Promoter directionality is controlled by U1 snRNP and polyadenylation signals in mouse embryonic stem cells

By

Albert E. Almada

B.S. Biological Sciences
University of California at Irvine, 2007

SUBMITTED TO THE DEPARTMENT OF BIOLOGY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTORATE OF PHILOSOPHY IN BIOLOGY

AT THE

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2013

Signature of Author…….....
Albert E. Almada
Department of Biology
August 29, 2013

Certified by……
Phillip A. Sharp
Institute Professor of Biology

Accepted by…….....
Stephen P. Bell
Professor of Biology
Chairman, Biology Graduate Committee

Promoter directionality is controlled by U1 snRNP and polyadenylation signals in mouse embryonic stem cells

By

Albert E. Almada

Submitted to the Department of Biology on August 29, 2013 in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy at the Massachusetts Institute of Technology

Abstract

RNA polymerase II (RNAPII) transcription is a tightly regulated process controlling cell type and state. Advancements in our understanding of how transcription is regulated will provide insight into the mechanisms controlling cell identity, cellular differentiation, and its misregulation in disease. It was generally presumed that RNAPII transcribed in a unidirectional manner to produce a coding mRNA. However, RNAPII has recently been found to initiate transcription upstream and antisense from active gene promoters in mammals and yeast. Although RNAPII initiates divergently from these promoters, efficient RNAPII elongation leading to the production of a full-length, stable, abundant RNA molecule is confined to the coding sense direction. These data suggest an unknown mechanism to suppress transcription from the upstream antisense region of divergent promoters.

In Chapter 2, we describe an analysis of uaRNA at a candidate set of divergent promoters in mouse embryonic stem cells (mESCs). We reveal that upstream antisense RNAs (uaRNAs) are less than 1 kb in size, 5'-capped, heterogeneous at their 3'-ends, and accumulate to 1-4 copies per cell at the steady state. In addition, uaRNA are transcribed with comparable kinetics as their linked mRNA and undergo RNAPII pausing and pause release via the recruitment and activity of P-TEFb. Furthermore, uaRNA have short half-lives (15-20 minutes), likely due to them being targeted for rapid degradation by the RNA exosome. Altogether, these data indicate that the mechanism regulating promoter directionality at divergent promoters occurs after P-TEFb recruitment.

In Chapter 3, we describe a genome-wide analysis to map the 3'-ends of polyadenylated RNAs in mESCs and reveal that uaRNAs terminate through a poly (A) site (PAS)-dependent mechanism shortly after being initiated. Interestingly, we find that an asymmetric distribution of encoded U1 snRNP binding sites (U1 sites or 5' splice sites) and PASs surrounding gene transcription start sites (TSSs) enforce promoter directionality by ensuring uaRNAs are prematurely terminated and likely subsequently degraded. Together, these studies highlight the importance of early splicing signals in producing a full-length coding mRNA, but more importantly, our data reveals that the genomic DNA contains the necessary instructions to read the gene in the correct orientation.

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

Acknowledgments

First and foremost, I would like to thank my advisor, Phil, for his mentorship over the past 5 years and his active involvement in my development as an independent scientist. Thank you for giving me the opportunity to review manuscripts, write grants, give frequent talks in lab meeting, travel to scientific conferences, and for providing a great environment to do "good" science. As I move forward in my scientific career, I will always cherish the time I spent in the Sharp Lab. You pushed me harder than I could have ever done on my own, thank you.

I would also like to thank my labmates, past and present, who have supported me over the years. I have learned a great deal from you and feel honored to have studied in the presence of such intelligent individuals. Amy Seila, for taking me under her wing the first few months in the lab and passing on to me a very fruitful project. Ryan Flynn, for our collaborations and providing a great example of a diligent, technically sound scientist. Jesse Zamudio, for your guidance on the work presented in Chapter 2. Xuebing Wu, you are an incredible scientist with unmatched computational prowess. I have learned a lot from you, about how to think about problems computationally as well as seeing the big picture. Mohini, from the beginning of graduate school you have always supported and encouraged me. Thank you for your friendship. I will miss you greatly as I move on. Allan, thank you for being a great example of a diligent, rigorous, and generous scientist. I have enjoyed our conversations about science and life over the years. Jeremy, thank you for your help on manuscripts, proposals, and presentations. Your feedback has been valuable and I have learned a lot from you. Lastly, I would like to thank my family. Amalia, thank you for all the love and support you have given me these past few years. To my second family, the Arudas, thank you for your encouragement and support. To my Mom, Dad, April, and Chris for being great examples of hard work, perseverance, and love.

Table of Contents

Chapter 1

Introduction

Transcription is a basic process that functions to copy the gene into a messenger RNA (mRNA) that can be transported to the cytoplasm and subsequently translated into a functional protein. Thus, transcriptional regulation is important in defining cell type and state, and its misregulation can lead to a diseased state. For the past 25 years, transcription was thought to proceed in a unidirectional manner from gene promoters. We now know that RNAPII initiates transcription in both directions from gene promoters (Core et al., 2008; Preker et al., 2008; Seila et al., 2008), producing short, low abundant, and unstable RNA in the antisense direction but full-length, stable mRNAs in the coding direction. These initial observations suggested an unknown mechanism to enforce promoter directionality through the suppression of upstream antisense RNA (uaRNA) transcription, which will be the main question addressed in this thesis.

In this introduction I will describe the mechanistic steps involved in transcribing a coding gene. Specifically, I will focus on a description of chromatin and its involvement in gene activation, transcription initiation, transcription elongation, transcription termination, pre-mRNA processing events that occur co-transcriptionally, and a description of the discovery of divergent transcription. Although we have learned a great deal about these processes from studying bacterial systems, I will focus here on eukaryotic transcription providing examples from both yeast and mammals. Prior to this thesis, little was known regarding the origin, structure, and biogenesis of uaRNA or why a full-length, stable, coding mRNA is not produced in the upstream antisense direction of gene promoters. Therefore, in Chapter 2, I will describe a set of experiments aimed to characterize the structure and sequence of uaRNA from a small cohort of divergent promoters. I found that uaRNA share similar characteristics as coding mRNA, in that they are 5'-capped, produced at a similar rate as sense pre-mRNA, regulated by RNAPII pausing, and are even elongated by P-TEFb activities. In contrast to mRNA, uaRNA are less than

1 kb in size, unstable, and can be targeted for degradation by the RNA exosome. In Chapter 3, I find that uaRNA are terminated in a PAS-dependent manner shortly after being initiated. Furthermore, we find that an asymmetric distribution of PAS and U1 sites in the DNA, surrounding gene TSSs, functions to enforce transcriptional directionality at gene promoters by regulating promoter-proximal cleavage and polyadenylation. Altogether, this thesis uncovers a mechanism to suppress upstream antisense transcription at gene promoters and we suggest that the U1-PAS axis may have a broader role in suppressing pervasive transcription outside coding genes.

Chromatin and gene activation

I will begin with a description of chromatin and its role in gene activation. Chromatin is the association between DNA and protein that function to compact the genomic information into the nuclear compartment of the eukaryotic cell. Early cytological studies in liverwort mosses and *Drosophila melanogaster* revealed two types of chromatin, euchromatin and heterochromatin, which can be distinguished under the microscope as less compacted and densely compacted chromatin, respectively (Heitz, 1928; Kornberg and Lorch, 1992). The fundamental repeating unit of chromatin is the nucleosome, which consists of 147 bases of DNA wrapped around an octamer of histone proteins. Each nucleosome consists of two copies of H2A, H2B, H3, and H4 (Kornberg and Lorch, 1992). In addition, each histone has an N-terminal extension, or histone "tail", that contain sites for post-translational modifications such as methylation, acetylation, ubiquitination, phosphorylation, sumoylation, and ribosylation (Suganuma and Workman, 2011). Histone modifications have been proposed to function as a "histone code" (Jenuwein and Allis, 2001) that may either structurally alter chromatin or be interpreted by effector proteins that

function to regulate gene expression (Rando and Chang, 2009). Although there has been advancements in our understanding of chromatin-related mechanisms of gene silencing (Beisel and Paro, 2011), I will focus on mechanisms of gene activation in this introduction.

The first step in initiating transcription, whether unidirectional or divergent, at a promoter is to make the template DNA accessible to RNAPII and general transcription factors (GTFs). This is accomplished by acetylating histone H3 (Brownell et al., 1996), which functions to weaken the association between histones and DNA (Graff and Tsai, 2013; Hebbes et al., 1994; Hebbes et al., 1988) but also serves as a binding site for ATP-dependent chromatin modifiers that act to remove nucleosomes from the template DNA (Clapier and Cairns, 2009; Smith and Peterson, 2005). Gene activation requires more than an accessible promoter free of nucleosomes. In fact, enhancers, or distal regulatory regions often located large distances away, bind multiple transcription factors and loop to contact the promoter with the aid of mediator and cohesin complexes (Kagey et al., 2010; Ong and Corces, 2011). These enhancer-promoter contacts function to provide an efficient platform to recruit transcription factors and RNAPII.

Transcription

In this section, I will describe the basic steps of transcription: initiation, elongation, and termination. In doing so, it will be important to consider each stage as a potential step to differentially regulate sense from antisense transcription at divergent promoters. I will first begin with a brief description on the structure and function of RNAPII, the DNA-dependent RNA polymerase that transcribes all coding mRNAs (Young, 1991).

RNAPII structure and the CTD

There are 3 known DNA-dependent RNA polymerases in mammals and yeast: RNAPI, RNAPII, and RNAPIII. RNAPI transcribes ribosomal RNA, RNAPII transcribes messenger RNA and various classes of noncoding RNAs, and RNAPIII transcribes transfer RNA and other small RNAs. We will focus on the function and activities of RNAPII. RNAPII is composed of a 12-member core enzyme with several members (RPB5, RPB6, RPB8, RPB10, RPB12) shared between all eukaryotic RNA polymerases. In addition, the largest subunit, RPB1, contains a carboxyl terminal domain (CTD) that consists of repeats (52 in mammals) of a heptapeptide sequence: Tyr-Ser-Pro-Thr-Ser-Pro-Ser. The residues in the CTD are frequently modified and these covalent changes are proposed to be read as a "code" that can influence gene expression (Buratowski, 2003).

The RNAPII CTD couples RNA processing to transcription

Current models propose that the CTD may act as a scaffold to link transcription to co-transcriptional processing of pre-mRNAs (Buratowski, 2009; Phatnani and Greenleaf, 2006). The importance of the CTD in RNA processing was first revealed when RNA processing steps such as capping, splicing, and polyadenylation were significantly reduced in the absence of a functional CTD domain (Cho et al., 1997; McCracken et al., 1997a; McCracken et al., 1997b), indicating a direct interaction between the CTD and the RNA processing machinery.

