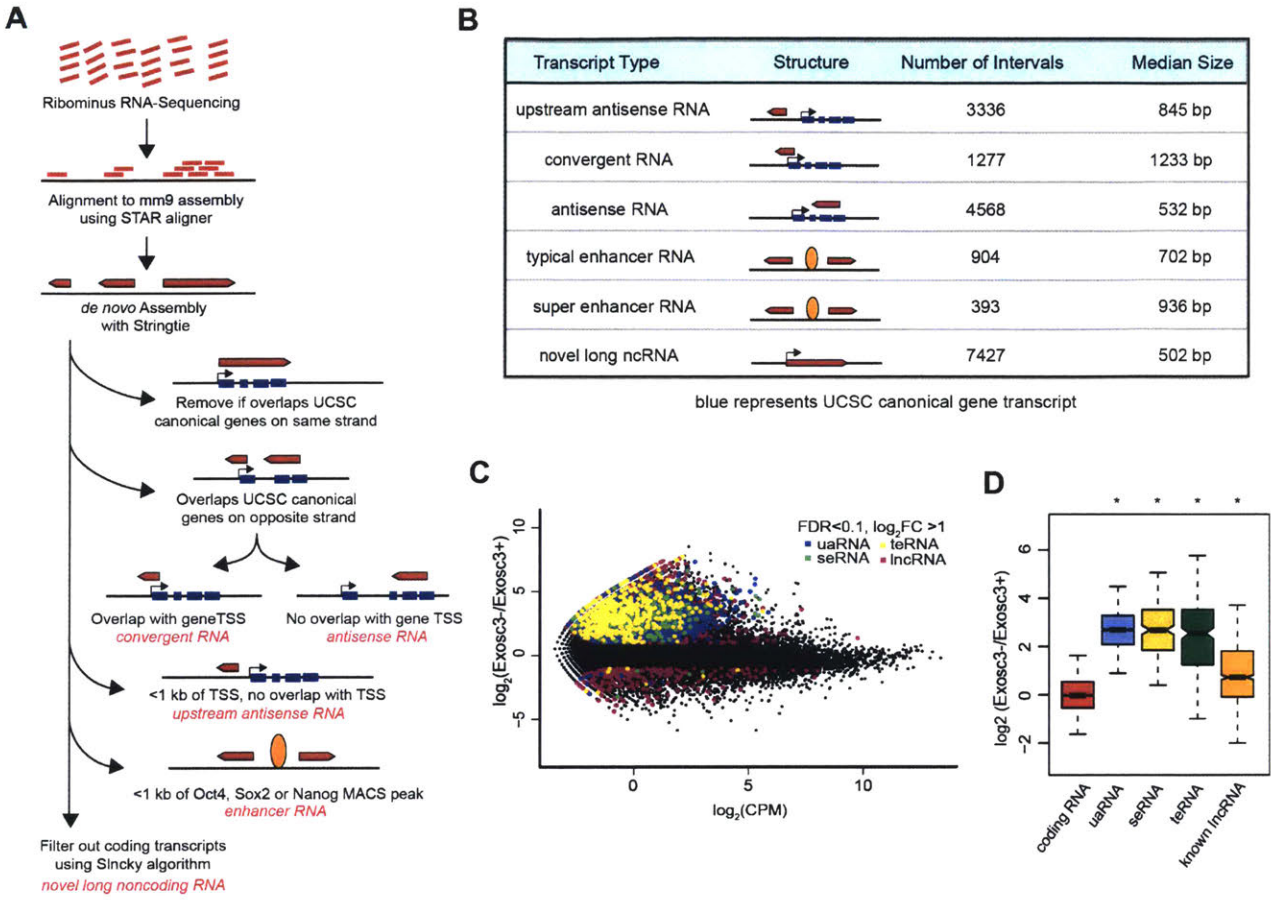

Figure 2. Low-Abundant Noncoding RNAs are Suppressed by the RNA Exosome

(A) Strategy for de novo transcriptome assembly and interval classification.
(B) Properties of intervals defined by de novo transcriptome assembly.
(C) MA Plot of Exosc3 CKO. Colored plots are statistically significant intervals, defined in edgeR with FDR<0.1 and log2fold change>1.
(D) Boxplot of $\log_2$(Exosc3-/Exosc3+) for various defined intervals. Asterisk represents statistically significantly different distributions (p-value <0.001) compared to coding RNAs using Wilcoxon signed-ranked sum test.

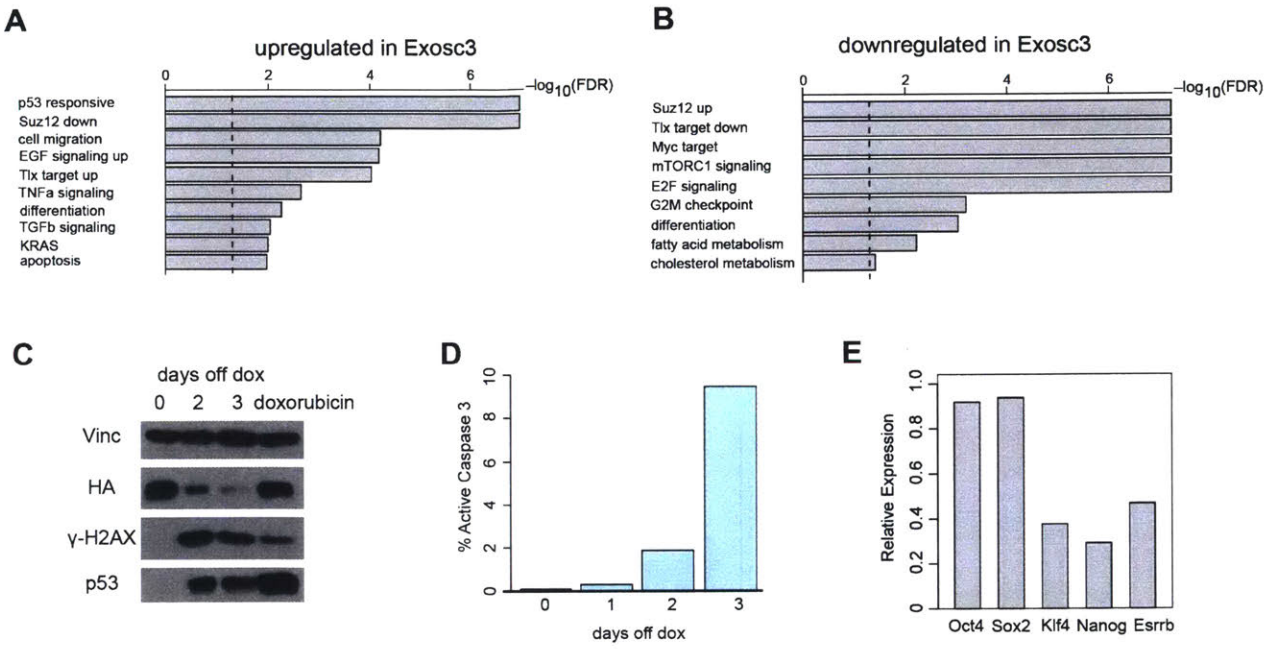

Figure 3. Differentially Expressed Genes in RNA Exosome CKO

(A-B) Boxplot showing select statistically significant pathways identified by gene set enrichment analysis (GSEA) for genes upregulated (A) or downregulated (B) upon Exosc3 loss. Dotted line represents false discovery rate of 0.05.
(C) Western Blot for vinculin, HA-tagged Exosc3, γ-H2AX and total p53 from protein lysate after 0 days, 2 days, 3 days off doxycycline or 7 hour treatment with 1 uM doxorubicin.
(D) Percent of cells that are positive for active caspase 3 by FACS analysis.
(E) Relative expression of pluripotency genes upon removal of doxycycline for 3 days, determined from RNA-seq data.

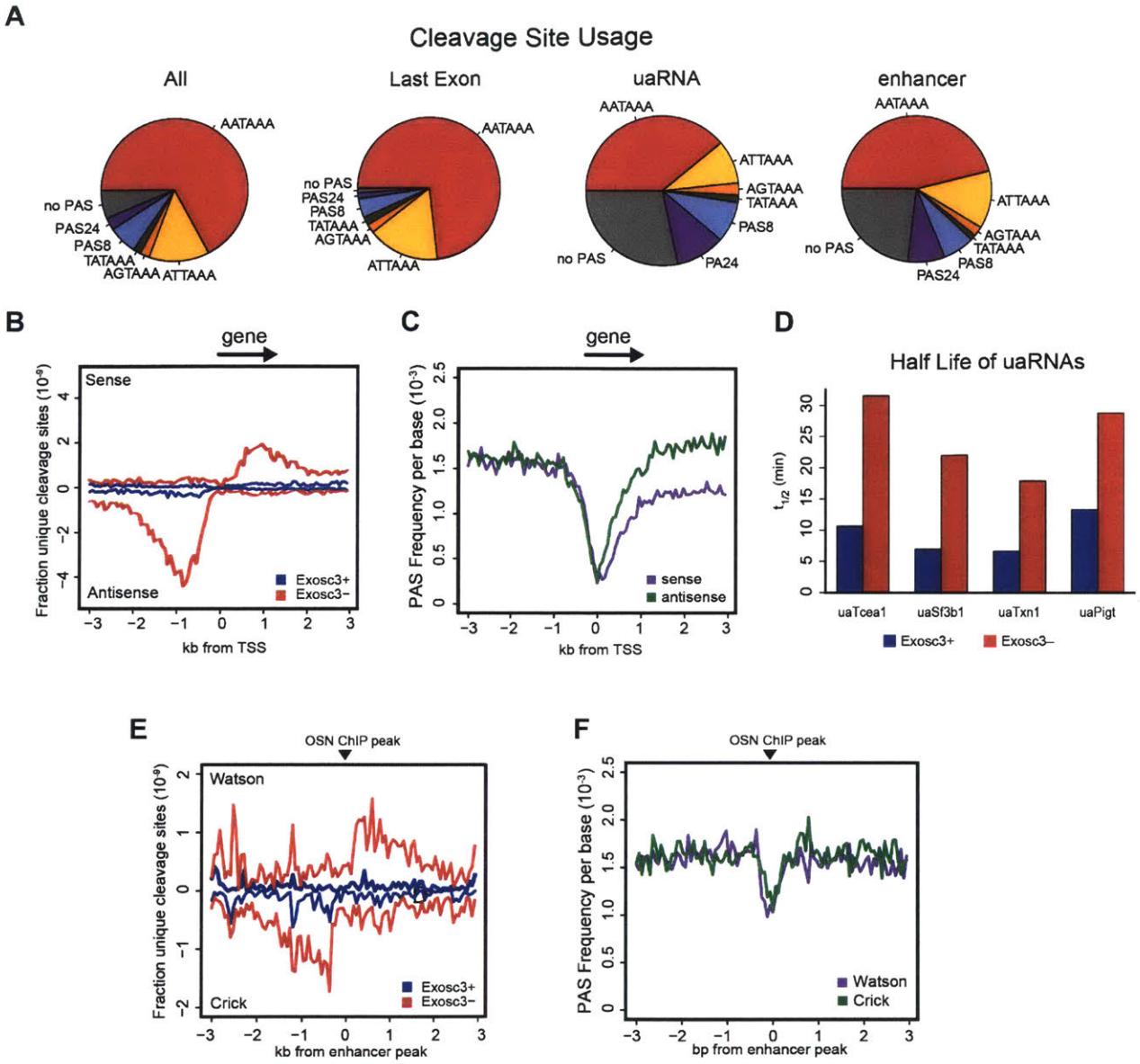

Figure 4. Polyadenylated uaRNAs and eRNAs are regulated by the RNA Exosome

(A) Pie chart showing distribution of various PAS motifs for all detected cleavage sites reads.
(B) Metaplot of mean unique cleavage sites with PAS motifs derived from 2P-seq libraries generated from Exosc3 CKO with (red) and without doxycycline (blue) for 3 days aligned to the nonoverlapping TSS of genes, normalized by library size.
(C) Frequency of predicted canonical PAS motifs (AATAAA/ATTAAA) flanking the TSS on sense (purple) or antisense strand (green).
(D) Half-lives of uaRNAs, from cells treated with 1 uM flavopiridol with or without dox.
(E) Mean signal of unique cleavage sites with canonical PAS motifs aligned around center of Oct4/Sox2/Nanog ChIP-seq peaks, normalized by library size.
(F) Frequency of predicted canonical PAS motifs (AATAAA/ATTAAA) flanking Oct4/Sox2/Nanog ChIP-seq peaks on Watson (purple) or Crick Strand (green).

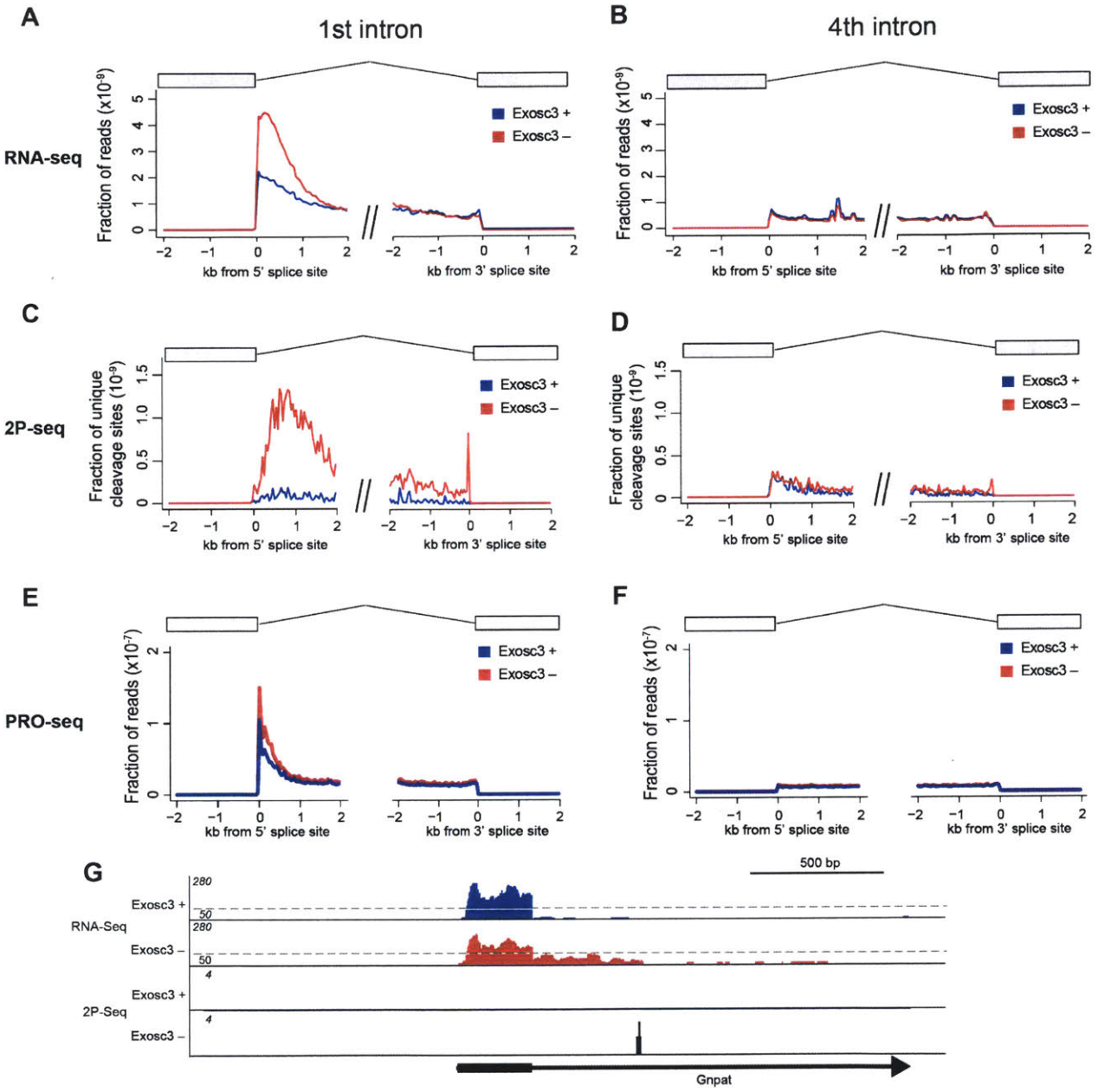

Figure 5. RNA Polymerase II Prematurely Terminates within the First Intron

(A-B) Mean fraction of exon-removed RNA-seq signal per intron flanking 5' or 3' splice sites for Exosc3+ or Exosc3- of introns at least 2 kb long, normalized by library depth.
(C-D) Mean fraction of unique cleavage sites per intron with PAS motifs around the first 5' or 3' splice site of introns at least 2 kb long, normalized by library depth.
(E-F) Mean fraction of PRO-seq signal per intron flanking 5' or 3' splice site of introns at least 2 kb long, normalized by library depth.
(G) Genome browser shot of *Gnpat* with RNA-seq and PAS-mediated cleavage sites. Scale changes at grey line, which represents a normalized signal of 50.

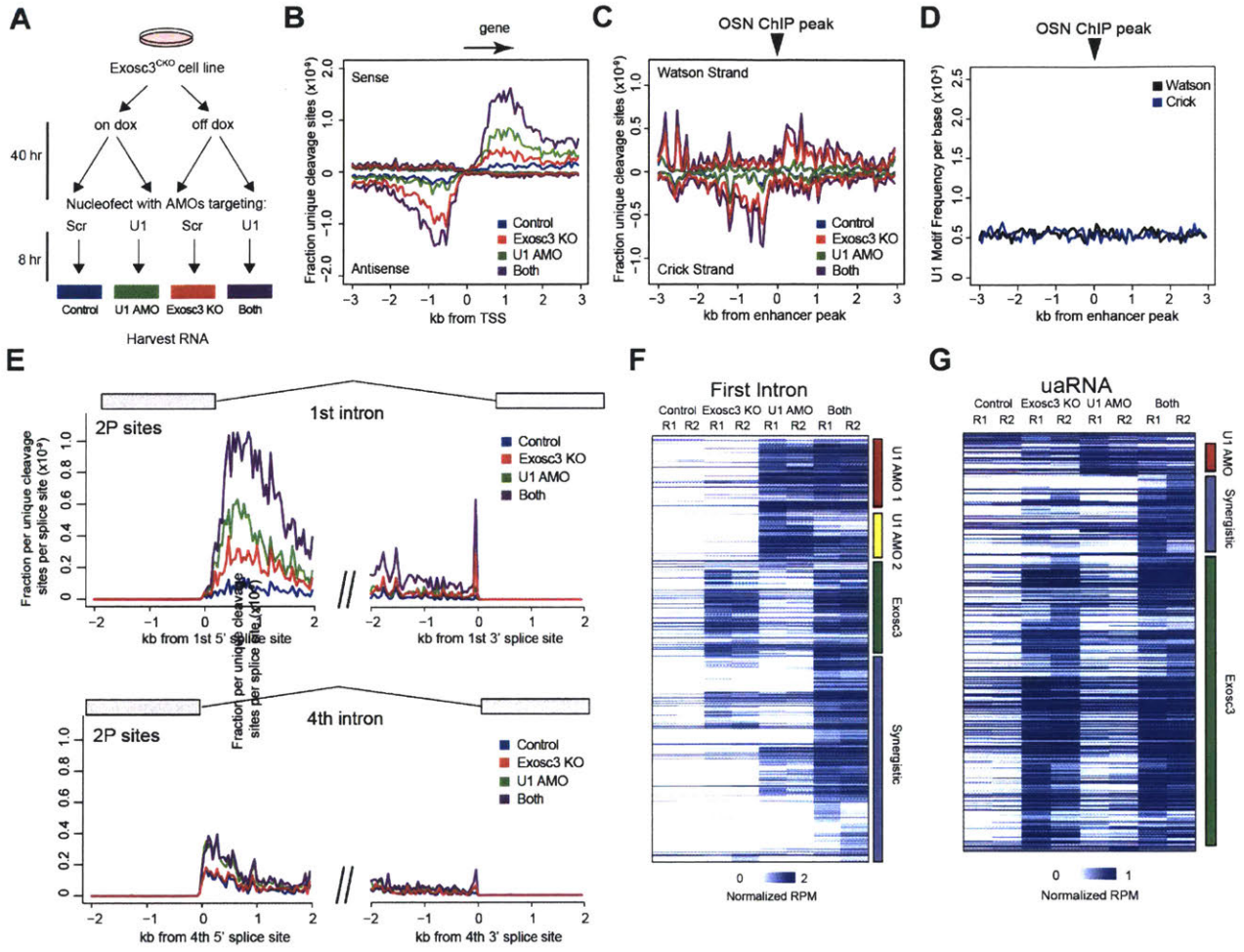

Figure 6. Regulation of Premature Termination by U1 AMO and Exosc3

(A) Experimental design for double treatment with U1 inhibition and Exosc3 depletion.
(B) Mean fraction unique cleavage site signal with PAS motifs around TSS, after Exosc3 depletion and/or U1 inhibition, normalized to mapped library size.
(C) Mean fraction unique cleavage site signal with PAS motifs around Oct4/Sox2/Nanog ChIP-seq peaks, after Exosc3 depletion and/or U1 inhibition, normalized to mapped library size.
(D) Predicted U1 signals around Oct4/Sox2/Nanog ChIP-Seq peaks.
(E) Mean unique cleavage site signal with PAS motifs aligned around 5' and 3' splice site of introns at least 2 kb in length, after Exosc3 depletion and/or U1 inhibition, normalized to mapped library size.
(F-G) Heatmap of library-size normalized RPM for hierarchically clustered PAS-linked 2P clusters within the first intron or uaRNA.

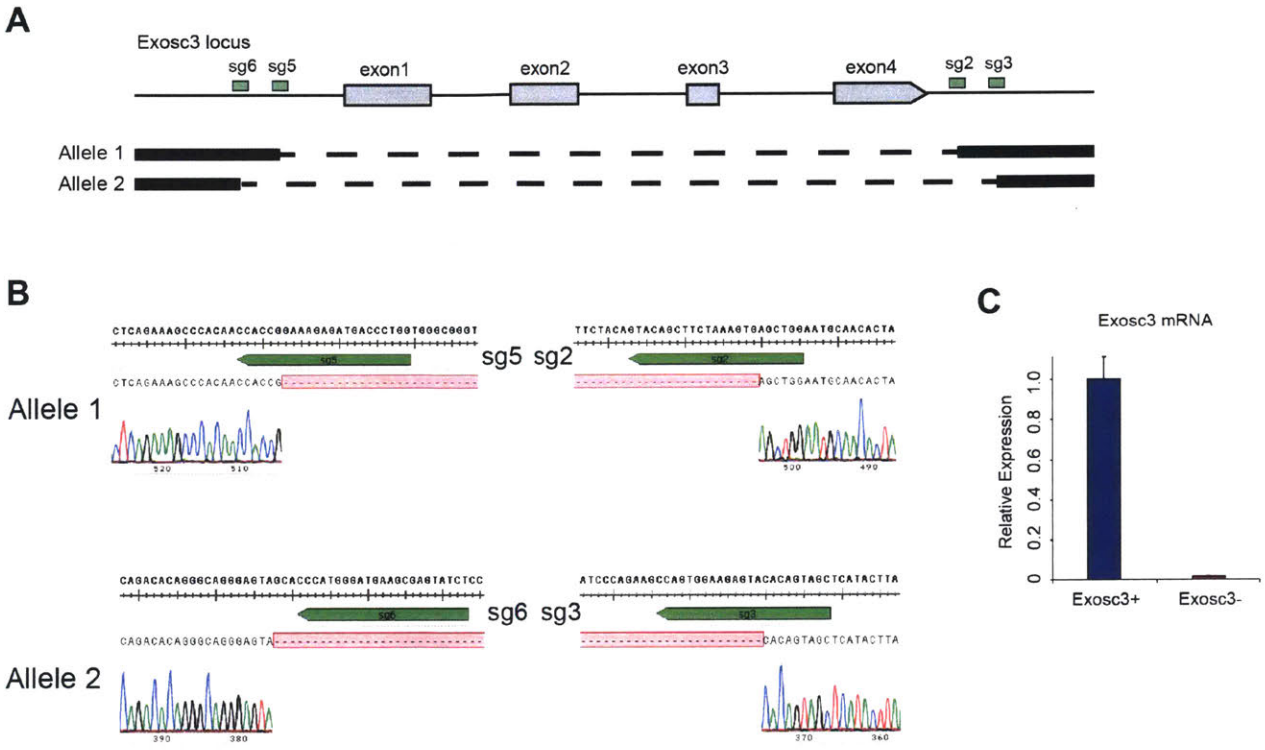

Figure S1. Validation of Exosc3 Knockout

(A) Allelic profiling of *Exosc3* CKO knockout.
(B) Sanger sequencing of PCR products across CRISPR target sites validating deletion of endogenous *Exosc3*.
(C) qRT-PCR of spliced Exosc3 from cDNA from Exosc3+ and Exosc3-.

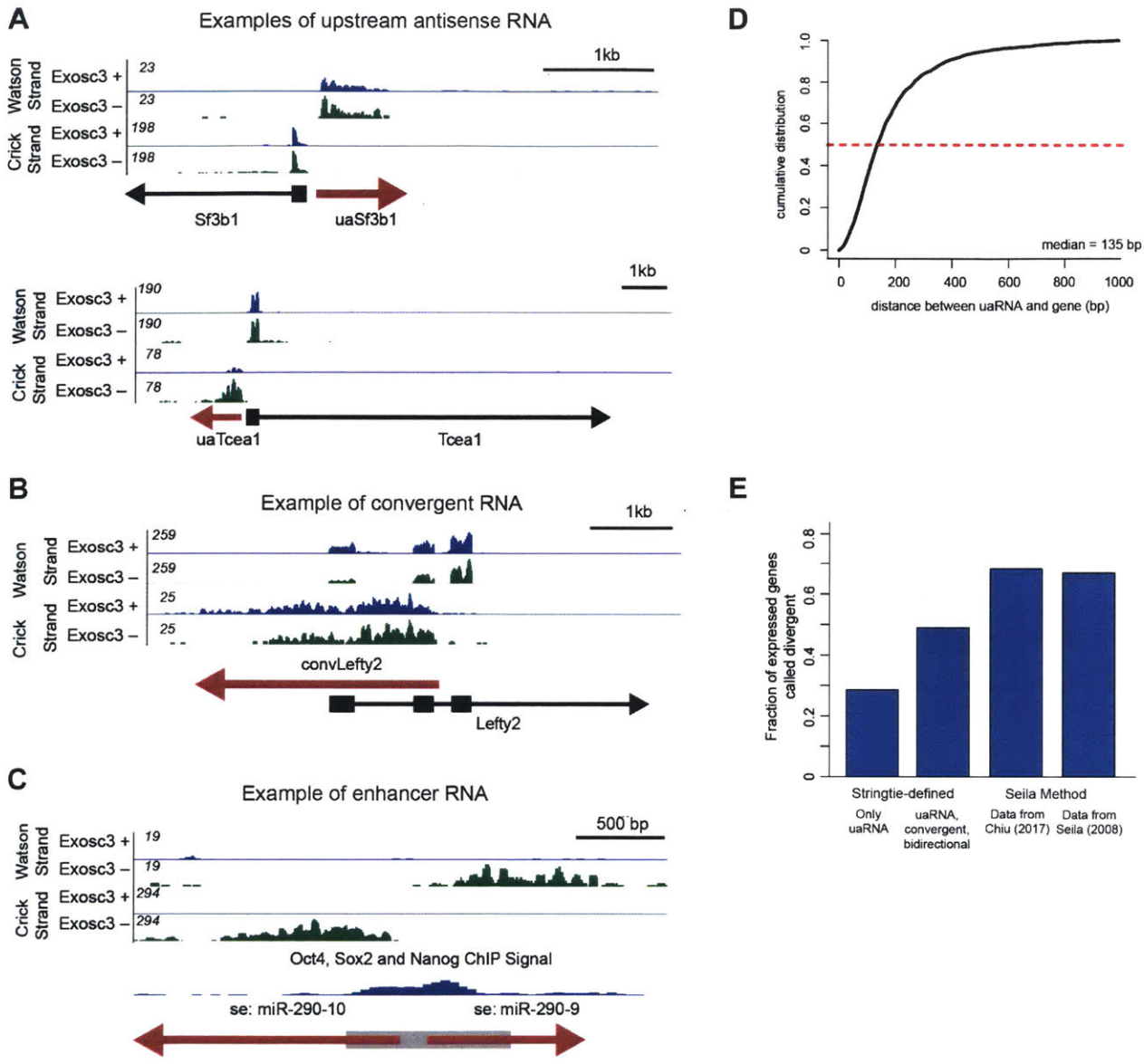

Figure S2. De Novo Identified Transcriptome Assembly

(A-C) Genome browser shots of representative examples of newly defined intervals (red) for an upstream antisense RNA (A), convergent RNA (B) and enhancer RNA (C). RNA-seq reads are illustrated as Exosc3+ (blue) and Exosc3- (green). For the enhancer RNA, the grey box represents the ChIP peaks of the combined Oct4, Sox2 and Nanog transcription factors by MACS.
(D) Cumulative distribution figure of the distance between the TSS of uaRNAs and the TSS of their corresponding mRNA. Red dashed line represents where the cumulative distribution would be the median.
(E) Fraction of expressed genes that are divergent using two metrics: statistically-defined uaRNA intervals from RNA-seq or identifying at least one read within 1.5kb of TSS. Seila (2008) refers to the value identified in previous paper.

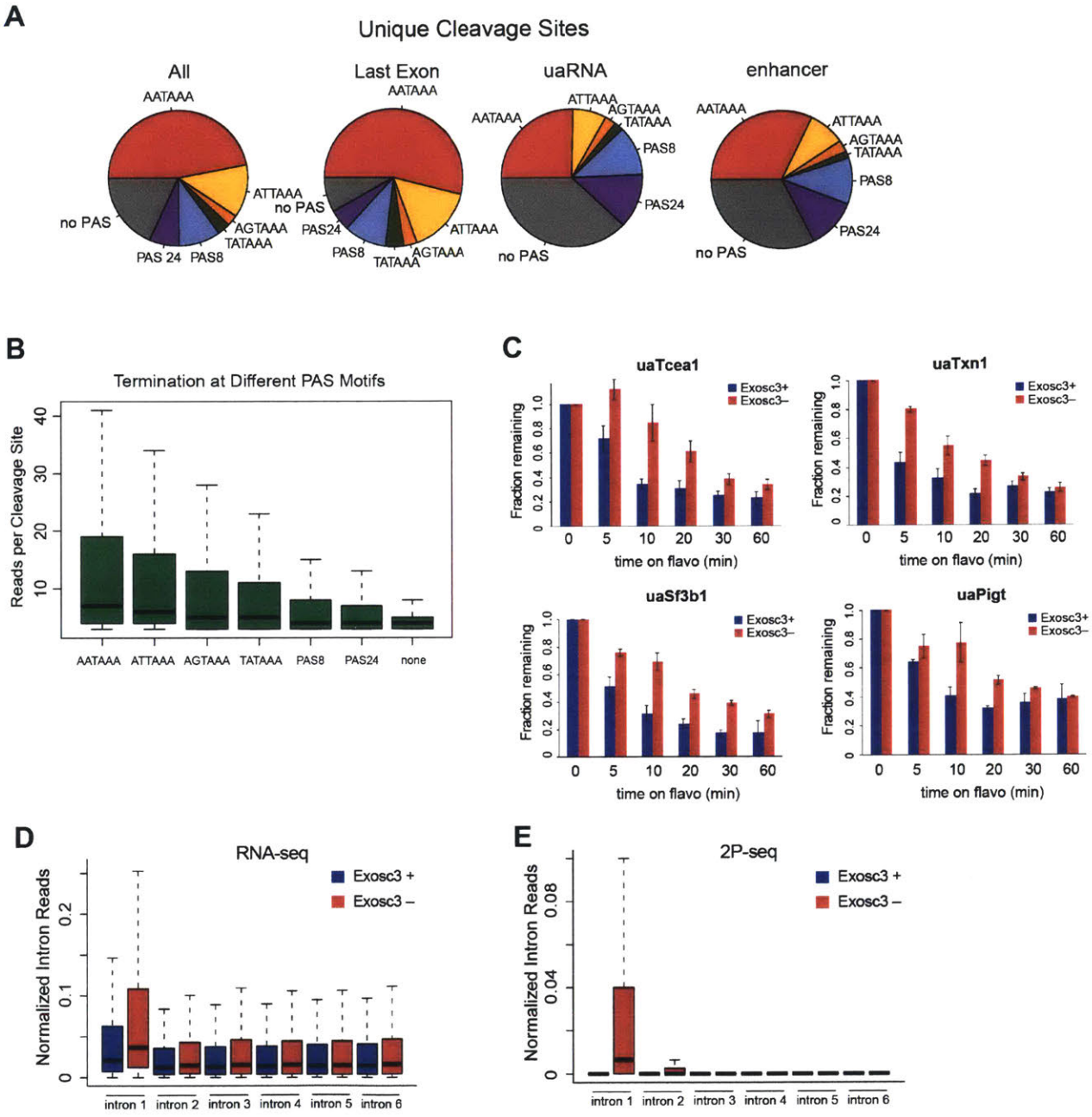

Figure S3. 2P-seq Profiling in Exosc3 CKO

(A) Distribution of PAS motifs for uniquely detected cleavage sites.
(B) Boxplot of number of reads for each cleavage sites sorted by type of PAS motif.
(C) Relative abundance of uaRNAs from oligo-dT primed cDNA after addition of 1 uM flavopiridol. Plotted is the mean from 3 biological replicates; error bars represent standard error.
(D) Boxplot of normalized RNA-seq reads ($FPKM_{\text{intron N reads}}/FPKM_{\text{mature transcript reads}}$) for UCSC canonical genes with at least 6 introns.
(E) Boxplot of normalized intronic cleavage site reads (Sum of 2P reads in intron N/Sum of 2P reads in last exon) for UCSC canonical genes with at least 6 introns.

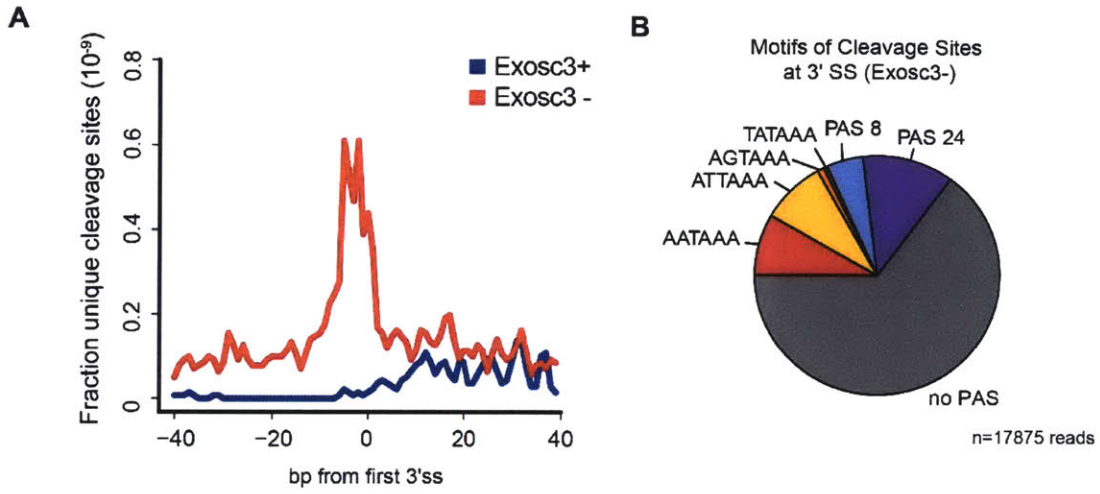

Figure S4. Degradation of Spliced Lariats at 3' Splice Sites

(A) Distribution of all unique cleavage sites around 3' splice site.
(B) Pie chart of PAS motif usage at unique cleavage events at 3' splice site.

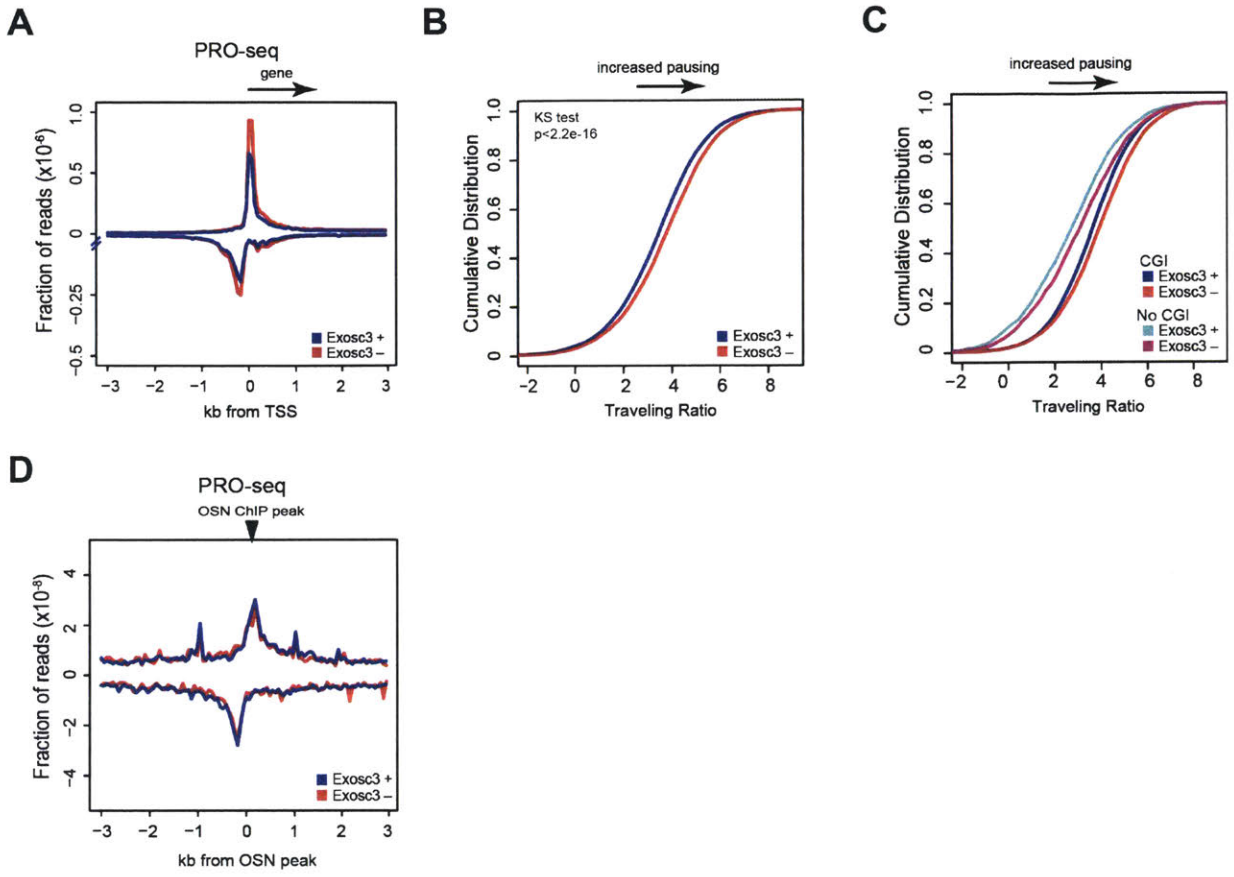

Figure S5. Increased Promoter Proximal Pausing upon Exosc3 Removal

(A) Metaplot of PRO-seq reads around a TSS of non-overlapping UCSC canonical genes.
(B) CDF of traveling ratios for all UCSC canonical nonoverlapping genes.
(B) CDF of traveling ratios for all UCSC canonical nonoverlapping genes, filtering those with a CpG Island (CGI) overlapping 0 to 100 bp from TSS.
(D) Metaplot of PRO-seq reads around a 3 kb window flanking centers of Oct4, Sox2 and Nanog ChIP-seq peaks, filtering out non-overlapping enhancers.

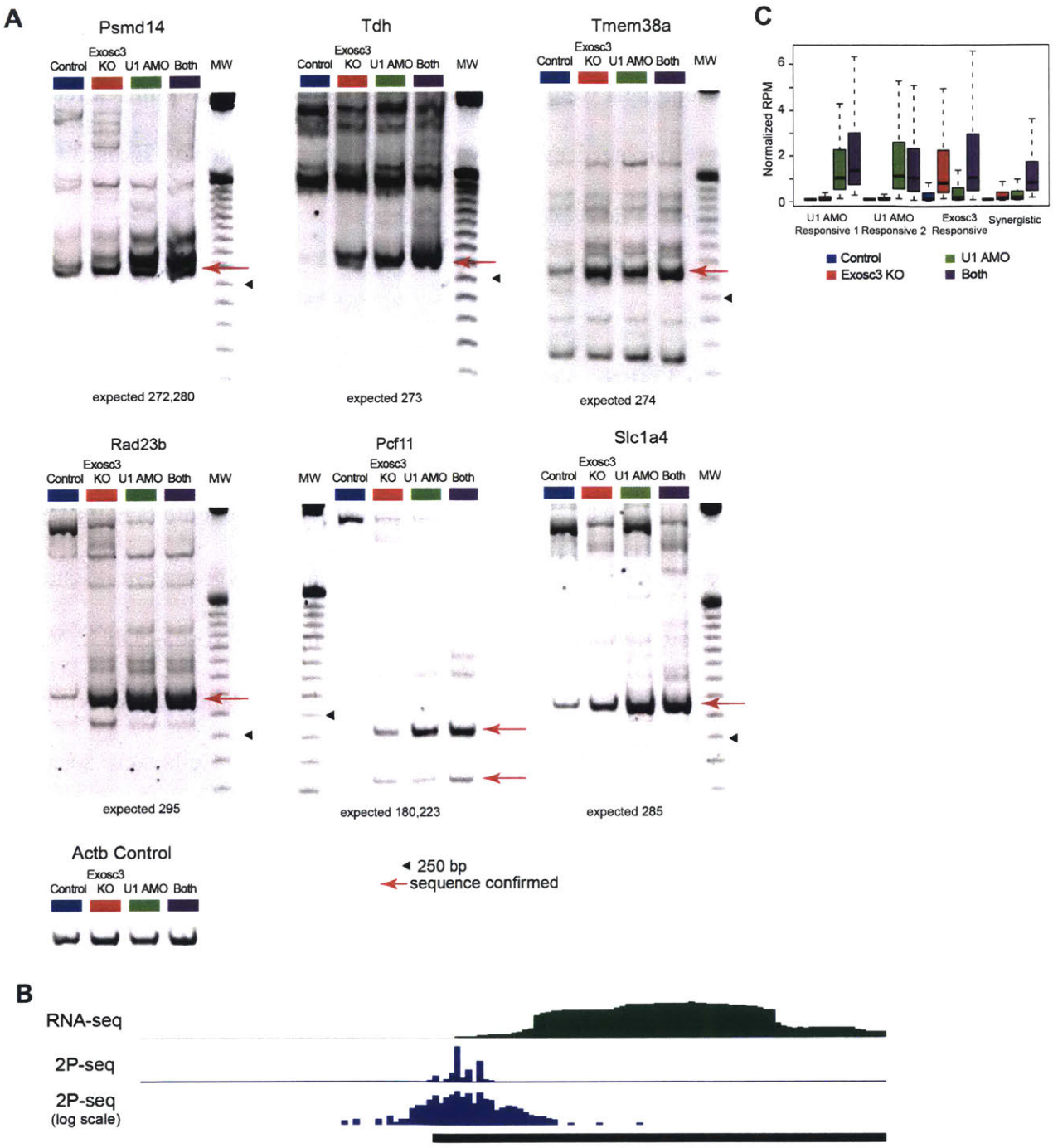

Figure S6. Validation of 2P-seq data

(A) Nested 3' RACE of 6 termination events. Red arrows indicate most frequent termination sites, which have been sequence validated. Molecular weight ladder is 25 bp ladder.
(B) Profile of 2P-seq at end of Actb. Top 2P track is linear scale, whereas bottom is log scale.
(C) Boxplot of normalized RPM for 4 identified clusters.

2.5 METHODS

Cell Culture

V6.5 mouse embryonic stem cells were grown under standard conditions without feeders (Almada et al., 2013). Cells were passaged every two days to avoid confluency. Exosc3 CKO clones were maintained in 0.1 μg/ml of doxycycline.

Generation of Conditional Exosc3 mESC Cell Lines

The knock-out clones were generated through two steps. To prepare lentivirus for conditional expression of Exosc3, HEK293T cells were transfected with packaging vectors VSV-G and dr8.91, as well as a lentivirus plasmid (pSLIK-Hygro) containing a hygromycin resistance cassette and doxycycline-inducible C-terminally tagged FLAG-HA mouse Exosc3 cDNA. Virus was collected days 2 and 3 post-transfection. V6.5 mESCs were seeded to be about 20% confluent and infected with virus and polybrene (1:2000 of 8 mg/ml polybrene stock). Cells were selected on 150 μg/ml of Hygromycin B, and single cell clones were isolated. Expression of FH-Exosc3 was validated using anti-HA antibodies (Roche 3F10).

Next deletion of endogenous Exosc3 gene was attempted by cotransfection of two CRISPR-Cas9 vectors (pX330) with sgRNAs (sgExosc3-2 and sgExosc3-5) flanking the Exosc3 gene. Heterozygotes were isolated and validated by PCR amplification across the deletion and subsequent sequencing. Heterozygotes were further transfected with two CRISPR-Cas9 vectors containing sgExosc3-3 and sgExosc3-6 to target the other allele under treatment with 0.1 ug/mL doxycycyline. Subsequent clones were screened for shortened PCR products across the entire gene (**Table S1**). The shortened PCR products were sequence confirmed. Finally, deletion of Exosc3

was further validated using qRT-PCR for the Exosc3 gene after 3 days of doxycycline removal. The sgRNA and primer sequences are described in Table S1.

Western Blotting

Protein lysate was run on 1.5% NuPAGE Bis-Tris Gels using the NuPAGE Western Blotting System (ThermoFisher Scientific). The gels were transferred in at 4°C in 10% Methanol and 1x NuPAGE Transfer Buffer onto PVDF membranes. Membranes were blocked with 5% skim milk and the incubated with primary antibody in 5 % milk overnight. Blots were washed in PBST and incubated with ECL HRP secondary antibody in milk for an hour at 1:10000 dilution. Blots were further washed in PBST before imaged using ECL. Antibodies used were HA (Roche, 3F10), Vinculin (Sigma, V9131), p53 (CST25245), and yH2AX (CST9718).

qRT-PCR

Total RNA was extracted using TRIzol Reagent (ThermoFisher Scientific) and genomic DNA was removed using DNase Turbo (Ambion AM2238). RNA was reverse-transcribed using random hexamers and SuperScript III First-Strand Synthesis System (ThermoFisher Scientific) according to the manufacturer's instructions. Quantitative PCR was performed with PowerUp SYBR Green Master Mix (Thermo Scientific) and the 7500 Fast Real-Time PCR System (Applied Biosystems). Sequences of PCR primers are described in **Table S1**.

RNA-seq Library Generation

Total RNA was isolated with TRIzol Reagent and treated with DNase Turbo (Ambion AM2238) to remove genomic DNA contamination. RNAs that passed a Bioanalyzer RIN score of

8.5 were subsequently used to prepare libraries. RNAs were depleted of ribosomal RNAs using the RiboZero rRNA removal kit (Epicentre MRZH116), converted into stranded RNA-Seq libraries with the Illumina Tru-Seq kit (Illumina RS-122-2101) and sequenced using the Illumina NEXT-Seq.

RNA-seq Processing

All analyses were carried out using UCSC (NCBI37/mm9) mouse gene annotations. Paired end reads were trimmed of adapters using Trimmomatic (Bolger et al., 2014). Reads were first mapped to ribosomal RNA and various repetitive sequences such as U1 snRNA using Bowtie2 (Langmead and Salzberg, 2012), and then subsequently mapped to the mouse UCSC transcriptome and genome using STAR aligner (Dobin et al., 2013). The ensuing reads were filtered for uniquely mapping, properly paired reads, and subsequently potential PCR duplicates were removed using the Picard Suite MARKDUP (http://broadinstitute.github.io/picard). In genome browser shots, the reads are displayed. For metaplot alignments, we further processed the reads by selecting read 2 of the paired-end read (same direction as the RNA), and filtered away any overlapping miRNAs, tRNAs, repeats from repeatMasker, or snoRNAs.

***de novo* Transcriptome Assembly**

To identify noncoding RNAs genomewide, the two doxycycline replicates were collapsed into one file. Subsequently, Stringtie was run on this using the parameters –f 0.1 –c 5 –g 10 (Pertea et al., 2015). The resulting candidate transcripts were first removed for any transcript that overlapped UCSC canonical genes, snoRNAs, and known miRNA genes. Any candidate transcripts were then aligned against the antisense version of UCSC canonical genes, and divided

into two categories: 1) *convergent RNAs*: those that started within the gene and was transcribed across the TSS of the canonical gene or 2) *antisense RNAs*: antisense transcripts that did not overlap the TSS. The remaining candidate transcripts were further analyzed for *uaRNAs*: transcripts that were antisense to the coding gene and started within 1 kb of the TSS. The remaining candidate transcripts were further segmented into *eRNAs*: transcripts that overlapped a flanking 1kb window of called Oct4, Sox2, and Nanog binding sites described in a previous report (Whyte et al., 2013). The remaining candidate RNAs were filtered for *de novo lncRNAs* by removing previously annotated lncRNAs followed by running the Slncky algorithm (Chen et al., 2016).

Differential Analysis and Gene Set Enrichment Analysis

The number of reads per transcript was counted by using intersectBed of the Bedtools suite (Quinlan and Hall, 2010), only allowing for exonic or spliced reads. After filtering out for intervals with low numbers, differential transcripts were called using edgeR, where we normalized libraries using UQ normalization. Statistically significant transcripts were those with at least a two-fold change and a false-discovery rate less than 0.10. For GSEA, genes were pre-ranked by $\log_2$(fold change) and the preranked algorithm was run against all gene sets (Subramanian et al., 2005).

Hierachical Clustering

The number of counts across robust clusters in the first intron or uaRNAs were counted and normalized by library size. Subsequently, the robust clusters were subjected to hierarchical clustering using the Pearson Correlation metric in Multiple Experiment Viewer.

Active Caspase 3 Assay

Exosc3 CKO cells were removed from doxycycline for 0, 1, 2 or 3 days. Subsequently, we labelled cells using the FITC Active Caspase-3 Apoptosis kit (BD Pharmingen) as per manufacturer's instructions, before FACS analysis for FITC positive cells.

Determination of uaRNA Half-Lives

Cells were maintained or removed from doxycycline for 2 days. Subsequently, cells were placed into mESC media containing 1 μg/ml flavopiridol dissolved in DMSO for 0 min, 5 min, 10 min, 20 min, 30 min or 1 hr before harvesting in TRIzol. cDNA was generated using oligo-dT_{20} and SuperScript III reverse transcriptase. qRT-PCR was performed using primers in **Table S1**. qRT-PCR was normalized to values at when time is 0. Averages across three experiments were used to determine fraction remaining. Half-lives were determined by fitting an exponential decay curve using R, starting with the formula: $y= e^{-bx}$, and then finding the point such that y=0.5.

U1 Inhibition Experiment

Exosc3 CKO cells were either kept in doxycycline or removed from doxycycline for 1 day and 16 hours. Cells were subsequently trypsinized, washed twice in PBS, and 5 million cells were nucleofected with 15 μM concentration of U1 AMO or Scr AMO (sequence in **Table S1**). Cells were seeded onto 10 cm dishes, and total RNA was harvested 8 hours later in TRIzol Reagent.

Metaplots

We filtered the intervals for metaplot as follows. For metaplots around TSS, UCSC canonical genes were filtered to remove any genes that overlapped within 5 kb of the TSS. For

metaplots at enhancers, we aligned against centers of all Oct4/Sox2/Nanog defined enhancers (typical enhancers and super-enhancers) according to a previous report (Suzuki et al., 2017; Whyte et al., 2013). Subsequently, we filtered out any overlapping enhancers peaks within a 3 kb window and also any that overlapped a UCSC canonical gene. For metaplots at splice sites, UCSC canonical genes with at least 4 introns were identified. We also removed any introns that had known snoRNAs and required introns be at least 2 kb long.

To create metaplots for RNA-seq or PRO-seq, we counted the number of overlapping reads across non-overlapping sub-intervals (bins) that span the aligned region. The one exception is for splice sites, we did an additional filter where we removed any reads that overlapped annotated exons. Bins were normalized by:

$$normalized\ bin = \frac{counts\ of\ filtered\ RNA\ Seq\ reads}{total\ mapped\ reads\ x\ number\ of\ aligned\ intervals}$$

For 2P-seq, we focused on unique PAS-linked cleavage sites rather than potential cleavage sites, because the low number of cleavage site positions created extremely spiky reads if we align uncollapsed reads. We counted the number of unique PAS-linked cleavage sites across non-overlapping bins that span the aligned region. Similar to RNA-seq, any cleavage sites that overlapped exons were removed if we were doing splice site alignments. Normalization for 2P-seq was challenging as we did not have spike ins. Normalization by number of detected unique sites is a challenge because a significant fraction of unique sites is located within genes, so any major shift (as expected with U1 inhibition) will misrepresent the number of unique cleavage sites. We chose to normalize by number of mapped 2P-sites which also factors in sequencing depth. In other words, bins were normalized by:

$$normalized\ bin = \frac{counts\ of\ unique\ filtered\ 2P\ sites}{total\ mapped\ reads\ x\ number\ of\ aligned\ intervals}$$

3' End Sequencing (or 2P-seq) Library Generation

2P-Seq was performed as described in (Spies et al., 2013). Briefly, total RNA is poly(A) selected using oligo-dT dynabeads. Subsequently, RNA is cleaved with trace levels of RNase T1 for 20 minutes at 22°C, inactivated and cleaned up with an ethanol precipitation. The resulting RNA is reverse transcribed using IW-RT1p and the size selected for 200-400 nts on a polyacrylamide gel. Next the cDNA is circularized using CircLigase II (Epicentre), PCR amplified with primers IW-PCR-F.1 and IW-PCR-RPI, and further size selected to remove adapters, before sequencing from the poly(A) tail using IW-Seq-PE1.1 on the Illumina NEXT-Seq.

2P-seq Read Processing

Reads were first quality filtered by trimming adapters using Trimmomatic and A stretches (>5 As) were removed if they were immediately downstream of first sequenced nucleotide. We interpreted these events as poly(A) tails that due to reverse transcription errors or biological reasons had a non-As added to the cDNA. Next, we mapped either filtered reads (set A) or filtered reads with the first 15 nts trimmed (set B) to the mm9 genome using STAR aligner, end-to-end mode. The trimming of first 15 nt was done to ensure that reads were not going to be lost due to mismatches at the 5' end, which may involve non-templated nucleotides (such as uridines), which are added to some termination events. For both sets, the first mapped nucleotide was considered the cleavage site.

The two mapped libraries were combined as follows. If the read only aligned in set A or set B, the cleavage site was used as is. If the read aligned in both set A and set B, we subjected the mapped site to one further test. If the mapped cleavage site in set A overlaps the mapped cleavage site minus 15 nucleotides in set B, the position in set A was used. However, if the mapped cleavage

site in set A differed substantially from the read in set B, we chose the site in set A as the mapped site. We attributed changes for this subset to the shorter read being harder to find exact matches, so preferred the mapped position of the longer read.

With the combined mapped cleavage sites, we then applied an internal priming filter, in which we removed reads with at least 7 adenosines in the 10 nucleotides 3' of the cleavage site, or 13 adenosines in the downstream 20 nucleotides. The remaining cleavage sites were filtered so that it must have at least 2 different reads mapping to it and also to not overlap B2 SINE elements. Finally, we scored reads as PAS containing or not PAS containing by surveying the 80 nucleotides upstream of the cleavage site for the presence of the top 36 PAS motifs, as described in (Almada et al., 2013). Specifically, the top 2 canonical PAS motifs are AATAAA or ATTAAA. Next, we also look for known variants, AGTAAA or TATAAA. We subsequently look for the next 8 most frequent sites or PAS8 (AATATA, AATACA, CATAAA, GATAAA, AATGAA, ACTAAA, AAGAAA, AATAGA). Finally we look for the remaining 24 PAS variants.

PRO-seq

PRO-seq was modified from the published PRO-seq protocol described in (Mahat et al., 2016). Briefly, nuclei were isolated from mESCs as described using cell permeabilization, followed by run on and biotin enrichment. Individual libraries were ligated with 3' barcoded adaptors and pooled into one tube, before completing addition enrichment for biotin and reverse transcription. Unlike regular PRO-seq which performs PCR amplification and size selection, we treated RNAs with a cocktail of RNase A and RNase H and phenol-chloroform extracted the ensuing single-stranded cDNA library. The library was sequenced on the NextSeq.

PRO-seq processing

Sequenced reads were aligned to mm9 using bowtie2 (Langmead and Salzberg, 2012) and options -D 15 -R 2 -N 0 -L 20 -i S,1,0.75. Processed reads were resized to the 3' most sequenced read, which represents the precise position of the RNA in the catalytic site.

3' RACE

DNase-treated RNA is reverse transcribed with Superscript III using the 3' RACE Adapter oligo. Subsequently, nested PCRs were performed using Phusion DNA Polymerase. In the first round, PCR Buffer conditions included GC Buffer, 3% DMSO, 1 mM dNTP and the 3' RACE Outer Primer and gene specific outer primers. In the next round, PCR Buffer conditions were similar but the 3' RACE inner Primer and gene specific inner primers were used instead. Products were run on a 5% nondenaturing polyacrylamide gel with 25 bp ladder (Life Technologies).

Traveling Ratio

Promoter proximal traveling ratios were calculated as described in (Rahl et al., 2010).

$$traveling\ ratio\ (TR) = \frac{\frac{promoter\ proximal\ reads\ (-30\ to\ +\ 300\ bp)}{promoterproximal\ length}}{\frac{gene\ body\ reads\ (+300\ bp\ to\ TES)}{gene\ body\ length}}$$

Cleavage Cluster Pipeline

Cleavage sites from biological replicates of 2P-seq datasets were collapsed. Sites within 25 nucleotides of each other on the same strand were merged using Bedtools mergeBed. The

tentative clusters were further merged across all 2P-seq datasets to create a combined cluster set with mergeBed, but this time only if they overlapped, creating a combined list of cleavage clusters.

Next we assigned whether the cleavage cluster was a PAS-linked or PAS independent cluster. To do this, the most abundant cleavage site within a cluster was called the max site. We looked up to 100 nucleotides upstream of the max site and looked for one of PAS 36 motifs. Those with PAS36 motifs were called PAS-linked clusters whereas those without were not PAS-independent clusters. Robust clusters were those where at least two independent libraries had non-zero reads; genes with premature clusters were defined as genes with robust clusters overlapped intron 1 of the gene.

2.7 SUPPLEMENTAL MATERIALS

Primer Sequences

sgRNA Primers

Name	Sequence (5'– 3')
sgExosc3-2 fw	CACCGGTACAGCTTCTAAAGTGAGC
sgExosc3-2 rv	AAACGCTCACTTTAGAAGCTGTACC
sgExo3sc3-3 fw	CACCGGCTACTGTGTACTCTTCCAC
sgExosc3-3 rv	AAACGTGGAAGAGTACACAGTAGCC
sgExosc3-5 fw	CACCGCCAGGGTCATCTCTTTCCGG
sgExosc3-5 rv	AAACCCGGAAAGAGATGACCCTGGC
sgExosc3-6 fw	CACCGAGATACTCGCTTCATCCCAT
sgExosc3-6 rv	AAACATGGGATGAAGCGAGTATCTC

qRT-PCR Primers

Name	Sequence (5'– 3')
qPCR-Actb fw	GACGAGGCCCAGAGCAAGAGAGG
qPCR-Actb rv	GGTGTTGAAGGTCTCAAACATG
qPCR-Exosc3 fw	TGATGTTGGAGGGAGTGAGC
qPCR-Exosc3 rv	CACACACTGGCCATAGATGAG
qPCR-uaSf3b1 fw	GCGGAAGAGGATGGCTACT
qPCR-uaSf3b1 rv	GTCTGTACAGCCCTGGCTTC
qPCR-uaTxn1 fw	GCCTCAAGGGCACTTTAACA
qPCR-uaTxn1 rv	GGTCTAGTTTGGGGCATGG
qPCR-uaTcea1 fw	CTATCCGGACTCGCGTTG
qPCR-uaTcea1 rv	CTTTAAGCCCTCGGCAATG
qPCR-uaPigt fw	GTGCTCGATATGCAGTGTGG

qRT-PCR Primers (cont.)

Name	Sequence (5'– 3')
qPCR-uaPigt rv	GGGCTAGGTTTTGAGCCAAG
qPCR-uaP4hb fw	TTGGGTGACGGACCCTAGTT
qPCR-uaP4hb rv	ATTCCGAATGGTGGACAGGA

3' RACE Primers

Name	Sequence (5'– 3')
3' RACE Adapter	GCGAGCACAGAATTAATACGACTCACTATAGGTTTTTTTTTTTTVN
3' RACE Adapter	GCGAGCACAGAATTAATACGACTCACTATAGGTTTTTTTTTTTTVN
3' RACE Outer Primer	GCGAGCACAGAATTAATACGACT
3' RACE Inner Primer	CGCGGATCCGAATTAATACGACTCACTATAGG
Slc1a4-pit-fw3	TCCGTTAGGTGGGATGTTAAG
Slc1a4-pit-fw4	CCTTACACTGGGCTCTCTCAG
Tmem38a-pit-fw1	TAGACAAGCTCCTTTACCAGCAG
Tmem38a-pit-fw2	AGCAGAGTCATCTTGCTGCTAC
Pcf11-pit-fw3	GATTGCAATTATCAGGATGAGC
Pcf11-pit-fw4	GATGAGCCTCCTTTTAGCAGAG
Psmd14-pit-fw1	TAGGGCTGGATGTCATCTCC
Psmd14-pit-fw2	TCTGCTTCCCTCTAGCTTGG
Rad23b-pit-fw1	AAATGCGTTCTTTCGGTCGT
Rad23b-pit-fw2	CTTTCGGTCGTCTTGGGAAC
Tdh-pit-fw1	CAGCAGGTGAAAGCAAGACA
Tdh-pit-fw2	CCAACCTCAGCAGGTGAAAG

Antisense Morpholinos

Name	Sequence (5'– 3')
U1 AMO	GGTATCTCCCCTGCCAGGTAAGTAT
Scr AMO	CCTCTTACCTCAGTTACAATTTATA

2.8 REFERENCES

Almada, A.E., Wu, X., Kriz, A.J., Burge, C.B., and Sharp, P.A. (2013). Promoter directionality is controlled by U1 snRNP and polyadenylation signals. Nature *499*, 360-363.

Andersen, P.K., Lykke-Andersen, S., and Jensen, T.H. (2012). Promoter-proximal polyadenylation sites reduce transcription activity. Genes Dev *26*, 2169-2179.

Andersen, P.R., Domanski, M., Kristiansen, M.S., Storvall, H., Ntini, E., Verheggen, C., Schein, A., Bunkenborg, J., Poser, I., Hallais, M., *et al.* (2013). The human cap-binding complex is functionally connected to the nuclear RNA exosome. Nat Struct Mol Biol *20*, 1367-1376.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., *et al.* (2014). An atlas of active enhancers across human cell types and tissues. Nature *507*, 455-461.

Andrulis, E.D., Werner, J., Nazarian, A., Erdjument-Bromage, H., Tempst, P., and Lis, J.T. (2002). The RNA processing exosome is linked to elongating RNA polymerase II in Drosophila. Nature *420*, 837-841.

Beaulieu, Y.B., Kleinman, C.L., Landry-Voyer, A.M., Majewski, J., and Bachand, F. (2012). Polyadenylation-dependent control of long noncoding RNA expression by the poly(A)-binding protein nuclear 1. PLoS Genet *8*, e1003078.

Berg, M.G., Singh, L.N., Younis, I., Liu, Q., Pinto, A.M., Kaida, D., Zhang, Z., Cho, S., Sherrill-Mix, S., Wan, L., *et al.* (2012). U1 snRNP determines mRNA length and regulates isoform expression. Cell *150*, 53-64.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114-2120.

Bousquet-Antonelli, C., Presutti, C., and Tollervey, D. (2000). Identification of a regulated pathway for nuclear pre-mRNA turnover. Cell *102*, 765-775.

Bresson, S.M., and Conrad, N.K. (2013). The human nuclear poly(a)-binding protein promotes RNA hyperadenylation and decay. PLoS Genet *9*, e1003893.

Bresson, S.M., Hunter, O.V., Hunter, A.C., and Conrad, N.K. (2015). Canonical Poly(A) Polymerase Activity Promotes the Decay of a Wide Variety of Mammalian Nuclear RNAs. PLoS Genet *11*, e1005610.

Chen, J., Shishkin, A.A., Zhu, X., Kadri, S., Maza, I., Guttman, M., Hanna, J.H., Regev, A., and Garber, M. (2016). Evolutionary analysis across mammals reveals distinct classes of long non-coding RNAs. Genome Biol *17*, 19.

Core, L.J., Waterfall, J.J., and Lis, J.T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science *322*, 1845-1848.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15-21.

Flynn, R.A., Almada, A.E., Zamudio, J.R., and Sharp, P.A. (2011). Antisense RNA polymerase II divergent transcripts are P-TEFb dependent and substrates for the RNA exosome. Proc Natl Acad Sci U S A *108*, 10460-10465.

Gardini, A., Baillat, D., Cesaroni, M., Hu, D., Marinis, J.M., Wagner, E.J., Lazar, M.A., Shilatifard, A., and Shiekhattar, R. (2014). Integrator regulates transcriptional initiation and pause release following activation. Mol Cell *56*, 128-139.

Ginno, P.A., Lott, P.L., Christensen, H.C., Korf, I., and Chedin, F. (2012). R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters. Mol Cell *45*, 814-825.

Hackett, J.A., and Surani, M.A. (2014). Regulatory principles of pluripotency: from the ground state up. Cell Stem Cell *15*, 416-430.

Hallais, M., Pontvianne, F., Andersen, P.R., Clerici, M., Lener, D., Benbahouche Nel, H., Gostan, T., Vandermoere, F., Robert, M.C., Cusack, S., *et al.* (2013). CBC-ARS2 stimulates 3'-end maturation of multiple RNA families and favors cap-proximal processing. Nat Struct Mol Biol *20*, 1358-1366.

Hilleren, P.J., and Parker, R. (2003). Cytoplasmic degradation of splice-defective pre-mRNAs and intermediates. Mol Cell *12*, 1453-1465.

Hsieh, C.L., Fei, T., Chen, Y., Li, T., Gao, Y., Wang, X., Sun, T., Sweeney, C.J., Lee, G.S., Chen, S., *et al.* (2014). Enhancer RNAs participate in androgen receptor-driven looping that selectively enhances gene activation. Proc Natl Acad Sci U S A *111*, 7319-7324.

Jonkers, I., Kwak, H., and Lis, J.T. (2014). Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. Elife *3*, e02407.

Kaida, D., Berg, M.G., Younis, I., Kasim, M., Singh, L.N., Wan, L., and Dreyfuss, G. (2010). U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. Nature *468*, 664-668.

Kim, T.K., Hemberg, M., Gray, J.M., Costa, A.M., Bear, D.M., Wu, J., Harmin, D.A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., *et al.* (2010). Widespread transcription at neuronal activity-regulated enhancers. Nature *465*, 182-187.

Kwak, H., Fuda, N.J., Core, L.J., and Lis, J.T. (2013). Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. Science *339*, 950-953.

Lai, F., Gardini, A., Zhang, A., and Shiekhattar, R. (2015). Integrator mediates the biogenesis of enhancer RNAs. Nature *525*, 399-403.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods *9*, 357-359.

Lemay, J.F., Larochelle, M., Marguerat, S., Atkinson, S., Bahler, J., and Bachand, F. (2014). The RNA exosome promotes transcription termination of backtracked RNA polymerase II. Nat Struct Mol Biol *21*, 919-926.

Li, M., He, Y., Dubois, W., Wu, X., Shi, J., and Huang, J. (2012). Distinct regulatory mechanisms and functions for p53-activated and p53-repressed DNA damage response genes in embryonic stem cells. Mol Cell *46*, 30-42.

Lin, T., Chao, C., Saito, S., Mazur, S.J., Murphy, M.E., Appella, E., and Xu, Y. (2005). p53 induces differentiation of mouse embryonic stem cells by suppressing Nanog expression. Nat Cell Biol *7*, 165-171.

Lubas, M., Andersen, P.R., Schein, A., Dziembowski, A., Kudla, G., and Jensen, T.H. (2015). The human nuclear exosome targeting complex is loaded onto newly synthesized RNA to direct early ribonucleolysis. Cell Rep *10*, 178-192.

Lubas, M., Christensen, M.S., Kristiansen, M.S., Domanski, M., Falkenby, L.G., Lykke-Andersen, S., Andersen, J.S., Dziembowski, A., and Jensen, T.H. (2011). Interaction profiling identifies the human nuclear exosome targeting complex. Mol Cell *43*, 624-637.

Lum, P.Y., Armour, C.D., Stepaniants, S.B., Cavet, G., Wolf, M.K., Butler, J.S., Hinshaw, J.C., Garnier, P., Prestwich, G.D., Leonardson, A., *et al.* (2004). Discovering modes of action for therapeutic compounds using a genome-wide screen of yeast heterozygotes. Cell *116*, 121-137.

Mahat, D.B., Kwak, H., Booth, G.T., Jonkers, I.H., Danko, C.G., Patel, R.K., Waters, C.T., Munson, K., Core, L.J., and Lis, J.T. (2016). Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). Nat Protoc *11*, 1455-1476.

Mayer, A., di Iulio, J., Maleri, S., Eser, U., Vierstra, J., Reynolds, A., Sandstrom, R., Stamatoyannopoulos, J.A., and Churchman, L.S. (2015). Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution. Cell *161*, 541-554.

Meola, N., Domanski, M., Karadoulama, E., Chen, Y., Gentil, C., Pultz, D., Vitting-Seerup, K., Lykke-Andersen, S., Andersen, J.S., Sandelin, A., *et al.* (2016). Identification of a Nuclear Exosome Decay Pathway for Processed Transcripts. Mol Cell *64*, 520-533.

Mitchell, P., Petfalski, E., Shevchenko, A., Mann, M., and Tollervey, D. (1997). The exosome: a conserved eukaryotic RNA processing complex containing multiple 3'-->5' exoribonucleases. Cell *91*, 457-466.

Mitchell, P., and Tollervey, D. (2003). An NMD pathway in yeast involving accelerated deadenylation and exosome-mediated 3'-->5' degradation. Mol Cell *11*, 1405-1413.

Muse, G.W., Gilchrist, D.A., Nechaev, S., Shah, R., Parker, J.S., Grissom, S.F., Zeitlinger, J., and Adelman, K. (2007). RNA polymerase is poised for activation across the genome. Nat Genet *39*, 1507-1511.

Pefanis, E., Wang, J., Rothschild, G., Lim, J., Chao, J., Rabadan, R., Economides, A.N., and Basu, U. (2014). Noncoding RNA transcription targets AID to divergently transcribed loci in B cells. Nature *514*, 389-393.

Pefanis, E., Wang, J., Rothschild, G., Lim, J., Kazadi, D., Sun, J., Federation, A., Chao, J., Elliott, O., Liu, Z.P., *et al.* (2015). RNA exosome-regulated long non-coding RNA transcription controls super-enhancer activity. Cell *161*, 774-789.

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T., and Salzberg, S.L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol *33*, 290-295.

Preker, P., Almvig, K., Christensen, M.S., Valen, E., Mapendano, C.K., Sandelin, A., and Jensen, T.H. (2011). PROMoter uPstream Transcripts share characteristics with mRNAs and are produced upstream of all three major types of mammalian promoters. Nucleic Acids Res *39*, 7179-7193.

Preker, P., Nielsen, J., Kammler, S., Lykke-Andersen, S., Christensen, M.S., Mapendano, C.K., Schierup, M.H., and Jensen, T.H. (2008). RNA exosome depletion reveals transcription upstream of active human promoters. Science *322*, 1851-1854.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842.

Rahl, P.B., Lin, C.Y., Seila, A.C., Flynn, R.A., McCuine, S., Burge, C.B., Sharp, P.A., and Young, R.A. (2010). c-Myc regulates transcriptional pause release. Cell *141*, 432-445.

Seila, A.C., Calabrese, J.M., Levine, S.S., Yeo, G.W., Rahl, P.B., Flynn, R.A., Young, R.A., and Sharp, P.A. (2008). Divergent transcription from active promoters. Science *322*, 1849-1851.

Spies, N., Burge, C.B., and Bartel, D.P. (2013). 3' UTR-isoform choice has limited influence on the stability and translational efficiency of most mRNAs in mouse fibroblasts. Genome Res *23*, 2078-2090.

Stadelmayer, B., Micas, G., Gamot, A., Martin, P., Malirat, N., Koval, S., Raffel, R., Sobhian, B., Severac, D., Rialle, S., *et al.* (2014). Integrator complex regulates NELF-mediated RNA polymerase II pause/release and processivity at coding genes. Nat Commun *5*, 5531.

Stefanska, B., Cheishvili, D., Suderman, M., Arakelian, A., Huang, J., Hallett, M., Han, Z.G., Al-Mahtab, M., Akbar, S.M., Khan, W.A., *et al.* (2014). Genome-wide study of hypomethylated and induced genes in patients with liver cancer unravels novel anticancer targets. Clin Cancer Res *20*, 3118-3132.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., *et al.* (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A *102*, 15545-15550.

Suzuki, H.I., Young, R.A., and Sharp, P.A. (2017). Super-Enhancer-Mediated RNA Processing Revealed by Integrative MicroRNA Network Analysis. Cell *168*, 1000-1014 e1015.

Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I., and Young, R.A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell *153*, 307-319.

Wyers, F., Rougemaille, M., Badis, G., Rousselle, J.C., Dufour, M.E., Boulay, J., Regnault, B., Devaux, F., Namane, A., Seraphin, B., *et al.* (2005). Cryptic pol II transcripts are degraded by a nuclear quality control pathway involving a new poly(A) polymerase. Cell *121*, 725-737.

Zeitlinger, J., Stark, A., Kellis, M., Hong, J.W., Nechaev, S., Adelman, K., Levine, M., and Young, R.A. (2007). RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo. Nat Genet *39*, 1512-1516.

Chapter 3

Pausing at the First Stable Nucleosome is Linked with Premature Termination

This chapter is adapted from the second half of the following manuscript as well as unpublished data:

Anthony C. Chiu, Hiroshi I. Suzuki, Xuebing Wu, Dig B. Mahat, Andrea J. Kriz, and Phillip A Sharp. U1 snRNP Suppresses Premature Polyadenylation at Transcription Pause Sites Associated with Stable Nucleosomes.

Contributions:

AC and HS designed the experiments, performed U1 inhibition and performed computational analyses. AC and XW generated 2P-seq data. AK generated the Exosc3 CKO cell line. AC generated RNA-seq libraries. AC and DM generated PRO-seq libraries.

3.1 ABSTRACT

Divergent transcription and RNA polymerase II (Pol II) pausing are two common features of vertebrate gene transcription, however, their link is unclear. By developing a conditional Exosc3 deletion mouse embryonic stem cell (mESC) system, we previously identified a large class of polyadenylated short transcripts in the sense mRNA direction that is degraded by the RNA exosome. These PAS termination events are enriched at the first few stable nucleosomes demarcated by CpG islands and suppressed by U1 snRNP. Interestingly, promoter-proximal Pol II pausing consists of two processes: TSS-proximal and +1 stable nucleosome pausing, and PAS termination coincides with the latter pausing. Furthermore, while pausing factors NELF/DSIF primarily function in the former step, flavopiridol-sensitive mechanism(s) and Myc are involved in both steps. Collectively, premature PAS termination near the nucleosome-associated pause site may represent a common transcriptional elongation checkpoint regulated by U1 snRNP.