It has become evident that specific post-translational modifications on the CTD, like Ser 5 and Ser 2, play critical roles in recruiting capping, splicing, and polyadenylation factors at the appropriate time during the transcription of pre-mRNA (Buratowski, 2009; Phatnani and Greenleaf, 2006). For example, Ser 5 of the CTD is phosphorylated at the start of transcription

initiation (Cho et al., 2001; Trigon et al., 1998), which leads to recruitment of the capping factors to the CTD (Cho et al., 1998; Ho and Shuman, 1999; Schroeder et al., 2000). Subsequently, Ser 5 is dephosphorylated and Ser 2 of the CTD is phosphorylated, which functions as a platform for splicing and cleavage and polyadenylation factors to bind (Buratowski, 2009; Phatnani and Greenleaf, 2006). More recently, phosphorylation at Ser 7 of the RNAPII CTD has been shown to be important for the removal of Ser 5 and recruitment of the integrator complex, which functions in the 3'-end processing at snRNA genes (Egloff, 2012; Egloff et al., 2012). However, a broader role for this modification in gene regulation is possible and will need to be further explored.

Transcription initiation

The first step in the transcription cycle involves the assembly of GTFs and RNAPII to cis-regulatory elements in the promoter sequence (Smale and Kadonaga, 2003). This process is enhanced by distal regulatory regions, known as enhancers, often located long distances from the genes that they act on (Ong and Corces, 2011). Early in vitro biochemical studies demonstrated that the GTFs, TFIIA, TFIIB, TFIID, TFIIE, TFIIF, TFIIH bind in a sequential manner leading to the recruitment of RNAPII to form the closed initiation complex (Buratowski et al., 1989; Conaway and Conaway, 1993; Zawel and Reinberg, 1993). The open initiation complex is formed when TFIIH begins to unwind the template DNA, forming a transcription bubble where RNAPII begins to synthesize short abortive transcripts (Dvir et al., 1997; Moreland et al., 1999; Tirode et al., 1999). Once RNAPII begins to synthesize RNA transcripts of sufficient length, thus breaking promoter contacts, RNAPII begins to transition to a more elongation competent form.

However, RNAPII undergoes additional barriers that must be overcome before productive elongation can lead to the production of a full-length transcript.

Transcription elongation

It was generally thought that recruitment of GTFs and RNAPII was the rate-limiting step in gene activation (Ptashne and Gann, 1997; Stargell and Struhl, 1996). However, current views now indicate that post-initiation modes of transcriptional regulation are very common and often exploited in cancer (Lin et al., 2012; Loven et al., 2013) and other diseases (Schwartz et al., 2012). In this section, we focus primarily on elongation control through a mechanism of RNAPII pausing, which has emerged as a very common mode of gene regulation in mammals and is a topic of study in this thesis.

Discovery and prevalence of RNAPII pausing Early studies in fruit flies demonstrated that RNAPII encounters barriers shortly downstream of the gene TSS. In particular, studies conducted by John Lis' laboratory demonstrated that RNAPII is associated with the 5'-end of the uninduced *hsp-70* heat shock gene. Specifically, in these experiments it was shown that RNAPII is paused 20-40 bases downstream of the TSS, a process referred to as RNAPII pausing (Gilmour and Lis, 1986; Rasmussen and Lis, 1993; Rougvie and Lis, 1988). Others subsequently found that RNAPII pausing occurs at a considerable number of other *Drosophila melanogaster* genes (Law et al., 1998; Muse et al., 2007; Rougvie and Lis, 1990; Zeitlinger et al., 2007). Insight into the prevalence of post-initiation modes of gene regulation in mammalian cells came from Richard Young's laboratory, who demonstrated that most genes in human cells undergo transcription initiation (Bernstein et al., 2002; Bernstein et al., 2005; Santos-Rosa et al., 2002),

yet only a fraction of these genes produced full-length transcripts and acquired chromatin histone marks indicative of transcription elongation (di- and tri- methylation of lysine 79 and 36 on histone 3, respectively, H3K79me^2 and H3K36me^3) (Bannister et al., 2005; Guenther et al., 2007; Strahl et al., 2002). More recently, the Lis Laboratory developed a high-throughput sequencing technique, Global Run-On sequencing (GRO-seq), to sequence nascent RNAs and found that most active human genes undergo RNAPII pausing downstream of the TSS (Core et al., 2012).

Mechanism of RNA polymerase II pausing and pause release Early studies in the mid-70's revealed 5,6-dichloro-1-B-D-ribofuranosylbenzimidazole (DRB) as a chemical compound capable of inhibiting the production of full-length messenger RNAs (Egyhazi, 1974, 1975, 1976; Sehgal et al., 1976). Interestingly, the effect of DRB on transcription *in vivo* was absent in *in vitro* transcription systems (Chodosh et al., 1989). These data indicated that factors that normally function to restrict full-length transcripts *in vivo* may be absent in these reconstituted transcription systems. Using this *in vitro* transcription system, Hiroshi Handa's group identified the DRB sensitivity-inducing factor (DSIF), composed of Supt4h and Supt5h, as the complex capable of inducing the production of short transcripts (28-32 bases in size) *in vitro* (Wada et al., 1998). Subsequently, the multi-subunit complex negative elongation factor (NELF) was observed to cooperate with DSIF to repress transcription elongation (Yamaguchi et al., 1999). Release from the paused state requires the action of the positive elongation factor (P-TEFb), which phosphorylates the DSIF-NELF complex, promotes the dissociation of NELF, and converts DSIF to an elongation-promoting factor (Marshall and Price, 1992, 1995; Wada et al., 1998). In addition, P-TEFb phosphorylates Ser 2 on the CTD of RNAPII, providing the

necessary signals to promote efficient transcription elongation and RNA processing (Peterlin and Price, 2006). In yeast, P-TEFb-like activities are split between two distinct protein: Ctk1 and Bur1 (Cho et al., 2001; Zhou et al., 2009). Ctk1 is responsible for the bulk of Ser 2 phosphorylation on the CTD, but recent data indicates Bur1 can also contribute to Ser 2 phosphorylation downstream of the promoter (Liu et al., 2009; Qiu et al., 2009). Bur1 mainly functions to phosphorylate DSIF (Zhou et al., 2009). More recently in mammals, two additional Ser 2 kinases (CDK12 and CDK13) have been found to phosphorylate Ser 2 on the CTD, but have been proposed to function downstream of P-TEFb (Bartkowiak et al., 2010). CDK12 and CDK13 have been linked to splicing (Berro et al., 2008; Even et al., 2006), genome stability (Blazek et al., 2011), and embryonic stem cell self-renewal and differentiation (Dai et al., 2012).

Transcription termination

It has become clear that transcription can be terminated through various mechanisms that depend on both signal sequences and the recruitment of specific factors to the 3'-end of RNA transcripts. In this section I will discuss two transcription termination mechanisms: canonical PAS-dependent and the non-canonical Nrd1-Nab3-Sen1 (NNS) pathway. PAS-dependent transcription termination is the most common mode of termination used at coding mRNAs in eukaryotes. PAS-dependent termination begins with first cleavage and polyadenylation of the nascent transcript followed by subsequent release of RNAPII from the genomic DNA template. I will focus this section on the latter step with a more thorough review of the mechanism of cleavage and polyadenylation in the section on co-transcriptional processing.

The importance of a functional PAS site for efficient transcription termination was illustrated in *in vitro* cleavage assays using substrates containing mutated PAS sites (Connelly

and Manley, 1988; Logan et al., 1987; Moore and Sharp, 1984, 1985). From these initial studies two models were proposed. The first model, known as the allosteric or anti-terminator model, proposed that upon transcription through a PAS, there is a conformational change in the elongation complex leading to the dissociation of the elongation factors and association of the termination factors (Logan et al., 1987). The second model, known as the torpedo model, proposed that cleavage of the nascent transcript downstream of the PAS site provided an entry site for a 5' to 3' exonuclease that degrads the tethered RNA and leads to its dissociation from the template DNA (Connelly and Manley, 1988). The torpedo model was strengthened when the 5'-3' exonuclease in yeast (Rat1) and human (Xrn2) was found to promote efficient termination (Kim et al., 2004; West et al., 2004). However, a recent study unifies both models by demonstrating that Rat1 and Xrn2 co-transcriptionally degrade the nascent RNA tethered to RNAPII but also recruit 3'-end formation factors. Interestingly, both activities are important for efficient termination (Luo et al., 2006). Further experimentation will be necessary to refine these models. However, it is clear that efficient transcription termination serves various functional roles in the cell, such as preventing read-through transcription between neighboring genes and promoting transcriptional recycling (Gilmour and Fan, 2008; Richard and Manley, 2009).

The non-canonical NNS pathway, first described in yeast, has been found to generate 3'-ends for snRNAs, snoRNAs, and other non-coding RNAs like cryptic unstable transcripts (CUTs) (Houseley et al., 2006). The NNS termination machinery is composed of Nrd1/Nab3 (RNA binding proteins) and the DNA helicase Sen1 (Kuehner et al., 2011). Termination is triggered by Nrd1 and Nab3 binding to GUA[A/G] and UCUU repeats at the 3'-end of the RNA, respectively (Carroll et al., 2004; Steinmetz and Brow, 1996, 1998). In addition, Nrd1 has been found to interact directly with Ser 5 on the CTD of RNAPII (Kubicek et al., 2012; Vasiljeva et

al., 2008). Recently, Sen1 has been shown to dissociate the RNAPII elongation complex by unwinding the DNA:RNA hybrid in an ATP-dependent manner (Porrua and Libri, 2013). Upon release of the nascent transcript from RNAPII and Sen1, the RNA transcript is targeted for transient polyadenylation by the Trf4/Air2/Mtr4p polyadenylation (TRAMP) and either 3'-end trimmed (sno/snRNAs) or completely degraded by the RNA exosome (CUTs) (Vasiljeva and Buratowski, 2006). How the NNS pathway decides between 3'-end trimming or complete degradation in an exosome-dependent manner is an ongoing question, but it has been proposed that RNA-binding proteins associated with the 3'-end may influence this decision (Houseley et al., 2006; Vasiljeva and Buratowski, 2006). Intriguingly, an Nrd1-Nab3 activity has yet to be described in higher eukaryotes, which may indicate an alternative mechanism to terminate the various classes of noncoding RNAs in these organisms.

Co-transcriptional processes

In the previous sections I have described the basic steps involved in transcription. I will now focus on several pre-mRNA processing steps that occur co-transcriptionally. These RNA processing steps function to stabilized the nascent RNA and promote the efficient transport of the mRNA to the cytoplasm where it can be translated. Any failure in one of these steps leads to rapid destruction of the pre-mRNA by the nuclear decay machineries (Fasken and Corbett, 2009).

5'-capping

The addition of a 7-methylguanosine (m^7G) cap to the 5'-end of the pre-mRNA is the first modification to occur shortly after the initiation of transcription, when RNAPII has begun to

synthesize a transcript roughly 15-20 nucleotides in size. The addition of a 5'-cap structure has been observed to enhance splicing (Konarska et al., 1984; Krainer et al., 1984; Noble et al., 1986), RNA stability, nuclear export, and translation (Moore and Proudfoot, 2009). Most of the work on characterizing the biochemical mechanism of 5'-cap formation has been conducted in *S. cerevisiae* and *S. pombe*. First, the m^7G cap is added to the growing nascent RNA in three enzymatic steps: removal of the gamma phosphate by an RNA triphosphatase, transfer of guanine monophosphate to the RNA di-phosphate end, and methylation at the 7-nitrogen of the guanosine cap by a methyl transferase. In *S. cerevisiae* and *S. pombe*, the three enzymatic steps are performed by 3 distinct proteins (Mao et al., 1995; Shibagaki et al., 1992; Tsukamoto et al., 1997), whereas in metazoans, including mammals, the triphosphatase and guanylytransferase activities are catalyzed by a single enzyme (Pillutla et al., 1998; Tsukamoto et al., 1998; Yamada-Okabe et al., 1998; Yue et al., 1997).