3.2 INTRODUCTION

Transcription is a highly regulated process in eukaryotes. While the majority of mammalian promoters generate divergent transcription from transcription start sites (TSSs), most upstream antisense RNAs are produced at much lower levels than the sense mRNA (Core et al., 2008; Preker et al., 2008; Seila et al., 2008). This is largely due to the activity of the RNA exosome, a complex with 3'-to-5'exonuclease activity linked with quality control of many classes of transcripts (Andersson et al., 2014; Flynn et al., 2011; Preker et al., 2011; Preker et al., 2008). It has been proposed that early termination by polyadenylation signal (PAS) motifs are a major driver of transcript instability (Andersen et al., 2012). Supporting this view, a subset of uaRNAs are polyadenylated close to the promoter and degraded (Almada et al., 2013; Ntini et al., 2013; Preker et al., 2008). Thus, we proposed that an asymmetric distribution of U1 binding sites and PAS motifs around the TSS (the U1-PAS axis) regulates early termination at uaRNAs and delays termination at mRNAs, resulting in differential transcript stability (Almada et al., 2013; Kaida et al., 2010; Ntini et al., 2013).

RNA Polymerase II (Pol II) frequently pauses during the transcription process. Immediately after transcription initiation, Pol II pauses in most metazoans around 30-60 nucleotides downstream of the transcription start site (TSS) (Muse et al., 2007; Rahl et al., 2010; Zeitlinger et al., 2007). The promoter proximal pause is established though binding of NELF and DSIF to Pol II (Wada et al., 1998a; Yamaguchi et al., 1999). Pause release is associated with phosphorylation of NELF, DSIF and Serine 2 of the carboxy-terminal domain (CTD) of Pol II by P-TEFb (Cheng and Price, 2007; Kim and Sharp, 2001; Wada et al., 1998b). Pausing also occurs when transcribing Pol II encounters nucleosomes, especially at the +1 nucleosome (Churchman and Weissman, 2011; Kwak et al., 2013; Weber et al., 2014). Chromatin remodeling and histone

modifications have been implicated in regulating transcription, and lowering the nucleosome barriers. The histone variant H2AZ typically is found near the TSS and lowers the +1 nucleosome barrier in Drosophila (Weber et al., 2014), likely because nucleosomes that contain H2A.Z are more unstable (Jin and Felsenfeld, 2007). In mESCs, incorporation of the histone variant H2AZ is essential for stabilization of uaRNAs upon knockdown of the RNA exosome (Rege et al., 2015). Additionally, Chd1 is highly biased towards the +1 nucleosome (de Dieuleveult et al., 2016) and is implicated in lowering the +1 nucleosome barrier in mammals (Skene et al., 2014). Other enzymes that act on histones have been implicated in regulating divergent transcription. In yeast, loss of histone chaperone CAF-I and other proteins in the H3K56ac chromatin-assembly pathway led to an increase in divergent transcription (Marquardt et al., 2014), and loss of RCO1 in the Rpd3S H4 deacetylase complex promoted divergent antisense transcription in yeast (Churchman and Weissman, 2011), suggesting chromatin structure can modulate divergent transcription.

We previously demonstrated that the RNA exosome destabilizes promoter proximal sense RNAs that terminate in the first intron, mirroring the activity of the RNA exosome on uaRNAs in the antisense direction. Here, we show that exosome-targeted PAS termination is dramatically enriched at the edges of promoter proximal regions devoid of stable nucleosomes, demarcated by CpG islands, and is associated with active regulation of chromatin remodeling and Pol II pausing. Our analysis further showed that these genomic domains mechanistically delineate stepwise Pol II pausing: TSS proximal pausing and +1 stable nucleosome pausing. Overall, this study proposes an elongation checkpoint involving the convergence of the U1-PAS axis, exosome activity, and Pol II pausing.

3.3 RESULTS

Enrichment of PAS-mediated Termination at the Edges of Nucleosome Free Regions

Previously, premature termination was proposed to occur within the first intron upon removal of the RNA exosome or inhibition of U1 activity. Manual inspection of several genes with detectable premature termination (Rad23b and Pcf11) suggested that the PAS-linked cleavage sites in the first intron are often at the periphery of a CpG island, rich in H2A.Z and H3K4me3, and close to the edge of a region of low nucleosome occupancy in MNase-seq (**Fig. 1A-B**). A genome-wide analysis revealed almost all expressed genes (FPKM > 0.5) with 2P clusters have promoters with a CpG island (p-value < 0.0001, hypergeometric test) (**Fig. 1C**). Since genes with CpG-islands are typically expressed at higher levels than other genes (Ramirez-Carrozzi et al., 2009), this correlation could reflect an expression bias. However, there was no clear relationship between expression levels and fraction of genes with 2P clusters above FPKM values of 1 (**Fig. 1D**). Moreover, after binning genes based on expression levels, the overlap with CGI promoters remains highly significant, arguing that gene expression is not the primary reason that premature clusters are detected on these genes (**Fig. 1E**).

In mammals, CpG islands are regions with unstable nucleosomes and are frequently flanked by more stable nucleosomes (Fenouil et al., 2012; Ramirez-Carrozzi et al., 2009). To further analyze the relationship between 2P clusters and stable nucleosomes, we generated a catalogue of the invariant nucleosomes in mESC. For this purpose, we utilized the recently developed NucTools algorithm, which integrates multiple MNase-seq datasets to define stable versus unstable nucleosomes using the relative error of nucleosome occupancy (Vainshtein et al., 2017). We further incorporated the information of precise nucleosome dyad centers recently defined by chemical mapping in mESC (Voong et al., 2016). The resulting +1 stable nucleosome

position correlated strongly with a dramatic increase in resistance to MNase digestion at the boundary of CpG islands in MNase-seq (**Fig. 2A**). Based on this stable nucleosome free regions[1] (SNFR), we compared the distribution of cleavage sites, CpG islands, PAS motifs, and nucleosomes (**Fig. 2A-C,S1A**). The annotated TSS was more biased to the upstream edge of the SNFR, and the majority of detected PAS termination events occurred as Pol II encounters stable nucleosomes in sense direction (**Fig. 2A**). Interestingly, this pattern was paralleled by uaRNAs, where antisense PAS termination events were enriched at the edge of the SNFR and CpG island.

Surprisingly, premature PAS termination events peaked immediately after the dyads of first stable nucleosomes, and extended through a downstream 1kb window spanning approximately 4 nucleosomes in the sense direction (**Fig. 2B**). We term this region where enhanced termination occurs the Stable Nucleosome Termination Area (or SNTA). Loss of U1 or Exosc3 resulted in substantial increase in detectable PAS cleavage sites in the SNTA. Similarly, an enrichment of termination signals is detectable midway through the upstream -1 stable nucleosome, defining the SNTA in the antisense direction (**Fig. 2C**). The PAS motif frequency strongly mirrors nucleosome positioning in both directions (**Fig. 2B-C**), primarily due to the high GC content in the SNFR since PAS motifs are enriched for A/Ts. As previously reported, the density of PAS motifs in the sense direction beyond the SNFR is lower than that in the antisense direction (Almada et al., 2013). In comparisons of wide and narrow SNFRs, we observed a similar trend, but the effects of U1 inhibition were more apparent for wide SNFR genes (**Fig. S1B**).

More noteworthy, while the frequency of PAS motifs remains constant across the gene body in the sense direction, premature PAS termination is restricted to the first few stable

[1] Stable nucleosome free region (SNFR) is different from the nucleosome free region (NFR). The NFR exists in between the TSSs on the antisense and sense TSS. In contrast, the SNFR is much broader. A comparison of various DNase-seq and MNase-seq by de Dieuleveult supports this distinction.

nucleosomes, i.e. SNTA (**Fig. 2B**), suggesting that the SNTA derives from a preference of Pol II to terminate at early PAS motifs in the absence of U1 suppression. Consistent with this, the most frequently used cluster in the first intron is often the first two predicted canonical PAS motifs, irrespective of U1 inhibition or Exosc3 depletion (**Fig. 2D**). This suggests that in normal cells, a subset of Pol II preferentially terminates at the first few PAS motifs rather than the main PAS site, perhaps reflecting that these elongation complexes have not fully matured and are incapable of transcribing through stable nucleosomes. Similar trends were observed for uaRNAs (**Fig. 2E**). While the core termination elements consist of the PAS hexamer and downstream elements, the most frequent premature termination events in the first intron occurred 50% of the time at the first two PAS motifs, or 60% of the time at the first two PAS hexamers at uaRNAs, suggesting that the PAS hexamer alone is sufficient to initiate termination.

Nucleosome positioning is strongly influenced by AA/TT/TA dinucleotide sequences phased at 10 base pairs intervals (Segal et al., 2006). We found that both canonical PAS motifs used in premature termination events within the gene body and predicted PAS motifs closely mimic the periodic AA/TT/TA dinucleotide patterns (**Fig. 2F,S1C**). These findings suggest an impact of sequence contexts on the functional relationship between nucleosome organization and PAS termination.

Association of PAS Termination and Chromatin Remodeling Factors at the +1 Stable Nucleosome

Nucleosome organization is influenced by various chromatin remodelers such as Chd1, Chd4, and Ep400 and is thought to influence the kinetics of Pol II elongation. Among them, Chd1

is reported to regulate transcription by modulating nucleosome turnover throughout the gene (Skene et al., 2014). Recently reported genome-wide remodeler-nucleosome interaction profiles have demonstrated specific accumulation of Chd1 and other chromatin remodelers at the edge of the SNFR and that Pol II navigates several hundreds of base pairs of regions with poorly defined nucleosomes before traversing remodeler-targeted nucleosomes towards the downstream gene body (de Dieuleveult et al., 2016). Using this MNase digestion-coupled ChIP-seq datasets (Table S1), we investigated the relationship between premature termination and chromatin remodeling. Most chromatin remodeling factors were enriched around the SNFR edges of genes with 2P clusters , aside from Chd2 being distributed across the gene body (**Fig. 3A-B,S2A**). In particular, Chd1 prefers +1 stable nucleosomes, whereas several factors such as Ep400 and Chd4 prefer -1 stable nucleosomes. We next compared binding patterns of chromatin remodelers for genes with and without 2P clusters. There was no major decrease in MNase-seq signal for the +1 stable nucleosome or -1 stable nucleosome between genes with 2P clusters or expression-matched genes without 2P clusters (**Fig. 3B**), indicating that genes with premature cleavage clusters do not have higher nucleosome occupancy at steady state, though they could potentially have different nucleosome turnover rates. Despite this, genes with 2P clusters were more strongly bound by several chromatin remodelers including Chd1, Chd2, and Chd9, suggesting that +1 stable nucleosomes associated with PAS termination are actively marked by chromatin remodelers (**Fig. 3B**). Both Chd1 and Chd2 were strongly biased towards the sense-coding direction.

In addition to nucleosome remodeling, changes to nucleosome composition can impact nucleosome dynamics. The incorporation of H2A.Z into nucleosomes is known to reduce the nucleosome barrier for Pol II transcription (Weber et al., 2014). At individual genes, many PAS termination sites occurred downstream of H2AZ and H3K4me3 (**Fig. 1A-B**).

Analysis of histone marks showed that H2A.Z is enriched over SNFR region upstream of cleavage signals overlapping the CpG island (**Fig. 3C**), suggesting that CpG islands may also promote H2A.Z association. However, there did not appear to be any difference in H2AZ occupancy between genes with or without premature termination (**Fig. S2B**). H3K4me3 also overlap the entire CpG island, consistent with reports that the H3K4 methyltransferase is recruited directly to CpG islands through Cfp1 (Thomson et al., 2010). As reported previously (Core et al., 2008; Seila et al., 2008), histone marks of productive elongation (H3K36me3 and H3K79me2) begin within the CpG island and are not uniquely associated with the SNFR (**Fig. 3C**), probably due to their association with splicing. The first 5' splice site frequently occurs upstream of the SNTA near the edges of the stable nucleosomes, consistent with the U1-PAS axis in which an upstream 5' splice site suppresses a downstream PAS.

Premature PAS Termination Correlates with Active Pol II Pause Regulation

Both H2AZ and Chd1 have been linked with regulating the stalling of Pol II at promoter proximal nucleosomes (Skene et al., 2014; Weber et al., 2014). Our analysis showed the ChIP signal for Pol II is primarily distributed close to the TSS for genes with 2P clusters with some signals within the SNFR (**Fig. 4A-B**). In contrast, global run-on (GRO)-seq signals (Jonkers et al., 2014), which detect transcribing Pol II, were abundant at both TSS-proximal regions and the edges of the SNFR in the sense direction. The bias of GROs-seq towards the edge of the SNFR was greater at genes with wider SNFRs. Antisense GRO-seq reads predominantly followed the edge of the -1 stable nucleosome. This GRO-seq pattern suggests that two types of Pol II pausing occurs in the sense direction especially for genes with wide SNFRs, where the two can be resolved: TSS-proximal pause and stable nucleosome pause.

Promoter-proximal pausing is enforced by the binding of NELF and DSIF (Wada et al., 1998a; Yamaguchi et al., 1999). Two pausing factors, NelfA subunit of NELF and Spt5 subunit of DSIF, were enriched at the site of TSS-proximal paused Pol II, consistent with their roles in promoting the promoter-proximal pause (**Fig. 4A**). Cdk9, a subunit of the P-TEFb complex that stimulates promoter-proximal pause release, accumulated at the TSS-proximal region in parallel with its substrates Pol II and DSIF, but was further distributed within the SNFR. Aff4 and Ell2, subunits of the Super Elongation Complex (SEC) associated with P-TEFb (Lin et al., 2010), were widely distributed from TSS to the SNFR edge. Interestingly, genes with premature PAS clusters had increased binding of Pol II, SEC components (Aff4 and Ell2), and NELF/DSIF, when compared to expression-matched controls (**Fig. 4B**), which suggests premature termination is associated with more active Pol II pause regulation at the edge of the SNFRs.

Modifications of the C-terminal repeat domain (CTD) of Pol II at Ser5 and Ser2 reflect the Pol II status during elongation. In order to better compare these modifications with respect to the site of PAS termination, we selected the most frequently used cleavage cluster in the sense direction for each gene and constructed metaplots (**Fig. 4C-E**). Similar to the SNFR view (**Fig. 3B**), Chd1 accumulated at the most frequently used 2P cluster, whereas the SEC signal, Aff4 and Ell2, diminished at the most frequently used 2P cluster (**Fig. 4C-D**). Though the density of Pol II reached a nadir at this point, the density of Ser2 phosphorylation increased, whereas that of Ser5 phosphorylation remained relatively constant (**Fig. 4E**). The increase in Ser2P is independent of the distance from TSS to the dyad axis, whereas the decrease in Ser5P is deeper as the distance from the TSS increases (**Fig. S2C-D**). This suggests that a Ser2 kinase, such as Cdk9, is likely active at these sites, as further evidenced by higher density of Cdk9 as compared to DSIF and NELF immediately upstream of this site (**Fig. 4E**).

PAS Termination is associated with a Flavopiridol-sensitive +1 Stable Nucleosome Pausing

Our previous results suggest that +1 stable nucleosomes associated with premature polyadenylation are marked by active chromatin remodeling and active Pol II pause regulation. Thus, PAS termination in the vicinity of the +1 stable nucleosome could represent a point of gene regulation. To further investigate a possible relationship between premature termination and Pol II pausing, we focused on genes with wide SNFRs (distance between TSS and +1 dyad axis > 600 bps) since it is difficult to distinguish between a TSS proximal pause and a +1 stable nucleosome pause at genes with narrow SNFRs. Alignments of Pol II ChIP-Seq around the TSS revealed a major pause immediately downstream of the TSS (**Fig. 5A**, blue bar), followed by a less steep ramp around 300 to 900 bp from the TSS (orange bar) representing the +1 stable nucleosome pause, followed by gene body signal (green bar). Genes with premature cluster events had an increased Pol II ChIP signal near the promoter relative to expression-matched gene sets without 2P clusters. Notably, in the +1 dyad-centric view, genes with premature clusters have a ramp of Pol II occupancy in front of the dyad and a peak of GRO-seq signal flanking the +1 stable nucleosome, and this phenomenon was less pronounced at genes without premature clusters (**Fig. 5B**). This suggests genes with premature clusters are more likely to be targets of active pausing at the +1 stable nucleosome.

We next closely compared differential sensitivity of Pol II pausing at genes with or without premature clusters to experimental modulation of Pol II pause regulators: treatment with flavopiridol (an inhibitor of Cdk9/Cdk12 kinase activity) and knockdown of DSIF and NELF, using previously published datasets (Rahl et al., 2010). Furthermore, to distinguish the effects on the TSS-proximal pause and +1 stable nucleosome-associated pause, we introduced two pausing

indices based on Pol II ChIP-seq: a TSS Pausing index and a +1 Nucleosome Pausing Index (**Fig. 5C**). The distances that defined each transition were determined by observing where the ramp started/ended in the Pol II ChIP alignments (**Fig. 5A-B**). In this analysis, a higher pausing index suggests increased pausing.

P-TEFb has a central role in promoter-proximal Pol II pausing kinetics by phosphorylating DSIF and NELF, and is inhibited by flavopiridol. Treatment with flavopiridol resulted in statistically-significant increases in mean Pol II signals at both the TSS-proximal region and the immediate upstream region from the dyads of +1 nucleosomes at genes with premature clusters (**Fig. 5D**). Comparisons of the TSS pausing index and the +1 nucleosome pausing index showed that flavopiridol-induced pausing is greater at the +1 nucleosome, and was even stronger at genes with premature clusters than genes without premature clusters (**Fig. 5E**). The pause factors DSIF and NELF bind to the transcription machinery and mediate TSS proximal pausing (Adelman and Lis, 2012). Knockdown of NELF component, NelfA, resulted in very modest effects on pause release at TSS-proximal regions for both gene sets (**Fig. S3A-B**). In contrast, knockdown of DSIF component Spt5 caused a substantial decrease in pausing only at TSS-proximal regions, though there was no apparent difference of TSS pausing and +1 nucleosome pausing indices between genes with and without premature clusters. These TSS-restricted effects are consistent with major accumulation of NelfA and Spt5 around TSS in ChIP-seq profiles (**Fig. 4A**). These analyses unexpectedly highlight differential contribution of DSIF/NELF and flavopiridol-sensitive mechanism(s) to two Pol II pausing steps.

In GRO-seq analysis with flavopiridol treatment (Jonkers et al., 2014), we confirmed that flavopiridol treatment results in a substantial increase in promoter proximal pausing (**Fig. 5F**), as previously described. Additionally, flavopiridol treatment induces a substantial drop in GRO-seq

signal near the +1 stable nucleosome for both genes with premature clusters and those without (**Fig. 5G**). Increased Pol II binding and increased GRO-seq signal upstream of the dyad upon flavopiridol treatment suggest that Cdk9 or other flavopiridol-sensitive kinase(s), e.g. Cdk12, may play a role in promoting transcription beyond stable nucleosomes.

Previously, we found that the RNA exosome regulated promoter proximal pausing, so the increased pausing upon Exosc3 loss could potentially result in easier recognition of PAS motifs and increased detection of premature termination. Depletion of Exosc3 elicited slight pausing effects at the +1 stable nucleosomes, but this effect did not differ between genes with or without 2P clusters (**Fig. S4A-C**). This suggests that the increase in PAS termination transcripts upon Exosc3 depletion is mainly attributable to RNA stabilization, and not from increased pausing.

Myc Regulates +1 Stable Nucleosome Pausing

In mESCs, gene regulation is governed by core transcriptional networks, including Oct4, Sox2, Nanog, and Myc. Myc has been reported to regulate the release of Pol II from the promoter region in mESC (Rahl et al., 2010). According to classification of mESC genes based on association with transcription factor binding (Chen et al., 2008), we found that over 60 % of genes with 2P clusters fall into gene classes with Myc binding (**Fig. 6A**, Class II and III). Myc-binding sites are preferentially found in CpG islands (Perna et al., 2012), consistent with a large overlap with genes sets with 2P clusters and CpG promoters (**Fig. 1C**). These data suggest that Myc may have an important role in regulating genes with premature PAS termination.

An examination of Pol II ChIP data upon treatment with a low-molecular-weight inhibitor of c-Myc/Max (Rahl et al., 2010) revealed that both genes with and without premature clusters

showed roughly a 2-fold increase in Pol II occupancy at the TSS following Myc inhibition (**Fig. 6B**). This was confirmed by an increase in the TSS pausing index for both gene sets, independently of whether there is a premature intron cluster (**Fig. 6C**). Strikingly, genes with premature clusters had an increase in +1 nucleosome pausing upon treatment with a Myc inhibitor, whereas there were much smaller changes at genes without premature clusters (**Fig. 6B-C**), suggesting that Myc preferentially regulates the +1 stable nucleosome pause at genes with premature clusters.

Finally, we analyzed the relationship between Myc-regulated Pol II pausing and Myc-dependent gene regulation. Myc regulates diverse synthetic and metabolic processes and double knockout of c-Myc and N-Myc in mESCs induces a pluripotent dormant state (Scognamiglio et al., 2016). For genes with/without 2P clusters, there is no statistical correlation between changes in gene expression upon Myc knockout and increases or decreases of the TSS pausing index upon Myc inhibition and flavopiridol (**Fig. 6D**). On the other hand, genes with 2P clusters and increased +1 nucleosome pausing upon treatment with Myc inhibitor and flavopiridol have a greater decrease in mRNA expression followinzg Myc knockout relative to other genes (**Fig. 6E**). Consistent with this, genes with increased +1 nucleosome pausing following Myc inhibition and flavopiridol treatment are strongly linked to biological processes characteristic to Myc target genes, including RNA processing, DNA metabolism, chromatin modification, and cell cycle (**Fig. 6F**). Interestingly, we previously observed that loss of Exosc3 results in reduced expression of Myc-regulated target genes (**Ch. 2, Fig. 3B**). These data suggest that the +1 stable nucleosome-associated pause site is an important regulatory point of Myc-dependent gene activation.

Premature Termination May be Conserved in Human

A study in humans using inhibitors of CDK9 suggested that some P-TEFb-regulated promoter-proximal pausing events occurred downstream of the usual TSS proximal pause region (Laitem et al., 2015). Suspecting some of these 'late' early elongation pauses may be linked with premature PAS termination, we examined reported examples of delayed promoter-proximal pauses. Examination of GRO-seq for DHX9 shows that not only does initiation occur within a CpG island accompanied by low MNase signal and high H2A.Z and H3K4me3 ChIP signal, inhibition of CDK9 resulted in a premature termination event with a nearby PAS variant motif (AGTAAA) (**Fig. 7A**). Premature PAS termination has been studied in humans using U1 inhibitors (Berg et al., 2012), so we investigated two genes where U1 inhibition is known to cause premature termination in the first intron. Premature termination occurred upon inhibition of P-TEFb near the edge of a CpG island (**Fig. 7B-C**). GNAI1 terminate close to a canonical AATAAA PAS motif, whereas CUL1 terminate at a canonical ATTAAA PAS motif. These examples strongly suggest there is an early premature termination checkpoint regulated by P-TEFb at the edge of SNFRs that is conserved from mice to human.

Taken together, these findings strongly suggest that promoter proximal pausing consists of at least two distinct processes differentially regulated by multiple pausing regulators: TSS-proximal pausing and +1 stable nucleosome-mediated pausing. NELF and DSIF primarily function in the former step, and flavopiridol-sensitive mechanism(s) and Myc have broader roles in stepwise pausing and are involved in the latter step. Furthermore, PAS termination is preferentially associated with active regulation of the latter step and this mechanism is likely conserved in humans.

3.4 DISCUSSION

Our analysis of premature PAS termination reveals that termination within the first intron or at uaRNAs is strongly associated with the edge of a stable nucleosome free region. This region is linked with a unique promoter structure: the presence of a H2A.Z rich, CpG island. Moreover, we provide evidence that these termination events are strongly associated with regulated pausing at the +1 stable nucleosome. These termination events are also detectable in human, suggesting conservation across species. Altogether, our work suggests that there is a major elongation checkpoint in mammals downstream of the TSS proximal pause.

Promoter proximal pausing of Pol II frequently occurs in metazoans near the TSS (Muse et al., 2007; Rahl et al., 2010; Zeitlinger et al., 2007). This pause 30-60 bp downstream of the TSS occurs due to binding of NELF and DSIF (Wada et al., 1998a; Yamaguchi et al., 1999), and is released by P-TEFb-mediated phosphorylation (Cheng and Price, 2007; Kim and Sharp, 2001; Wada et al., 1998b). More recently, there is growing recognition of a separate pause when Pol II encounters the +1 nucleosome barrier, which can be modulated by chromatin remodelers Chd1 and H2A.Z (Skene et al., 2014; Weber et al., 2014). We now provide evidence that promoter-proximal pausing is made up of two pausing events: a TSS-proximal pause and a +1 stable nucleosome pause. For wide SNFR genes where we can resolve the two, the latter is downstream of the previously described pause site that is regulated by NELF and DSIF. At promoters with short NFRs, it is difficult to resolve the previously well characterized TSS proximal pausing of Pol II and the paused state associated with the +1 nucleosome. Since U1 snRNP suppression of PAS termination is observed at both short and long SNFRs, it is likely that TSS-proximal pausing and nucleosome dependent pausing both occur at short SNFR promoters as well.

Most premature PAS termination events occur at boundaries of the SNFR in the SNTA. We conjecture that premature PAS termination at the SNTA represents an intriguing checkpoint of Pol II elongation in the sense direction (**Fig. 8**). The frequency of this premature termination is suppressed by U1 snRNP presumably through recognition of 5′ splice site sequences near the TSS (Almada et al., 2013; Kaida et al., 2010). Importantly, in both directions, termination predominantly occurs at the edges of the SNFR as defined by micrococcal nuclease digestion. While previous reports described relationships between nucleosome organization (+1 nucleosome) and Pol II pausing, our findings indicate that each of the +1 and -1 stable nucleosomes demarcated by CpG islands represents a key feature of this elongation checkpoint, which is also marked by accumulation of chromatin remodelers.

This checkpoint may be a product of DNA sequence elements. CG-rich segments known as CpG islands overlap about 60-70 % of mammalian promoters (Saxonov et al., 2006) and are regions of low nucleosome occupancy (Fenouil et al., 2012; Ramirez-Carrozzi et al., 2009). The unstable nucleosomes within the CpG island contain H3K4me3 and H2A.Z, a histone variant linked with destabilizing nucleosomes and the +1 nucleosome barrier (Jin and Felsenfeld, 2007; Weber et al., 2014). This region is bracketed by the -1 and +1 stable nucleosome whose precise positions are created by AA/TT/TA dinucleotides spaced at 10 bps that kink the DNA around the nucleosome (Segal et al., 2006). Both in the upstream antisense direction and in the sense direction, PAS sequences are present in these stable nucleosome-bound AT-rich sequences and are utilized to direct cleavage and terminate transcription. We picture these stable nucleosomes forming a barrier to the elongating polymerase after it traverses unstable nucleosomes. This barrier pauses it and enhance the rate of cleavage within the first few unstable nucleosomes, unless the transcription elongation complex has matured to a processive form competent to elongate beyond the SNTA.

The increase in Ser2 phosphorylation density near these PAS termination sites and sensitivity to flavopiridol suggest that Cdk9 or Cdk12 acts to regulate this maturation. U1 snRNP may be important to bypass this checkpoint by suppressing the rate of cleavage and perhaps by recruiting factors such as Chd1 and/or by generating a processive polymerase complex through the coupling elongation with splicing of nascent RNA. Early termination and degradation of uaRNAs may be due to the inability of Pol II to bypass this elongation checkpoint in the antisense direction.

It is currently unknown why CpG islands have been actively selected for over time, where older genes are more likely to have wider CpG islands (Almada et al., 2013). One reason may be that CpG islands facilitate transcription initiation through reduced nucleosome occupancy (Ramirez-Carrozzi et al., 2009) and enhanced H3K4 trimethylation mediated by Cfp1 (Thomson et al., 2010), which in turn recruits the chromatin remodeler Chd1 (Sims et al., 2007) and association of TFIID (Vermeulen et al., 2007). Based on this study, we propose another reason for the evolution of CpG islands is to delay entry into the first stable nucleosome, giving more time for Pol II to mature rather than being forced to encounter a strong +1 stable nucleosome barrier. Pol II gradually increase its processivity the farther it transcribes into the gene (Jonkers et al., 2014), so the observation that early-termination by PAS leads to transcript instability (Andersen et al., 2012) may reflect an inefficiently elongating Pol II. This pathway also appears to function in humans, which have CpG islands, selection for a U1-PAS axis (Ntini et al., 2013) and a secondary CDK9 checkpoint (Laitem et al., 2015) that we found to be associated with premature termination.

Myc promotes promoter-proximal pause release at many promoters in mESC by recruiting P-TEFb (Rahl et al., 2010). We found promoters with premature PAS termination have increased nucleosome pausing and higher sensitivity to flavopiridol treatment and Myc inhibition. Increases

in Myc activity correlate with enhanced cell division and genes with CpG islands are enriched for housekeeping proteins critical for the cell's bio-synthetic capacity (Ramirez-Carrozzi et al., 2009; Saxonov et al., 2006). Thus, Myc-mediated regulation of PAS termination at CpG island promoters could be important for cell growth and other processes critical for tumorigenesis. It is likely that Myc collaborates with Cdk9 in regulation of PAS termination, but probably not through a process requiring NELF and DSIF. Further analysis of this checkpoint would expand our understanding of transcriptional regulation and offer a possibility to target transcriptional perturbation in diseases including Myc-dependent cancers.

3.5 FIGURES

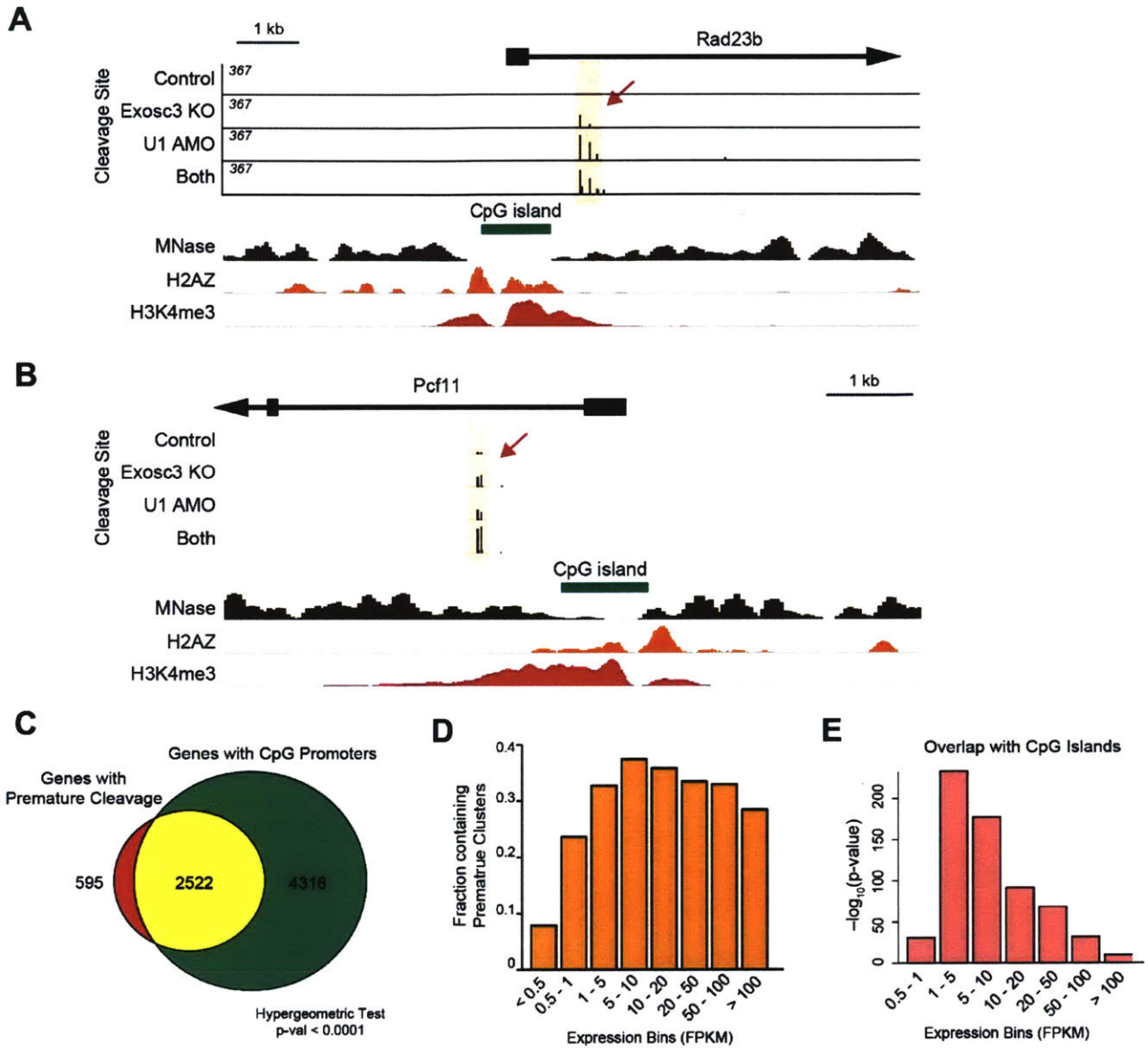

Fig 1. Premature Termination is Associated with CpG Islands

(A-B) Genome browser shots of Rad23b and Pcf11 showing PAS-linked cleavage sites (top,orange shade), annotated CpG island (green), MNase-seq (brown), H2AZ ChIP-seq (orange) and H3K4me3 ChIP-seq (red).
(C) Venn diagram demonstrating significant overlap of genes with premature 2P cleavage and genes with CpG promoters. Expressed genes are analyzed (FPKM > 0.5).
(D) Fraction of genes in different expression bins with detectable premature cleavage events
(E) P-values of hypergeometric test for different expression classes, showing that the overlap between CpG islands and clusters is highly significant, independent of expression.

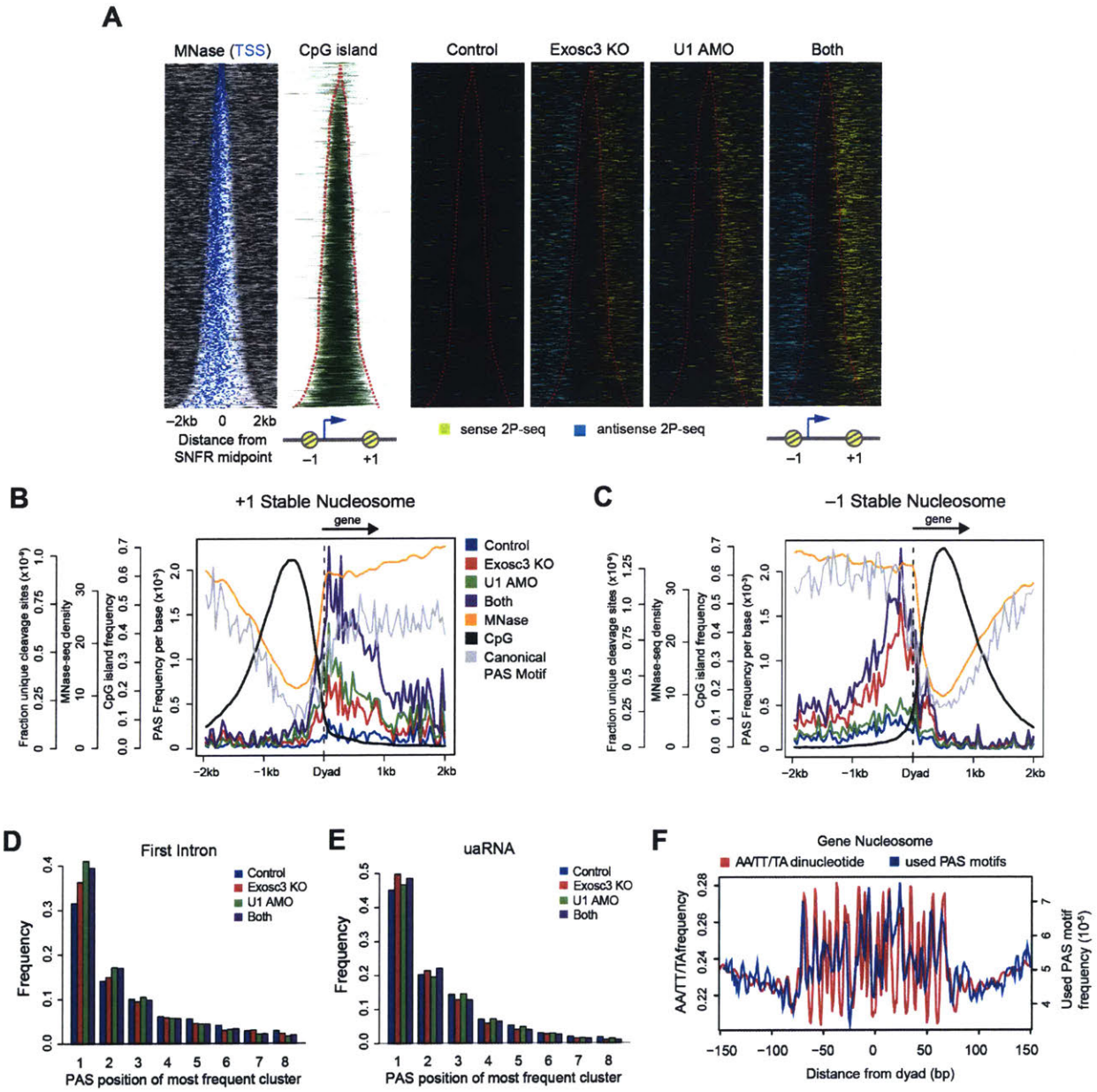

Figure 2. Premature Termination occurs at Edge of Nucleosome Free Region

(A) Heatmap of MNase-seq, CpG islands and PAS-linked cleavage sites (yellow: sense 2P-seq reads, light blue: antisense 2P-seq reads) flanking the SNFR midpoint for nonoverlapping expressed genes, sorted by increasing SNFR width. Red lines indicate SNFR edges. Reads were normalized to mapped library size.
(B-C) Metaplot of PAS-filtered cleavage sites, MNase-seq, CpG Islands and predicted canonical PAS motifs around the dyad axis of the +1 (B) or -1 (C) stable nucleosome.
(D,E) Frequency of the PAS position for the most frequently used cluster with canonical PAS motifs in the first intron (D) or uaRNA (E).
(F) AA/TT/TA dinucleotide frequency (red) and frequency of unique used PAS motifs at cleavage clusters (blue) per gene body nucleosome in a 150 bp window from chemical mapping-derived dyad axis. Gene body nucleosomes are between TSS and 2kb upstream of the transcription end site (TES) of genes.

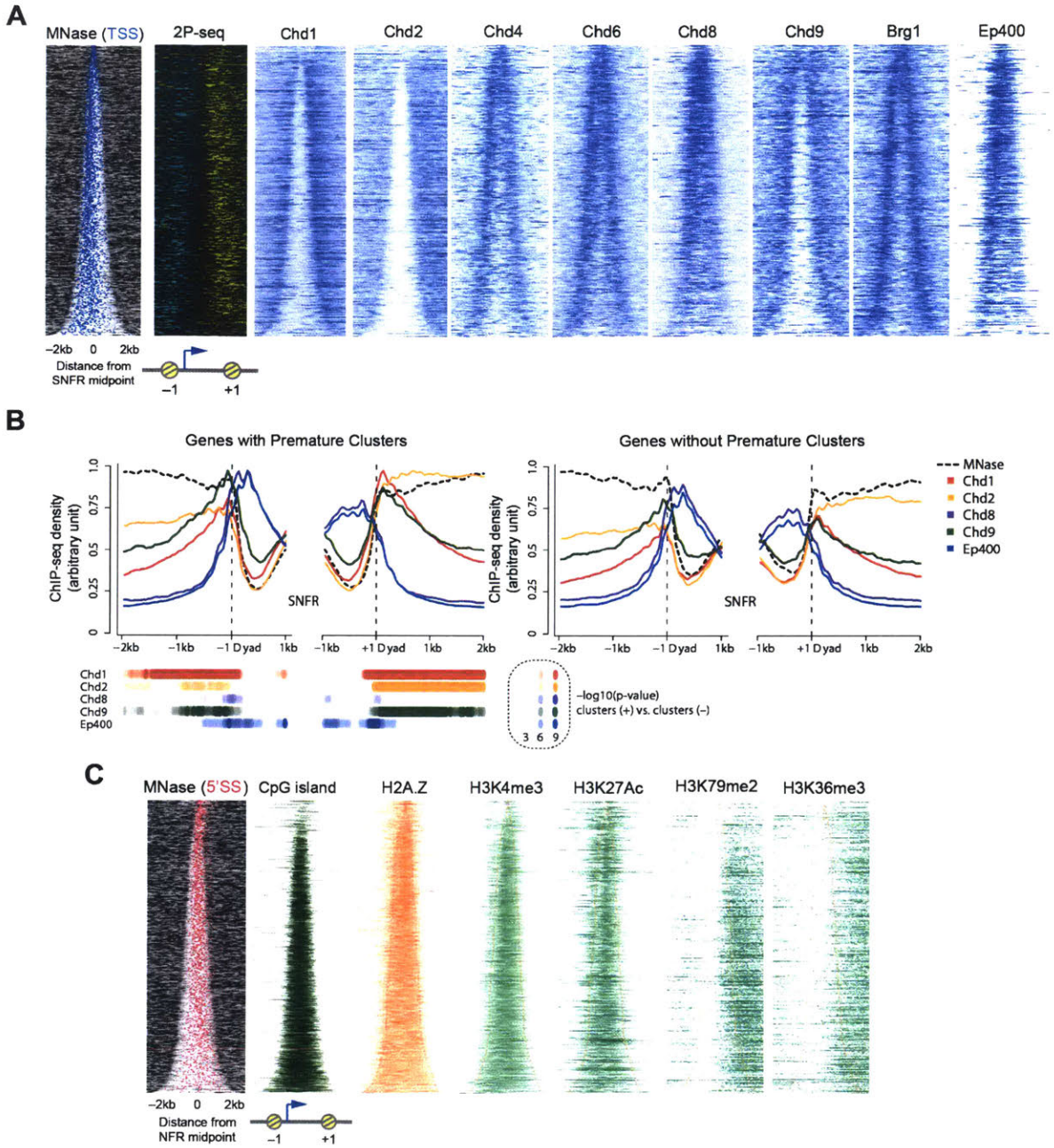

Figure 3. Premature Termination is Associated with Reduced Recruitment of Chromatin Remodellers

(A) Heatmap of ChIP-seq signal for various chromatin remodelers in a 2kb window flanking the SNFR midpoint for non-overlapping expressed genes, ranked by increasing SNFR width.
(B) Read coverage of MNase-seq and MNase digestion-coupled ChIP-seq of various chromatin remodelers in a -2 kb to 1 kb window around the -1 stable nucleosome dyad axis and -1 kb to +2 kb window around the + 1 stable nucleosome dyad axis, separated for genes with premature intron clusters (left) and expression-matched genes without premature intron clusters (right). P values with K-S test at each bin are displayed.
(C) Heatmap of histone modifications around the midpoint of the SNFR, sorted by SNFR width. Red lines indicate SNFR edges.

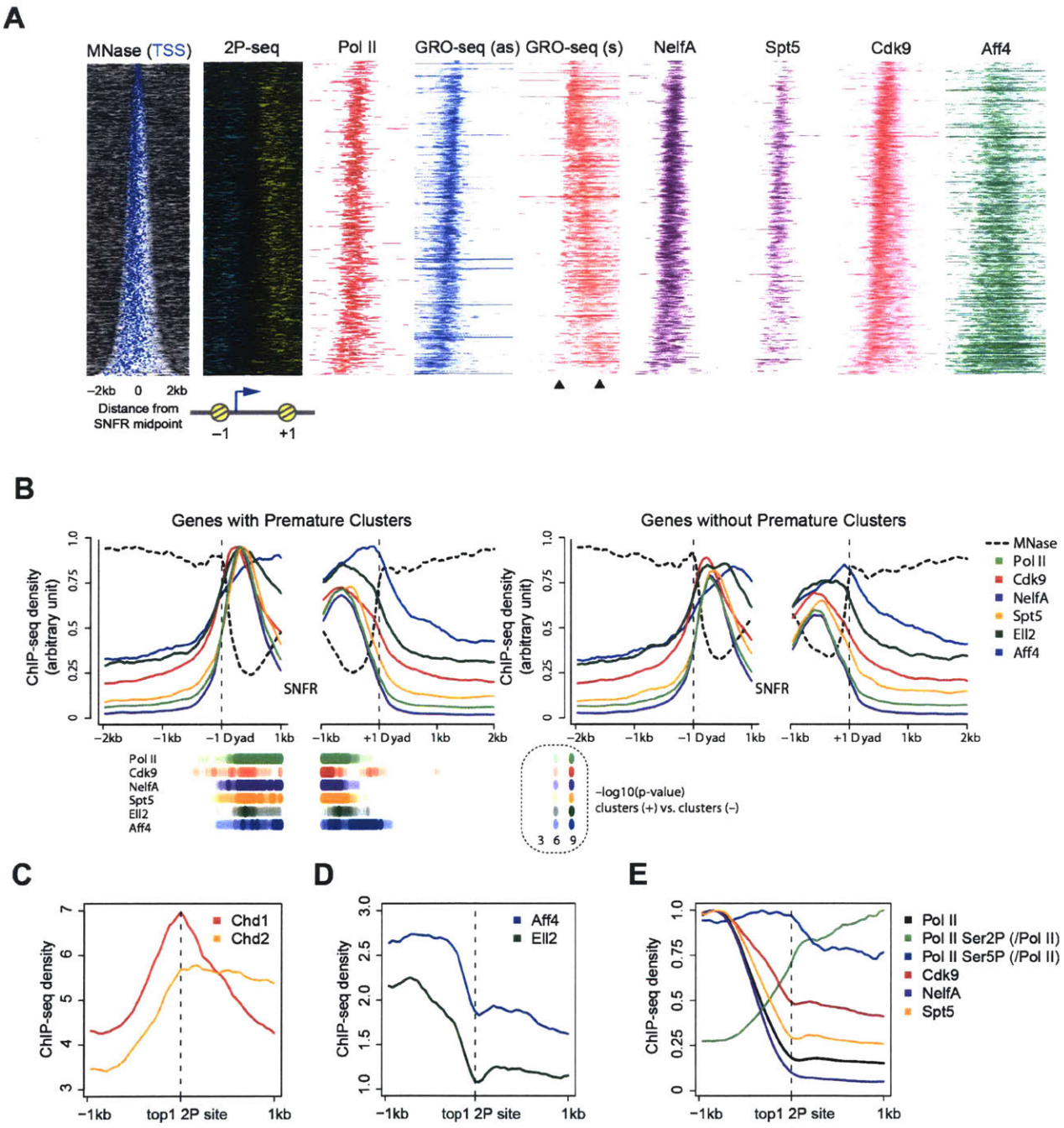

Figure 4. Premature Termination is Associated with Reduced Recruitment of Pausing Factors

(A) Heatmap of MNase-seq, GRO-seq, and ChIP-seq for Pol II and various pausing and elongation factors in a 2 kb window flanking the SNFR midpoint for non-overlapping expressed genes with 2P clusters, ranked by increasing SNFR width.
(B) Read coverage of MNase-seq and ChIP-seq for Pol II and various pausing and elongation factors around -1 and +1 stable nucleosomes. P values with K-S test at each bin are displayed.
(C-E) Metaplots of Chd1, Chd2, SEC components, and Pol II/pausing factors in a 1 kb window around the most frequent PAS-linked termination clusters.

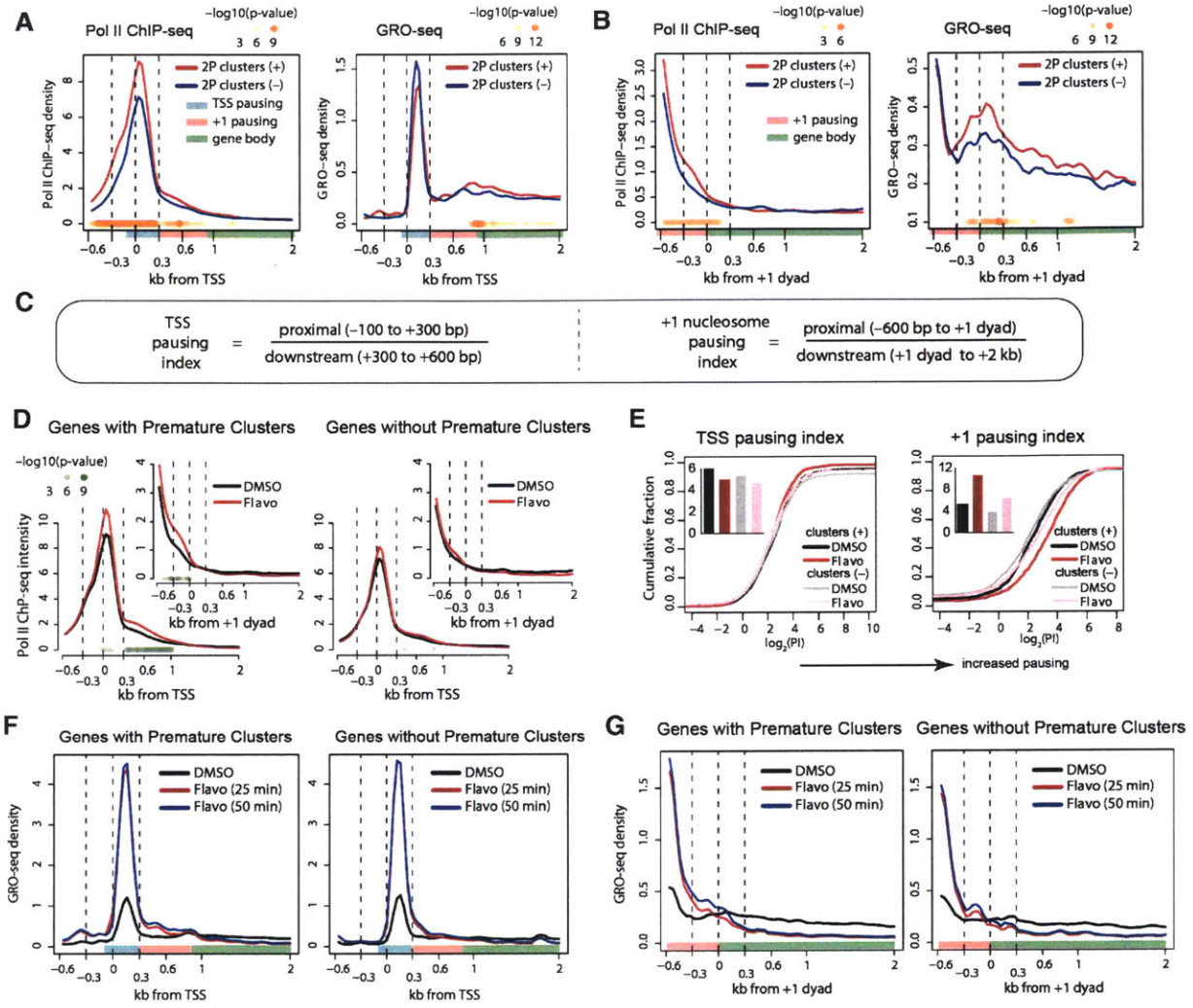

Figure 5. PAS termination and +1 stable nucleosome-associated Pol II pause regulation.

(A-B) Metaplots of mean Pol II ChIP-seq (left) or GRO-seq (right) read density around the TSS (A) or the +1 dyad (B) for wide SNFR genes with 2P clusters (red) and expression-matched wide SNFR genes without 2P clusters (blue). P values with K-S test at each bin are displayed in panels (A), (B), (D), and (F).
(C) Formulas for the two pausing indices.
(D) Metaplots of Pol II ChIP-seq density around the TSS or +1 dyad (inset) of wide SNFR genes with DMSO or flavopiridol treatment.
(E) Cumulative distribution plot of $\log_2$(pausing index) of the TSS proximal (left) and +1 stable nucleosome pause (right) for wide SNFR genes with 2P clusters and expression-matched wide SNFR genes without 2P clusters under DMSO or flavopiridol treatment. Inset shows the median pausing indices in a raw scale.
(F-G) Metaplots of GRO-seq density around the TSS (F) or +1 dyad (G) upon flavopiridol treatment.

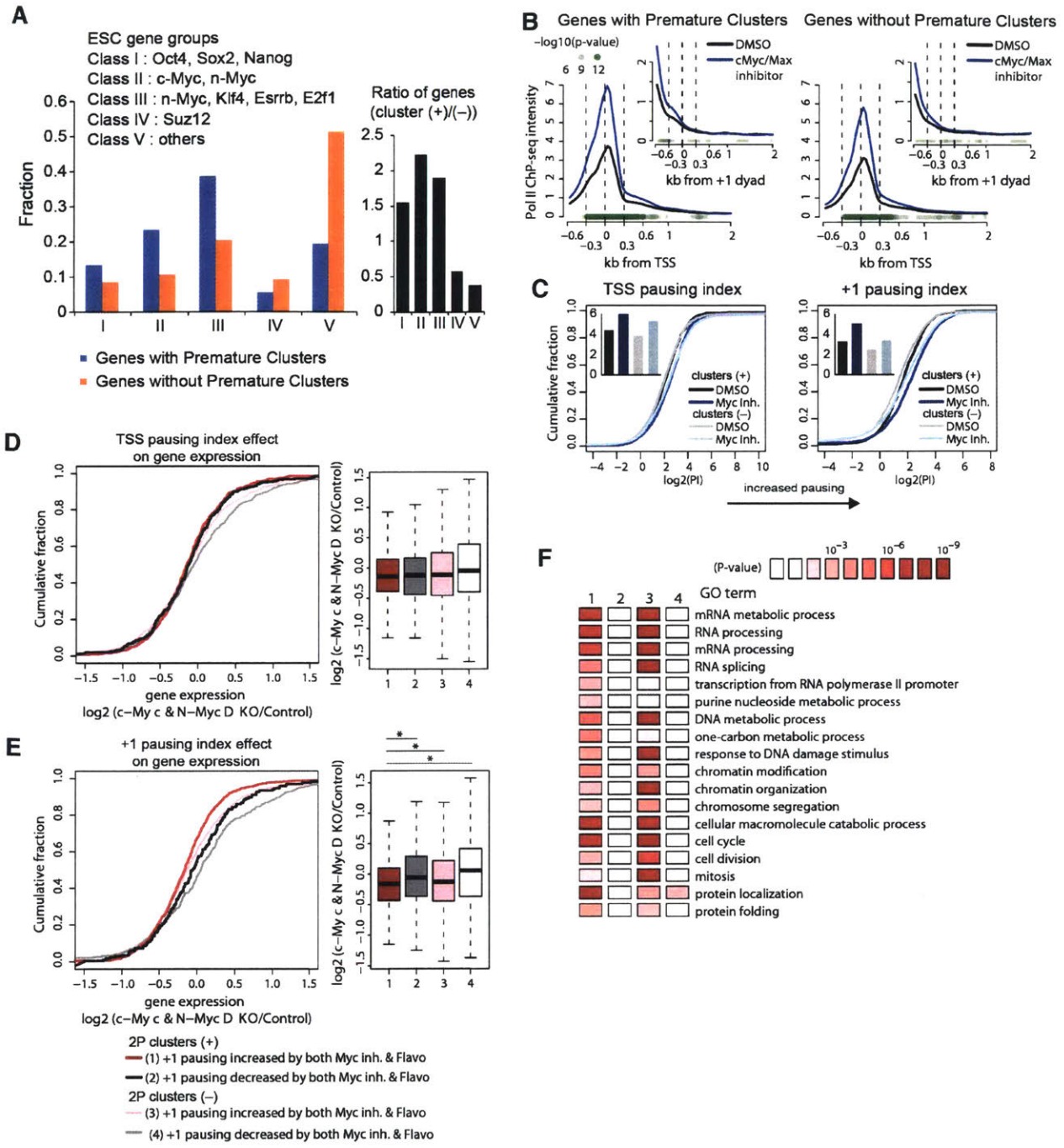

Figure 6. Myc regulates genes with PAS termination and +1 stable nucleosome Pol II pause

(A) Fraction of ESC gene groups for genes with premature intronic clusters (blue) or without premature intronic clusters (orange). Bar graph in black represents fold change.
(B-C) Metaplots of Pol II ChIP-seq density around the TSS or +1 dyad (B) and cumulative distribution plot of $\log_2$(pausing index) of the TSS proximal or +1 stable nucleosome pause (C) for wide SNFR genes with/without 2P clusters upon treatment with DMSO or Myc inhibitor.

Figure 6 (cont).

(D) Effects of TSS pause on Myc-dependent gene regulation. Cumulative distribution of log2 fold change of RNA expression in c-Myc and N-Myc double knockout (DKO) mESC is shown for wide SNFR genes with/without PAS termination and flavopiridol/Myc-sensitive TSS pausing.
(E) Cumulative distribution plot is shown as in panel (D) using +1 stable nucleosome pausing indices. * $p < 0.001$ with K-S test.
(F) Gene ontology terms enriched in each gene sets as defined in panel (E). All expressed genes were analyzed.

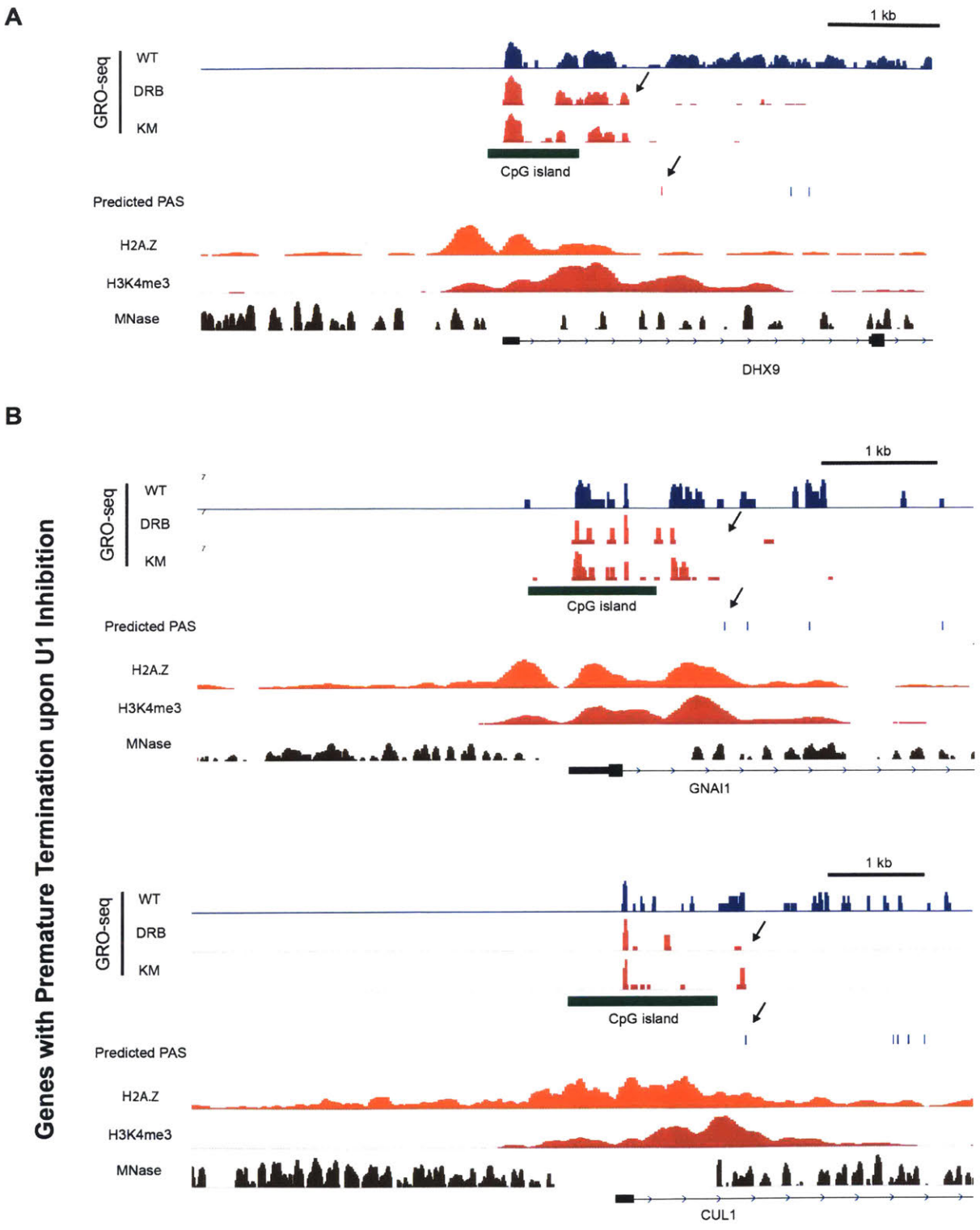

Figure 7. Premature Termination in Humans are Regulated by Flavopiridol

(A-C) IGV browser shot of GRO-seq in humans upon inhibition with DRB or KM05283. Predicted PAS is the canonical PAS hexamer, except for DHX9, where the red line reflects an AGTAAA.

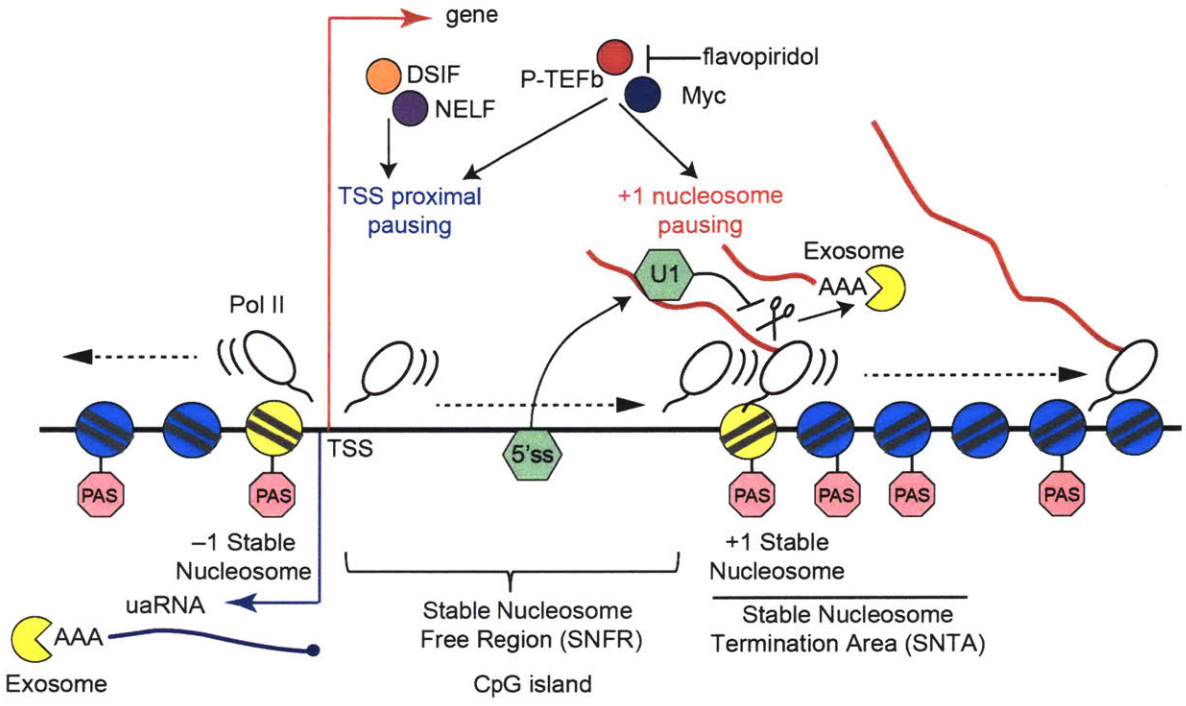

Figure 8. Model of Premature Termination.

In the sense direction, Pol II experiences a NELF/DSIF dependent TSS proximal pause. Upon pause release by P-TEFb, Pol II transcribes through the stable nucleosome free region (SNFR), which is promoted by the CpG island. It encounters a 5'SS, recruiting U1 snRNA to the RNA. The transcription machinery hits the first +1 stable nucleosome and pauses. Pause release here is regulated by MYC and CDK9 activity. If the Pol II is not processive, it will terminate at the first PAS and be degraded by the RNA exosome, which creates a stable nucleosome termination area (SNTA). In the antisense direction, Pol II encounters the -1 stable nucleosome, pauses and then terminates proximal to the promoter, inducing exosome decay.

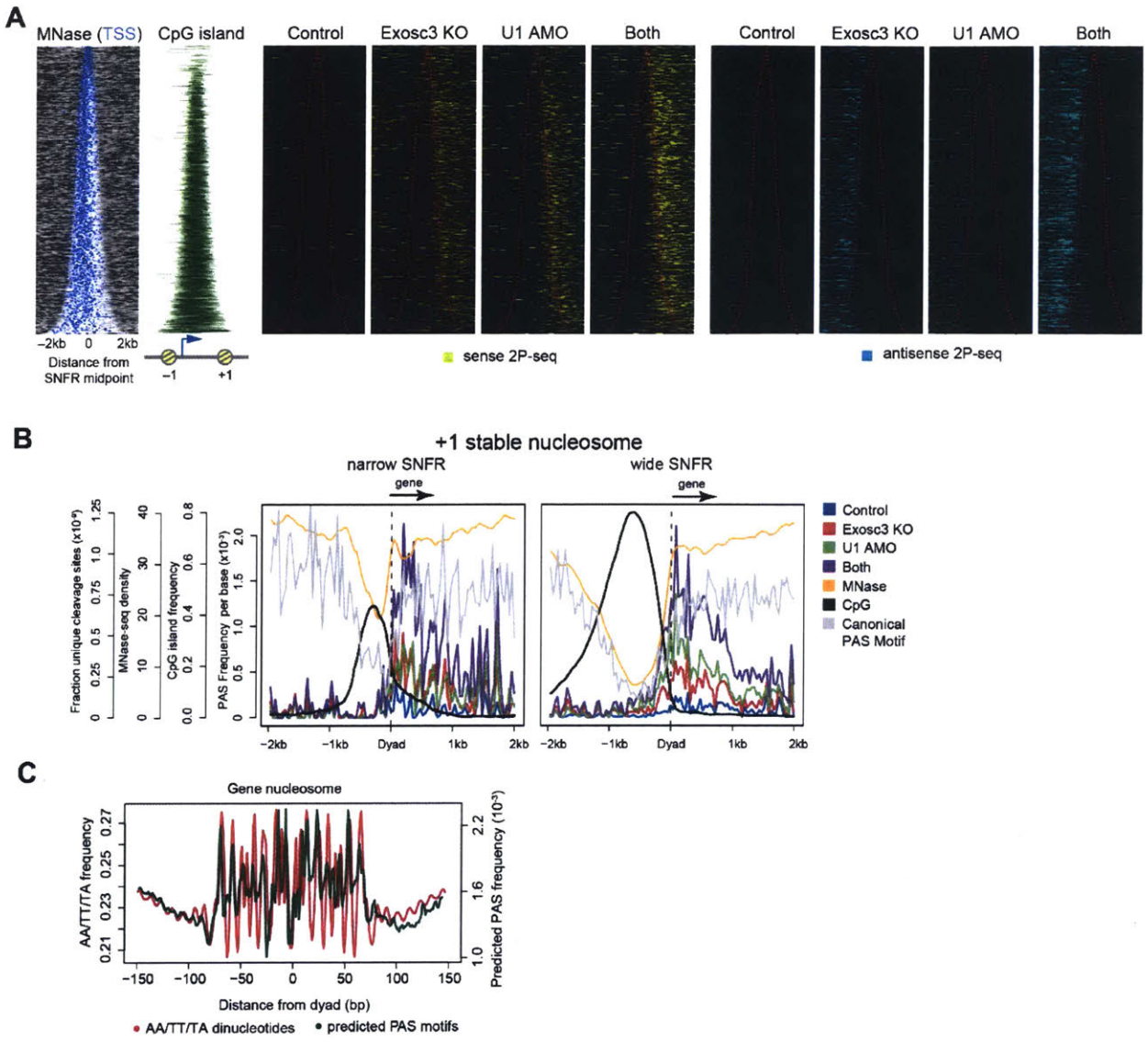

Figure S1. PAS termination and -1/+1 stable nucleosomes

(A) Heatmap of MNase-seq, CpG islands, and PAS-linked cleavage sites around a 2kb window flanking the SNFR midpoint for nonoverlapping expressed genes with 2P clusters, ranked by increasing SNFR width.
(B) Metaplots of PAS-linked cleavage sites, MNase-seq, CpG islands, and predicted canonical PAS motifs around the dyad axis of the +1 stable nucleosome of genes with narrow SNFR (<750 bp, left) and wide SNFR (>750 bp, right).
(C) AA/TT/TA dinucleotide frequency (red) and frequency of predicted canonical PAS motifs (green) per gene body nucleosome in a 150 bp window from chemical mapping-defined dyad axis. Gene body nucleosomes are between TSS and 2kb upstream from the TES of genes.

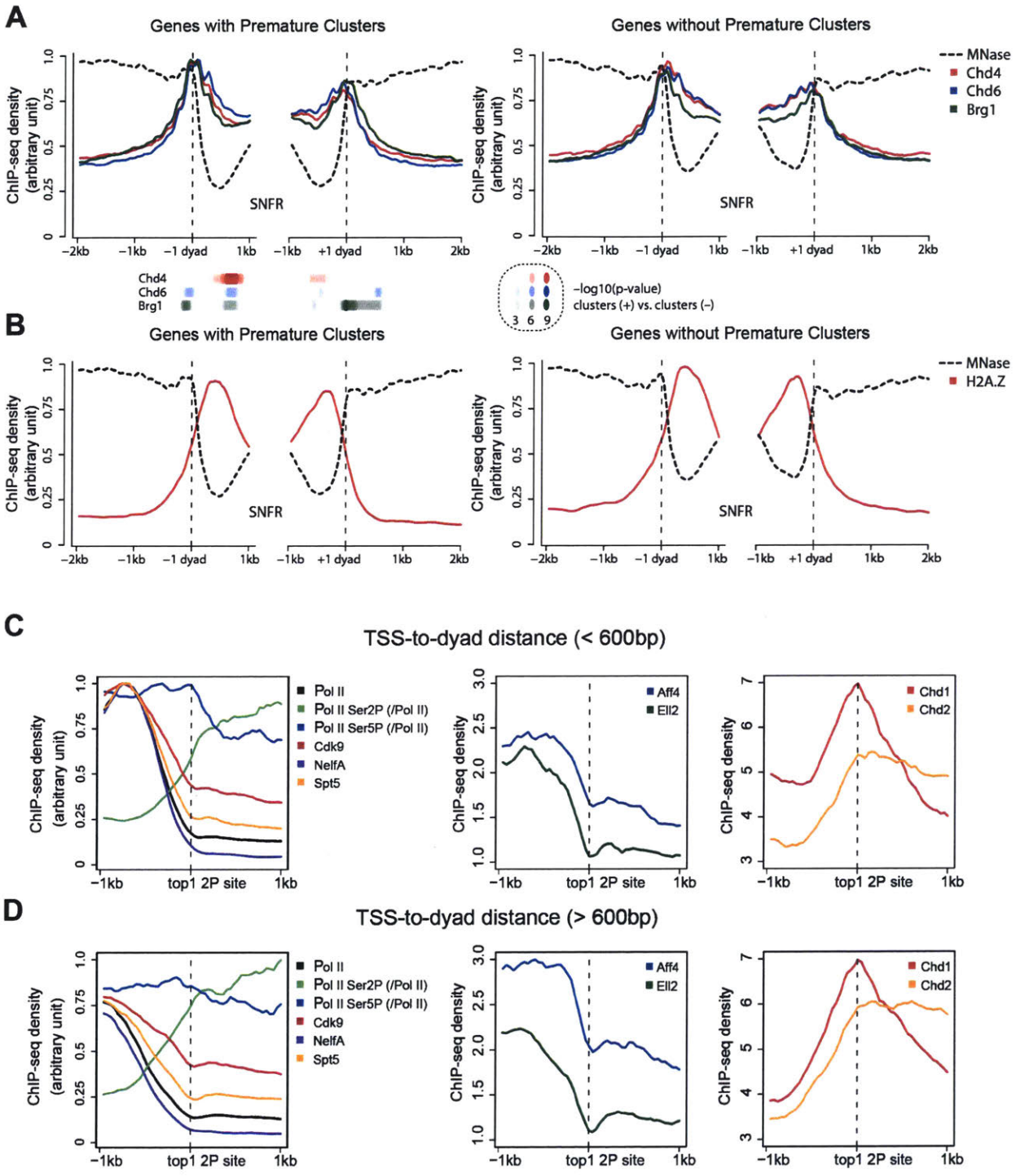

Figure S2. Distribution of additional around the -1/+1 stable nucleosomes

(A-B) Alignment of chromatin remodeling factors or H2A.Z flanking the -1/+1 stable nucleosome dyad, separated for genes with or without premature intron clusters. P values with K-S test at each bin for each factor are displayed.

(C-D) Metaplot of transcription machinery flanking the most frequently used 2P site, filtered for genes with short TSS-to-dyad distances (<600 bp) or wide TSS-to-dyad distances (>600 bp)

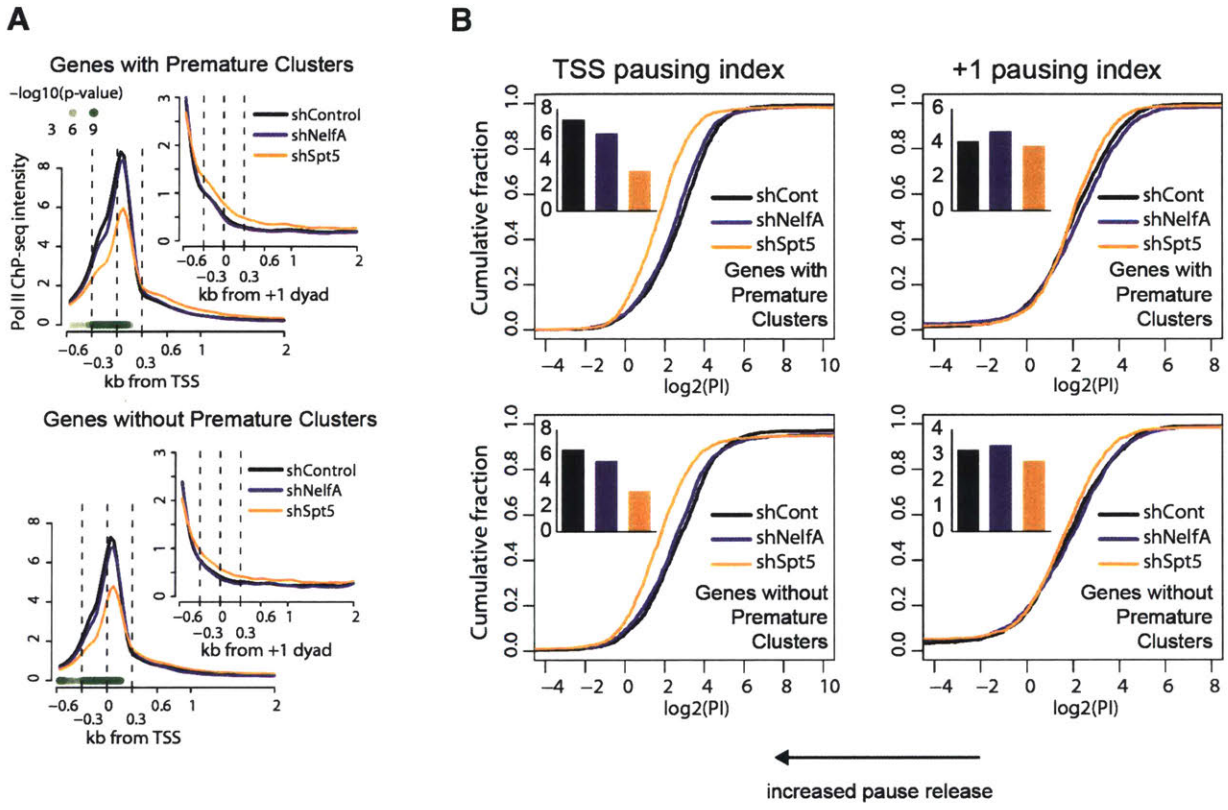

Figure S3. Impact of NELF and DSIF on +1 nucleosome pausing.