RNA Splicing

RNAPII transcribes primary transcripts that are composed of both coding exons and intervening noncoding introns. It is necessary for these intronic sequences to be removed from the RNA transcript for the mRNA to mature and encode a functional protein. From the initial discovery of split genes in adenovirus (Berget et al., 1977; Chow et al., 1977) a model was proposed for the existence of a sophisticated mechanism and machinery that functions in pre-mRNA splicing. In this section we will describe two pathways utilized to splice pre-mRNA transcripts: the major and minor splicing pathways.

5' and 3' splice sites at major class introns A comparison of genomic and cDNA sequences from the ovalbumin locus revealed common, short sequence elements at the exon/intron boundaries, indicating that cis-elements within the pre-mRNA may promote the splicing reaction (Breathnach et al., 1978). From an analysis that compiled all known splice-junction sequences and calculated the occurrence of each nucleotide at each position, a consensus 5' splice site (5'SS) was determined: (C or A)AG **GU**(A or G)AGU (Mount, 1982) (Figure 1a). The 5'SS consensus sequence involves the last 3 nucleotides in the exon and the first 6 nucleotides in the downstream intron. Almost all higher eukaryotic introns have an invariant GU at their 5'-end. Because the GU positions are the only invariant nucleotides in the 9 base sequence motif (denoted in bold), there are a number of 5'SS derivative sequences that can actively be used in splicing. In fact, computational algorithms have recently been developed to predict the strength of a given 5' splice site variant based on their ability to promote splicing in human pre-mRNAs (Yeo and Burge, 2004). Furthermore, algorithms to predict 5'SS sequences genome-wide have been a valuable tool to predict gene structure across the genome (Faustino and Cooper, 2003).

The functional significance of the 5'SS sequence was determined through mutational analysis and subsequent assaying of the impact on pre-mRNA processing. Experiments using the β-globin gene demonstrated that the 5' most 6 nucleotides in the intron were necessary for splicing (Wieringa et al., 1984). Specifically, mutations in the invariant GU (positions 4 and 5 in the motif) completely inhibited the splicing event (Treisman et al., 1983; Wieringa et al., 1983), whereas mutations at other positions in the 5'SS motif were still capable of splicing to varying degrees (Solnick, 1981; Treisman et al., 1983). Although not as conserved as the 5'SS, mammalian 3' splice site (3'SS) sequences are composed of the following: an invariable AG at the 3' most nucleotides of the intron, an upstream polypyrimidine tract, and a branch point

(Figure 1a). Likewise, mutational analysis demonstrated the functional importance of the 3'SS in catalyzing the splicing reaction (van Santen and Spritz, 1985; Wieringa et al., 1984). In recent years, it has become clear that mutations that disrupt cis- splicing signals can result in alternative protein isoforms or completely inactivate protein products leading to disease (Baralle and Baralle, 2005; Faustino and Cooper, 2003).

The splicing reaction and ribonucleoproteins involved in the major pathway The development of *in vitro* splicing reactions not only provided the initial proof-of concept for the existence of an endogenous splicing activity (Hernandez and Keller, 1983; Krainer et al., 1984) but it also provided the controllable systems to determine the order of steps in the splicing reaction through the isolation of splicing intermediates and products (Padgett et al., 1984; Ruskin et al., 1984). Collectively, these studies described a two-step process of pre-mRNA splicing. First, the pre-mRNA is cleaved at the 5'SS producing two splicing intermediates: a product corresponding to the upstream first exon and another species representing the intervening intron and downstream RNA in a lariat structure. This is due to the 2'-5' phosphodiester linkage between the guanosine at the 5'SS and an adenosine near the 3'-end of the intron. The second step involves cleavage at the 3'SS and subsequent joining of the two exons (Padgett et al., 1984; Ruskin et al., 1984).

The major spliceosome, a complex composed of multiple small ribonucleoproteins (snRNPs), catalyzes the splicing reaction by binding in a sequential manner to the nascent pre-mRNA (Figure 1b). An intron is first recognized, or defined by binding of the U1 snRNP to the 5'SS (Bindereif and Green, 1987; Chabot and Steitz, 1987; Krainer et al., 1984; Mount, 1983; Ruby and Abelson, 1988; Seraphin et al., 1988; Seraphin and Rosbash, 1989), a step that is

greatly enhanced through the binding of SR proteins to upstream exonic splicing enhancer sequences (Kuo et al., 1991; Robberson et al., 1990; Talerico and Berget, 1990; Zhong et al., 2009). The second step involves the binding of the U2 snRNP at the branch point of the 3'SS with the aid of an extrinsic factor known at U2AF, which binds the polypyrimidine tract (Ruskin et al., 1988; Zamore and Green, 1989; Zamore et al., 1992). Subsequently, the U4, U6, and U5 snRNP associate with the complex through interactions with the U1 and U2 snRNP (Cheng and Abelson, 1987; Konarska and Sharp, 1987). Next, a conformational change in the complex results in the destabilization of U1 and U4 from the complex prior to a U6:U2 interaction to form the active site, which facilitates the two successive cleavage steps described above. After the second catalytic step, the pre-mRNA is released from the spliceosome and U2, U5, U6 snRNPs (bound to the lariat) are recycled from subsequent splicing (Wahl et al., 2009).

Minor splicing pathway Most higher eukaryotic pre-mRNA introns are spliced by the major spliceosome described above. However, minor class introns, which represent a small proportion of all introns, are spliced using an alternative spliceosome complex (Patel and Steitz, 2003). Minor class introns are characterized by a highly conserved 5'SS and branch point sequence, distinct from those at major introns, as well as the lack of a polypyrimidine tract at the 3'-end of the intron (Hall and Padgett, 1994). Furthermore, minor class introns require a unique set of ribonucleoprotein complexes to catalyze the splicing reaction. For example, the minor spliceosome is composed of U11 and U12 (functionally similar to the U1 and U2 snRNPs), U4atac and U6atac (functionally similar to the U4 and U6 snRNPs), and the U5 snRNP, which is shared between the minor and major pathways (Hall and Padgett, 1994; Tarn and Steitz, 1996a,

b). Although minor introns are less frequent in the genome, their conservation among metazoans indicates important cellular functions.

Cleavage and polyadenylation

Evidence and function of a poly (A) tail Studies in the early '70s uncovered a unique feature of mammalian mRNAs in that they contained, at the 3'-end terminal sequences, a stretch of poly (A) (Adesnik et al., 1972; Birnboim et al., 1973; Edmonds et al., 1971; Lim and Canellakis, 1970; Mendecki et al., 1972). The poly (A) polymerase was subsequently discovered and shown to catalyze the addition of poly (A) to the end of mRNA (Winters and Edmonds, 1973a, b). Given the unique property of mRNA, Phillip Leder's group devised a method to isolate poly (A) globin mRNA from mammalian red blood cells using chromatography on oligothymidylic acid-cellulose and derivatives of this method would prove useful in isolating various other mRNAs (Brownlee et al., 1973; Mathews et al., 1971; Rosen et al., 1975). In fact, most coding mRNAs contain poly (A) tails, except a subset of histone mRNAs that are rapidly expressed at the beginning of S phase (Marzluff et al., 2008). The poly (A) tail has been shown to be important for RNA stability, mRNA export, and translation (Proudfoot et al., 2002).

Identification of the poly (A) site (PAS) and GU-rich motif Seminal work conducted by Proudfoot and Brownlee described the sequencing of the first six mRNA 3'-ends from rabbit, human, mouse, and chicken (Proudfoot and Brownlee, 1976). These studies revealed a conserved AAUAAA hexamer PAS roughly 20-30 nucleotides from the 3'-terminal poly (A) tail. From these findings, they proposed that the PAS was necessary for cleavage and polyadenylation and likely the first step in promoting efficient transcription termination. Recent studies analyzing

hexamers upstream of expressed sequence tags (ESTs) (Beaudoing et al., 2000; Gautheret et al., 1998; Tian et al., 2005) and 3'-ends tags generated from high-throughput sequencing (Derti et al., 2012; Hoque et al., 2013; Shepard et al., 2011) indicate that most coding mRNAs contain a canonical PAS, AAUAAA, or the common variant, AUUAAA, upstream of the cleavage site. However, 9 additional PAS variant hexamers that are enriched at the expected 21 nucleotides upstream the cleavage site have been identified, suggesting their ability to function as a signal for cleavage and polyadenylation (Beaudoing et al., 2000; Tian et al., 2005). In addition to the PAS, a GU-rich sequence element downstream of the cleavage site and an upstream UGUA motif can enhance 3'-end processing (Brown and Gilmartin, 2003; Gil and Proudfoot, 1984; McDevitt et al., 1984). It is likely that additional cis-elements are necessary for proper 3'-end formation as indicated by a recent study (Hu et al., 2005) and their identification will be important in understanding how cleavage and polyadenylation events are controlled and regulated.

Protein components that catalyze the cleavage and polyadenylation reaction 3'-end maturation involves a two-step process: cleavage of the nascent pre-mRNA downstream of the PAS and subsequent addition of approximately 200 adenines to the 3'-end of the mRNA. Cleavage and polyadenylation is a tightly coupled process in a living cell and the machinery catalyzing these reactions are composed of multiple core components: cleavage and polyadenylation specificity factor (CPSF), cleavage stimulation factor (CstF), cleavage factor I (CFI), cleavage factor II (CFII), and the poly (A) polymerase (PAP). Most of our understanding on the specific functions for each component has been determined using biochemical assays that reconstitute cleavage and polyadenylation *in vitro* (Hart et al., 1985; Moore and Sharp, 1984, 1985). For example, early studies to uncouple these processes demonstrated that an intact

AAUAAA sequence was necessary for cleavage and polyadenylation (Manley et al., 1985; Zarkower et al., 1986) and, through UV-crosslinking experiments, shown to be bound by the cleavage and polyadenylation specificity factor (CPSF) (Keller et al., 1991). In addition, binding of CPSF is greatly enhanced by the binding of the cleavage stimulatory factor (CstF) (Gilmartin and Nevins, 1989; Weiss et al., 1991), which interacts directly with the downstream GU-rich region (MacDonald et al., 1994; Takagaki and Manley, 1997; Wilusz and Shenk, 1988). CFI and CFII are less defined but CFI was initially identified as a factor that is required for the cleavage step (Takagaki et al., 1989), perhaps acting by stabilizing CPSF binding to the PAS (Ruegsegger et al., 1996). More recently, using SELEX technology, CFI was demonstrated to bind aptamers enriched for the UGUA motif (located upstream of the cleavage site as described above) and upon depletion of CFI, PAS-dependent cleavage was inhibited in *in vitro* cleavage assays (Brown and Gilmartin, 2003). The poly (A) polymerase, or PAP, has little poly (A) activity *in vitro* alone, but in the presence of CPSF and poly (A) binding protein nuclear 1 (also known as Pab2), PAP catalyzes the addition of approximately 200 adenines to the 3'-end of pre-mRNA substrates (Christofori and Keller, 1988, 1989; Takagaki et al., 1988; Wahle, 1991; Wahle and Ruegsegger, 1999). Although the mechanism and key players involved in cleavage and polyadenylation have been well characterized, it was not until recently that the component responsible for the endonucleolytic cleavage event (CPSF-73) was identified (Mandel et al., 2006). Furthermore, that a recent study reported over 80 different proteins that interact with the core cleavage and polyadenylation machinery (Shi et al., 2009), highlights the need for future experiments to test the function of these additional factors. It seems probable that some of these factors may link polyadenylation to other processes such as transcription, splicing, gene looping, and mRNA export (Richard and Manley, 2009).