(A) Metaplots of Pol II ChIP-seq density around the TSS or +1 dyad (inset) in shControl, shSpt5, and shNelfA mESCs.
(B) Cumulative distribution plot of $\log_2$(pausing index) of either the TSS or +1 stable nucleosome pause for genes with 2P clusters (top) and genes without 2P clusters (bottom) in shControl, shSpt5, and shNelfA mESCs.

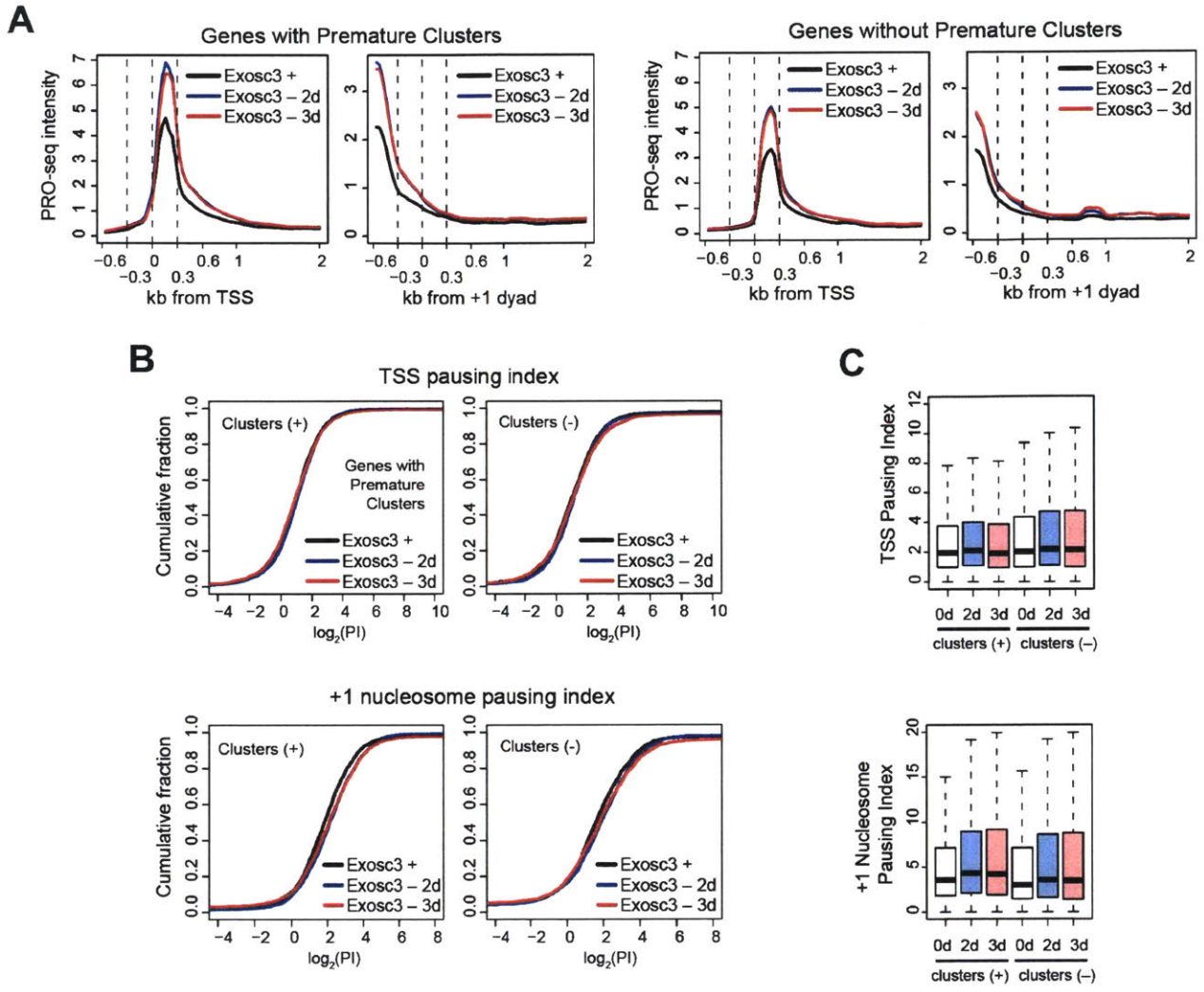

Figure S4. Genes with Premature Clusters are Sensitive to Inhibition of P-TEFb and Myc inhibition

(A) Metaplots of PRO-seq density around the TSS (left) or +1 dyad (right) for genes with 2P clusters and expression-matched genes without 2P clusters in Exosc3+ and Exosc3 – mESCs (dox off 2 and 3 days).
(B) Cumulative distribution plot of $\log_2$(pausing index) of either the TSS (top) or +1 stable nucleosome pause (bottom) for genes with 2P clusters (left) and expression-matched genes without 2P clusters (right) in Exosc3+ and Exosc3 – mESCs (dox off 2 and 3 days).
(C) Boxplots show distribution of pausing indices.

3.6 METHODS

CpG Analysis

Annotations of CpG islands were downloaded from UCSC genome browser, and genes with CpG island promoters were defined as genes where a CpG island overlaps the TSS to +100 bp of UCSC canonical genes. FPKM was calculated by taking total number of RNA-seq reads overlapping exons, dividing by gene length (kb), and normalizing to a library size of 1 million. Replicates were averaged to calculate FPKM for Exosc3+ and Exosc3- states. Expressed genes are those where FPKM>0.5 in the Exosc3+ condition. Statistical significance of overlaps with CpG island promoters was determined using the hypergeometric distribution in R.

ChIP-seq/GRO-seq Analysis and Pausing Indices

Data was downloaded from GEO database. GEO accession identifiers for the datasets used in this study are described in Table S2. Reads were aligned to the mouse genome build mm9 using bowtie as described previously (Suzuki et al., 2017).

Pausing indices were defined as follows:

$$TSS\ pausing\ index = \frac{reads\ from\ (TSS - 100)\ to\ (TSS + 300)}{reads\ from\ (TSS + 300) to\ (TSS + 600)}$$

$$+1\ stable\ nucleosome\ pausing\ index = \frac{reads\ from\ (dyad - 600)\ to\ (dyad + 0)}{reads\ from\ (dyad + 0)\ to\ (dyad + 2000)}$$

The widths of intervals used to calculate pausing indices were determined from analysis of the Pol II ChIP-seq alignments in Figures 5, taking into account the widths of a Pol II ramp and a

flavopiridol-affected region upstream of the +1 dyad. In GRO-seq analysis, normalization between datasets was done with uniquely aligned *A thaliana* spike-in RNA reads (Jonkers et al., 2014).

MNase-seq analysis

A stable nucleosome is defined as a nucleosome with low variance across multiple MNase-seq libraries. The information of precise nucleosome dyads in mESCs were downloaded from (Voong et al., 2016). To identify regions with stable nucleosomes, we analyzed five mESC MNase-seq datasets using NucTools (see Supplemental) (Vainshtein et al., 2017). We determined stable nucleosome regions using stable_nucs_replicates.pl. A sliding window of 50 bp was used and stable regions were selected based on the relative error based on five replicates < 0.5. The chemically-defined dyads that lie within NucTools-defined stable regions that were the most proximal to the TSS were defined as the +1/-1 stable nucleosome dyads and used for subsequent analysis. Wide stable nucleosome regions (SNFRs) were genes where the distance between the +1 and -1 stable nucleosome is greater than 600 bp. Narrow SNFRs are those where the distance between the +1 and -1 stable nucleosome is less than 600 bp.

Heatmap

Reads from various datasets were assigned to nonoverlapping bins in a 2 kb window flanking the SNFR for each gene filtered for a) no overlapping genes within the 2 kb window and b) containing a robust uaRNA or premature cleavage cluster. The intervals were sorted by increasing SNFR width and visualized in R.

Identification of Most Frequently Used Cluster

To identify the most frequent cleavage cluster at both the first intron and at uaRNAs, we overlapped robust cleavage clusters found in replicate 1 of each of 4 2P-seq libraries (Control, U1 AMO, Exosc3 KO, Both) to either the first intron of nonoverlapping UCSC canonical genes or to a 3 kb window upstream of the TSS for nonoverlapping UCSC canonical genes. The cluster with the most 2P-seq reads for each first intron or each "uaRNA" window was defined as the most frequently used cleavage cluster.

All predicted canonical PAS motifs (A[A/T]TAAA) were identified in the genome, and mapped across nonoverlapping UCSC canonical genes or uaRNAs. PAS site were spatially ranked by position, whereby position 1 is the one closest to the TSS. The most frequently used cleavage clusters were filtered for those associated with canonical PAS motifs (A[A/T]TAAA) and then assigned each cluster to a position based on the ranked PAS motifs.

Metaplots

Genes with premature clusters here are defined as those where FPKM>0.5 and overlaps a robust cluster within intron 1. Genes without premature clusters have been expressed matched using the R-package MatchIt. ChIP-seq, MNase-seq, PAS motif frequencies or 2P-seq reads were aligned in a 2kb window flanking the +1/-1 nucleosome dyad. 2P-seq or PAS motifs were aligned so that they are on the same strand as the sense RNA or uaRNA. Counts were normalized by library size and by number of aligned intervals to permit comparison between figures. Alignments against TSS or the +1 nucleosome dyad similarly involve aligning reads across the TSS as defined by UCSC nonoverlapping canonical genes or chemical mapping plus NucTools defined center. For

the +1 nucleosome dyad alignments, genes were also filtered so the distance between the TSS and the +1 nucleosome dyad was at least 600 bps, to spatially separate out TSS proximal pausing and +1 nucleosome pausing.

Dinucleotide Frequency Analysis

Gene body nucleosomes were defined as nucleosome that was between the dyad and TES-2kb, because it is known that there is a nucleosome free window upstream of the termination site. The number of AA/TT/TA dinucleotides was counted in a 2 base pair sliding window along a 150 bp window flanking the dyad axis and divided by the total number of gene body nucleosomes. The predicted PAS frequency was identified by searching for A[A/T]TAAA on the same strand of the gene, using a sliding 6 bp window along the 150 bp window flanking the dyad axis, divided by total number of gene body nucleosomes. The used PAS frequency was identified by counting the number of PAS motif assigned to robust clusters in a sliding 6 bp window along the 150 bp window flanking the dyad axis, divided by total number of gene body nucleosomes.

Myc DKO mESC RNA-seq Analysis

RNA-seq in c-Myc and N-Myc double knockout mESCs was previously reported (Scognamiglio et al., 2016). Sample 2 (Control, c-Myc$^{\Delta/\Delta}$ and N-Myc$^{\Delta/fl}$) and sample 6 (DKO, 96 hours) were compared. Gene ontology (GO) analyses were performed using Database for Annotation, Visualization, and Integrated Discovery (DAVID; https://david.ncifcrf.gov) and GO BP (Biological Process) terms.

Statistical Analysis

In Fig. 1C, statistical significance for Venn diagram overlaps was evaluated using the hypergeometric test ($P < 0.0001$).

In Fig. 3B (genes with 2P clusters vs. genes without 2P clusters) and 4B (genes with 2P clusters vs. genes without 2P clusters), Kolmogorov–Smirnov (K-S) test across all bins showed that each factor in these panels shows increased binding for genes with 2P clusters relative to expression-matched genes without 2P clusters.

In Fig. 4A, 4B, 5A, 5B, 5D, S3A, 6B, and S2A, P values with Kolmogorov–Smirnov test at each bin are displayed (Fig. 4A, 4B, and S2A: one-sided test for increases in genes with 2P clusters; Fig. 5A and 5B: two-sided test between genes with 2P clusters and genes without 2P clusters; Fig. 5D: one-sided test for increases upon flavopiridol treatment; Fig. S3A: one-sided test for decreases in shSpt5 relative to shControl; and Fig. 6B: one-sided test for increases upon Myc inhibitor treatment).

In Fig. 5E, statistical significance for flavopiridol-mediated pause effects in +1 nucleosome pausing index was evaluated with Kolmogorov–Smirnov test, showing $P < 0.01$ in both gene sets with/without 2P clusters. In Fig. S3B, statistical significance for pause release effects in TSS pausing index was evaluated with Kolmogorov–Smirnov test, showing $P < 0.01$ in shNelfA and shSpt5 samples in both gene sets. In Fig. 6C, statistical significance for Myc-inhibition-mediated pause effects in TSS pausing index or +1 nucleosome pausing index was evaluated with Kolmogorov–Smirnov test, showing $P < 0.01$ in both gene sets with/without 2P clusters. In addition, in Fig. 6C, TSS pausing index and +1 nucleosome pausing index of genes with 2P clusters upon Myc inhibition were significantly higher than those of genes without 2P clusters upon Myc

inhibition ($P < 0.01$ with Kolmogorov–Smirnov test), supporting that Myc preferentially regulates the +1 stable nucleosome pause at genes with premature clusters.

In Fig. 6E, statistical significance was evaluated with Kolmogorov–Smirnov test and displayed as $P < 0.01$ with asterisks.

3.7 SUPPLEMENTAL MATERIALS

DATASETS USED

Mus musculus

a) RNA-seq

Library	Cell Type	Lab	Authors	GEO Accession
Exosc3 + dox rep.1	Exosc3 CKO	Sharp	Chiu et al., 2017	*this paper*
Exosc3 + dox rep.2	Exosc3 CKO	Sharp	Chiu et al., 2017	*this paper*
Exosc3 - dox rep.1 (3 d)	Exosc3 CKO	Sharp	Chiu et al., 2017	*this paper*
Exosc3 - dox rep.2 (3 d)	Exosc3 CKO	Sharp	Chiu et al., 2017	*this paper*
Myc DKO	mESC	Trumpp	Scognamiglio et al., 2016	E-MTAB-3386

b) 2P-seq

Library	Cell Type	Lab	Authors	GEO Accession
Exosc3 + dox rep.1	Exosc3 CKO	Sharp	Chiu et al., 2017	*this paper*
Exosc3 + dox rep.2	Exosc3 CKO	Sharp	Chiu et al., 2017	*this paper*
Exosc3 - dox rep.1 (3 d)	Exosc3 CKO	Sharp	Chiu et al., 2017	*this paper*
Exosc3 - dox rep.2 (3 d)	Exosc3 CKO	Sharp	Chiu et al., 2017	*this paper*
Scr AMO + dox rep.1	Exosc3 CKO	Sharp	Chiu et al., 2017	*this paper*
Scr AMO + dox rep.2	Exosc3 CKO	Sharp	Chiu et al., 2017	*this paper*
Scr AMO - dox rep.1 (2 d)	Exosc3 CKO	Sharp	Chiu et al., 2017	*this paper*
Scr AMO - dox rep.2 (2 d)	Exosc3 CKO	Sharp	Chiu et al., 2017	*this paper*
U1 AMO + dox rep.1	Exosc3 CKO	Sharp	Chiu et al., 2017	*this paper*
U1 AMO + dox rep.2	Exosc3 CKO	Sharp	Chiu et al., 2017	*this paper*
U1 AMO - dox rep.1 (2 day)	Exosc3 CKO	Sharp	Chiu et al., 2017	*this paper*
U1 AMO - dox rep.2 (2 day)	Exosc3 CKO	Sharp	Chiu et al., 2017	*this paper*

c) PRO-seq

Library	Cell Type	Lab	Authors	GEO Accession
Exosc3 + dox rep.1	Exosc3 CKO	Sharp	Chiu et al., 2017	*this paper*
Exosc3 + dox rep.2	Exosc3 CKO	Sharp	Chiu et al., 2017	*this paper*
Exosc3 - dox rep.1 (3day)	Exosc3 CKO	Sharp	Chiu et al., 2017	*this paper*

Exosc3 - dox rep.2 (3day)	Exosc3 CKO	Sharp	Chiu et al., 2017	*this paper*

d) MNase-seq

Library	Cell Type	Lab	Authors	GEO Accession
ESC merged mononucleosome	129P2/Ola	Rippe	Teif et al., 2012	GSM1004653
ESC nucleosome 2	ESC.1	Kingston	West et al., 2014	GSM1425441
ESC nucleosome 3	ESC.2	Kingston	West et al., 2014	GSM1425442
ESC nucleosome 4	J1	Liu	Zhang et al., 2014	GSM1252095
ESC nucleosome 5	J1	Liu	Zhang et al., 2014	GSM1252095

e) ChIP-seq

Library	Cell Type	Lab	Authors	GEO Accession
Oct4	V6.5	Young	Whyte et al., 2013	GSM1082340
Sox2	V6.5	Young	Whyte et al., 2013	GSM1082341
Nanog	V6.5	Young	Whyte et al., 2013	GSM1082342
Chd1	46C (129/Ola)	Gérard	Dieuleveult et al., 2016	GSM1581288
Chd2	46C (129/Ola)	Gérard	Dieuleveult et al., 2016	GSM1581290
Chd4	46C (129/Ola)	Gérard	Dieuleveult et al., 2016	GSM1581292
Chd6	46C (129/Ola)	Gérard	Dieuleveult et al., 2016	GSM1581294
Chd8	46C (129/Ola)	Gérard	Dieuleveult et al., 2016	GSM1581296
Chd9	46C (129/Ola)	Gérard	Dieuleveult et al., 2016	GSM1581298
Ep400	46C (129/Ola)	Gérard	Dieuleveult et al., 2016	GSM1581300
Brg1	46C (129/Ola)	Gérard	Dieuleveult et al., 2016	GSM1581286
Cdk9	V6.5	Young	Whyte et al., 2013	GSM1082347
NelfA	V6.5	Young	Rahl et al., 2010	GSM515664
Spt5	V6.5	Young	Rahl et al., 2010	GSM515665
Ell2	KH2	Shilatifard	Lin et al., 2011	GSM749809
Aff4	KH2	Shilatifard	Lin et al., 2011	GSM749810
GRO-Seq	V6.5	Lis	Jonkers et al., 2014	GSM1186440
H3K4me3	V6.5	Young	Ji et al., 2015	GSM1526288
H3K27ac	V6.5	Young	Ji et al., 2015	GSM1526287
H3K36me3	V6.5	Young	Ji et al., 2015	GSM1526290
H3K79me2	V6.5	Young	Ji et al., 2015	GSM1526289

H2A.Z	V6.5	Boyer	Subramanian et al., 2013	GSM984544
Pol II	V6.5	Young	Seila et al., 2008	GSM318444
Ser2P	V6.5	Young	Rahl et al., 2010	GSM515663
Ser5P	V6.5	Young	Rahl et al., 2010	GSM515662
Pol II - DMSO 1hr	V6.5	Young	Rahl et al., 2010	GSM515670
Pol II - Flavopiridol 1hr	V6.5	Young	Rahl et al., 2010	GSM515671
Pol II - DMSO 6hr	V6.5	Young	Rahl et al., 2010	GSM515672
Pol II - Myc inhibitor 6hr	V6.5	Young	Rahl et al., 2010	GSM515673
Pol II - shControl	V6.5	Young	Rahl et al., 2010	GSM515667
Pol II - shNelfA	V6.5	Young	Rahl et al., 2010	GSM515668
Pol II - shSpt5	V6.5	Young	Rahl et al., 2010	GSM515669

Homo sapiens

f) GRO-seq

Library	Cell Type	Lab	Authors	GEO Accession
none	HeLa	Murphy	Laitem et al., 2015	*E-MTAB-3360*
KM05382	HeLa	Murphy	Laitem et al., 2015	*E-MTAB-3360*
DRB	HeLa	Murphy	Laitem et al., 2015	*E-MTAB-3360*

3.8 REFERENCES

Almada, A.E., Wu, X., Kriz, A.J., Burge, C.B., and Sharp, P.A. (2013). Promoter directionality is controlled by U1 snRNP and polyadenylation signals. Nature *499*, 360-363.

Andersen, P.K., Lykke-Andersen, S., and Jensen, T.H. (2012). Promoter-proximal polyadenylation sites reduce transcription activity. Genes Dev *26*, 2169-2179.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., *et al.* (2014). An atlas of active enhancers across human cell types and tissues. Nature *507*, 455-461.

Berg, M.G., Singh, L.N., Younis, I., Liu, Q., Pinto, A.M., Kaida, D., Zhang, Z., Cho, S., Sherrill-Mix, S., Wan, L., *et al.* (2012). U1 snRNP determines mRNA length and regulates isoform expression. Cell *150*, 53-64.

Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V.B., Wong, E., Orlov, Y.L., Zhang, W., Jiang, J., *et al.* (2008). Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell *133*, 1106-1117.

Cheng, B., and Price, D.H. (2007). Properties of RNA polymerase II elongation complexes before and after the P-TEFb-mediated transition into productive elongation. J Biol Chem *282*, 21901-21912.

Churchman, L.S., and Weissman, J.S. (2011). Nascent transcript sequencing visualizes transcription at nucleotide resolution. Nature *469*, 368-373.

Core, L.J., Waterfall, J.J., and Lis, J.T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science *322*, 1845-1848.

de Dieuleveult, M., Yen, K., Hmitou, I., Depaux, A., Boussouar, F., Bou Dargham, D., Jounier, S., Humbertclaude, H., Ribierre, F., Baulard, C., *et al.* (2016). Genome-wide nucleosome specificity and function of chromatin remodellers in ES cells. Nature *530*, 113-116.

Fenouil, R., Cauchy, P., Koch, F., Descostes, N., Cabeza, J.Z., Innocenti, C., Ferrier, P., Spicuglia, S., Gut, M., Gut, I., *et al.* (2012). CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters. Genome Res *22*, 2399-2408.

Flynn, R.A., Almada, A.E., Zamudio, J.R., and Sharp, P.A. (2011). Antisense RNA polymerase II divergent transcripts are P-TEFb dependent and substrates for the RNA exosome. Proc Natl Acad Sci U S A *108*, 10460-10465.

Jin, C., and Felsenfeld, G. (2007). Nucleosome stability mediated by histone variants H3.3 and H2A.Z. Genes Dev *21*, 1519-1529.

Jonkers, I., Kwak, H., and Lis, J.T. (2014). Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. Elife *3*, e02407.

Kaida, D., Berg, M.G., Younis, I., Kasim, M., Singh, L.N., Wan, L., and Dreyfuss, G. (2010). U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. Nature *468*, 664-668.

Kim, J.B., and Sharp, P.A. (2001). Positive transcription elongation factor B phosphorylates hSPT5 and RNA polymerase II carboxyl-terminal domain independently of cyclin-dependent kinase-activating kinase. J Biol Chem *276*, 12317-12323.

Kwak, H., Fuda, N.J., Core, L.J., and Lis, J.T. (2013). Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. Science *339*, 950-953.

Laitem, C., Zaborowska, J., Isa, N.F., Kufs, J., Dienstbier, M., and Murphy, S. (2015). CDK9 inhibitors define elongation checkpoints at both ends of RNA polymerase II-transcribed genes. Nat Struct Mol Biol *22*, 396-403.

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

Lin, C., Smith, E.R., Takahashi, H., Lai, K.C., Martin-Brown, S., Florens, L., Washburn, M.P., Conaway, J.W., Conaway, R.C., and Shilatifard, A. (2010). AFF4, a component of the ELL/P-TEFb elongation complex and a shared subunit of MLL chimeras, can link transcription elongation to leukemia. Mol Cell *37*, 429-437.

Marquardt, S., Escalante-Chong, R., Pho, N., Wang, J., Churchman, L.S., Springer, M., and Buratowski, S. (2014). A chromatin-based mechanism for limiting divergent noncoding transcription. Cell *157*, 1712-1723.

Muse, G.W., Gilchrist, D.A., Nechaev, S., Shah, R., Parker, J.S., Grissom, S.F., Zeitlinger, J., and Adelman, K. (2007). RNA polymerase is poised for activation across the genome. Nat Genet *39*, 1507-1511.

Ntini, E., Jarvelin, A.I., Bornholdt, J., Chen, Y., Boyd, M., Jorgensen, M., Andersson, R., Hoof, I., Schein, A., Andersen, P.R., *et al.* (2013). Polyadenylation site-induced decay of upstream transcripts enforces promoter directionality. Nat Struct Mol Biol *20*, 923-928.

Preker, P., Almvig, K., Christensen, M.S., Valen, E., Mapendano, C.K., Sandelin, A., and Jensen, T.H. (2011). PROMoter uPstream Transcripts share characteristics with mRNAs and are produced upstream of all three major types of mammalian promoters. Nucleic Acids Res *39*, 7179-7193.

Preker, P., Nielsen, J., Kammler, S., Lykke-Andersen, S., Christensen, M.S., Mapendano, C.K., Schierup, M.H., and Jensen, T.H. (2008). RNA exosome depletion reveals transcription upstream of active human promoters. Science *322*, 1851-1854.

Rahl, P.B., Lin, C.Y., Seila, A.C., Flynn, R.A., McCuine, S., Burge, C.B., Sharp, P.A., and Young, R.A. (2010). c-Myc regulates transcriptional pause release. Cell *141*, 432-445.

Ramirez-Carrozzi, V.R., Braas, D., Bhatt, D.M., Cheng, C.S., Hong, C., Doty, K.R., Black, J.C., Hoffmann, A., Carey, M., and Smale, S.T. (2009). A unifying model for the selective regulation of inducible transcription by CpG islands and nucleosome remodeling. Cell *138*, 114-128.

Rege, M., Subramanian, V., Zhu, C., Hsieh, T.H., Weiner, A., Friedman, N., Clauder-Munster, S., Steinmetz, L.M., Rando, O.J., Boyer, L.A., *et al.* (2015). Chromatin Dynamics and the RNA Exosome Function in Concert to Regulate Transcriptional Homeostasis. Cell Rep *13*, 1610-1622.

Saxonov, S., Berg, P., and Brutlag, D.L. (2006). A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc Natl Acad Sci U S A *103*, 1412-1417.

Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thastrom, A., Field, Y., Moore, I.K., Wang, J.P., and Widom, J. (2006). A genomic code for nucleosome positioning. Nature *442*, 772-778.

Seila, A.C., Calabrese, J.M., Levine, S.S., Yeo, G.W., Rahl, P.B., Flynn, R.A., Young, R.A., and Sharp, P.A. (2008). Divergent transcription from active promoters. Science *322*, 1849-1851.

Sims, R.J., 3rd, Millhouse, S., Chen, C.F., Lewis, B.A., Erdjument-Bromage, H., Tempst, P., Manley, J.L., and Reinberg, D. (2007). Recognition of trimethylated histone H3 lysine 4 facilitates the recruitment of transcription postinitiation factors and pre-mRNA splicing. Mol Cell *28*, 665-676.

Skene, P.J., Hernandez, A.E., Groudine, M., and Henikoff, S. (2014). The nucleosomal barrier to promoter escape by RNA polymerase II is overcome by the chromatin remodeler Chd1. Elife *3*, e02042.

Suzuki, H.I., Young, R.A., and Sharp, P.A. (2017). Super-Enhancer-Mediated RNA Processing Revealed by Integrative MicroRNA Network Analysis. Cell *168*, 1000-1014 e1015.

Thomson, J.P., Skene, P.J., Selfridge, J., Clouaire, T., Guy, J., Webb, S., Kerr, A.R., Deaton, A., Andrews, R., James, K.D., *et al.* (2010). CpG islands influence chromatin structure via the CpG-binding protein Cfp1. Nature *464*, 1082-1086.

Vainshtein, Y., Rippe, K., and Teif, V.B. (2017). NucTools: analysis of chromatin feature occupancy profiles from high-throughput sequencing data. BMC Genomics *18*, 158.

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

Voong, L.N., Xi, L., Sebeson, A.C., Xiong, B., Wang, J.P., and Wang, X. (2016). Insights into Nucleosome Organization in Mouse Embryonic Stem Cells through Chemical Mapping. Cell *167*, 1555-1570 e1515.

Wada, T., Takagi, T., Yamaguchi, Y., Ferdous, A., Imai, T., Hirose, S., Sugimoto, S., Yano, K., Hartzog, G.A., Winston, F., *et al.* (1998a). DSIF, a novel transcription elongation factor that regulates RNA polymerase II processivity, is composed of human Spt4 and Spt5 homologs. Genes Dev *12*, 343-356.

Wada, T., Takagi, T., Yamaguchi, Y., Watanabe, D., and Handa, H. (1998b). Evidence that P-TEFb alleviates the negative effect of DSIF on RNA polymerase II-dependent transcription in vitro. EMBO J *17*, 7395-7403.

Weber, C.M., Ramachandran, S., and Henikoff, S. (2014). Nucleosomes are context-specific, H2A.Z-modulated barriers to RNA polymerase. Mol Cell *53*, 819-830.

Yamaguchi, Y., Takagi, T., Wada, T., Yano, K., Furuya, A., Sugimoto, S., Hasegawa, J., and Handa, H. (1999). NELF, a multisubunit complex containing RD, cooperates with DSIF to repress RNA polymerase II elongation. Cell *97*, 41-51.

Zeitlinger, J., Stark, A., Kellis, M., Hong, J.W., Nechaev, S., Adelman, K., Levine, M., and Young, R.A. (2007). RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo. Nat Genet *39*, 1512-1516.

Chapter 4

Pabpn1 Suppresses Early PAS Termination Transcripts

This chapter is based on preliminary, unpublished work performed in parallel as the Exosc3 experiments.

Contributions:

AC designed the experiments and performed most of the analyses. AC generated the RNA-seq library and AC and XW generated the Pabpn1 2P-seq library. HS wrote some of the scripts used in these figures.

4.1 ABSTRACT

Promoters of most active mammalian genes produce divergent transcription, but productive elongation is limited to the sense direction. The instability of upstream antisense RNAs (uaRNAs) is regulated by promoter proximal PAS termination, which is influenced by the frequency of PAS motifs and U1 snRNP binding sites. The nuclear poly(A) binding protein Pabpn1 has been reported to repress expression of polyadenylated noncoding RNAs. Here, we show that Pabpn1 knockout upregulates a subset of uaRNAs, but more modestly compared to Exosc3 knockout. Pabpn1 regulated substrates are polyadenylated substrates that terminate at PAS motifs. Sensitivity to Pabpn1 removal is inversely proportional to the distance of the termination site from the TSS. Interestingly, Pabpn1-regulated termination for uaRNAs occurs at the edge of stable nucleosome free regions. Collectively, these results suggest that Pabpn1 functions in the U1-PAS axis along with the RNA exosome.

4.2 INTRODUCTION

Divergent transcription is a common feature of mammalian genomes, whereby 70% of promoters transcribe divergently (Andersson et al., 2014; Core et al., 2008; Preker et al., 2008; Seila et al., 2008). However, the upstream antisense RNA (uaRNA) is frequently found at lower steady state levels compared to the corresponding mRNA, due to active decay by the RNA exosome (Flynn et al., 2011; Preker et al., 2011; Preker et al., 2008). While initiation and promoter proximal pausing occur in the antisense direction (Flynn et al., 2011; Preker et al., 2011), uaRNAs prematurely terminate before marks of productive elongation are deposited (Core et al., 2008; Preker et al., 2008; Seila et al., 2008) suggesting premature termination may be the trigger for transcript instability.

Supporting this hypothesis, early termination through a U1-PAS axis drives differential expression between sense and antisense transcription (Almada et al., 2013; Ntini et al., 2013). Most initiation events are unproductive due to a high frequency of polyadenylation signal (PAS) motifs, resulting in early termination and exosome decay (Andersen et al., 2012(Andersen et al., 2012). In contrast, a selective depletion of PAS motifs in the sense direction results in later termination, evading decay. Moreover, enrichment for U1 snRNP binding sites near the transcription start site (TSS) of the sense transcript further suppresses early termination, resulting in the production of stable transcripts (Kaida et al., 2010). Components of the cleavage and polyadenylation machinery bind to uaRNAs as well as to promoter proximal regions in the sense transcript (Nojima et al., 2015), suggesting this pathway uses the same termination machinery involved in 3' end processing of mature RNAs. In addition, termination at early PAS motifs occurs at the edge of stable nucleosome free regions (SNFRs) created by CpG islands, and is linked with genes that have greater pausing at the +1 stable nucleosome.

Since polyadenylated transcripts are usually linked with production of stable cytoplasmic mRNAs, it is unclear how polyadenylated transcripts are being degraded. The presence of a poly(A) tail suggests that a poly(A) binding protein may be a critical player. In *S pombe*, polyadenylated meiotic RNAs are normally suppressed through the activity of the RNA binding protein Mmi1 and the nuclear poly(A) binding protein, Pab2 (Harigaya et al., 2006; Yamanaka et al., 2010). Similarly, knockdown of the human homolog of Pab2, PABPN1, promotes stabilization of a subset of polyadenylated lncRNAs, including NEAT1, snoRNA host genes (SNHG) and a subset of divergent lincRNAs (Beaulieu et al., 2012), in a process that also involves the RNA exosome. In addition, PABPN1 promotes degradation of a polyadenylated viral transcript that deleted for a triple helix termination signal (PANΔENE) as well as improperly spliced transcripts (Bresson and Conrad, 2013; Bresson et al., 2015). This activity is associated with hyperadenylation and binding of Pabpn1 to the poly(A) tail.

Here, we report that Pabpn1 is involved in degrading a subset of polyadenylated uaRNAs. Through the generation of a conditional knockout Pabpn1 cell line, we found Pabpn1 responsive uaRNAs are enriched for polyadenylation at PAS motifs. Most Pabpn1 responsive uaRNAs are substrates of the RNA exosome, but the RNA exosome also degrades an extended uaRNA transcript. Sensitivity to Pabpn1 is linked with the distance between the termination site and the TSS, suggesting this may be the mechanism that degrades promoter proximal polyadenylated transcripts. Interestingly, Pabpn1 responsive uaRNAs often occur at the edge of stable nucleosome free regions, similar to Exosc3 responsive uaRNAs. Altogether, this suggests that there are at least two modes of degrading uaRNAs, both involving the activity of the RNA exosome, whereby Pabpn1 plays an important role in degrading polyadenylated uaRNAs.

4.3 RESULTS

Generation of PABPN1 Knockout Cell Lines

A previous report suggested that Pabpn1 may regulate divergent lncRNAs, but only detected 20 significantly changing divergent lncRNAs (Beaulieu et al., 2012). We ascribe the small amount to their use of siRNAs to knock down a protein critical in promoting 3' end processing. Consequently, we created a knockout of Pabpn1, using a similar strategy as our Exosc3 CKO cell line (**Fig. 1A**). The cDNA for Pabpn1 was subcloned into a dox-inducible piggyBac transposable vector and inserted into the genome, creating FH-Pabpn1. Subsequently, the first exon of endogenous Pabpn1 was deleted using CRISPR-Cas9 technology, and validated by both Western blotting after 3 days of doxycycline removal (**Fig. 1B**) and sequencing the PCR product across the deletion (data not shown). One of the clones with the best knockdown by qRT-PCR (**Fig. 1C**) was selected to be analyzed, and will henceforth be called Pabpn1 CKO. In addition, one representative uaRNA (uaP4hb) was found to be upregulated upon doxycycline removal. Unlike Exosc3 CKO, doxycycline withdrawal did not result in cell death during the time frame of these experiments despite the protein being gone. Removing doxycycline for a longer period eventually resulted in cell death.

Pabpn1 Stabilizes a Subset of uaRNAs Genome-wide

RNA-seq was performed on ribosomal RNA-depleted RNAs from Pabpn1 CKO after 3 days of dox withdrawal, and confirmed that Pabpn1 RNA was lost (**Fig. 1D**). In addition, known substrates of Pabpn1, snoRNA host genes (Beaulieu et al., 2012; Lemay et al., 2010), were upregulated upon Pabpn1 withdrawal (**Fig. S1**), confirming functional knockdown of Pabpn1. Alignments of RNA-seq reads around the TSS demonstrate that uaRNAs were stabilized upon

Pabpn1 loss, supporting a role for Pabpn1 at destabilizing uaRNAs (**Fig. 2A**). In contrast, minor changes were detected around enhancer peaks (**Fig. 2B**). These changes were quantified more precisely using previously identified intervals, revealing that uaRNAs were stabilized 2-fold in the absence of Pabpn1 compared to and 8-fold stabilization upon loss of Exosc3 (**Fig. 2C,S2A**). This pattern was shared across other ncRNAs, where Pabpn1 knockout had a weaker phenotype than Exosc3 knockout.

Next, we identified high-confidence, differentially expressed intervals (fold-change>2, FDR<0.1) from both Exosc3 CKO and Pabpn1 CKO cell lines (**Fig. 2D-E**). Most expressed uaRNAs were upregulated upon Exosc3 loss, whereas a third of uaRNAs (655 of 2032 expressed uaRNAs) were Pabpn1 substrates. Importantly, most of the transcripts significantly changing upon Pabpn1 loss were also targeted for degradation by the RNA exosome (**Fig. 2F,S2B**), suggesting that Pabpn1 and Exosc3 cooperate in a similar pathway to degrade uaRNAs.

Pabpn1 substrates terminate close to the TSS

To determine what distinguishes uaRNAs that were more Pabpn1 sensitive from those that were less, more abundant uaRNAs were split into the top 250 differentially expressed uaRNAs (Pabpn1-responsive), and 250 expression matched controls that changed the least (Pabpn1-nonresponsive) (**Fig. S3A**). There is a modest but significant decrease in PAS density for Pabpn1-responsive uaRNAs compared to Pabpn1-nonresponsive uaRNAs ($p<0.003$, KS test) (**Fig. S3B**), largely because Pabpn1 responsive uaRNAs were more enriched for PAS motifs proximal to the TSS than Pabpn1-nonrespive uaRNAs. (**Fig. 3A**). This suggests encountering early PAS motifs may induce sensitivity to Pabpn1-linked decay. In contrast, we did not detect any difference in U1 motif frequency between the two.

To determine whether these PAS motifs were being used, the 3' ends of poly(A) RNAs were sequenced from the Pabpn1 CKO cell line using poly(A)-primed sequencing (2P-seq) (Spies et al., 2013). This technique identifies precise cleavage site by sequencing from the poly(A) tail, which is computationally processed for upstream PAS motifs (Almada et al., 2013). Pabpn1-responsive uaRNAs had a significantly higher amounts of PAS-terminated 3' ends in the wild-type state (**Fig. 3B**). Furthermore, loss of Pabpn1 results in even more detectable PAS-terminated poly(A) ends, suggesting Pabpn1 specifically destabilizes polyadenylated uaRNAs.