Alternative cleavage and polyadenylation (APA) Given that an appreciable number of human mRNAs contain multiple PAS signals at their 3' terminal ends (Beaudoing et al., 2000; Tian et al., 2005), its not surprising that APA is a major mode of regulating gene expression (Di Giammartino et al., 2011; Elkon et al., 2013; Tian and Manley, 2013). APA refers to the utilization of an alternative PAS in the UTR leading to mRNAs with the same coding region but with different length 3'-UTRs. APA's impact on cell growth and development became apparent in a set of studies that discovered an intimate connection between proliferative status and usage of either a proximal or distal PAS signal in the 3'-UTR of coding genes (Mayr and Bartel, 2009; Sandberg et al., 2008). For example, a seminal study performed by Sandberg and colleagues observed a widespread shortening of 3'-UTR's upon activation of highly proliferative murine $CD4^+$ T lymphocytes and that 3'UTR shortening has global impacts on gene expression (Sandberg et al., 2008). Subsequently, this finding was echoed in additional studies comparing normal and transformed cancer cells in various tissues (Mayr and Bartel, 2009; Morris et al., 2012; Singh et al., 2009), myoblasts to differentiated myotubes (Ji et al., 2009), and even fibroblast to induced pluripotent stem cells (iPSCs) (Ji and Tian, 2009).

It is unclear how APA contributes to a change in the proliferative status of a cell. However, it is clear that regulating the length of the 3'-UTR could impact gene expression since mRNAs contain destabilizing sequences like microRNA binding sites (Bartel, 2009), AU-rich elements (AREs), GU-rich elements (GREs), and Puf protein binding elements (Garneau et al., 2007) in their 3'-UTRs. Consistent with this, through correlations revealed by genome-wide analysis and experimental testing of specific examples in mini-gene constructs, it was determined that mRNAs containing shorter 3'-UTRs evaded microRNA-mediated repression (Mayr and Bartel, 2009; Sandberg et al., 2008). Aside from its impact in altering the length of

UTRs, APA can also result in qualitative changes by activating PAS sites in introns or coding exons that can result in the production of a different protein isoform. For example, intronic cleavage and polyadenylation produces dominant-negative, secreted receptor tyrosine kinases, which influences angiogenesis (Vorlova et al., 2011).

Regulating cleavage and polyadenylation Initial bioinformatic analysis suggested that for genes containing multiple PAS sites in the 3'-UTR, the strongest PAS, in terms of ability to induce cleavage, was often the most distal 3' site (Beaudoing et al., 2000; Tian et al., 2005). This observation led to the hypothesis that regulating the levels of the canonical cleavage and polyadenylation factors may influence whether the proximal (weak) or distal (strong) PAS at the 3'-end is utilized. For example, in the case where cleavage factors are limiting, stronger distal PAS sites would be predicted to be favored. However, when cleavage factor levels are in excess, weaker proximal PAS sites may be utilized. Indeed, this hypothesis is supported by various studies that find a direct correlation between expression of core cleavage and polyadenylation factors and the proliferative status (shorter 3'-UTRs), and that these factors are down-regulated upon differentiation when cells proliferate less and acquire longer 3'-UTR's (Ji et al., 2009; Ji and Tian, 2009). Although the mechanisms controlling core cleavage factor expression are unclear under these cellular conditions, a recent study shows that many core cleavage factors contain binding motifs for E2F family members in their promoters; E2Fs have an established role in controlling proliferation (Elkon et al., 2012).

Recently several alternative mechanisms have been proposed as regulators of APA. For example, knockdown of the poly (A) binding protein nuclear 1 (Pab2) resulted in a shift from the usage of a distal PAS to a more proximal PAS. Pab2 was shown to bind the proximal PAS and

directly compete with the core cleavage machinery under normal conditions (Jenal et al., 2012). Second, the cleavage factor 1_m68 (CFI_m68) has been observed to promote the usage of the more distal PAS, since upon its knockdown there was a widespread shift to the utilization of the more proximal PAS signals (Martin et al., 2012). These results may indicate that CFI_m68 activates the distal PAS through its affinity to the strongest PAS signal in the 3'-UTR. Lastly, slight reductions in the level of functional U1 snRNP in the cell can result in a switch from the proximal to more distal PAS sites at coding genes (Berg et al., 2012), indicating an intimate relationship between U1 snRNP binding and cleavage and polyadenylation. Altogether, there has been an increasing amount of excitement to further define the mechanisms controlling APA, but more studies are needed in order to fully appreciate the complexities of APA in development and disease.

U1 snRNP impacts cleavage and polyadenylation In this section, I plan to elaborate in more detail on the recent discovery linking U1 snRNP binding to the control of cleavage and polyadenylation. The first indication of this mechanism was described in experiments studying the expression of late genes in bovine papillomavirus. Specifically, it was found that a 5'SS (bound by U1 snRNP) upstream of the late gene PAS functioned to inhibit polyadenylation of the late gene transcript (Furth et al., 1994), likely through a direct interaction with U1-70K (a U1 snRNP-associated protein) and PAP (Gunderson et al., 1998). Further support for this mechanism was obtained when U1 snRNP was modified to target mRNA 3'-ends (upstream of PAS in 3'-UTR) leading to gene silencing (Beckley et al., 2001; Goraczniak et al., 2009), presumably by increasing target instability through the inhibition of the polyadenylation process. However, a recent set of studies indicate U1 snRNP binding at 5'SS's at exon/intron boundaries

(or cryptic 5'SS within introns) can function to inhibit cleavage and polyadenylation at intronic PAS sites (Berg et al., 2012; Kaida et al., 2010). For example, seminal work from Gideon Dreyfuss' laboratory revealed extensive premature cleavage and polyadenylation (PCPA) within the first intron of coding genes upon U1 inhibition (Kaida et al., 2010). Subsequent studies from the same group show that PCPA is a conserved feature of metazoans, as it is detected in human, mouse, and flies (Berg et al., 2012). Furthermore, they demonstrate at a single gene that U1 snRNP can suppress a downstream PAS at least 1 kb away (Berg et al., 2012). Together, these data indicate that regulating the levels of functional U1 in the cell may have profound effects on RNA length, transcript isoform, and expression (Berg et al., 2012).

Discovery of divergent transcription

The recent advancements in the development of high-throughput RNA sequencing technologies allowed for the detection of rare transcripts and led to the realization that RNAPII transcribes pervasively throughout the eukaryotic genome to produce both intergenic and genic-associated noncoding RNAs (Berretta and Morillon, 2009; Dinger et al., 2009; Jacquier, 2009; Kapranov et al., 2007). I will focus this section on a class of genic-associated noncoding RNA that result from divergent transcription at mammalian gene promoters.

From the initial description of the basic elements of a gene many years ago, it was presumed that RNAPII was recruited to the gene promoter and began transcription in a unidirectional manner to transcribe a protein-coding gene using the various mechanisms described in the previous sections of this introduction. However, RNAPII has recently been discovered to initiate transcription in the antisense orientation from the promoter of most active coding genes in mammals and yeast, a process referred to as divergent transcription (Core et al.,

2008; Neil et al., 2009; Preker et al., 2008; Seila et al., 2008; Xu et al., 2009). We focus here on the observations described in mammals.

In one study, high-throughput sequencing to profile small RNAs in mouse embryonic stem cells (mESCs) detected a new class of small RNA that are approximately 20 nucleotides in size, contain a 5' phosphate and 3' hydroxyl, and map within 1500 bps from gene TSSs in non-overlapping peaks (separated by roughly 250 bps) on the sense and antisense strand (Seila et al., 2008). These promoter-proximal small RNAs that mapped on the sense and antisense strand were referred to as transcription start site-associated RNAs (TSSa-RNAs). Various other promoter-proximal sense and antisense (with respect to gene TSSs) small RNAs, with similar features as TSSa-RNAs, were described in subsequent studies (Fejes-Toth, 2009; Taft et al., 2009). Northern blot analysis revealed that antisense TSSa-RNAs were low abundant and represented a subset of an RNA population roughly 20-90 nucleotides in size (Seila et al., 2008). Curiously, it was unclear whether antisense TSSa-RNAs represented the 5'-end or an internal sequence of larger precursor RNA products, a topic that will be addressed in Chapter 2.

As a first test to determine whether antisense TSSa-RNAs were synthesized by RNAPII transcribing in the opposite orientation from promoters, Seila and colleagues performed chromatin immunoprecipitation sequencing (ChIP-Seq) to construct genome-wide binding profiles for RNAPII, H3K4me^3 (histone H3 methylation modification at lysine 4 of the histone tail, initiation mark), and H3K79me^2 (histone H3 methylation modification at lysine 79 of the histone tail, productive elongation mark)(Seila et al., 2008). These experiments revealed a peak of RNAPII and H3K4me^3 upstream of the TSS that co-aligned with the peak of antisense TSSa-RNAs, which suggested the existence of an antisense RNAPII transcription event. Direct evidence for an upstream antisense polymerase complex engaged in transcription was revealed in

a concurrent study using a novel technique, Global Run-On sequencing (GRO-Seq), to sequence nascent RNAs. They found evidence for widespread divergent transcription at gene promoters in human lung fibroblasts (Core et al., 2008). These studies provided support for an RNAPII transcription event on the upstream antisense strand of divergent promoters. However, despite RNAPII initiating transcription divergently, productive elongation is confined to the downstream sense direction. This is supported by a lack of histone H3 modifications at lysine 36 and lysine 79 (histone marks indicative of transcription elongation) in the upstream antisense region of divergent promoters (Guenther et al., 2007; Barski et al., 2007; Seila et al., 2008). These data provided the first indication that productive elongation over long distances may be suppressed in the upstream antisense direction of divergent promoters.

A concurrent study from Torben Jenson's laboratory demonstrated that in the absence of the RNA exosome, there was a stabilization of polyadenylated sense and antisense PROMoter uPstream Transcripts (PROMPTs) 1 kb upstream of promoters in human cells (Preker et al., 2011). PROMPTs were suggested to be polyadenylated noncoding transcripts that are roughly 200-600 nucleotides in size, indicating PROMPTs were inefficiently elongated (compared to coding mRNA) prior to their termination. However, in these studies it was unclear as to how PROMPTs were generated, the mechanism leading to their early termination, and their relationship to TSSa-RNAs, if any, since they both arose in distinct locations upstream of promoters (Figure 2).

In this thesis, I aimed to define the mechanism that suppresses the production of full-length, stable mRNAs in the upstream antisense direction of divergent gene promoters in mESCs. To address this question, we first performed a thorough biochemical analysis of the structure and sequence of divergently transcribed upstream antisense RNA and tested whether

uaRNA are transcribed using the same transcriptional mechanisms (described in this introduction) as sense mRNA. In Chapter 2, we found that antisense TSSa-RNA mapped 15-40 bases downstream of the upstream antisense RNA (uaRNA) TSS. uaRNA were less than 1 kb in size, 5'-capped, and contained non-polyadenylated heterogeneous 3'-ends. Furthermore, we found that uaRNA are produced with comparable kinetics as coding mRNA and even undergo RNAPII pausing and pause release. Lastly, we found that uaRNA are targeted for rapid degradation by the RNA exosome. We suggest that the uaRNA cloned and sequenced in this experiment may represent degradation products, given that uaRNA are exosome substrates and that uaRNA ends are heterogeneous and non-polyadenylated. In Chapter 3, we performed poly (A) 3'-end deep sequencing and find that uaRNA are cleavage and polyadenylated at their 3'-end using PAS-dependent mechanisms shortly after being initiated. Given the known role for the U1 snRNP in suppressing downstream PAS signals (Kaida et al., 2009; Berg et al., 2012), we find that an asymmetric distribution of U1 and PAS signals in the DNA sequence flanking gene TSS's function to promote premature cleavage and polyadenylation (and likely subsequent degradation) in the upstream antisense direction of gene promoters. Altogether, these findings indicate promoter directionality is encoded in the DNA-sequence as a U1-PAS axis and may explain why transcription outside coding genes is suppressed.

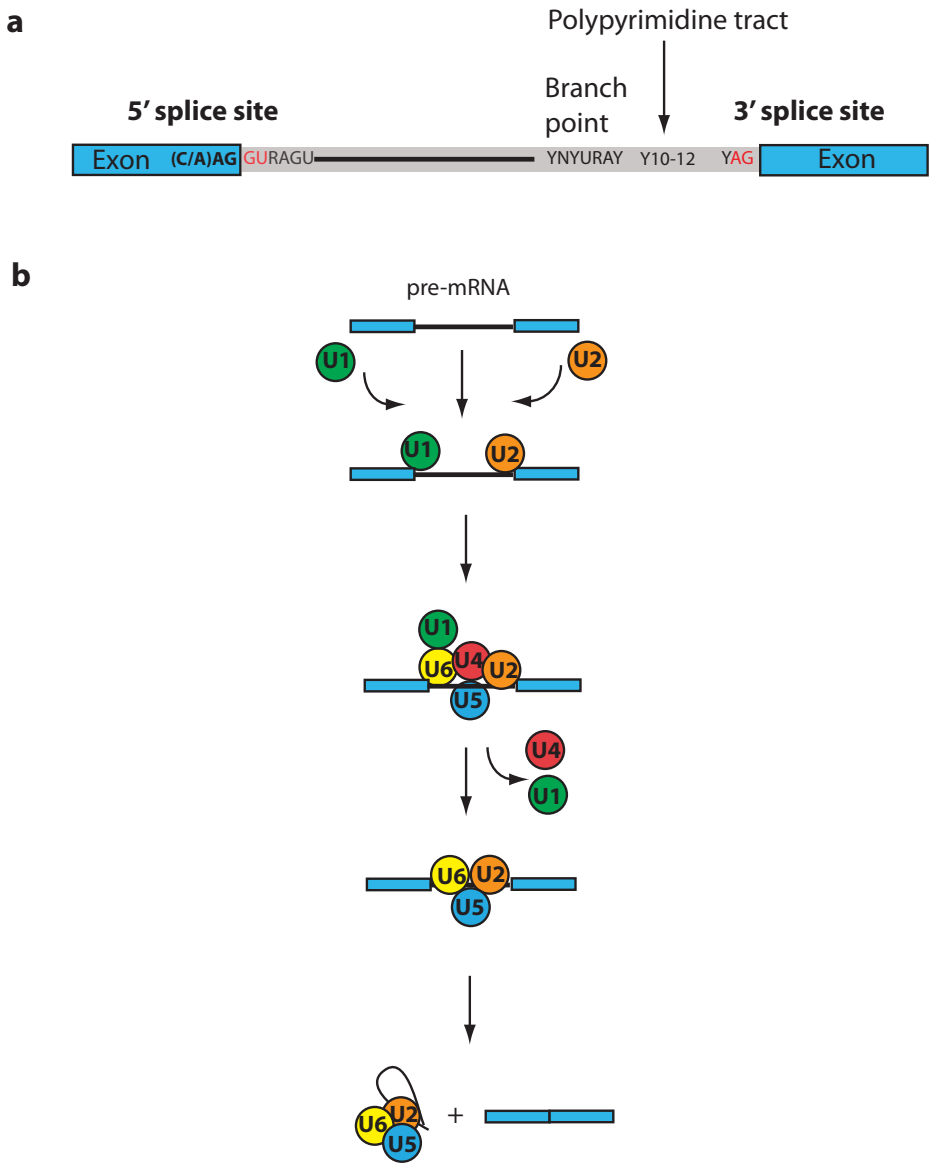


Figure 1. Mechanism of pre-mRNA splicing at a major intron. a. Diagram showing the 5' splice site, branch point, and 3' splice site consensus sequences, where N is any nucleotide, R is a purine, and Y is a pyrimidine. The invariant GU and AG are depicted in red at the 5' splice site and 3' splice site, respectively. The polypyrimidine tract is a pyrimidine-rich stretch between the branch point and the 3' splice site sequence. b. The mechanism of spliceosome assembly at a major intron of a pre-mRNA. U1 and U2 snRNPs bind to the 5' splice site and branch point of the 3' splice site, respectively. Then, U4, U5, and U6 snRNPs assemble to the complex through interactions with the U1 and U2 snRNPs. A conformational change in the complex leads to the U1 and U4 snRNPs leaving the complex and U2, U5, and U6 snRNPs make direct contacts to form the active spliceosome complex. Lastly, two successive cleavage steps result in the joining of the two exons and release of the lariat bound U2, U5, and U6 snRNPs for subsequent rounds of splicing.

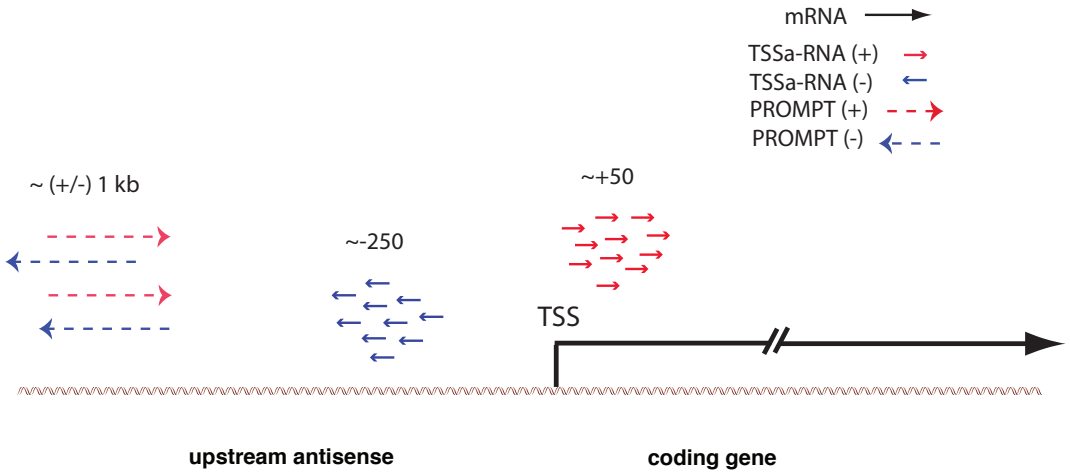


Figure 2. Promoter-proximal noncoding RNAs at divergent promoters. Displaying promoter-proximal sense and antisense TSSa-RNAs (red and blue, respectively), sense and antisense PROMPTs (red and blue, respectively), and the coding mRNA shown with a black arrow pointing towards the right. TSSa-RNAs are approximately 20 nucleotides in size, contain a 5' phosphate and 3' hydroxyl, and map within 1500 bps from gene TSSs in non-overlapping peaks (separated by roughly 250 bps) on the sense and antisense strand (Seila et al., 2008) . In these studies it was unclear whether antisense TSSa-RNAs represented the 5'-end or internal fragments of larger precurosor products, a topic addressed in Chapter 2. PROMPTs are several hundred nucleotides long, contain 3'-adenylated tails (of unknown size), and are stabilized in the absence of the RNA exosome approximately 1 kb upstream of human gene TSSs (Preker et al., 2008). From these intitial studies, it was unclear whether PROMPTs were related to TSSa-RNAs given their distinct locations upstream of gene promoters.

References

Adesnik, M., Salditt, M., Thomas, W., and Darnell, J.E. (1972). Evidence that all messenger RNA molecules (except histone messenger RNA) contain Poly (A) sequences and that the Poly(A) has a nuclear function. J Mol Biol *71*, 21-30.

Bannister, A.J., Schneider, R., Myers, F.A., Thorne, A.W., Crane-Robinson, C., and Kouzarides, T. (2005). Spatial distribution of di- and tri-methyl lysine 36 of histone H3 at active genes. J Biol Chem *280*, 17732-17736.

Baralle, D., and Baralle, M. (2005). Splicing in action: assessing disease causing sequence changes. J Med Genet *42*, 737-748.

Bartel, D.P. (2009). MicroRNAs: target recognition and regulatory functions. Cell *136*, 215-233.

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.

Beaudoing, E., Freier, S., Wyatt, J.R., Claverie, J.M., and Gautheret, D. (2000). Patterns of variant polyadenylation signal usage in human genes. Genome Res *10*, 1001-1010.

Beckley, S.A., Liu, P., Stover, M.L., Gunderson, S.I., Lichtler, A.C., and Rowe, D.W. (2001). Reduction of target gene expression by a modified U1 snRNA. Mol Cell Biol *21*, 2815-2825.

Beisel, C., and Paro, R. (2011). Silencing chromatin: comparing modes and mechanisms. Nat Rev Genet *12*, 123-135.

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.

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., Humphrey, E.L., Erlich, R.L., Schneider, R., Bouman, P., Liu, J.S., Kouzarides, T., and Schreiber, S.L. (2002). Methylation of histone H3 Lys 4 in coding regions of active genes. Proc Natl Acad Sci U S A *99*, 8695-8700.

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.

Berretta, J., and Morillon, A. (2009). Pervasive transcription constitutes a new level of eukaryotic genome regulation. EMBO Rep *10*, 973-982.

Berro, R., Pedati, C., Kehn-Hall, K., Wu, W., Klase, Z., Even, Y., Geneviere, A.M., Ammosova, T., Nekhai, S., and Kashanchi, F. (2008). CDK13, a new potential human immunodeficiency virus type 1 inhibitory factor regulating viral mRNA splicing. J Virol *82*, 7155-7166.

Bindereif, A., and Green, M.R. (1987). An ordered pathway of snRNP binding during mammalian pre-mRNA splicing complex assembly. EMBO J *6*, 2415-2424.