A metagene analysis of unique cleavage sites across defined uaRNA demonstrates that the majority of novel unique PAS cleavage sites upon Pabpn1 removal occur at the 5' end of the uaRNA (**Fig. 3C**), suggesting that Pabpn1 predominantly affects promoter proximal PAS termination events. In contrast, loss of the RNA exosome for 2 days results in a small bias towards increased unique PAS cleavage sites at the 5' end of the uaRNA, but also targets sites further into the gene body (**Fig. 3D**). This suggests that Exosc3 and Pabpn1 may be working together in a similar pathway to regulate a subset of termination sites. This was confirmed by hierarchical clustering of relative expression for cleavage clusters within uaRNAs, where a substantial fraction of termination sites that increase upon Pabpn1 loss were also regulated by Exosc3 (**Fig. 3E**).

Loss of the RNA exosome for 3 days resulted in a general increase in PAS termination across the entire uaRNA region (**Fig. S3C**), likely due to two non-mutually exclusive reasons. First, there is less Exosc3 protein at 3 days, resulting in a stronger phenotype. Additionally, the increase may reflect the stressed state of the cells, because deletion of Exosc3 stimulates activation of p53 and apoptosis (Pefanis et al., 2015). Loss of Pabpn1 correlates better with Exosc3 CKO after 2 days off dox, than Exosc3 CKO 3 days off dox (**Fig. S3D**). As a control, cleavage clusters upon Exosc3 loss for 2 and 3 days correlate relatively well (**Fig. S3E**).

To directly observe whether proximity to the TSS is critical for Pabpn1 sensitivity, we examined all PAS-containing cleavage clusters within uaRNAs. PAS cleavage clusters that terminate closer to the TSS of the uaRNA were stabilized to a greater degree upon Pabpn1 knockout (Pearson correlation=-0.32) (**Fig. 3F**). Exosc3 knockout for 2 days also had a negative slope, but the slope was less steep (**Fig. 3G**), likely because Exosc3 can target RNAs that terminate both close to the TSS as well as those further in the gene body (**Fig. 3D**). In conjunction with observations that Pabpn1-responsive uaRNAs encounter early PAS sites (**Fig. 3A**), we conclude that Pabpn1 downregulates transcripts that terminate at PAS motifs proximal to the TSS, in conjunction with the RNA exosome.

Pabpn1 substrates terminate close to the edge of stable nucleosome free regions

Since Pabpn1 removal appeared to elicit slightly different phenotypes than Exosc3 removal (**Fig. 3C-D**), we looked at 4 model uaRNAs. Loss of Pabpn1 stabilized transcripts that terminate precisely at a cleavage site (**Fig. 4**), arguing that Pabpn1 specifically promotes degradation of a polyadenylated transcript. To our surprise, loss of Exosc3 results in the stabilization of an extended transcript. Exosc3 is a 3'-to-5' exonuclease that is not known to stabilize the 3' cleavage product at mRNA ends; rather the 3' cleavage product is normally degraded by a 5'-to-3' exonuclease (Kim et al., 2004; West et al., 2004). Hence, the RNA exosome degrades multiple transcript types within these uaRNAs: one terminating early at a PAS motif and one (or more) that terminates later. Termination also occurs at the first canonical PAS motif (**Fig. 4**), after the CpG island beyond a region of H2A.Z enrichment and beyond a region of MNase depletion, suggesting Pabpn1 may be associated with the previously described elongation checkpoint.

The U1-PAS axis regulates termination in both the sense and antisense direction (Almada et al., 2013; Ntini et al., 2013). Alignments of unique cleavage sites around the TSS demonstrate that termination in the sense direction is also regulated by Pabpn1 (**Fig. S4A**). Additionally, loss of Pabpn1 results in an sharp increase in TSS-proximal termination signal within 1 kb from the TSS, whereby Exosc3 knockout upregulated both TSS-proximal as well as termination events more distal to the TSS (Chapter 2). Exosome-regulated PAS termination was previously suggested to occur primarily at the edge of stable nucleosome free regions (SNFRs); similarly, Pabpn1 knockdown increases PAS termination signals at the edge of the SNFR, perhaps by promoting the degradation of uaRNAs in collaboration with the RNA exosome (**Fig. 5A**). This can be seen more precisely with alignments at the -1 nucleosome demonstrating that Pabpn1-regulated termination events peak at the first nucleosome dyad (**Fig. 5B**). Interestingly, this pattern is similar to the stabilization upon loss of Exosc3, though the RNA exosome also degrades longer transcripts.

Termination signals at the edge of the SNFR also increases in the sense direction upon Pabpn1 depletion (**Fig. 5A**). In the absence of Pabpn1, additional PAS cleavage sites are used near the +1 stable nucleosome, mimicking the Exosc3 depletion (**Fig. 5C**), suggesting that Pabpn1 regulates degradation of premature transcripts in the sense direction. However, the effect on uaRNA is not very substantial; alignment of intronic RNA-seq reads around the first 5' splice site observes a very modest stabilization (25%) after Pabpn1 removal (**Fig. S4B**), in contrast to a 2-fold increase after Exosc3 removal (Chapter 2). We suspect this smaller increase is because a subset of Exosc3-regulated premature clusters in the first intron are Pabpn1 targets, similar to what is observed at uaRNAs (**Fig. S4C**). For example, Nudt2 and Pcf11 both have a Pabpn1-sensitive premature cluster (**Fig. 5D, S4**), whereas some of the other previously identified premature events such as Rad23b appeared not to be sensitive to Pabpn1 (data not shown). Pabpn1 sensitivity in the

first intron is likely based on a combination of the strength of the 5' splice site and proximity of the termination site to the TSS of gene (approximately less than 1.5 kb) (**Fig. S4D**). Additionally, Exosc3 loss removes the degradation enzyme whereas Pabpn1 presumably removes a protein that recruits the degradation enzyme; hence we would expect the former to have the most stabilization. Further work needs to be done to clarify what is happening in the sense direction.

4.4 DISCUSSION

Our analysis into the degradation of polyadenylated uaRNAs reveals that Pabpn1 also regulates divergent transcription. A subset of uaRNAs were upregulated upon knockout of Pabpn1, but the effect is smaller than knockout of the RNA exosome. These Pabpn1-responsive uaRNAs are polyadenylated and preferentially terminate closer to the TSS. Pabpn1-responsive uaRNAs also terminate at the edge of stable nucleosome free regions, suggesting that Pabpn1 and the RNA exosome may collaborate to regulate the U1-PAS axis.

Previous studies suggest that Pabpn1 collaborates with the RNA exosome to degrade various classes of RNAs, including divergent lincRNAs. (Beaulieu et al., 2012), polyadenylated viral transcripts (Bresson and Conrad, 2013), unspliced transcripts (Bresson and Conrad, 2013; Bresson et al., 2015), specific mRNAs (Bergeron et al., 2015; Lee et al., 2013; Yamanaka et al., 2010) and snoRNA host genes (Bresson et al., 2015). We now describe more generally that Pabpn1 targets a subset of uaRNAs, similar to other recent studies (Bresson et al., 2015; Meola et al., 2016). This pathway require polyadenylation of the transcript, as the activity of the canonical poly(A) polymerases is essential (Bresson and Conrad, 2013; Bresson et al., 2015), likely because the poly(A) tail creates a platform for Pabpn1 to bind. Previous studies showed that promoter-

proximal PAS termination at model genes promotes degradation of transcripts through the RNA exosome (Andersen et al., 2012). Here, we find promoter-proximal termination stimulates transcript degradation by Pabpn1, where the amount of suppression is dependent on the length of the transcript (approximately less than 1.5 kb). The dependence on the gene length may be due to a series of proteins that physically link the RNA exosome to the 5' cap of RNAs (Meola et al., 2016). The RNA helicase hMTR4 associates with the CBCA complex that binds to the 5'-methyl cap of RNAs (Andersen et al., 2013; Hallais et al., 2013) and with the RNA exosome though ZFC3H1 (Meola et al., 2016). PABPN1 can associate with the ZFC3H1 subunit in this PAXT connection. Hence, as RNAs are transcribed and polyadenylated, PABPN1 binds to the poly(A) tail. If the transcript is short, PABPN1 can bind to ZFC3H1 and stabilize an association that brings its binding partner, the poly(A) tail, in close proximity to the ZFC3H1-associated RNA exosome (**Fig. 6**). In contrast, when transcripts are long, PABPN1 can bind the poly(A) tail, but finds it much harder to interact with the 5' end of transcripts, so RNAs are not degraded. PABPN1 can still encounter the RNA exosome at low frequency, so if a transcript is not exported from the nucleus due to incomplete splicing, PABPN1 eventually may encounter the RNA exosome and target the unspliced RNA for decay.

This polyadenylation pathway is highly conserved all the way to yeast. Polyadenylated meiotic transcripts are selectively degraded when *S pombe* switches from meiotic growth to vegetative growth, due to the binding of Mmi1 which recruits a complex containing Pab2 to promote exosome-mediated decay (Harigaya et al., 2006; Lee et al., 2013; Sugiyama and Sugioka-Sugiyama, 2011; Yamanaka et al., 2010). Homologs of most components of the yeast pathway are found in the mammalian PABPN1 pathway to degrade polyadenylated transcripts (Andersen et al., 2013; Hallais et al., 2013; Lubas et al., 2011; Meola et al., 2016). Interestingly, untemplated

adenosines are a critical component of degrading bacterial transcripts and triggers degradation by exosome homologs in bacteria and Achaea (Houseley et al., 2006). We also observed significant levels of oligoadenylated degradation intermediates from the murine mitochondrial transcriptome. Hence, poly(A) tails are the ancestral marker to degrade transcripts, and eukaryotes chose to retain this degradation machinery while also developing a method to bypass it. Most likely, bypass of this pathway coincides with the evolution of the nucleus, as mutations in export pathways or processes that deposit export proteins such as splicing result in increased association of transcripts in the nucleus, hyperadenylation and exosome-mediated decay (Bresson and Conrad, 2013; Bresson et al., 2015; Hilleren et al., 2001).

Curiously, we found the RNA exosome degrades not only the poly(A) transcript but also ones that terminate beyond the PAS signal. A recent study suggested that most uaRNAs and enhancer RNAs do not have poly(A) tails and are recruited to the RNA exosome through the NEXT complex (Andersson et al., 2014; Lubas et al., 2015; Lubas et al., 2011; Meola et al., 2016). Similar to PABPN1, the NEXT complex is held near the promoter through a physical link to the 5' methyl cap. Hence, cells uses numerous non-redundant pathways to ensure unwanted RNAs are suppressed, potentially because noncoding RNAs function non-specifically in various biological pathways including pause release (Schaukowitch et al., 2014), polycomb repression (Zhao et al., 2010) or transcription factor recruitment (Sigova et al., 2015). Since the RNA exosome requires an accessible 3' end, these non-polyadenylated transcripts may originate from separate cleavage pathways. One pathway that creates free 3' ends is the Integrator complex, a cleavage machinery that is an orthologs of the CPSF machinery that performs PAS cleavage (Baillat et al., 2005). Recently, Integrator was shown to promote termination of non-polyadenylated enhancer RNAs (Lai et al., 2015). Moreover, knockdown of Integrator subunits resulted in a stabilization of

uaRNAs in humans (Stadelmayer et al., 2014). It is unclear how Integrator might target uaRNAs, but it binds to the Pol II CTD (Baillat et al., 2005), suggesting it may be co-transcriptionally recruited.

More generally, CpG islands are frequently found in actively transcribed genes in mammals. As genes age, evolution has selected for CpG islands and promoter proximal 5' splice sites, and selected against promoter proximal PAS motifs (Almada et al., 2013). CpG island genes tend to be higher expressed and encode housekeeping genes required for maintenance of cell function (Ramirez-Carrozzi et al., 2009; Saxonov et al., 2006). Hence, our work suggests the selection for CpG islands may originate from a need to evade the ancestral Pabpn1 pathway that degrades prematurely-terminated polyadenylated transcripts. In contrast, the majority of Pabpn1-sensitive termination events occur close to the TSS of uaRNAs, suggesting they have not yet evolved to evade this pathway. It is unknown whether uaRNAs as a class have broad functions, but uaRNAs are nuclear, low-abundant and noncoding, arguing that most likely they are nonfunctional (Preker et al., 2011; Preker et al., 2008). The majority of lncRNAs originate from divergent transcription (Sigova et al., 2013), are higher expressed than uaRNAs and have intermediate depletion for PAS motifs/enrichment for 5' splicing signals (Almada et al., 2013), suggesting they are evolutionary intermediates on the way to becoming a stable gene.

4.5 FIGURES

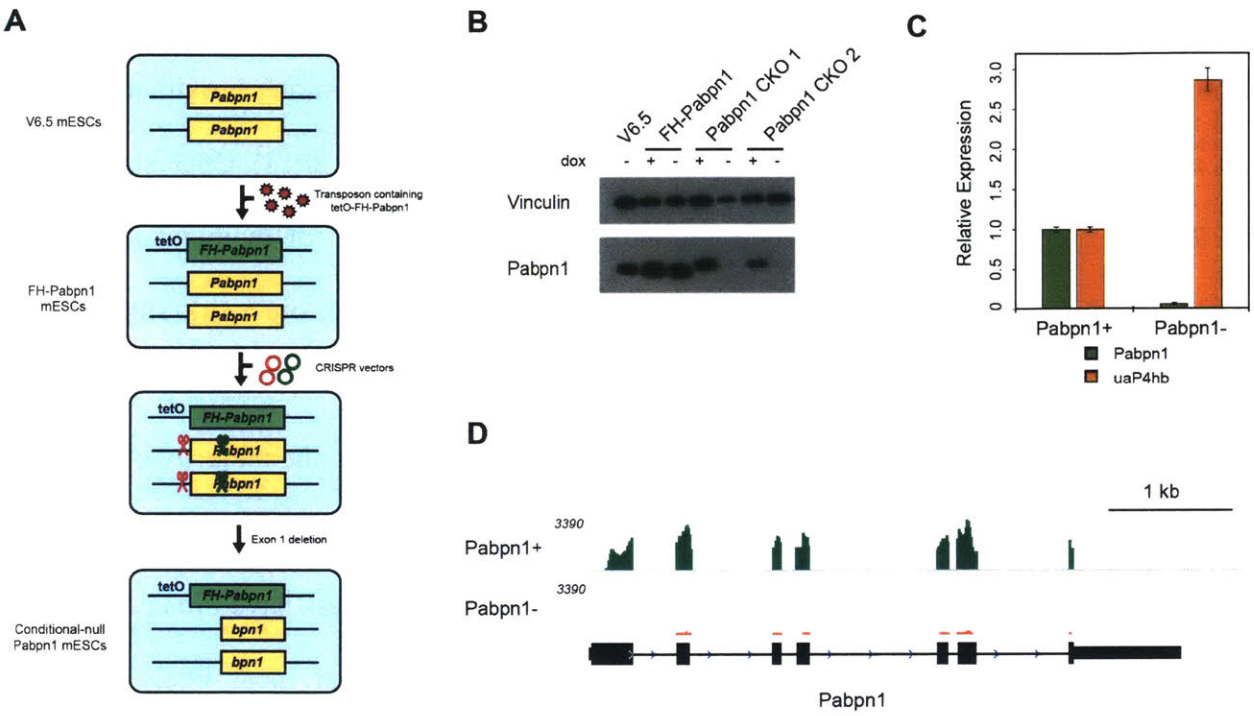

Figure 1. Generation of Pabpn1 CKO

(A) Strategy to knockout exon 1 of PAbpn1.
(B) Western blotting for Pabpn1, HA and Vinculin in the presence or absence of doxycyline.
(C) qRT-PCR for Pabpn1 and uaP4hb.
(D) Genome browser shot demonstrating knockout of Pabpn1 in the RNA-seq.

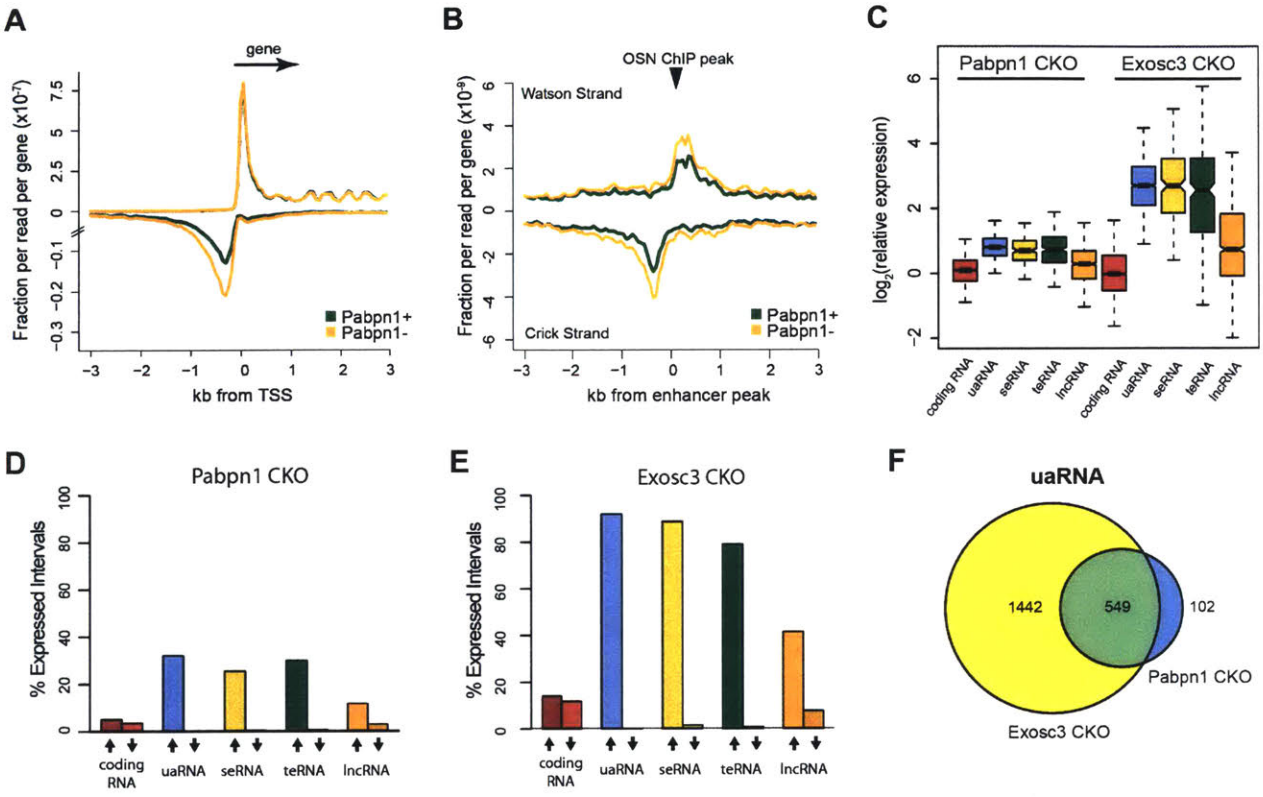

Figure 2. Transcription changes upon Pabpn1 KO.

(A) Alignment of RNA-seq reads around the TSS of UCSC canonical genes, filtered for nonoverlapping genes in Pabpn1+ or Pabpn1- conditions.
(B) Alignment of RNA-seq reads around the Oct4/Sox2/Nanog ChIP peaks, filtered for nonoverlapping enhancers in Pabpn1+ or Pabpn1- conditions.
(C) Boxplot of fold-change upon loss of Pabpn1 or Exosc3.
(D-E) Percent significantly changing intervals (fold-change>2, FDR<0.1), upregulated or downregulated, upon loss of Pabpn1 or Exosc3.
(F) Pie chart of statistically-significantly changing uaRNAs by RNA-seq under Exosc3 or Pabpn1 loss.

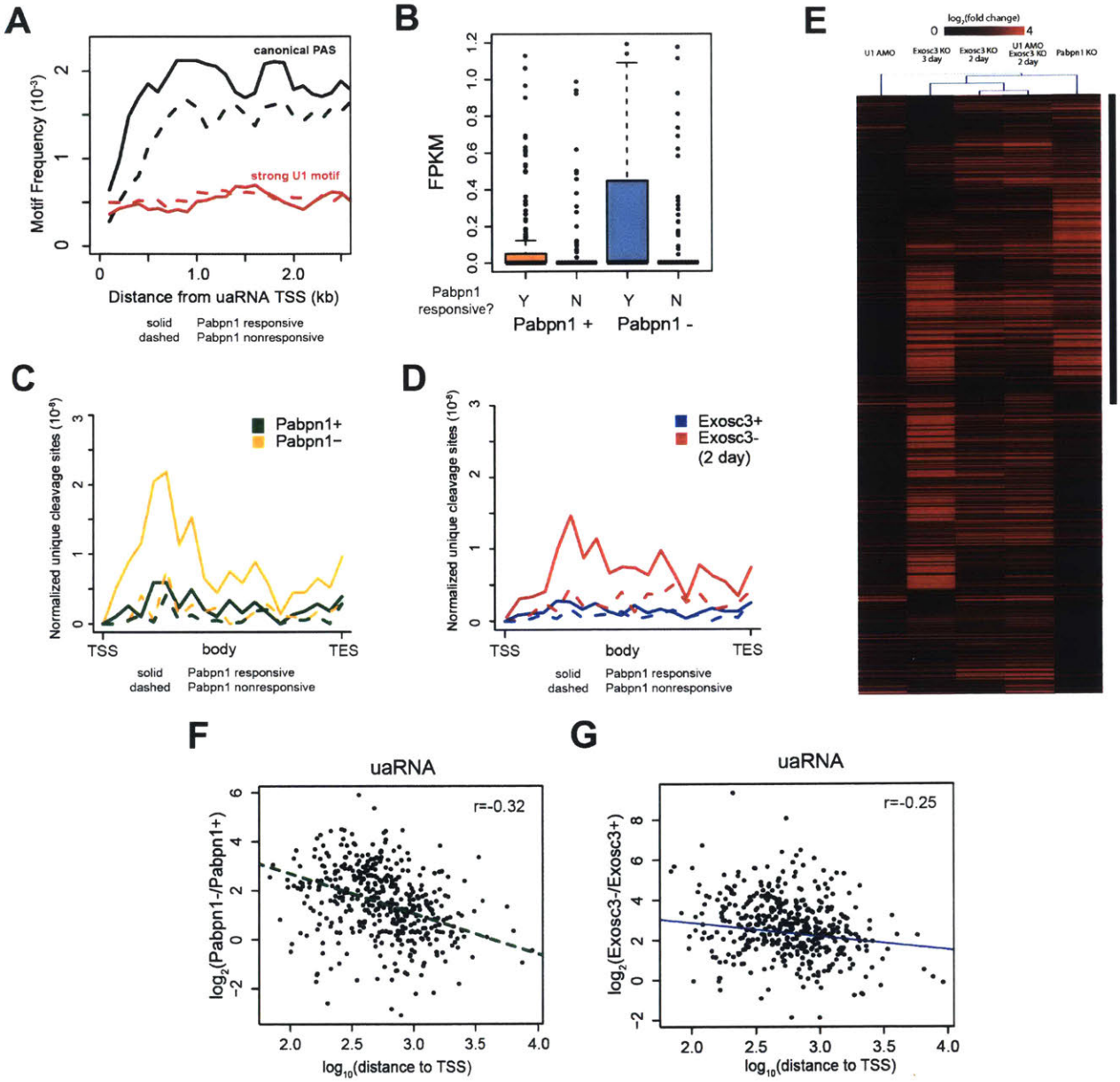

Figure 3. PAS Termination near the promoter induces Pabpn1 sensitivity.

(A) Frequency of canonical PAS motifs (AATAAA/ATTAAA) or strong U1 splice sites across the first 2.5 kb from the TSS of defined uaRNAs.
(B) Boxplot of density of total 2P reads across defined uaRNA intervals for matched Pabpn1 responsive or Pabpn1 nonresponsive uaRNAs, under Pabpn1+ or Pabpn1- conditions.
(C-D) Metagene profile of unique cleavage sites across scaled uaRNAs for Pabpn1 or 2 days off Exosc3.
(E) Hierarchical clustering of $\log_2$(fold change) for cleavage sites at uaRNAs upon loss of Exosc3 (2 or 3 day), U1 inhibition, or loss of Pabpn1. Black bar represents major cluster of upregulated Pabpn1.
(F-G) Scatterplot of $\log_2$(fold change) upon loss of Pabpn1 or Exosc3 (2 days) for cleavage clusters in a 3kb window upstream of the uaRNA TSS compared to $\log_{10}$(distance from uaRNA TSS). Pearson correlation.

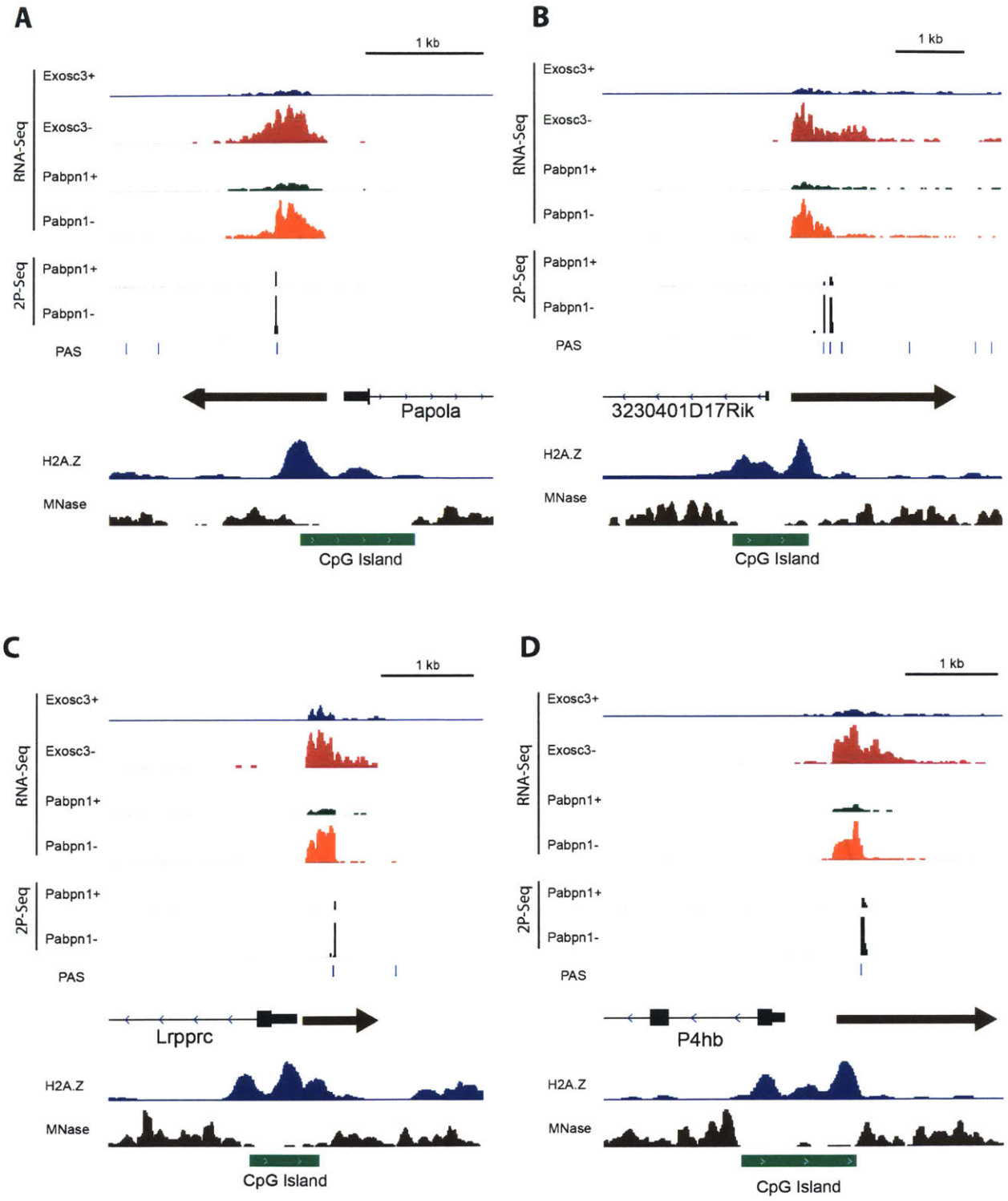

Figure 4. Pabpn1 specifically targets polyadenylated transcripts.

(A-D) Genome-browser shots of RNA-seq from Exosc3 CKO, Pabpn1 CKO and 2P-seq from Pabpn1 CKO, scaled for each direct comparison. Other tracks are predicted canonical PAS motifs, previously defined uaRNA, ChIP-seq of H2A.Z, MNase-seq and CpG islands from UCSC genome browser.

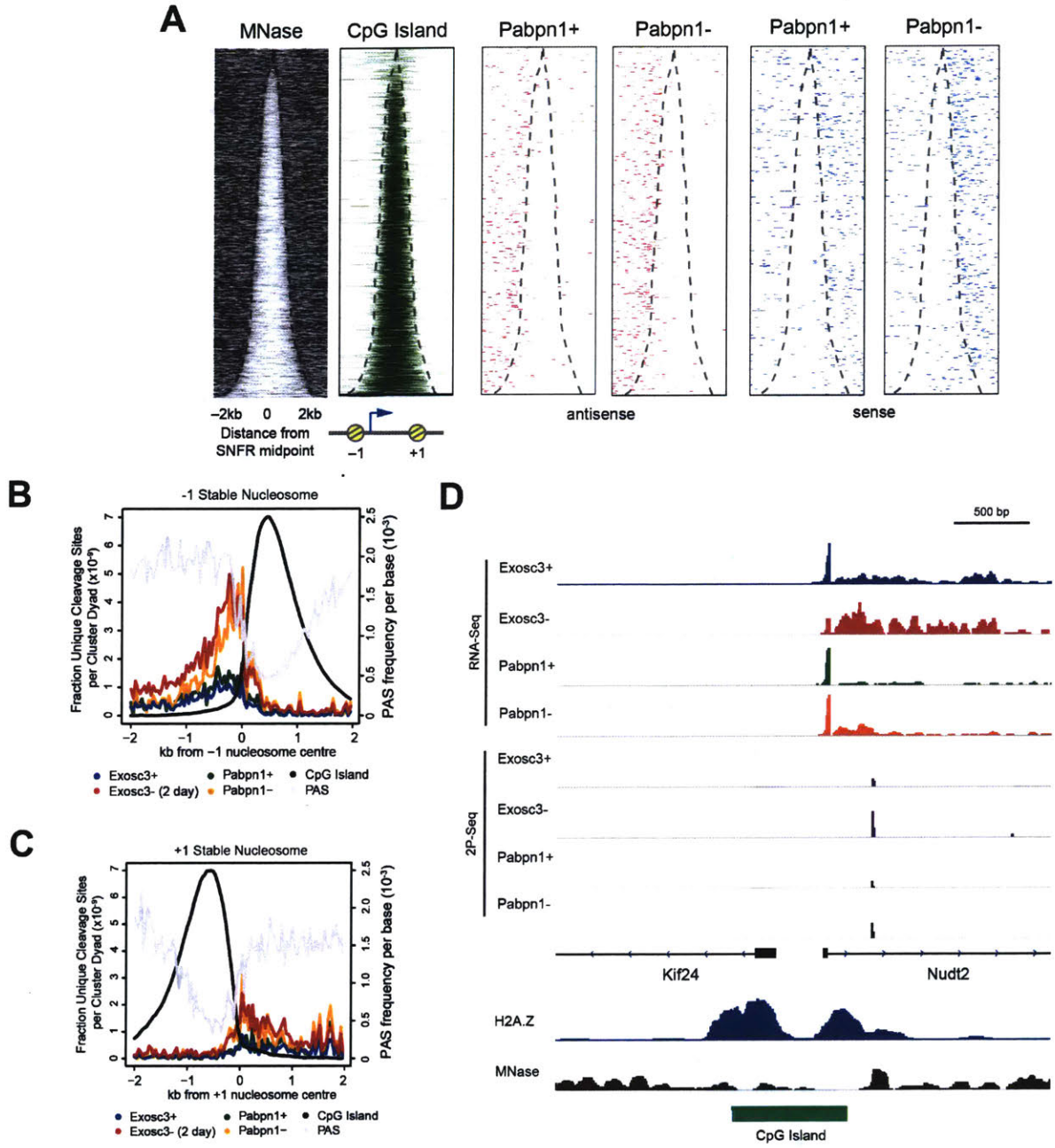

Figure 5. Pabpn1 may regulate termination at the -1/+1 stable nucleosome.

(A) Alignment of Pabpn1 2P-seq reads around the center of stable nucleosome free regions, for regions containing premature clusters in first intron.
(B-C) Metaplot of unique Pabpn1 2P-seq sites flanking the -1/+1 stable nucleosome.
(D) Genome-browser shots of Nudt2 demonstrating premature termination in the first intron.

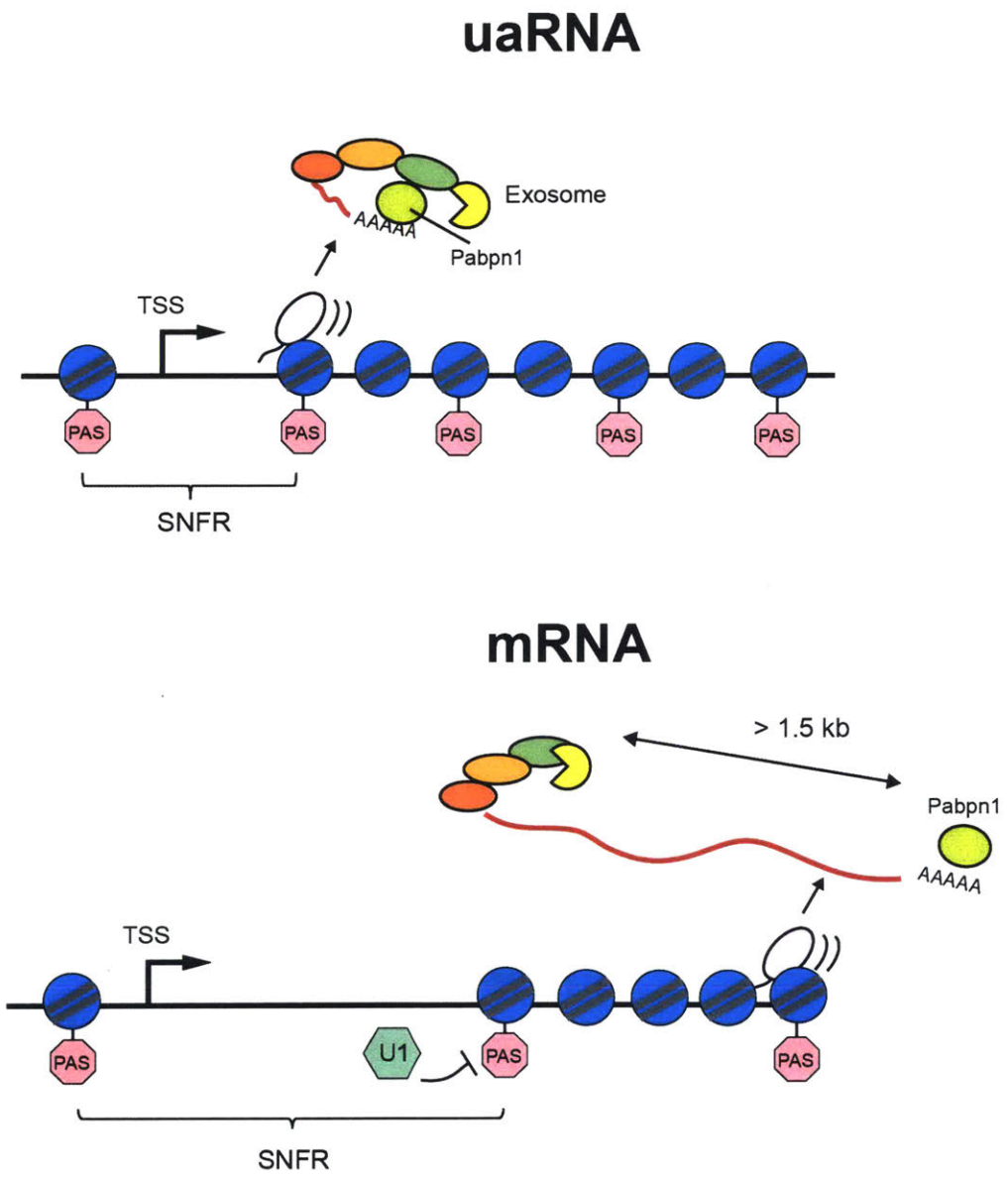

Figure 6. Model of Pabpn1 Sensitivity.

At uaRNAs, terminating close to the promoter by PAS motifs results in binding of Pabpn1 to the poly(A) tail, which can associate with RNA exosome to promote transcript decay. At mRNAs, terminating far from the promoter by PAS motifs still results in Pabpn1to poly(A) tail, but is now too far to interact with 5'-cap-bound RNA exosome, so transcripts evade degradation. U1 snRNP also inhibits early PAS motifs, especially at the edge of SNFRs.