Birnboim, H.C., Mitchel, R.E., and Straus, N.A. (1973). Analysis of long pyrimidine polynucleotides in HeLa cell nuclear DNA: absence of polydeoxythymidylate. Proc Natl Acad Sci U S A *70*, 2189-2192.

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.

Breathnach, R., Benoist, C., O'Hare, K., Gannon, F., and Chambon, P. (1978). Ovalbumin gene: evidence for a leader sequence in mRNA and DNA sequences at the exon-intron boundaries. Proc Natl Acad Sci U S A *75*, 4853-4857.

Brown, K.M., and Gilmartin, G.M. (2003). A mechanism for the regulation of pre-mRNA 3' processing by human cleavage factor Im. Mol Cell *12*, 1467-1476.

Brownell, J.E., Zhou, J., Ranalli, T., Kobayashi, R., Edmondson, D.G., Roth, S.Y., and Allis, C.D. (1996). Tetrahymena histone acetyltransferase A: a homolog to yeast Gcn5p linking histone acetylation to gene activation. Cell *84*, 843-851.

Brownlee, G.G., Cartwright, E.M., Cowan, N.J., Jarvis, J.M., and Milstein, C. (1973). Purification and sequence of messenger RNA for immunoglobulin light chains. Nat New Biol *244*, 236-240.

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

Buratowski, S. (2009). Progression through the RNA polymerase II CTD cycle. Mol Cell *36*, 541-546.

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

Carroll, K.L., Pradhan, D.A., Granek, J.A., Clarke, N.D., and Corden, J.L. (2004). Identification of cis elements directing termination of yeast nonpolyadenylated snoRNA transcripts. Mol Cell Biol *24*, 6241-6252.

Chabot, B., and Steitz, J.A. (1987). Multiple interactions between the splicing substrate and small nuclear ribonucleoproteins in spliceosomes. Mol Cell Biol *7*, 281-293.

Cheng, S.C., and Abelson, J. (1987). Spliceosome assembly in yeast. Genes Dev *1*, 1014-1027.

Cho, E.J., Kobor, M.S., Kim, M., Greenblatt, J., and Buratowski, S. (2001). Opposing effects of Ctk1 kinase and Fcp1 phosphatase at Ser 2 of the RNA polymerase II C-terminal domain. Genes Dev *15*, 3319-3329.

Cho, E.J., Rodriguez, C.R., Takagi, T., and Buratowski, S. (1998). Allosteric interactions between capping enzyme subunits and the RNA polymerase II carboxy-terminal domain. Genes Dev *12*, 3482-3487.

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.

Chodosh, L.A., Fire, A., Samuels, M., and Sharp, P.A. (1989). 5,6-Dichloro-1-beta-D-ribofuranosylbenzimidazole inhibits transcription elongation by RNA polymerase II in vitro. J Biol Chem *264*, 2250-2257.

Chow, L.T., Gelinas, R.E., Broker, T.R., and Roberts, R.J. (1977). An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. Cell *12*, 1-8.

Christofori, G., and Keller, W. (1988). 3' cleavage and polyadenylation of mRNA precursors in vitro requires a poly(A) polymerase, a cleavage factor, and a snRNP. Cell *54*, 875-889.

Christofori, G., and Keller, W. (1989). Poly(A) polymerase purified from HeLa cell nuclear extract is required for both cleavage and polyadenylation of pre-mRNA in vitro. Mol Cell Biol *9*, 193-203.

Clapier, C.R., and Cairns, B.R. (2009). The biology of chromatin remodeling complexes. Annu Rev Biochem *78*, 273-304.

Conaway, R.C., and Conaway, J.W. (1993). General initiation factors for RNA polymerase II. Annu Rev Biochem *62*, 161-190.

Connelly, S., and Manley, J.L. (1988). A functional mRNA polyadenylation signal is required for transcription termination by RNA polymerase II. Genes Dev *2*, 440-452.

Core, L.J., Waterfall, J.J., Gilchrist, D.A., Fargo, D.C., Kwak, H., Adelman, K., and Lis, J.T. (2012). Defining the status of RNA polymerase at promoters. Cell Rep *2*, 1025-1035.

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.

Dai, Q., Lei, T., Zhao, C., Zhong, J., Tang, Y.Z., Chen, B., Yang, J., Li, C., Wang, S., Song, X., *et al.* (2012). Cyclin K-containing kinase complexes maintain self-renewal in murine embryonic stem cells. J Biol Chem *287*, 25344-25352.

Derti, A., Garrett-Engele, P., Macisaac, K.D., Stevens, R.C., Sriram, S., Chen, R., Rohl, C.A., Johnson, J.M., and Babak, T. (2012). A quantitative atlas of polyadenylation in five mammals. Genome Res *22*, 1173-1183.

Di Giammartino, D.C., Nishida, K., and Manley, J.L. (2011). Mechanisms and consequences of alternative polyadenylation. Mol Cell *43*, 853-866.

Dinger, M.E., Amaral, P.P., Mercer, T.R., and Mattick, J.S. (2009). Pervasive transcription of the eukaryotic genome: functional indices and conceptual implications. Brief Funct Genomic Proteomic *8*, 407-423.

Dvir, A., Conaway, R.C., and Conaway, J.W. (1997). A role for TFIIH in controlling the activity of early RNA polymerase II elongation complexes. Proc Natl Acad Sci U S A *94*, 9006-9010.

Edmonds, M., Vaughan, M.H., Jr., and Nakazato, H. (1971). Polyadenylic acid sequences in the heterogeneous nuclear RNA and rapidly-labeled polyribosomal RNA of HeLa cells: possible evidence for a precursor relationship. Proc Natl Acad Sci U S A *68*, 1336-1340.

Egloff, S. (2012). Role of Ser7 phosphorylation of the CTD during transcription of snRNA genes. RNA Biol *9*, 1033-1038.

Egloff, S., Zaborowska, J., Laitem, C., Kiss, T., and Murphy, S. (2012). Ser7 phosphorylation of the CTD recruits the RPAP2 Ser5 phosphatase to snRNA genes. Mol Cell *45*, 111-122.

Egyhazi, E. (1974). A tentative initiation inhibitor of chromosomal heterogeneous RNA synthesis. J Mol Biol *84*, 173-183.

Egyhazi, E. (1975). Inhibition of Balbiani ring RNA synthesis at the initiation level. Proc Natl Acad Sci U S A *72*, 947-950.

Egyhazi, E. (1976). Initiation inhibition and reinitiation of the synthesis of heterogenous nuclear RNA in living cells. Nature *262*, 319-321.

Elkon, R., Drost, J., van Haaften, G., Jenal, M., Schrier, M., Vrielink, J.A., and Agami, R. (2012). E2F mediates enhanced alternative polyadenylation in proliferation. Genome Biol *13*, R59.

Elkon, R., Ugalde, A.P., and Agami, R. (2013). Alternative cleavage and polyadenylation: extent, regulation and function. Nat Rev Genet *14*, 496-506.

Even, Y., Durieux, S., Escande, M.L., Lozano, J.C., Peaucellier, G., Weil, D., and Geneviere, A.M. (2006). CDC2L5, a Cdk-like kinase with RS domain, interacts with the ASF/SF2-associated protein p32 and affects splicing in vivo. J Cell Biochem *99*, 890-904.

Fasken, M.B., and Corbett, A.H. (2009). Mechanisms of nuclear mRNA quality control. RNA Biol *6*, 237-241.

Faustino, N.A., and Cooper, T.A. (2003). Pre-mRNA splicing and human disease. Genes Dev *17*, 419-437.

Fejes-Toth, K. (2009). Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs. Nature *457*, 1028-1032.

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.

Garneau, N.L., Wilusz, J., and Wilusz, C.J. (2007). The highways and byways of mRNA decay. Nat Rev Mol Cell Biol *8*, 113-126.

Gautheret, D., Poirot, O., Lopez, F., Audic, S., and Claverie, J.M. (1998). Alternate polyadenylation in human mRNAs: a large-scale analysis by EST clustering. Genome Res *8*, 524-530.

Gil, A., and Proudfoot, N.J. (1984). A sequence downstream of AAUAAA is required for rabbit beta-globin mRNA 3'-end formation. Nature *312*, 473-474.

Gilmartin, G.M., and Nevins, J.R. (1989). An ordered pathway of assembly of components required for polyadenylation site recognition and processing. Genes Dev *3*, 2180-2190.

Gilmour, D.S., and Fan, R. (2008). Derailing the locomotive: transcription termination. J Biol Chem *283*, 661-664.

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.

Goraczniak, R., Behlke, M.A., and Gunderson, S.I. (2009). Gene silencing by synthetic U1 adaptors. Nat Biotechnol *27*, 257-263.

Graff, J., and Tsai, L.H. (2013). Histone acetylation: molecular mnemonics on the chromatin. Nat Rev Neurosci *14*, 97-111.

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.

Hall, S.L., and Padgett, R.A. (1994). Conserved sequences in a class of rare eukaryotic nuclear introns with non-consensus splice sites. J Mol Biol *239*, 357-365.

Hart, R.P., McDevitt, M.A., and Nevins, J.R. (1985). Poly(A) site cleavage in a HeLa nuclear extract is dependent on downstream sequences. Cell *43*, 677-683.

Hebbes, T.R., Clayton, A.L., Thorne, A.W., and Crane-Robinson, C. (1994). Core histone hyperacetylation co-maps with generalized DNase I sensitivity in the chicken beta-globin chromosomal domain. EMBO J *13*, 1823-1830.

Hebbes, T.R., Thorne, A.W., and Crane-Robinson, C. (1988). A direct link between core histone acetylation and transcriptionally active chromatin. EMBO J *7*, 1395-1402.

Heitz, E. (1928). Das Heterochromatin der Moose. Jahrd Wiss Botanik *69*, 762-818.
Hernandez, N., and Keller, W. (1983). Splicing of in vitro synthesized messenger RNA precursors in HeLa cell extracts. Cell *35*, 89-99.

Ho, C.K., and Shuman, S. (1999). Distinct roles for CTD Ser-2 and Ser-5 phosphorylation in the recruitment and allosteric activation of mammalian mRNA capping enzyme. Mol Cell *3*, 405-411.

Hoque, M., Ji, Z., Zheng, D., Luo, W., Li, W., You, B., Park, J.Y., Yehia, G., and Tian, B. (2013). Analysis of alternative cleavage and polyadenylation by 3' region extraction and deep sequencing. Nat Methods *10*, 133-139.

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

Hu, J., Lutz, C.S., Wilusz, J., and Tian, B. (2005). Bioinformatic identification of candidate cis-regulatory elements involved in human mRNA polyadenylation. RNA *11*, 1485-1493.

Jacquier, A. (2009). The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. Nat Rev Genet *10*, 833-844.

Jenal, M., Elkon, R., Loayza-Puch, F., van Haaften, G., Kuhn, U., Menzies, F.M., Oude Vrielink, J.A., Bos, A.J., Drost, J., Rooijers, K.*, et al.* (2012). The poly(A)-binding protein nuclear 1 suppresses alternative cleavage and polyadenylation sites. Cell *149*, 538-553.