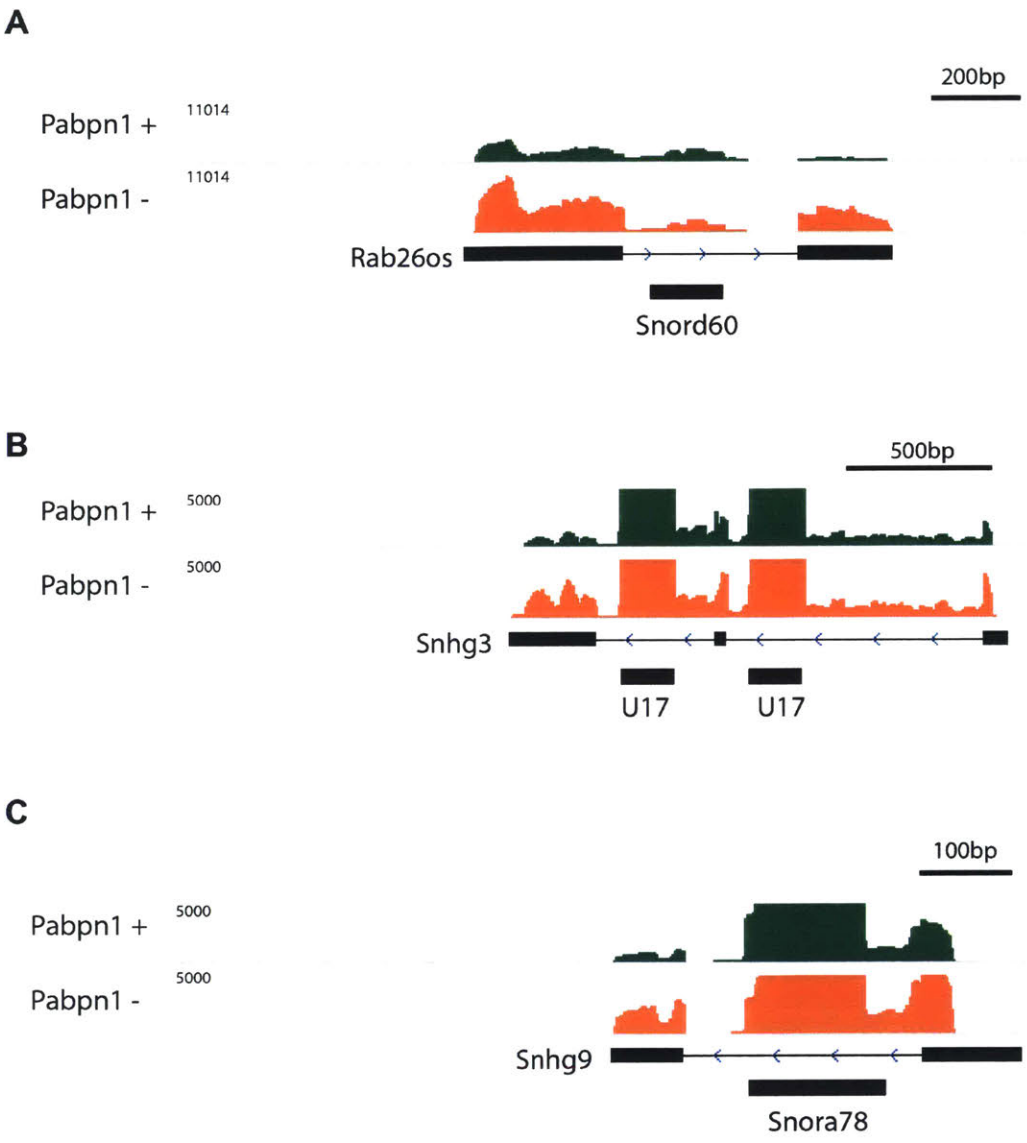

Figure S1. Pabpn1 knockout upregulates snoRNA host genes.

(A-C) Genome-browser shots of 3 snoRNA host genes, as a control for Pabpn1 KO.

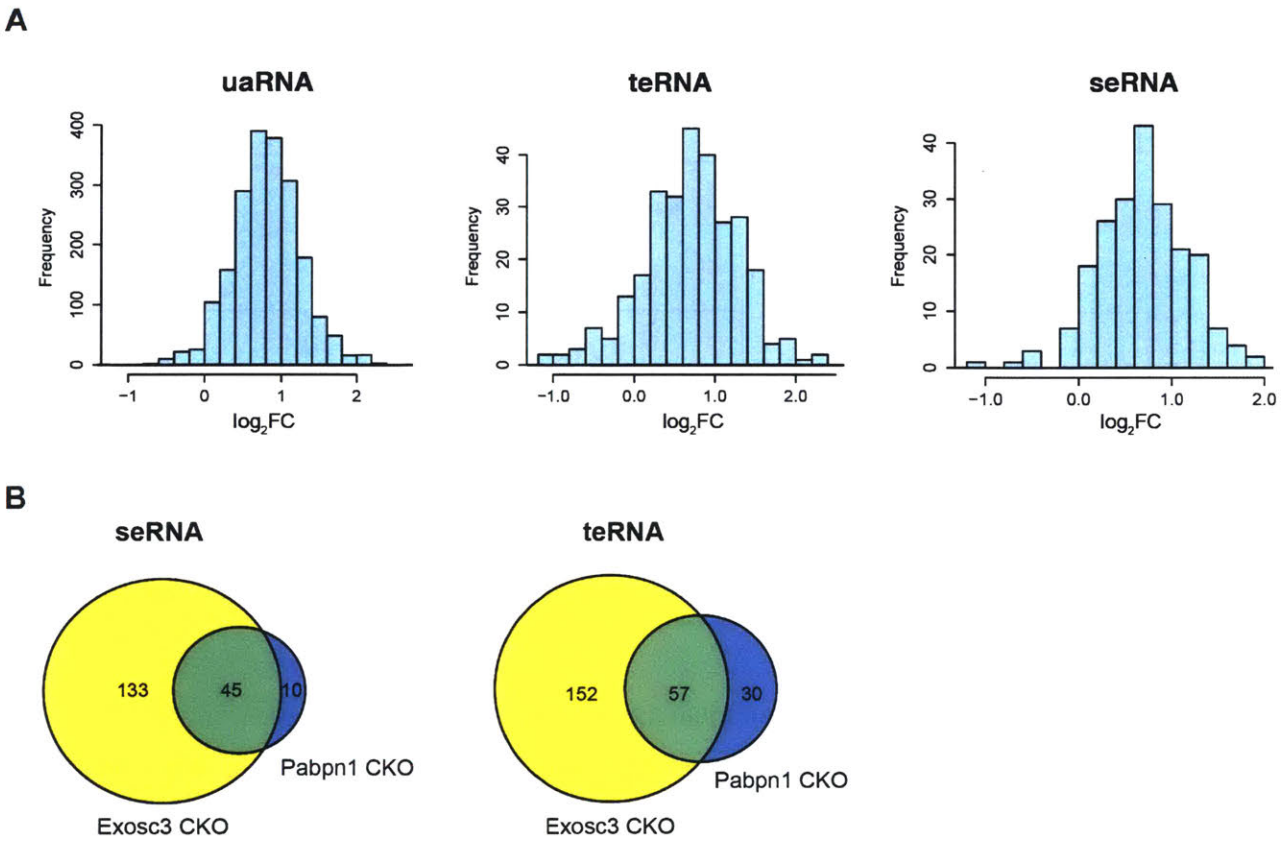

Figure S2. Properties of transcripts upon Pabpn1 KO.

(A) Histogram of fold-change of expression changes of noncoding RNA classes upon Pabpn1 knockout. (B) Pie chart of statistically significantly changing super-enhancer associated RNA (seRNA) and typical-enhancer associated RNA (teRNA) from Pabpn1 CKO or Exosc3 CKO upon dox withdrawal for 3 days.

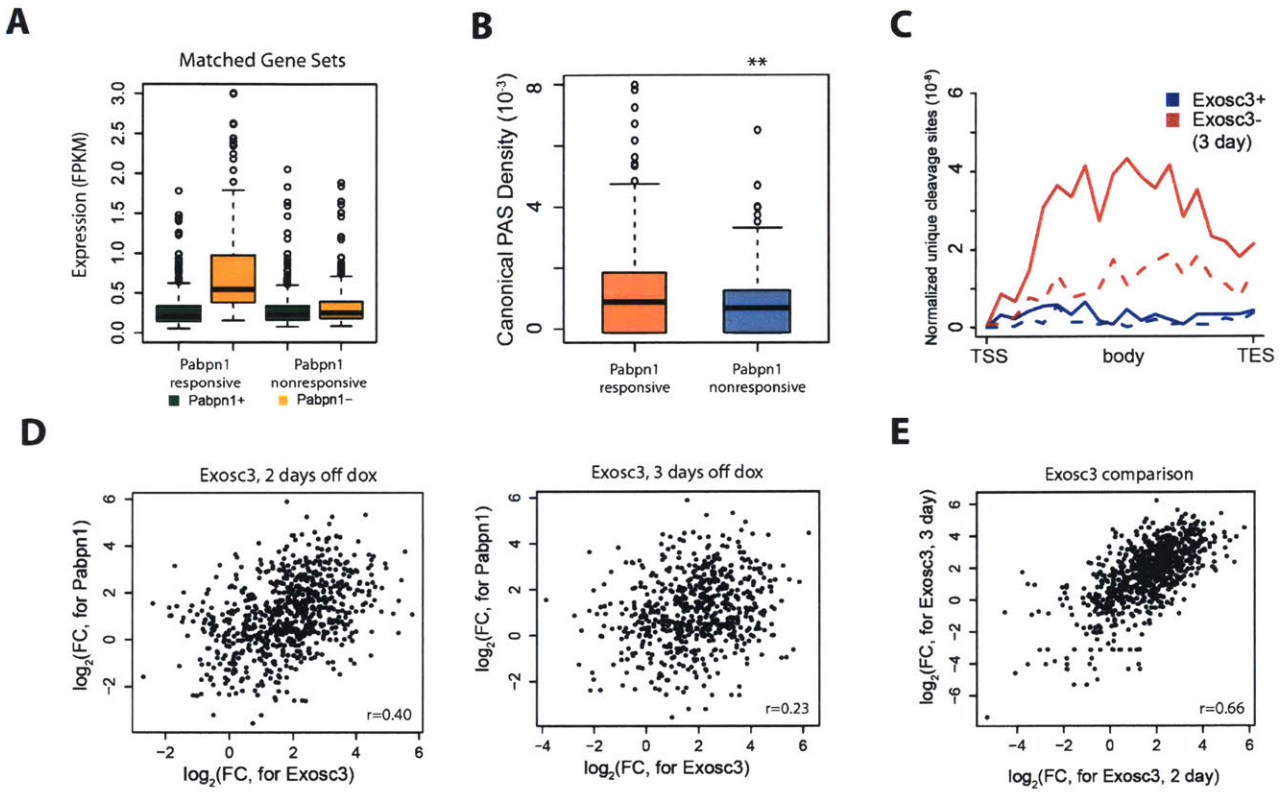

Figure S3. Effects of loss of Pabpn1 on cleavage clusters in uaRNAs.

(A) Expression (FPKM) of matched gene sets for Pabpn1 responsive compare to Pabpn1 nonresponsive.
(B) Density of canonical PAS motifs (A[AT]TAAA) across defined uaRNAs ($p<0.003$, KS test)
(C) Metagene profile of unique cleavage sites across uaRNAs for 3 days of Exosc3 loss. Solid represents Pabpn1-responsive uaRNAs whereas dashed are Pabpn1-nonresponsive uaRNAs.
(D) Scatterplot of log fold-change of cleavage clusters between Pabpn1 CKO and Exosc3 CKO, under 2 or 3 days off dox. (Pearson Correlation)
(E) Scatterplot of log fold-change of cleavage clusters between Exosc3CKO for 2 days off dox and 3 days off dox. (Pearson Correlation)

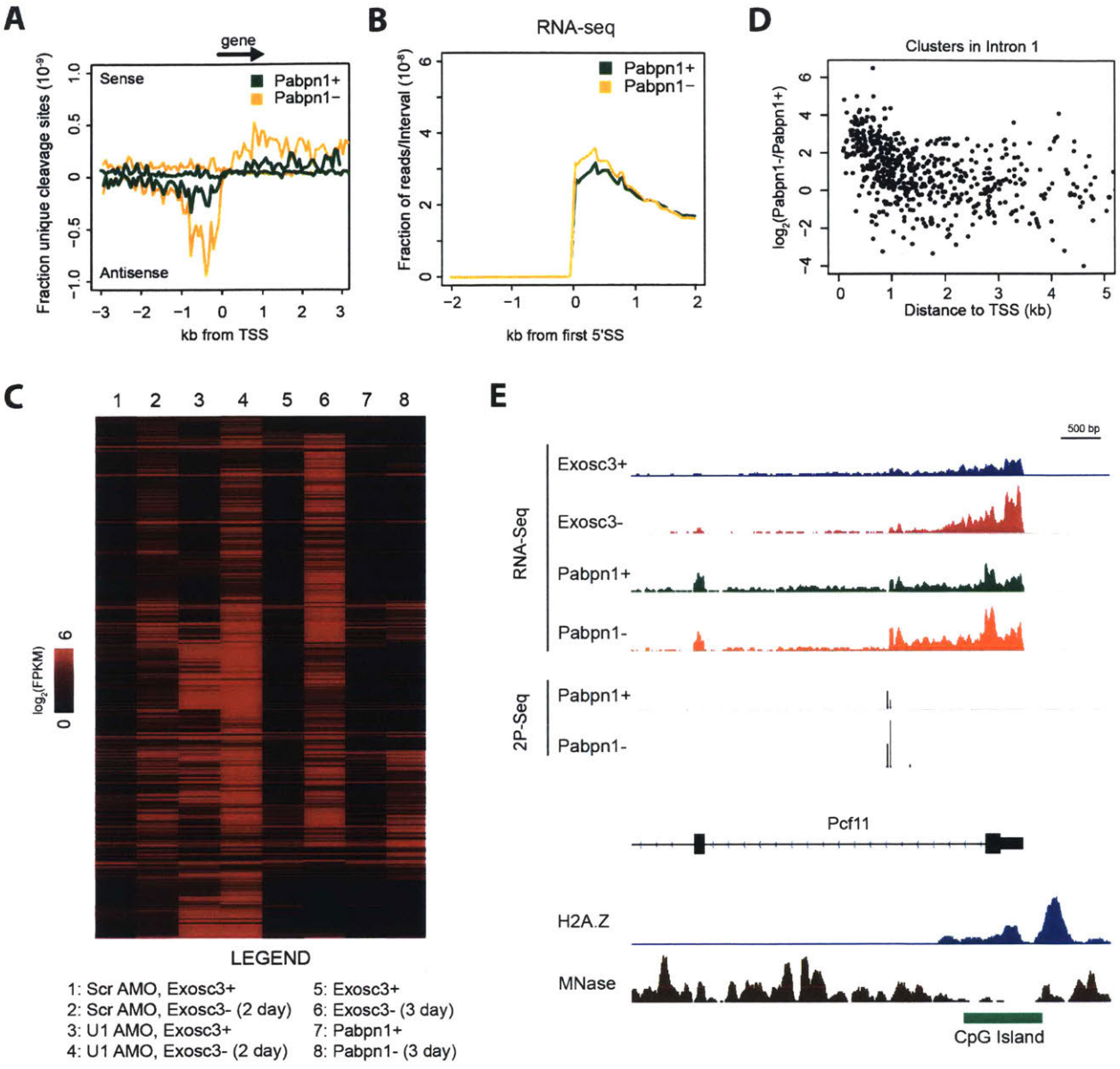

Figure S4. Pabpn1 knockout effect on sense transcription.

(A) Alignment of Pabpn1+ and Pabpn1- unique 2P-seq reads around the TSS of UCSC canonical genes, filtered for nonoverlapping genes, normalized by library size.
(B) Alignment of Pabpn1 RNA-seq reads around the first 5'SS, filtering for introns at least 2 kb long.
(C) Hierarchical clustering of FPKM of cleavage clusters in a 1 kb window flanking the +1 stable nucleosome, clustered using Pearson correlation.
(D) Scatterplot of $\log_2$(fold-change) upon loss of Pabpn1 for cleavage clusters across the first intron compared to $\log_{10}$(distance from the annotated TSS). (Pearson correlation). Plot was trimmed to 5 kb.
(E) Genome browser shot of Pcf11.

4.6 METHODS

Generation of Conditional Pabpn1 mESC Cell Lines

First, we created a dox-inducible Flag-HA tagged Pabpn1cell line. cDNA of Pabpn1 was 5' tagged with Flag-HA and cloned into piggyBac tetO-PB-Neo. The transposon was cotransfected with piggyBac transposase into V6.5 mouse embryonic stem cells, and selected on 500 μg/mL G418. Stable Pabpn1 expression of FH-Pabpn1was validated using anti-HA antibody. Next we deleted endogenous Pabpn1 gene by cotransfecting two CRISPR-Cas9 vectors (pX458) with sgRNAs (sgPabKO-2 and sgPabKO-3) flanking the first exon of Pabpn1. Clones were FACS sorted for GFP signal onto individual wells of a 96 well plate and maintained on 1 μg/ml of doxycycline. Subsequent clones were screened for shortened PCR product across the entire gene (**Table S1**). The shortened PCR product were sequence confirmed. Finally, deletion of Pabpn1 was further confirmed using qRT-PCR for the Pabpn1 gene and western blotting after 3 days of doxycycline removal. The sgRNA and primer sequences are described in Table S1. Cells maintained on gelatin under standard conditions (Almada et al., 2013) in 1 μg/ml of doxycycline.

RNA-seq

Total RNA was isolated with TRIzol Reagent and treated with DNase Turbo (Ambion AM2238) to remove genomic DNA contamination. RNA was quality checked, requiring a Bioanalyzer RIN score of at least 8.5 for library prep. RNAs were depleted of ribosomal RNAs using the RiboZero rRNA removal kit (Epicentre MRZH116), converted into stranded RNA-Seq libraries with the Illumina Tru-Seq kit (Illumina RS-122-2101), and sequenced in paired end read mode using the Illumina NEXT-Seq500.

RNA-seq Analysis

All analyses were carried out using UCSC (NCBI37/mm9) mouse gene annotations. Paired end reads were first mapped to ribosomal RNA and various repetitive sequences such as U1 snRNA using Bowtie2 (Langmead and Salzberg, 2012), and then subsequently mapped to the mouse UCSC transcriptome and genome using STAR aligner (Dobin et al., 2013). The resulting reads were filtered for uniquely mapping, properly paired reads. Potential PCR duplicates were removed using the Picard Suite MARKDUP (http://broadinstitute.github.io/picard). In genome browser shots, the reads are displayed. For metaplot alignments, we further processed the reads by selecting read 2 of the paired-end read (same direction as the RNA), and filtered away any overlapping miRNAs, tRNAs, repeats from repeatMasker, or snoRNA.

3' End Sequencing (2P-seq)

2P-Seq was performed as described in (Spies et al., 2013). Total RNA is poly(A) selected using oligo-dT dynabeads and cleaved with trace amounts of RNase T1 for 20 minutes at 22°C, inactivated, and cleaned up with an ethanol precipitation. The resulting RNA was reverse transcribed using IW-RT1p and the size selected for 200-400 nts on a polyacrylamide gel. Next the cDNA was circularized using CircLigase II (Epicentre), PCR amplified with primers IW-PCR-F.1 and IW-PCR-RPI, and further size selected to remove adapters, before sequencing from the poly(A) tail using IW-Seq-PE1.1 in single end read mode on the Illumina NEXT-Seq500.

2P-seq Analysis

Reads were first quality filtered by adapter trimming with Trimmomatic (Bolger et al., 2014), and oligo(A) stretches (>5 As) were removed if they were immediately downstream of first sequenced nucleotide. We interpreted these events as poly(A) tails that due to reverse transcription errors or biological reasons had a non-As added to the cDNA. Next, we mapped either filtered reads (set A) or filtered reads with the first 15 nts trimmed (set B) to the mm9 genome using STAR aligner, end-to-end mode. The trimming of first 15 nt was done to ensure that reads were not going to be lost due to mismatches at the 5' end, which may involve non-templated nucleotides (such as uridines), which are added to some termination events. For both sets, the first mapped nucleotide was considered the cleavage site.

The two mapped libraries were combined as follows. If the read only aligned in set A or set B, the cleavage site was used as is. If the read aligned in both set A and set B, we subjected the mapped site to one further test. If the mapped cleavage site in set A overlaps the mapped cleavage site minus 15 nucleotides in set B, the position in set A was used. However, if the mapped cleavage site in set A differed substantially from the read in set B, we chose the site in set A as the mapped site. We attributed changes for this subset to the shorter read being harder to find exact matches, so preferred the mapped position of the longer read.

With the combined mapped cleavage sites, we next removed reads with at least 7 adenosines in the 10 nucleotides 3' of the cleavage site, or 13 adenosines in the downstream 20 nucleotides, to remove any internal priming. The remaining cleavage sites were filtered so that it must have at least 2 different reads mapping to it and also to not overlap B2 SINE elements. Finally, we scored reads as PAS containing or not PAS containing by surveying the 80 nucleotides upstream of the cleavage site for the presence of the top 36 PAS motifs, as described in (Almada

et al., 2013). Specifically, the top 2 canonical PAS motifs are AATAAA or ATTAAA. Next, we also look for known variants, AGTAAA or TATAAA. We subsequently look for the next 8 most frequent sites or PAS8 (AATATA, AATACA, CATAAA, GATAAA, AATGAA, ACTAAA, AAGAAA, AATAGA). Finally we look for the remaining 24 PAS variants.

Differential Expression Analysis and Expression Matched Controls

The number of reads per transcript was counted by using intersectBed of the Bedtools suite (Quinlan and Hall, 2010) across defined uaRNAs. We filtered away intervals with low numbers, defining robustly expressed uaRNAs were those where there was at least 0.1 CPM across all 4 libraries. Differential transcripts were called using the R package edgeR, where we normalized libraries using UQ normalization. Statistically significantly changing uaRNAs were those with fold-change>2 and FDR<0.1.

The top 250 uaRNAs that changed the most after upper-quartile normalization were defined as the Pabpn1-sensitive set. All of these were statistically significant as well. For the negative set, we pulled out the bottom 395 ($|\log_2(\text{fold change})| < 0.47$). We subsequently used the R package MatchIt to expression match the least-changing uaRNAs.

Hierarchical Clustering

Raw counts of PAS cleavage clusters were calculated using intersectBed, requiring at least 4 libraries with non-zeros. Subsequently, counts were normalized by 2P-seq library depth, and biological replicates were averaged. Finally, we hierarchically clustered libraries based on Pearson

correlations, using Complete Linkage and Sorting by Cluster ID using Multi Experiment Viewer. Either log(fold-change) or log(normalized counts) were depicted using R.

Metaplots

For TSS plots, intervals were prefiltered to remove overlapping UCSC canonical genes within 5 kb of the TSS. In enhancer plots, we selected peaks of Oct4/Sox2/Nanog defined enhancers according to a previous report (Suzuki et al., 2017; Whyte et al., 2013), filtering away intragenic enhancers (overlapping UCSC canonical genes) and those that overlapped enhancers within a 3 kb window. For the 5' splice site alignment UCSC, we pulled out first introns of genes with at least 4 introns and at least 2kb long.

To make metaplots, the number of reads (RNA-seq) or unique cleavage sites (2P-seq) were counted across non-overlapping bins spanning the aligned region. Splice site alignments were also modified to filter away exonic reads.

Bins were normalized either using the following for RNA-seq:

$$normalized\ bin = \frac{counts\ of\ filtered\ RNA\ Seq\ reads}{total\ mapped\ reads\ x\ number\ of\ aligned\ intervals}$$

Or alternatively, the following for 2P-seq:

$$normalized\ bin = \frac{counts\ of\ unique\ filtered\ 2P\ sites}{total\ mapped\ reads\ x\ number\ of\ aligned\ intervals}$$

Scaled Metagene Plots

Previously defined uaRNA intervals were split into 40 nonoverlapping bins. Unique cleavage sites for Exosc3 CKO datasets (Scr AMO or Exosc3 3 day) or Pabpn1 CKO were aligned against them.

Distance from TSS Analysis

For each cleavage cluster overlapping the uaRNA or first intron, we counted the total number of PAS filtered 2P-seq reads across replicate 1 to estimate the frequency of use of each site. Counts were normalized using total library size before calculating log fold change. The distance from the corresponding TSS was tabulated. Subsequently, plots were visualized in R. Correlations were defined using Pearson correlation.

4.6 SUPPLEMENTAL MATERIALS

Primers for Cloning

Name	Sequence (5'– 3')
Clon-5'HA-Pabpn1 (+)	GCGACTAGTCCACCATGGACTACAAGGACGACGATGACAAGT ACCCTTATGACGTGCCCGATTACGCTGCGGCGGCGGCGGCGG CGGCAG
Clon-Pabpn1 (-)	GCCGCGGCCGCTTAGTAAGGGGAATACCATGATG

sgRNA Primers

Name	Sequence (5'– 3')
sgPabKO-2 fw	CACCGGTACAGCTTCTAAAGTGAGC
sgPabKO-2 rv	AAACGCTCACTTTAGAAGCTGTACC
sgPabKO-3 fw	CACCGGCTACTGTGTACTCTTCCAC
sgPabKO-3 rv	AAACGTGGAAGAGTACACAGTAGCC

qRT-PCR Primers

Name	Sequence (5'– 3')
qPCR-Actb fw	GACGAGGCCCAGAGCAAGAGAGG
qPCR-Actb rv	GGTGTTGAAGGTCTCAAACATG
qPCR-Pabpn1 fw	TCGGACAAAGAGTCAGTGAGG
qPCR-Pabpn1 rv	CTGATGCCTGGTCTGTTGG
qPCR-uaP4hb fw	TTGGGTGACGGACCCTAGTT
qPCR-uaP4hb rv	ATTCCGAATGGTGGACAGGA

Antibodies	Company	ID
Vinculin	Sigma	V9131
HA	Roche	3F10
Pabpn1	Abcam	ab75855

DATASETS USED

a) RNA-seq

Library	Cell Type	Lab	Authors	GEO Accession
Pabpn1 + dox rep.1	Pabpn1 CKO	Sharp	Chiu et al., 2017	*this paper*
Pabpn1 + dox rep.2	Pabpn1 CKO	Sharp	Chiu et al., 2017	*this paper*
Pabpn1 - dox rep.1 (3 day)	Pabpn1 CKO	Sharp	Chiu et al., 2017	*this paper*
Pabpn1 - dox rep.2 (3 day)	Pabpn1 CKO	Sharp	Chiu et al., 2017	*this paper*
Exosc3 + dox rep.1	Exosc3 CKO	Sharp	Chiu et al., 2017	*this paper*
Exosc3 + dox rep.2	Exosc3 CKO	Sharp	Chiu et al., 2017	*this paper*
Exosc3 - dox rep.1 (3 day)	Exosc3 CKO	Sharp	Chiu et al., 2017	*this paper*
Exosc3 - dox rep.2 (3 day)	Exosc3 CKO	Sharp	Chiu et al., 2017	*this paper*

b) 2P-seq

Library	Cell Type	Lab	Authors	GEO Accession
Pabpn1 + dox rep.1	Pabpn1 CKO	Sharp	Chiu et al., 2017	*this paper*
Pabpn1 + dox rep.2	Pabpn1 CKO	Sharp	Chiu et al., 2017	*this paper*
Pabpn1 - dox rep.1 (3 day)	Pabpn1 CKO	Sharp	Chiu et al., 2017	*this paper*
Pabpn1 - dox rep.2 (3 day)	Pabpn1 CKO	Sharp	Chiu et al., 2017	*this paper*
Exosc3 + dox rep.1	Exosc3 CKO	Sharp	Chiu et al., 2017	*this paper*
Exosc3 + dox rep.2	Exosc3 CKO	Sharp	Chiu et al., 2017	*this paper*
Exosc3 - dox rep.1 (3 day)	Exosc3 CKO	Sharp	Chiu et al., 2017	*this paper*
Exosc3 - dox rep.2 (3 day)	Exosc3 CKO	Sharp	Chiu et al., 2017	*this paper*
Scr AMO + dox rep.1	Exosc3 CKO	Sharp	Chiu et al., 2017	*this paper*
Scr AMO + dox rep.2	Exosc3 CKO	Sharp	Chiu et al., 2017	*this paper*
Scr AMO - dox rep.1 (2 day)	Exosc3 CKO	Sharp	Chiu et al., 2017	*this paper*
Scr AMO - dox rep.2 (2 day)	Exosc3 CKO	Sharp	Chiu et al., 2017	*this paper*
U1 AMO + dox rep.1	Exosc3 CKO	Sharp	Chiu et al., 2017	*this paper*
U1 AMO + dox rep.2	Exosc3 CKO	Sharp	Chiu et al., 2017	*this paper*
U1 AMO - dox rep.1 (2 day)	Exosc3 CKO	Sharp	Chiu et al., 2017	*this paper*
U1 AMO - dox rep.2 (2 day)	Exosc3 CKO	Sharp	Chiu et al., 2017	*this paper*

c) MNase-seq

Library	Cell Type	Lab	Authors	GEO Accession
ESC merged mononucleosome	129P2/Ola	Rippe	Teif et al., 2012	GSM1004653

d) ChIP-seq

Library	Cell Type	Lab	Authors	GEO Accession
H2A.Z	V6.5	Boyer	Subramanian et al., 2013	GSM984544

4.7 REFERENCES

Almada, A.E., Wu, X., Kriz, A.J., Burge, C.B., and Sharp, P.A. (2013). Promoter directionality is controlled by U1 snRNP and polyadenylation signals. Nature *499*, 360-363.

Andersen, P.K., Lykke-Andersen, S., and Jensen, T.H. (2012). Promoter-proximal polyadenylation sites reduce transcription activity. Genes Dev *26*, 2169-2179.

Andersen, P.R., Domanski, M., Kristiansen, M.S., Storvall, H., Ntini, E., Verheggen, C., Schein, A., Bunkenborg, J., Poser, I., Hallais, M., *et al.* (2013). The human cap-binding complex is functionally connected to the nuclear RNA exosome. Nat Struct Mol Biol *20*, 1367-1376.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., *et al.* (2014). An atlas of active enhancers across human cell types and tissues. Nature *507*, 455-461.

Baillat, D., Hakimi, M.A., Naar, A.M., Shilatifard, A., Cooch, N., and Shiekhattar, R. (2005). Integrator, a multiprotein mediator of small nuclear RNA processing, associates with the C-terminal repeat of RNA polymerase II. Cell *123*, 265-276.

Beaulieu, Y.B., Kleinman, C.L., Landry-Voyer, A.M., Majewski, J., and Bachand, F. (2012). Polyadenylation-dependent control of long noncoding RNA expression by the poly(A)-binding protein nuclear 1. PLoS Genet *8*, e1003078.

Bergeron, D., Pal, G., Beaulieu, Y.B., Chabot, B., and Bachand, F. (2015). Regulated Intron Retention and Nuclear Pre-mRNA Decay Contribute to PABPN1 Autoregulation. Mol Cell Biol *35*, 2503-2517.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114-2120.

Bresson, S.M., and Conrad, N.K. (2013). The human nuclear poly(a)-binding protein promotes RNA hyperadenylation and decay. PLoS Genet *9*, e1003893.

Bresson, S.M., Hunter, O.V., Hunter, A.C., and Conrad, N.K. (2015). Canonical Poly(A) Polymerase Activity Promotes the Decay of a Wide Variety of Mammalian Nuclear RNAs. PLoS Genet *11*, e1005610.

Core, L.J., Waterfall, J.J., and Lis, J.T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science *322*, 1845-1848.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15-21.

Flynn, R.A., Almada, A.E., Zamudio, J.R., and Sharp, P.A. (2011). Antisense RNA polymerase II divergent transcripts are P-TEFb dependent and substrates for the RNA exosome. Proc Natl Acad Sci U S A *108*, 10460-10465.

Hallais, M., Pontvianne, F., Andersen, P.R., Clerici, M., Lener, D., Benbahouche Nel, H., Gostan, T., Vandermoere, F., Robert, M.C., Cusack, S., *et al.* (2013). CBC-ARS2 stimulates 3'-end maturation of multiple RNA families and favors cap-proximal processing. Nat Struct Mol Biol *20*, 1358-1366.

Harigaya, Y., Tanaka, H., Yamanaka, S., Tanaka, K., Watanabe, Y., Tsutsumi, C., Chikashige, Y., Hiraoka, Y., Yamashita, A., and Yamamoto, M. (2006). Selective elimination of messenger RNA prevents an incidence of untimely meiosis. Nature *442*, 45-50.

Hilleren, P., McCarthy, T., Rosbash, M., Parker, R., and Jensen, T.H. (2001). Quality control of mRNA 3'-end processing is linked to the nuclear exosome. Nature *413*, 538-542.

Houseley, J., LaCava, J., and Tollervey, D. (2006). RNA-quality control by the exosome. Nat Rev Mol Cell Biol *7*, 529-539.

Kaida, D., Berg, M.G., Younis, I., Kasim, M., Singh, L.N., Wan, L., and Dreyfuss, G. (2010). U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. Nature *468*, 664-668.

Kim, M., Krogan, N.J., Vasiljeva, L., Rando, O.J., Nedea, E., Greenblatt, J.F., and Buratowski, S. (2004). The yeast Rat1 exonuclease promotes transcription termination by RNA polymerase II. Nature *432*, 517-522.

Lai, F., Gardini, A., Zhang, A., and Shiekhattar, R. (2015). Integrator mediates the biogenesis of enhancer RNAs. Nature *525*, 399-403.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods *9*, 357-359.

Lee, N.N., Chalamcharla, V.R., Reyes-Turcu, F., Mehta, S., Zofall, M., Balachandran, V., Dhakshnamoorthy, J., Taneja, N., Yamanaka, S., Zhou, M., *et al.* (2013). Mtr4-like protein coordinates nuclear RNA processing for heterochromatin assembly and for telomere maintenance. Cell *155*, 1061-1074.

Lemay, J.F., D'Amours, A., Lemieux, C., Lackner, D.H., St-Sauveur, V.G., Bahler, J., and Bachand, F. (2010). The nuclear poly(A)-binding protein interacts with the exosome to promote synthesis of noncoding small nucleolar RNAs. Mol Cell *37*, 34-45.

Lubas, M., Andersen, P.R., Schein, A., Dziembowski, A., Kudla, G., and Jensen, T.H. (2015). The human nuclear exosome targeting complex is loaded onto newly synthesized RNA to direct early ribonucleolysis. Cell Rep *10*, 178-192.

Lubas, M., Christensen, M.S., Kristiansen, M.S., Domanski, M., Falkenby, L.G., Lykke-Andersen, S., Andersen, J.S., Dziembowski, A., and Jensen, T.H. (2011). Interaction profiling identifies the human nuclear exosome targeting complex. Mol Cell *43*, 624-637.

Meola, N., Domanski, M., Karadoulama, E., Chen, Y., Gentil, C., Pultz, D., Vitting-Seerup, K., Lykke-Andersen, S., Andersen, J.S., Sandelin, A., *et al.* (2016). Identification of a Nuclear Exosome Decay Pathway for Processed Transcripts. Mol Cell *64*, 520-533.

Nojima, T., Gomes, T., Grosso, A.R., Kimura, H., Dye, M.J., Dhir, S., Carmo-Fonseca, M., and Proudfoot, N.J. (2015). Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing. Cell *161*, 526-540.

Ntini, E., Jarvelin, A.I., Bornholdt, J., Chen, Y., Boyd, M., Jorgensen, M., Andersson, R., Hoof, I., Schein, A., Andersen, P.R., *et al.* (2013). Polyadenylation site-induced decay of upstream transcripts enforces promoter directionality. Nat Struct Mol Biol *20*, 923-928.

Pefanis, E., Wang, J., Rothschild, G., Lim, J., Kazadi, D., Sun, J., Federation, A., Chao, J., Elliott, O., Liu, Z.P., *et al.* (2015). RNA exosome-regulated long non-coding RNA transcription controls super-enhancer activity. Cell *161*, 774-789.

Preker, P., Almvig, K., Christensen, M.S., Valen, E., Mapendano, C.K., Sandelin, A., and Jensen, T.H. (2011). PROMoter uPstream Transcripts share characteristics with mRNAs and are produced upstream of all three major types of mammalian promoters. Nucleic Acids Res *39*, 7179-7193.

Preker, P., Nielsen, J., Kammler, S., Lykke-Andersen, S., Christensen, M.S., Mapendano, C.K., Schierup, M.H., and Jensen, T.H. (2008). RNA exosome depletion reveals transcription upstream of active human promoters. Science *322*, 1851-1854.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842.

Ramirez-Carrozzi, V.R., Braas, D., Bhatt, D.M., Cheng, C.S., Hong, C., Doty, K.R., Black, J.C., Hoffmann, A., Carey, M., and Smale, S.T. (2009). A unifying model for the selective regulation of inducible transcription by CpG islands and nucleosome remodeling. Cell *138*, 114-128.

Saxonov, S., Berg, P., and Brutlag, D.L. (2006). A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc Natl Acad Sci U S A *103*, 1412-1417.

Schaukowitch, K., Joo, J.Y., Liu, X., Watts, J.K., Martinez, C., and Kim, T.K. (2014). Enhancer RNA facilitates NELF release from immediate early genes. Mol Cell *56*, 29-42.

Seila, A.C., Calabrese, J.M., Levine, S.S., Yeo, G.W., Rahl, P.B., Flynn, R.A., Young, R.A., and Sharp, P.A. (2008). Divergent transcription from active promoters. Science *322*, 1849-1851.

Sigova, A.A., Abraham, B.J., Ji, X., Molinie, B., Hannett, N.M., Guo, Y.E., Jangi, M., Giallourakis, C.C., Sharp, P.A., and Young, R.A. (2015). Transcription factor trapping by RNA in gene regulatory elements. Science *350*, 978-981.

Sigova, A.A., Mullen, A.C., Molinie, B., Gupta, S., Orlando, D.A., Guenther, M.G., Almada, A.E., Lin, C., Sharp, P.A., Giallourakis, C.C., *et al.* (2013). Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells. Proc Natl Acad Sci U S A *110*, 2876-2881.

Spies, N., Burge, C.B., and Bartel, D.P. (2013). 3' UTR-isoform choice has limited influence on the stability and translational efficiency of most mRNAs in mouse fibroblasts. Genome Res *23*, 2078-2090.

Stadelmayer, B., Micas, G., Gamot, A., Martin, P., Malirat, N., Koval, S., Raffel, R., Sobhian, B., Severac, D., Rialle, S., *et al.* (2014). Integrator complex regulates NELF-mediated RNA polymerase II pause/release and processivity at coding genes. Nat Commun *5*, 5531.

Sugiyama, T., and Sugioka-Sugiyama, R. (2011). Red1 promotes the elimination of meiosis-specific mRNAs in vegetatively growing fission yeast. EMBO J *30*, 1027-1039.

Suzuki, H.I., Young, R.A., and Sharp, P.A. (2017). Super-Enhancer-Mediated RNA Processing Revealed by Integrative MicroRNA Network Analysis. Cell *168*, 1000-1014 e1015.

West, S., Gromak, N., and Proudfoot, N.J. (2004). Human 5' --> 3' exonuclease Xrn2 promotes transcription termination at co-transcriptional cleavage sites. Nature *432*, 522-525.

Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I., and Young, R.A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell *153*, 307-319.