Jenuwein, T., and Allis, C.D. (2001). Translating the histone code. Science *293*, 1074-1080.

Ji, Z., Lee, J.Y., Pan, Z., Jiang, B., and Tian, B. (2009). Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. Proc Natl Acad Sci U S A *106*, 7028-7033.

Ji, Z., and Tian, B. (2009). Reprogramming of 3' untranslated regions of mRNAs by alternative polyadenylation in generation of pluripotent stem cells from different cell types. PLoS One *4*, e8419.

Kagey, M.H., Newman, J.J., Bilodeau, S., Zhan, Y., Orlando, D.A., van Berkum, N.L., Ebmeier, C.C., Goossens, J., Rahl, P.B., Levine, S.S.*, et al.* (2010). Mediator and cohesin connect gene expression and chromatin architecture. Nature *467*, 430-435.

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.

Kapranov, P., Willingham, A.T., and Gingeras, T.R. (2007). Genome-wide transcription and the implications for genomic organization. Nat Rev Genet *8*, 413-423.

Keller, W., Bienroth, S., Lang, K.M., and Christofori, G. (1991). Cleavage and polyadenylation factor CPF specifically interacts with the pre-mRNA 3' processing signal AAUAAA. EMBO J *10*, 4241-4249.

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.

Konarska, M.M., Padgett, R.A., and Sharp, P.A. (1984). Recognition of cap structure in splicing in vitro of mRNA precursors. Cell *38*, 731-736.

Konarska, M.M., and Sharp, P.A. (1987). Interactions between small nuclear ribonucleoprotein particles in formation of spliceosomes. Cell *49*, 763-774.

Kornberg, R.D., and Lorch, Y. (1992). Chromatin structure and transcription. Annu Rev Cell Biol *8*, 563-587.

Krainer, A.R., Maniatis, T., Ruskin, B., and Green, M.R. (1984). Normal and mutant human beta-globin pre-mRNAs are faithfully and efficiently spliced in vitro. Cell *36*, 993-1005.

Kubicek, K., Cerna, H., Holub, P., Pasulka, J., Hrossova, D., Loehr, F., Hofr, C., Vanacova, S., and Stefl, R. (2012). Serine phosphorylation and proline isomerization in RNAP II CTD control recruitment of Nrd1. Genes Dev *26*, 1891-1896.

Kuehner, J.N., Pearson, E.L., and Moore, C. (2011). Unravelling the means to an end: RNA polymerase II transcription termination. Nat Rev Mol Cell Biol *12*, 283-294.

Kuo, H.C., Nasim, F.H., and Grabowski, P.J. (1991). Control of alternative splicing by the differential binding of U1 small nuclear ribonucleoprotein particle. Science *251*, 1045-1050.

Law, A., Hirayoshi, K., O'Brien, T., and Lis, J.T. (1998). Direct cloning of DNA that interacts in vivo with a specific protein: application to RNA polymerase II and sites of pausing in Drosophila. Nucleic Acids Res *26*, 919-924.

Lim, L., and Canellakis, E.S. (1970). Adenine-rich polymer associated with rabbit reticulocyte messenger RNA. Nature *227*, 710-712.

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.

Logan, J., Falck-Pedersen, E., Darnell, J.E., Jr., and Shenk, T. (1987). A poly(A) addition site and a downstream termination region are required for efficient cessation of transcription by RNA polymerase II in the mouse beta maj-globin gene. Proc Natl Acad Sci U S A *84*, 8306-8310.

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.

Luo, W., Johnson, A.W., and Bentley, D.L. (2006). The role of Rat1 in coupling mRNA 3'-end processing to transcription termination: implications for a unified allosteric-torpedo model. Genes Dev *20*, 954-965.

MacDonald, C.C., Wilusz, J., and Shenk, T. (1994). The 64-kilodalton subunit of the CstF polyadenylation factor binds to pre-mRNAs downstream of the cleavage site and influences cleavage site location. Mol Cell Biol *14*, 6647-6654.

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.

Manley, J.L., Yu, H., and Ryner, L. (1985). RNA sequence containing hexanucleotide AAUAAA directs efficient mRNA polyadenylation in vitro. Mol Cell Biol *5*, 373-379.

Mao, X., Schwer, B., and Shuman, S. (1995). Yeast mRNA cap methyltransferase is a 50-kilodalton protein encoded by an essential gene. Mol Cell Biol *15*, 4167-4174.

Marshall, N.F., and Price, D.H. (1992). Control of formation of two distinct classes of RNA polymerase II elongation complexes. Mol Cell Biol *12*, 2078-2090.

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.

Martin, G., Gruber, A.R., Keller, W., and Zavolan, M. (2012). Genome-wide analysis of pre-mRNA 3' end processing reveals a decisive role of human cleavage factor I in the regulation of 3' UTR length. Cell Rep *1*, 753-763.

Marzluff, W.F., Wagner, E.J., and Duronio, R.J. (2008). Metabolism and regulation of canonical histone mRNAs: life without a poly(A) tail. Nat Rev Genet *9*, 843-854.

Mathews, M.B., Osborn, M., and Lingrel, J.B. (1971). Translation of globin messenger RNA in a heterologous cell-free system. Nature *233*, 206-209.

Mayr, C., and Bartel, D.P. (2009). Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. Cell *138*, 673-684.

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.

McDevitt, M.A., Imperiale, M.J., Ali, H., and Nevins, J.R. (1984). Requirement of a downstream sequence for generation of a poly(A) addition site. Cell *37*, 993-999.

Mendecki, J., Lee, S.Y., and Brawerman, G. (1972). Characteristics of the polyadenylic acid segment associated with messenger ribonucleic acid in mouse sarcoma 180 ascites cells. Biochemistry *11*, 792-798.

Moore, C.L., and Sharp, P.A. (1984). Site-specific polyadenylation in a cell-free reaction. Cell *36*, 581-591.

Moore, C.L., and Sharp, P.A. (1985). Accurate cleavage and polyadenylation of exogenous RNA substrate. Cell *41*, 845-855.

Moore, M.J., and Proudfoot, N.J. (2009). Pre-mRNA processing reaches back to transcription and ahead to translation. Cell *136*, 688-700.

Moreland, R.J., Tirode, F., Yan, Q., Conaway, J.W., Egly, J.M., and Conaway, R.C. (1999). A role for the TFIIH XPB DNA helicase in promoter escape by RNA polymerase II. J Biol Chem *274*, 22127-22130.

Morris, A.R., Bos, A., Diosdado, B., Rooijers, K., Elkon, R., Bolijn, A.S., Carvalho, B., Meijer, G.A., and Agami, R. (2012). Alternative cleavage and polyadenylation during colorectal cancer development. Clin Cancer Res *18*, 5256-5266.

Mount, S.M. (1982). A catalogue of splice junction sequences. Nucleic Acids Res *10*, 459-472.

Mount, S.M. (1983). RNA processing. Sequences that signal where to splice. Nature *304*, 309-310.

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.

Noble, J.C., Prives, C., and Manley, J.L. (1986). In vitro splicing of simian virus 40 early pre mRNA. Nucleic Acids Res *14*, 1219-1235.

Ong, C.T., and Corces, V.G. (2011). Enhancer function: new insights into the regulation of tissue-specific gene expression. Nat Rev Genet *12*, 283-293.

Padgett, R.A., Konarska, M.M., Grabowski, P.J., Hardy, S.F., and Sharp, P.A. (1984). Lariat RNA's as intermediates and products in the splicing of messenger RNA precursors. Science *225*, 898-903.

Patel, A.A., and Steitz, J.A. (2003). Splicing double: insights from the second spliceosome. Nat Rev Mol Cell Biol *4*, 960-970.

Peterlin, B.M., and Price, D.H. (2006). Controlling the elongation phase of transcription with P-TEFb. Mol Cell *23*, 297-305.

Phatnani, H.P., and Greenleaf, A.L. (2006). Phosphorylation and functions of the RNA polymerase II CTD. Genes Dev *20*, 2922-2936.

Pillutla, R.C., Shimamoto, A., Furuichi, Y., and Shatkin, A.J. (1998). Human mRNA capping enzyme (RNGTT) and cap methyltransferase (RNMT) map to 6q16 and 18p11.22-p11.23, respectively. Genomics *54*, 351-353.

Porrua, O., and Libri, D. (2013). A bacterial-like mechanism for transcription termination by the Sen1p helicase in budding yeast. Nat Struct Mol Biol *20*, 884-891.

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.

Proudfoot, N.J., Furger, A., and Dye, M.J. (2002). Integrating mRNA processing with transcription. Cell *108*, 501-512.

Ptashne, M., and Gann, A. (1997). Transcriptional activation by recruitment. Nature *386*, 569-577.

Qiu, H., Hu, C., and Hinnebusch, A.G. (2009). Phosphorylation of the Pol II CTD by KIN28 enhances BUR1/BUR2 recruitment and Ser2 CTD phosphorylation near promoters. Mol Cell *33*, 752-762.

Rando, O.J., and Chang, H.Y. (2009). Genome-wide views of chromatin structure. Annu Rev Biochem *78*, 245-271.

Rasmussen, E.B., and Lis, J.T. (1993). In vivo transcriptional pausing and cap formation on three Drosophila heat shock genes. Proc Natl Acad Sci U S A *90*, 7923-7927.

Richard, P., and Manley, J.L. (2009). Transcription termination by nuclear RNA polymerases. Genes Dev *23*, 1247-1269.

Robberson, B.L., Cote, G.J., and Berget, S.M. (1990). Exon definition may facilitate splice site selection in RNAs with multiple exons. Mol Cell Biol *10*, 84-94.

Rosen, J.M., Woo, S.L., Holder, J.W., Means, A.R., and O'Malley, B.W. (1975). Preparation and preliminary characterization of purified ovalbumin messenger RNA from the hen oviduct. Biochemistry *14*, 69-78.

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.

Rougvie, A.E., and Lis, J.T. (1990). Postinitiation transcriptional control in Drosophila melanogaster. Mol Cell Biol *10*, 6041-6045.

Ruby, S.W., and Abelson, J. (1988). An early hierarchic role of U1 small nuclear ribonucleoprotein in spliceosome assembly. Science *242*, 1028-1035.

Ruegsegger, U., Beyer, K., and Keller, W. (1996). Purification and characterization of human cleavage factor Im involved in the 3' end processing of messenger RNA precursors. J Biol Chem *271*, 6107-6113.

Ruskin, B., Krainer, A.R., Maniatis, T., and Green, M.R. (1984). Excision of an intact intron as a novel lariat structure during pre-mRNA splicing in vitro. Cell *38*, 317-331.

Ruskin, B., Zamore, P.D., and Green, M.R. (1988). A factor, U2AF, is required for U2 snRNP binding and splicing complex assembly. Cell *52*, 207-219.

Sandberg, R., Neilson, J.R., Sarma, A., Sharp, P.A., and Burge, C.B. (2008). Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. Science *320*, 1643-1647.

Santos-Rosa, H., Schneider, R., Bannister, A.J., Sherriff, J., Bernstein, B.E., Emre, N.C., Schreiber, S.L., Mellor, J., and Kouzarides, T. (2002). Active genes are tri-methylated at K4 of histone H3. Nature *419*, 407-411.