Yamanaka, S., Yamashita, A., Harigaya, Y., Iwata, R., and Yamamoto, M. (2010). Importance of polyadenylation in the selective elimination of meiotic mRNAs in growing S. pombe cells. EMBO J *29*, 2173-2181.

Zhao, J., Ohsumi, T.K., Kung, J.T., Ogawa, Y., Grau, D.J., Sarma, K., Song, J.J., Kingston, R.E., Borowsky, M., and Lee, J.T. (2010). Genome-wide identification of polycomb-associated RNAs by RIP-seq. Mol Cell *40*, 939-953.

Chapter 5

Conclusions and Future Directions

This chapter briefly summarizes our findings, as well as propose some areas to explore.

5. 1 Summary

In conclusion, this thesis builds on pre-existing work in the field of transcription. Previous studies suggested that the U1-PAS axis was a key decision point that determined whether RNAs that terminated early were degraded (Almada et al., 2013; Ntini et al., 2013). In addition, the RNA exosome was known to regulate uaRNA transcript stability (Flynn et al., 2011; Preker et al., 2011; Preker et al., 2008) and there were suggestions that Pabpn1 may also be linked (Beaulieu et al., 2012). Through the use of recent CRISPR-Cas9 technologies (Cong et al., 2013), we generated a system that enabled us to further dissect these questions. A brief summary of five general observations:

First, the RNA exosome broadly degrades polyadenylated noncoding RNAs in mammals, including uaRNAs, eRNAs and premature termination transcripts. uaRNAs have very short half lives (8-12 mins), which increase 2-3x in the absence of the RNA exosome. These polyadenylated ends are found in 40% of defined uaRNA intervals, suggesting that an alternative pathway collaborates with the RNA exosome to degrade non-polyadenylated RNAs, likely the NEXT complex (Lubas et al., 2011).

Secondly, both the RNA exosome and U1 snRNP regulate PAS-linked termination within the first intron. Inhibition of either the RNA exosome or U1 snRNA results in increased termination signals within the first 4-5 stable nucleosomes, in a region termed the stable nucleosome termination area (SNTA). A SNTA is found at both the -1 and +1 stable nucleosome, which flanks a stable nucleosome free region (SNFR) marked by a CpG island. The SNFR is distinct from previously studied NFRs, which are narrower and defined by DNase seq or low levels of MNase (de Dieuleveult et al., 2016).

Thirdly, there are two types of pausing near the promoter: a TSS-proximal pause and a +1 stable nucleosome pause. These pauses can be distinguished when promoters are split for those

with wide SNFRs, as narrow SNFRs do not have enough resolution. Genes with detectable premature termination have a stronger pause at both the TSS-proximal region and the +1 stable nucleosome, compared to expression-matched genes without premature termination. Consistent with this, genes with premature termination tend to be targets of cMyc and P-TEFb, and have increased association of pausing factors.

Fourthly, similar to the TSS pause, the +1 nucleosome pause is actively regulated. Inhibition of Myc or P-TEFb results in reduced pause release from the +1 stable nucleosome. Importantly, genes with premature termination have a stronger response to flavopiridol or Myc inhibitor, arguing that these genes have higher pausing. Moreover, they have higher association of chromatin remodelers such as Chd1, which can regulate the nucleosome barrier. Additionally the RNA exosome itself promotes pause release, since loss of Exosc3 results in increased pausing.

Lastly, degradation of polyadenylated transcripts involves the activity of Pabpn1. This activity depends on the distance between the TSS and the termination site being under 1.5-2 kb, and specifically targets polyadenylated transcripts in conjunction with the RNA exosome. The RNA exosome also has broader roles and can can degrade transcripts that terminate further from promoter-proximal PAS.

Combined, these results suggest that there is an additional step in the mammalian transcription cycle during transcription elongation at the +1 stable nucleosome Aspects of this flavopiridol-sensitive checkpoint have been observed in human cells, suggesting evolutionary conservation (Laitem et al., 2015). After promoter proximal pause release, Pol II frequently transcribes through a CpG island in mammals. While these CpG islands contain nucleosomes, these nucleosomes are less stable due to a combination of weaker association with CpG islands (Fenouil et al., 2012; Ramirez-Carrozzi et al., 2009) and incorporation of the histone variant

H2A.Z, which destabilizes nucleosomes (Jin and Felsenfeld, 2007; Weber et al., 2014). After it passes through the CpG island, Pol II encounter stable nucleosomes and is paused. This pause can be regulated by the activity of a flavopiridol-sensitive kinase (likely P-TEFb). If Pol II transforms into a processive transcription mode, it can bypass this barrier. However, if Pol II is not processive, it recognizes early PAS motifs especially in the absence of U1 snRNP, promoting early termination and degradation by the RNA exosome and Pabpn1.

There are several questions that arise from these studies, so we will discuss them and propose experiments to test them.

5.2 Stable Nucleosome Pausing and Premature Termination

Pausing occurs frequently throughout transcription, as a checkpoint mechanism to ensure appropriate cotranscriptional events have occurred. For instance, the promoter-proximal pause is thought to promote capping of the 5' end of RNAs. Phosphorylation of Ser2 and Spt5 of the Pol II CTD by CDK9 accompanies the release from the promoter-proximal pause. Recently, P-TEFb activity or phosphorylation of Ser2 has been shown to function in regulating additional pausing events throughout transcription. In yeast, there is a major pausing event at the 3' splice site of introns (Alexander et al., 2010). Pol II associated with that pause has both Ser5P and Ser2P. However Ser2P signals mostly begin at the 3' splice site, suggesting that Ser2P may be linked with escape from the pause. Additionally, Pol II frequently pauses at the 3' end of mammalian RNAs over G-rich sequences, which promotes recognition of termination signals (Yonaha and Proudfoot, 1999). Recent work has found that CDK9 activity is important for regulating Pol II escape from a pause at the 3' ends of genes (Laitem et al., 2015). These results suggest a major area of research in the future will be understanding how other types of pausing are regulated. It is possible that a

second Ser2 kinase CDK12 (a homolog of Ctk1) recently described in mammals may also be associated with these Ser2P-linked pauses (Bartkowiak et al., 2010; Blazek et al., 2011). Interestingly, nucleosomes are enriched after the 3' splice site over exons (Schwartz et al., 2009; Spies et al., 2009) and also immediately after the PAS termination signals at the 3' end of genes (Spies et al., 2009). One model that could explain this correlation is Ser2 phosphorylation may promotes transcription beyond nucleosome-mediated pauses, perhaps by recruiting specific factors, but additional experiments are required to test this hypothesis.

Our work found that there is a checkpoint at the +1 stable nucleosome regulated by P-TEFb or another kinase sensitive to flavopiridol such as CDK12. This pause is strongly linked with premature termination by PAS motifs. How is pausing at the +1 stable nucleosome related to PAS-mediated termination? One might speculate that early PAS-termination is a readout of increased +1 nucleosome pausing, as pausing would provide more time for the termination machinery to recognize PAS motifs. However, these motifs typically occur downstream of the +1 stable nucleosome, so the PAS motif would not have been transcribed while the pause is happening.

We envision that two forms of elongating Pol II function in cells: productive elongation complexes and unproductive elongation complexes. Productive elongation complexes are able to transcribe through stable nucleosomes, rapidly bypassing the PAS motif and allowing U1 snRNP to suppress usage of PAS motifs. In contrast, unproductive elongation complexes have difficulty elongating through the stable nucleosomes. Since the cleavage and polyadenylation (CPA) machinery tethered to transcribing Pol II through an association with the Pol II CTD (Ahn et al., 2004), unproductive elongation complexes grant more time for the CPA machinery to recognize early PAS motifs, resulting in early termination and RNA exosome degradation. The role of transcription elongation kinetics has been implicated in regulating alternative polyadenylation, in

which a slowly transcribing Pol II promoted the use of an early PAS motif (Pinto et al., 2011). The +1 nucleosome pause likely reflects the difficulty of Pol II in switching from H2A.Z-containing nucleosomes to the first non-H2A.Z nucleosome, of which the former are substantially less stable and have a lower transcription barrier than the latter (Jin and Felsenfeld, 2007; Weber et al., 2014). Thus, we hypothesize genes that experience a greater +1 nucleosome pause have more unproductive elongation complexes, thus are more likely to have premature PAS termination.

If this model is correct, modulating the nucleosome barrier should alter the amount of premature termination. First, H2A.Z and Chd1 have been implicated in lowering the +1 nucleosome barrier near the TSS (Skene et al., 2014; Weber et al., 2014). We plan to adjust levels of SRCAP components (the complex that deposits H2A.Z) or Chd1 to modulate the nucleosome barriers through dox-inducible cell lines. If our hypothesis is correct, removing these factors would promote the use of early PAS signals compared to the 3' end of genes since nucleosomes are harder to traverse. In contrast, overexpressing these factors would increase nucleosome instability, allowing Pol II to travel faster and evade early termination. Due to the difficulty of quantitatively detecting low abundant 3' ends, it is critical to pick a model gene that has substantial premature termination events in the control state, so that an increase or a decrease would be easily detectable.

Alternatively, one could modulate the +1 nucleosome barrier by using CRISPR-mediated approaches to adjust the location of the +1 stable nucleosome. Mutating the +1 stable nucleosome binding site so that the sequence better matches the previously described phased AT/TA/TT dinucleotide pattern would increase the nucleosome barrier (Segal et al., 2006) and may result in slower Pol II transcription and more early termination. Alternatively, one could alter the size of the CpG island to shift the positioning of the +1 stable nucleosome. Moving the +1 stable nucleosome later may expose the PAS motif, promoting more early termination.

5.3 Properties of Stable Elongation Complexes

In addition to this line of experimentation, one broader question is why are productive elongation complexes productive? Since the barrier to elongation is largely a nucleosomal barrier, the key components are likely histone chaperones or chromatin remodelers that are recruited later during elongation. One key candidate is the histone chaperone FACT, which is essential for in vitro transcription of chromatinized templates through the exchange of H2A/H2B dimers (Belotserkovskaya et al., 2003; Orphanides et al., 1998). In addition, the activity of P-TEFb likely plays a critical role this transition (Laitem et al., 2015), so the key candidates will be recruited after P-TEFb phosphorylation

We plan to use a dual-prong approach to investigate this question. First, we will analyze published ChIP-seq datasets for transcription elongation factors across mammalian species to determine where the factors associate spatially during transcription. Given the factor is necessary for transcribing through stable nucleosomes, it is likely enriched downstream of the +1 stable nucleosome. Moreover, this factor probably does not associate with uaRNA regions since uaRNAs frequently terminate at the -1 stable nucleosome. One complication is this factor could be recruited earlier in transcription and be primed to be used later when needed.

Many factors have not yet been studied in detail, so we might have to perform ChIP-seq on additional factors, such as phosphorylated Spt5 (DSIF). Similar to the CTD of Pol II, Spt5 has a 5-amino acid repeat region in the C-terminal repeat (CTR) domain which is phosphorylated by P-TEFb to promote transcription elongation (Yamada et al., 2006). Given there are multiple repeats, DSIF may require time for multiple phosphorylation events by P-TEFb to become fully activated. Spatially deciphering the phosphorylation status of Spt5 will provide mechanistic insights in P-TEFb activity and may demonstrate that it is fully phosphorylated after the +1 stable

nucleosome. In addition, our preliminary work in examining published datasets suggest the PAF complex is mostly recruited after the +1 stable nucleosome in mouse embryonic stem cells. PAF is an adapter complex that associates with many transcription elongation factors, including the histone chaperone FACT (Squazzo et al., 2002). As a result, other candidates to investigate will include PAF subunits, PAF-interacting proteins as well as proteins that affect nucleosomes. Ideally, we would use high-resolution ChIP approaches because we are looking for proteins that specifically transition around the +1 stable nucleosome. For instance, ChIP-nexus combines the nucleotide resolution of ChIP-exo with recent advances in barcoding libraries (He et al., 2015). One caveat is that transcription elongation factors do not bind directly to DNA so it is unknown whether this approach would work.

In parallel, we plan to perform a high-throughput shRNA screen to see which factors promote premature termination. shRNAs will be used instead of CRISPR since the key protein may be essential; known proteins that regulate premature termination such as the RNA exosome and Pabpn1 are essential. Initial targets will include chromatin remodelers, histone chaperones, known transcription elongation factors as well as other candidates that interact with the transcription machinery in mammalian protein-protein interaction screens.

A key step in the screen is introducing the proper model gene to an endogenous locus, such as the *Rosa26* promoter. Our proposed design uses two fluorescent proteins in tandem, allowing the use of fluorescence to quickly screen candidates in a 24-well format. The first protein will be the shortest GFP variant that we can find, after which we will perform synonymous mutations so that it is highly CpG rich so the entire protein is in an artificial CpG island. Subsequent to this, we will embed a canonical PAS motif, so that it lies within the CpG island. Several hundred nucleotides after, we will introduce a strong nucleosome positioning sequence (AA/TT/TA, spaced

10 bps) while transitioning to AT-rich segments that will contain an IRES to drive the expression of mCherry. Hence, increase in the GFP-to-mCherry signal would suggest that premature termination is being used more frequently. Knockdown of the RNA exosome or Pabpn1 will be used to validate the artificial gene. If this model gene is unsuccessful, an alternate approach is to use quantitative real-time PCR to compare an amplicon prior to a known early termination site to an amplicon after. However, this approach is not amenable does not easily scale to large numbers of candidate proteins.

Both approaches coupled with choosing appropriate candidates based on the literature will provide insights into the mechanisms behind establishing productive elongation complexes and the regulation of promoter proximal cleavage. Moreover, this may provide interesting insights into physiological disease. A significant amount of cancers are linked with defects in factors that regulate pausing or mutations that bypass pausing, such as MYC (Rahl et al., 2010), NF-KB (Barboric et al., 2001), MLL and ELL (Luo et al., 2012), and the RNA exosome, which we found promoted promoter-proximal pausing. Mutations in pathways that reduce pausing in the stable nucleosome regions are likely to increase gene expression and promote tumorigenesis.

5.4 U1 snRNA and Nucleosome Turnover

Promoter proximal 5' splice sites have been selected for during evolution (Almada et al., 2013), likely because they are critical for promoting gene expression (Furger et al., 2002). Crosslinking studies show that U1 snRNP binds throughout the gene body, as well as the 5' splice sites marking 5' ends of introns (Engreitz et al., 2014), suggesting U1 has broad roles beyond splicing. Various studies have identified potential non-splicing roles for 5' splice sites in initiation including promoting TFIIH-dependent transcription reinitiation (Kwek et al., 2002) and recruiting the basal transcription machinery (Damgaard et al., 2008). We suggested that U1 may also function

to preventing promoter-proximal termination and decay, through the suppression of PAS motifs within the gene body (Berg et al., 2012; Kaida et al., 2010), potentially through a direct interaction with the polyadenylation machinery (Gunderson et al., 1998; Lutz et al., 1996). However, another study found functional 5' splice sites increase H3K4me3 signal proximal to the promoter (Damgaard et al., 2008), a mark known to promote the association of the chromatin remodeler Chd1through its chromodomain (Flanagan et al., 2005). Moreover, U1 tends to bind before the +1 stable nucleosomes. Hence, we speculate U1 snRNA may also function to reduce the nucleosome barrier. Such an activity would also explain why U1 inhibition promotes early termination, since blocking this activity prevents recruitment of chromatin remodeling complexes, resulting in increased nucleosome stability, slower transcription and more time for Pol II to recognize termination signals.

We propose examining the impact of U1 inhibition on nucleosome occupancy using MNase-seq or nucleosome turnover using assays such as CATCH-IT (Deal et al., 2010). If U1 regulates nucleosome accessibility, there would be increases in nucleosome turnover or MNase sensitivity at genes that prematurely terminate upon U1 inhibition. Moreover, changes in active Pol II occupancy across the gene body can be directly assayed using 3'NT or mNET-seq (Mayer et al., 2015; Weber et al., 2014). Unlike GRO-seq, those two techniques do not depend on a 3'OH of RNA being in the active site of Pol II, so backtracked Pol II can be observed. These assays may reveal a correlation between nascent termination and increased nucleosome occupancy, supporting the role of U1 at regulating the nucleosome barrier.

One caveat is the transcription process normally promotes nucleosome turnover, by evicting nucleosomes as Pol II passes through and then re-depositing it behind. Hence, we would not be able to distinguish between one scenario where a more stable nucleosome promotes Pol II

pausing and an alternate scenario where reduced transcription results in more stable nucleosome association. We found that about two-thirds of actively transcribed genes do not have premature termination upon U1 inhibition. Hence, an examination of changes in nucleosome occupancy for genes without U1-regulated premature termination may demonstrate that U1 snRNA can promote nucleosome turnover. Additionally, segmentation of genes where the distance between the first 5'SS and the PAS motif is large may provide sufficient resolution to determine whether the nucleosomes become more stable prior to the PAS motif being reached. This would further argue that U1 can modulate nucleosome occupancy.

5.5 Mechanisms of uaRNA Degradation

Our work adds additional evidence that multiple pathways converge through the RNA exosome to degrade unwanted transcripts. Pabpn1 knockout resulted in stabilization of uaRNAs that precisely terminate at a PAS motif, in comparison to a broader stabilization of extended uaRNAs in the Exosc3 knockout. These results are consistent with descriptions of two separate pathways to degrade low-abundant RNAs (Meola et al., 2016). In one pathway, transcripts with short A-tails are degraded through the activity of the NEXT complex, comprising hMTR4, ZCCHC8 and RBM7 (Lubas et al., 2011). Alternatively, polyadenylated transcripts are degraded through the PAXT connection involving hMTR4, ZFC3H1 and PABPN1 (Meola et al., 2016). PABPN1 is recruited by the poly(A) tail itself, and targets premature terminated transcripts for exosome decay (Bresson and Conrad, 2013; Bresson et al., 2015; Meola et al., 2016).

Pabpn1 promotes formation of the proper 3' end as well as promoter proximal decay. An important question is why promoter-proximal polyadenylated transcripts are degraded whereas those at the 3' end of transcripts are not, despite both involving Pabpn1. We found sensitivity of transcripts to Pabpn1 removal correlates with the proximity of termination to the TSS. To

determine whether this is a causal relationship, we propose to directly move PAS motifs using CRISPR in the Pabpn1 CKO background. RNA sequences often have stability elements, so the PAS sequences will be replaced rather than deleting/inserting sequences between the PAS and the TSS. In addition, the exosome-interacting RNA helicase hMTR4 associates with the cap binding complex (CBCA) through the adaptor protein ZC3H18 (Andersen et al., 2013; Hallais et al., 2013). This interaction network may be the mechanism that physically holds the RNA exosome in close proximity to the poly(A) tail of prematurely terminated transcripts. Hence, changing the length between the RNA exosome and the cap binding protein might change the sensitivity of uaRNAs to Pabpn1. For instance, overexpressing a fusion of ZFC3H1 with the cap binding protein would shorten the distance and stabilize more early termination products.

Promoter proximity is not the sole criteria for Pabpn1 sensitivity, since specific mRNAs can also be targeted for degradation such as snoRNA host genes or Pabpn1 itself. Pabpn1 autoregulates its own production by binding to a genomically-encoded A tract near the 3' end, which inhibits splicing of the terminal exon and promotes exosome-mediated transcript degradation (Bergeron et al., 2015). One explanation is that there is a kinetic dependence for RNA exosome activity. The RNA exosome mainly functions at the promoter, due to protein-protein associations with the cap binding protein (Andersen et al., 2013; Hallais et al., 2013). At the 3' end, the RNA exosome may be present at lower concentrations. Defects in mRNA processing such as improper splicing or mRNA export defects result in hyperadenylated transcripts retained on chromatin, which may have more time to associate with the RNA exosome in conjunction with poly(A)-binding proteins (Bresson and Conrad, 2013; Hilleren et al., 2001; Lemieux et al., 2011).

We described general properties of genes with detectable premature cleavage clusters, but they arose from different cellular perturbations. Some were dependent on Exosc3 removal, others

upon U1 inhibition, and some only appeared when both were inhibited. Moreover, Pabpn1 knockout also upregulated a subset of clusters, yet many of the clusters upregulated in the Exosc3 knockout that were at the first PAS motif were not upregulated in the Pabpn1 knockout. Hence, we plan to elucidate what may differentiate between these different termination classes. For instance, clusters associated with Pabpn1 responsiveness may be more effectively polyadenylated. In this future study, we plan to incorporate the strength of the PAS motifs, rather than viewing PAS-containing clusters as one class. Moreover, mammalian PAS motifs are comprised of a core hexamer (A[AT]TAAA) and a downstream GU-rich element (Chan et al., 2011). We previously ignored the latter element because the core hexamer is the main contributor to PAS-mediated termination and the GU-rich element is degenerate. We also ignored the contribution of the U1 binding sites; Pabpn1-responsive cleavage clusters in the sense direction may have weaker or more distal 5' splice sites. To investigate the impact of PAS strength and U1 binding on these different types of clusters, PAS motif strengths will be assigned to observed cleavage sites using a computational model and then analyzed using techniques like Principle Component Analysis to determine if there are rules that dictate sensitivity to U1 inhibition, Exosc3 loss and Pabpn1 loss. An analysis of datasets of U1 crosslinking to nascent RNA may also provide important insights (Engreitz et al., 2014).

Exosc3 knockout stabilized both transcripts terminating at a PAS site as well as extended, non-poly(A) transcripts. The RNA exosome degrades RNAs from the 3' end, so these non-polyadenylated uaRNAs must be cleaved prior to degradation by the RNA exosome. One possibility is an alternative cleavage factor known as the Integrator may release RNAs and promote transcript decay. The Integrator promotes the cleavage of enhancer RNAs, whereby knockdown results in accumulation of eRNAs on chromatin (Lai et al., 2015). In addition, depletion of

Integrator subunits result in increased RNA-seq signal upstream of the TSS, suggesting that it may function to cleave uaRNAs (Stadelmayer et al., 2014). Hence, we propose investigating whether Integrator functions to cleave uaRNAs that have not terminated at PAS motifs. One way to study this is to use a 3' end sequencing technique that is independent of poly(A)-tails to assay the differences between knockout of Ints11 (catalytic subunit) compared with knockout of Pabpn1 and Exosc3. We envision this technique would involve ligating a pre-adenylated barcode to the 3' end of capped, non-fragmented RNAs. Additionally, cells will be pulsed with 4-thioU prior to harvesting, to enrich for nascent RNAs rather than cytoplasmic degradation intermediates. Phenotypically, Exosc3 knockout would result in an increase in termination sites with PAS motifs as well as those without PAS motifs. Moreover, termination sites at the PAS motifs should increase in the Pabpn1 knockout but little-to-no effect on other termination sites. Lastly, we predict that the Ints11 knockout see a reduction in the number of non-PAS cleavage events, but have little impact on those at PAS motifs. It may be useful to sequence active transcription (ie. mNET-seq) in Ints11 knockouts or alternatively perform CLIP-seq on Ints11 to validate the use of these cleavage sites in uaRNAs. In addition to Ints11, non-polyadenylated 3' ends can also be generated by Pol II backtracking (Lemay et al., 2014). These may occur within stable-nucleosome containing regions in the antisense direction, so knockdowns of TFIIS may also reveal insights into PABPN1-independent pathways.

5.6 Conclusion

Transcription is one of the most fundamental processes in cells, governing the conversion of genetic information to RNA. Central to this is the use of transcriptional checkpoints that not only ensures proper steps have occurred, they also serve to downregulate unwanted transcription events. The RNA exosome collaborates with U1 snRNP to promote early termination of polyadenylated RNAs at the edge of Stable Nucleosome Free Regions, some of which are also substrates of Pabpn1. These are part of a larger, regulated transcriptional checkpoint in the sense direction, whereby pausing of RNA polymerase is associated with transcription through stable nucleosome-bound chromatin. This event is also associated with Myc binding, suggesting there may be implications in diseases of Myc dysregulation such as cancer. Hence, premature PAS termination near the nucleosome-associated pause site may represent a previously undescribed transcriptional elongation checkpoint regulated by U1 snRNP.

5.7 REFERENCES

Ahn, S.H., Kim, M., and Buratowski, S. (2004). Phosphorylation of serine 2 within the RNA polymerase II C-terminal domain couples transcription and 3' end processing. Mol Cell *13*, 67-76.

Alexander, R.D., Innocente, S.A., Barrass, J.D., and Beggs, J.D. (2010). Splicing-dependent RNA polymerase pausing in yeast. Mol Cell *40*, 582-593.

Almada, A.E., Wu, X., Kriz, A.J., Burge, C.B., and Sharp, P.A. (2013). Promoter directionality is controlled by U1 snRNP and polyadenylation signals. Nature *499*, 360-363.

Andersen, P.R., Domanski, M., Kristiansen, M.S., Storvall, H., Ntini, E., Verheggen, C., Schein, A., Bunkenborg, J., Poser, I., Hallais, M., *et al.* (2013). The human cap-binding complex is functionally connected to the nuclear RNA exosome. Nat Struct Mol Biol *20*, 1367-1376.

Barboric, M., Nissen, R.M., Kanazawa, S., Jabrane-Ferrat, N., and Peterlin, B.M. (2001). NF-kappaB binds P-TEFb to stimulate transcriptional elongation by RNA polymerase II. Mol Cell *8*, 327-337.

Bartkowiak, B., Liu, P., Phatnani, H.P., Fuda, N.J., Cooper, J.J., Price, D.H., Adelman, K., Lis, J.T., and Greenleaf, A.L. (2010). CDK12 is a transcription elongation-associated CTD kinase, the metazoan ortholog of yeast Ctk1. Genes Dev *24*, 2303-2316.

Beaulieu, Y.B., Kleinman, C.L., Landry-Voyer, A.M., Majewski, J., and Bachand, F. (2012). Polyadenylation-dependent control of long noncoding RNA expression by the poly(A)-binding protein nuclear 1. PLoS Genet *8*, e1003078.

Belotserkovskaya, R., Oh, S., Bondarenko, V.A., Orphanides, G., Studitsky, V.M., and Reinberg, D. (2003). FACT facilitates transcription-dependent nucleosome alteration. Science *301*, 1090-1093.

Berg, M.G., Singh, L.N., Younis, I., Liu, Q., Pinto, A.M., Kaida, D., Zhang, Z., Cho, S., Sherrill-Mix, S., Wan, L., *et al.* (2012). U1 snRNP determines mRNA length and regulates isoform expression. Cell *150*, 53-64.

Bergeron, D., Pal, G., Beaulieu, Y.B., Chabot, B., and Bachand, F. (2015). Regulated Intron Retention and Nuclear Pre-mRNA Decay Contribute to PABPN1 Autoregulation. Mol Cell Biol *35*, 2503-2517.

Blazek, D., Kohoutek, J., Bartholomeeusen, K., Johansen, E., Hulinkova, P., Luo, Z., Cimermancic, P., Ule, J., and Peterlin, B.M. (2011). The Cyclin K/Cdk12 complex maintains genomic stability via regulation of expression of DNA damage response genes. Genes Dev *25*, 2158-2172.

Bresson, S.M., and Conrad, N.K. (2013). The human nuclear poly(a)-binding protein promotes RNA hyperadenylation and decay. PLoS Genet *9*, e1003893.

Bresson, S.M., Hunter, O.V., Hunter, A.C., and Conrad, N.K. (2015). Canonical Poly(A) Polymerase Activity Promotes the Decay of a Wide Variety of Mammalian Nuclear RNAs. PLoS Genet *11*, e1005610.

Chan, S., Choi, E.A., and Shi, Y. (2011). Pre-mRNA 3'-end processing complex assembly and function. Wiley Interdiscip Rev RNA *2*, 321-335.

Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A., *et al.* (2013). Multiplex genome engineering using CRISPR/Cas systems. Science *339*, 819-823.

Damgaard, C.K., Kahns, S., Lykke-Andersen, S., Nielsen, A.L., Jensen, T.H., and Kjems, J. (2008). A 5' splice site enhances the recruitment of basal transcription initiation factors in vivo. Mol Cell *29*, 271-278.

Dantonel, J.C., Murthy, K.G., Manley, J.L., and Tora, L. (1997). Transcription factor TFIID recruits factor CPSF for formation of 3' end of mRNA. Nature *389*, 399-402.

de Dieuleveult, M., Yen, K., Hmitou, I., Depaux, A., Boussouar, F., Bou Dargham, D., Jounier, S., Humbertclaude, H., Ribierre, F., Baulard, C., *et al.* (2016). Genome-wide nucleosome specificity and function of chromatin remodellers in ES cells. Nature *530*, 113-116.

Deal, R.B., Henikoff, J.G., and Henikoff, S. (2010). Genome-wide kinetics of nucleosome turnover determined by metabolic labeling of histones. Science *328*, 1161-1164.

Engreitz, J.M., Sirokman, K., McDonel, P., Shishkin, A.A., Surka, C., Russell, P., Grossman, S.R., Chow, A.Y., Guttman, M., and Lander, E.S. (2014). RNA-RNA interactions enable specific targeting of noncoding RNAs to nascent Pre-mRNAs and chromatin sites. Cell *159*, 188-199.

Fenouil, R., Cauchy, P., Koch, F., Descostes, N., Cabeza, J.Z., Innocenti, C., Ferrier, P., Spicuglia, S., Gut, M., Gut, I., *et al.* (2012). CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters. Genome Res *22*, 2399-2408.

Flanagan, J.F., Mi, L.Z., Chruszcz, M., Cymborowski, M., Clines, K.L., Kim, Y., Minor, W., Rastinejad, F., and Khorasanizadeh, S. (2005). Double chromodomains cooperate to recognize the methylated histone H3 tail. Nature *438*, 1181-1185.

Flynn, R.A., Almada, A.E., Zamudio, J.R., and Sharp, P.A. (2011). Antisense RNA polymerase II divergent transcripts are P-TEFb dependent and substrates for the RNA exosome. Proc Natl Acad Sci U S A *108*, 10460-10465.

Furger, A., O'Sullivan, J.M., Binnie, A., Lee, B.A., and Proudfoot, N.J. (2002). Promoter proximal splice sites enhance transcription. Genes Dev *16*, 2792-2799.

Gunderson, S.I., Polycarpou-Schwarz, M., and Mattaj, I.W. (1998). U1 snRNP inhibits pre-mRNA polyadenylation through a direct interaction between U1 70K and poly(A) polymerase. Mol Cell *1*, 255-264.

Hallais, M., Pontvianne, F., Andersen, P.R., Clerici, M., Lener, D., Benbahouche Nel, H., Gostan, T., Vandermoere, F., Robert, M.C., Cusack, S., *et al.* (2013). CBC-ARS2 stimulates 3'-end maturation of multiple RNA families and favors cap-proximal processing. Nat Struct Mol Biol *20*, 1358-1366.

He, Q., Johnston, J., and Zeitlinger, J. (2015). ChIP-nexus enables improved detection of in vivo transcription factor binding footprints. Nat Biotechnol *33*, 395-401.

Hilleren, P., McCarthy, T., Rosbash, M., Parker, R., and Jensen, T.H. (2001). Quality control of mRNA 3'-end processing is linked to the nuclear exosome. Nature *413*, 538-542.

Jin, C., and Felsenfeld, G. (2007). Nucleosome stability mediated by histone variants H3.3 and H2A.Z. Genes Dev *21*, 1519-1529.

Kaida, D., Berg, M.G., Younis, I., Kasim, M., Singh, L.N., Wan, L., and Dreyfuss, G. (2010). U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. Nature *468*, 664-668.

Kwek, K.Y., Murphy, S., Furger, A., Thomas, B., O'Gorman, W., Kimura, H., Proudfoot, N.J., and Akoulitchev, A. (2002). U1 snRNA associates with TFIIH and regulates transcriptional initiation. Nat Struct Biol *9*, 800-805.

Lai, F., Gardini, A., Zhang, A., and Shiekhattar, R. (2015). Integrator mediates the biogenesis of enhancer RNAs. Nature *525*, 399-403.

Laitem, C., Zaborowska, J., Isa, N.F., Kufs, J., Dienstbier, M., and Murphy, S. (2015). CDK9 inhibitors define elongation checkpoints at both ends of RNA polymerase II-transcribed genes. Nat Struct Mol Biol *22*, 396-403.

Lemay, J.F., Larochelle, M., Marguerat, S., Atkinson, S., Bahler, J., and Bachand, F. (2014). The RNA exosome promotes transcription termination of backtracked RNA polymerase II. Nat Struct Mol Biol *21*, 919-926.

Lemieux, C., Marguerat, S., Lafontaine, J., Barbezier, N., Bahler, J., and Bachand, F. (2011). A Pre-mRNA degradation pathway that selectively targets intron-containing genes requires the nuclear poly(A)-binding protein. Mol Cell *44*, 108-119.

Lubas, M., Christensen, M.S., Kristiansen, M.S., Domanski, M., Falkenby, L.G., Lykke-Andersen, S., Andersen, J.S., Dziembowski, A., and Jensen, T.H. (2011). Interaction profiling identifies the human nuclear exosome targeting complex. Mol Cell *43*, 624-637.

Luo, Z., Lin, C., and Shilatifard, A. (2012). The super elongation complex (SEC) family in transcriptional control. Nat Rev Mol Cell Biol *13*, 543-547.

Lutz, C.S., Murthy, K.G., Schek, N., O'Connor, J.P., Manley, J.L., and Alwine, J.C. (1996). Interaction between the U1 snRNP-A protein and the 160-kD subunit of cleavage-polyadenylation specificity factor increases polyadenylation efficiency in vitro. Genes Dev *10*, 325-337.

Mayer, A., di Iulio, J., Maleri, S., Eser, U., Vierstra, J., Reynolds, A., Sandstrom, R., Stamatoyannopoulos, J.A., and Churchman, L.S. (2015). Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution. Cell *161*, 541-554.

Meola, N., Domanski, M., Karadoulama, E., Chen, Y., Gentil, C., Pultz, D., Vitting-Seerup, K., Lykke-Andersen, S., Andersen, J.S., Sandelin, A., *et al.* (2016). Identification of a Nuclear Exosome Decay Pathway for Processed Transcripts. Mol Cell *64*, 520-533.

Ntini, E., Jarvelin, A.I., Bornholdt, J., Chen, Y., Boyd, M., Jorgensen, M., Andersson, R., Hoof, I., Schein, A., Andersen, P.R., *et al.* (2013). Polyadenylation site-induced decay of upstream transcripts enforces promoter directionality. Nat Struct Mol Biol *20*, 923-928.

Orphanides, G., LeRoy, G., Chang, C.H., Luse, D.S., and Reinberg, D. (1998). FACT, a factor that facilitates transcript elongation through nucleosomes. Cell *92*, 105-116.

Pinto, P.A., Henriques, T., Freitas, M.O., Martins, T., Domingues, R.G., Wyrzykowska, P.S., Coelho, P.A., Carmo, A.M., Sunkel, C.E., Proudfoot, N.J., *et al.* (2011). RNA polymerase II kinetics in polo polyadenylation signal selection. EMBO J *30*, 2431-2444.

Preker, P., Almvig, K., Christensen, M.S., Valen, E., Mapendano, C.K., Sandelin, A., and Jensen, T.H. (2011). PROMoter uPstream Transcripts share characteristics with mRNAs and are produced upstream of all three major types of mammalian promoters. Nucleic Acids Res *39*, 7179-7193.

Preker, P., Nielsen, J., Kammler, S., Lykke-Andersen, S., Christensen, M.S., Mapendano, C.K., Schierup, M.H., and Jensen, T.H. (2008). RNA exosome depletion reveals transcription upstream of active human promoters. Science *322*, 1851-1854.

Rahl, P.B., Lin, C.Y., Seila, A.C., Flynn, R.A., McCuine, S., Burge, C.B., Sharp, P.A., and Young, R.A. (2010). c-Myc regulates transcriptional pause release. Cell *141*, 432-445.

Ramirez-Carrozzi, V.R., Braas, D., Bhatt, D.M., Cheng, C.S., Hong, C., Doty, K.R., Black, J.C., Hoffmann, A., Carey, M., and Smale, S.T. (2009). A unifying model for the selective regulation of inducible transcription by CpG islands and nucleosome remodeling. Cell *138*, 114-128.

Schwartz, S., Meshorer, E., and Ast, G. (2009). Chromatin organization marks exon-intron structure. Nat Struct Mol Biol *16*, 990-995.

Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thastrom, A., Field, Y., Moore, I.K., Wang, J.P., and Widom, J. (2006). A genomic code for nucleosome positioning. Nature *442*, 772-778.

Skene, P.J., Hernandez, A.E., Groudine, M., and Henikoff, S. (2014). The nucleosomal barrier to promoter escape by RNA polymerase II is overcome by the chromatin remodeler Chd1. Elife *3*, e02042.

Spies, N., Nielsen, C.B., Padgett, R.A., and Burge, C.B. (2009). Biased chromatin signatures around polyadenylation sites and exons. Mol Cell *36*, 245-254.

Squazzo, S.L., Costa, P.J., Lindstrom, D.L., Kumer, K.E., Simic, R., Jennings, J.L., Link, A.J., Arndt, K.M., and Hartzog, G.A. (2002). The Paf1 complex physically and functionally associates with transcription elongation factors in vivo. EMBO J *21*, 1764-1774.

Stadelmayer, B., Micas, G., Gamot, A., Martin, P., Malirat, N., Koval, S., Raffel, R., Sobhian, B., Severac, D., Rialle, S., *et al.* (2014). Integrator complex regulates NELF-mediated RNA polymerase II pause/release and processivity at coding genes. Nat Commun *5*, 5531.

Weber, C.M., Ramachandran, S., and Henikoff, S. (2014). Nucleosomes are context-specific, H2A.Z-modulated barriers to RNA polymerase. Mol Cell *53*, 819-830.

Yamada, T., Yamaguchi, Y., Inukai, N., Okamoto, S., Mura, T., and Handa, H. (2006). P-TEFb-mediated phosphorylation of hSpt5 C-terminal repeats is critical for processive transcription elongation. Mol Cell *21*, 227-237.

Yonaha, M., and Proudfoot, N.J. (1999). Specific transcriptional pausing activates polyadenylation in a coupled in vitro system. Mol Cell *3*, 593-600.

Thank you for reading my thesis.

Anthony

It is good to have an end to journey toward;
but it is the journey that matters, in the end.

- Ursula K. Le Guin