Schroeder, S.C., Schwer, B., Shuman, S., and Bentley, D. (2000). Dynamic association of capping enzymes with transcribing RNA polymerase II. Genes Dev *14*, 2435-2440.

Schwartz, J.C., Ebmeier, C.C., Podell, E.R., Heimiller, J., Taatjes, D.J., and Cech, T.R. (2012). FUS binds the CTD of RNA polymerase II and regulates its phosphorylation at Ser2. Genes Dev *26*, 2690-2695.

Sehgal, P.B., Tamm, I., and Vilcek, J. (1976). Regulation of human interferon production. II. Inhibition of interferon messenger RNA synthesis by 5, 6-dichloro-1-beta-D-ribofuranosylbenzimidazole. Virology *70*, 542-544.

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.

Seraphin, B., Kretzner, L., and Rosbash, M. (1988). A U1 snRNA:pre-mRNA base pairing interaction is required early in yeast spliceosome assembly but does not uniquely define the 5' cleavage site. EMBO J *7*, 2533-2538.

Seraphin, B., and Rosbash, M. (1989). Identification of functional U1 snRNA-pre-mRNA complexes committed to spliceosome assembly and splicing. Cell *59*, 349-358.

Shepard, P.J., Choi, E.A., Lu, J., Flanagan, L.A., Hertel, K.J., and Shi, Y. (2011). Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. RNA *17*, 761-772.

Shi, Y., Di Giammartino, D.C., Taylor, D., Sarkeshik, A., Rice, W.J., Yates, J.R., 3rd, Frank, J., and Manley, J.L. (2009). Molecular architecture of the human pre-mRNA 3' processing complex. Mol Cell *33*, 365-376.

Shibagaki, Y., Itoh, N., Yamada, H., Nagata, S., and Mizumoto, K. (1992). mRNA capping enzyme. Isolation and characterization of the gene encoding mRNA guanylytransferase subunit from Saccharomyces cerevisiae. J Biol Chem *267*, 9521-9528.

Singh, P., Alley, T.L., Wright, S.M., Kamdar, S., Schott, W., Wilpan, R.Y., Mills, K.D., and Graber, J.H. (2009). Global changes in processing of mRNA 3' untranslated regions characterize clinically distinct cancer subtypes. Cancer Res *69*, 9422-9430.

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

Smith, C.L., and Peterson, C.L. (2005). ATP-dependent chromatin remodeling. Curr Top Dev Biol *65*, 115-148.

Solnick, D. (1981). An adenovirus mutant defective in splicing RNA from early region 1A. Nature *291*, 508-510.

Stargell, L.A., and Struhl, K. (1996). A new class of activation-defective TATA-binding protein mutants: evidence for two steps of transcriptional activation in vivo. Mol Cell Biol *16*, 4456-4464.

Steinmetz, E.J., and Brow, D.A. (1996). Repression of gene expression by an exogenous sequence element acting in concert with a heterogeneous nuclear ribonucleoprotein-like protein, Nrd1, and the putative helicase Sen1. Mol Cell Biol *16*, 6993-7003.

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.

Strahl, B.D., Grant, P.A., Briggs, S.D., Sun, Z.W., Bone, J.R., Caldwell, J.A., Mollah, S., Cook, R.G., Shabanowitz, J., Hunt, D.F.*, et al.* (2002). Set2 is a nucleosomal histone H3-selective methyltransferase that mediates transcriptional repression. Mol Cell Biol *22*, 1298-1306.

Suganuma, T., and Workman, J.L. (2011). Signals and combinatorial functions of histone modifications. Annu Rev Biochem *80*, 473-499.

Taft, R.J., Kaplan, C.D., Simons, C., and Mattick, J.S. (2009). Evolution, biogenesis and function of promoter-associated RNAs. Cell Cycle *8*, 2332-2338.

Takagaki, Y., and Manley, J.L. (1997). RNA recognition by the human polyadenylation factor CstF. Mol Cell Biol *17*, 3907-3914.

Takagaki, Y., Ryner, L.C., and Manley, J.L. (1988). Separation and characterization of a poly(A) polymerase and a cleavage/specificity factor required for pre-mRNA polyadenylation. Cell *52*, 731-742.

Takagaki, Y., Ryner, L.C., and Manley, J.L. (1989). Four factors are required for 3'-end cleavage of pre-mRNAs. Genes Dev *3*, 1711-1724.

Talerico, M., and Berget, S.M. (1990). Effect of 5' splice site mutations on splicing of the preceding intron. Mol Cell Biol *10*, 6299-6305.

Tarn, W.Y., and Steitz, J.A. (1996a). Highly diverged U4 and U6 small nuclear RNAs required for splicing rare AT-AC introns. Science *273*, 1824-1832.

Tarn, W.Y., and Steitz, J.A. (1996b). A novel spliceosome containing U11, U12, and U5 snRNPs excises a minor class (AT-AC) intron in vitro. Cell *84*, 801-811.

Tian, B., Hu, J., Zhang, H., and Lutz, C.S. (2005). A large-scale analysis of mRNA polyadenylation of human and mouse genes. Nucleic Acids Res *33*, 201-212.

Tian, B., and Manley, J.L. (2013). Alternative cleavage and polyadenylation: the long and short of it. Trends Biochem Sci *38*, 312-320.

Tirode, F., Busso, D., Coin, F., and Egly, J.M. (1999). Reconstitution of the transcription factor TFIIH: assignment of functions for the three enzymatic subunits, XPB, XPD, and cdk7. Mol Cell *3*, 87-95.

Treisman, R., Orkin, S.H., and Maniatis, T. (1983). Specific transcription and RNA splicing defects in five cloned beta-thalassaemia genes. Nature *302*, 591-596.

Trigon, S., Serizawa, H., Conaway, J.W., Conaway, R.C., Jackson, S.P., and Morange, M. (1998). Characterization of the residues phosphorylated in vitro by different C-terminal domain kinases. J Biol Chem *273*, 6769-6775.

Tsukamoto, T., Shibagaki, Y., Imajoh-Ohmi, S., Murakoshi, T., Suzuki, M., Nakamura, A., Gotoh, H., and Mizumoto, K. (1997). Isolation and characterization of the yeast mRNA capping enzyme beta subunit gene encoding RNA 5'-triphosphatase, which is essential for cell viability. Biochem Biophys Res Commun *239*, 116-122.

Tsukamoto, T., Shibagaki, Y., Niikura, Y., and Mizumoto, K. (1998). Cloning and characterization of three human cDNAs encoding mRNA (guanine-7-)-methyltransferase, an mRNA cap methylase. Biochem Biophys Res Commun *251*, 27-34.

van Santen, V.L., and Spritz, R.A. (1985). mRNA precursor splicing in vivo: sequence requirements determined by deletion analysis of an intervening sequence. Proc Natl Acad Sci U S A *82*, 2885-2889.

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.

Vasiljeva, L., Kim, M., Mutschler, H., Buratowski, S., and Meinhart, A. (2008). The Nrd1-Nab3-Sen1 termination complex interacts with the Ser5-phosphorylated RNA polymerase II C-terminal domain. Nat Struct Mol Biol *15*, 795-804.

Vorlova, S., Rocco, G., Lefave, C.V., Jodelka, F.M., Hess, K., Hastings, M.L., Henke, E., and Cartegni, L. (2011). Induction of antagonistic soluble decoy receptor tyrosine kinases by intronic polyA activation. Mol Cell *43*, 927-939.

Wada, T., Takagi, T., Yamaguchi, Y., Ferdous, A., Imai, T., Hirose, S., Sugimoto, S., Yano, K., Hartzog, G.A., Winston, F.*, et al.* (1998). 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.

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

Wahle, E. (1991). A novel poly(A)-binding protein acts as a specificity factor in the second phase of messenger RNA polyadenylation. Cell *66*, 759-768.

Wahle, E., and Ruegsegger, U. (1999). 3'-End processing of pre-mRNA in eukaryotes. FEMS Microbiol Rev *23*, 277-295.

Weiss, E.A., Gilmartin, G.M., and Nevins, J.R. (1991). Poly(A) site efficiency reflects the stability of complex formation involving the downstream element. EMBO J *10*, 215-219.

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.

Wieringa, B., Hofer, E., and Weissmann, C. (1984). A minimal intron length but no specific internal sequence is required for splicing the large rabbit beta-globin intron. Cell *37*, 915-925.

Wieringa, B., Meyer, F., Reiser, J., and Weissmann, C. (1983). Unusual splice sites revealed by mutagenic inactivation of an authentic splice site of the rabbit beta-globin gene. Nature *301*, 38-43.

Wilusz, J., and Shenk, T. (1988). A 64 kd nuclear protein binds to RNA segments that include the AAUAAA polyadenylation motif. Cell *52*, 221-228.

Winters, M.A., and Edmonds, M. (1973a). A poly(A) polymerase from calf thymus. Characterization of the reaction product and the primer requirement. J Biol Chem *248*, 4763-4768.

Winters, M.A., and Edmonds, M. (1973b). A poly(A) polymerase from calf thymus. Purification and properities of the enzyme. J Biol Chem *248*, 4756-4762.

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-Okabe, T., Doi, R., Shimmi, O., Arisawa, M., and Yamada-Okabe, H. (1998). Isolation and characterization of a human cDNA for mRNA 5'-capping enzyme. Nucleic Acids Res *26*, 1700-1706.

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.

Yeo, G., and Burge, C.B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J Comput Biol *11*, 377-394.

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

Yue, Z., Maldonado, E., Pillutla, R., Cho, H., Reinberg, D., and Shatkin, A.J. (1997). Mammalian capping enzyme complements mutant Saccharomyces cerevisiae lacking mRNA guanylyltransferase and selectively binds the elongating form of RNA polymerase II. Proc Natl Acad Sci U S A *94*, 12898-12903.

Zamore, P.D., and Green, M.R. (1989). Identification, purification, and biochemical characterization of U2 small nuclear ribonucleoprotein auxiliary factor. Proc Natl Acad Sci U S A *86*, 9243-9247.

Zamore, P.D., Patton, J.G., and Green, M.R. (1992). Cloning and domain structure of the mammalian splicing factor U2AF. Nature *355*, 609-614.

Zarkower, D., Stephenson, P., Sheets, M., and Wickens, M. (1986). The AAUAAA sequence is required both for cleavage and for polyadenylation of simian virus 40 pre-mRNA in vitro. Mol Cell Biol *6*, 2317-2323.

Zawel, L., and Reinberg, D. (1993). Initiation of transcription by RNA polymerase II: a multi-step process. Prog Nucleic Acid Res Mol Biol *44*, 67-108.

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.

Zhong, X.Y., Wang, P., Han, J., Rosenfeld, M.G., and Fu, X.D. (2009). SR proteins in vertical integration of gene expression from transcription to RNA processing to translation. Mol Cell *35*, 1-10.

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.

CHAPTER 2

Antisense RNA polymerase II divergent transcripts are P-TEFb dependent and substrates for the RNA exosome

The material in this chapter was adapted with permission from the following publication:

Flynn RA*, Almada AE*, Zamudio JR, Sharp PA. (2011) Antisense RNA polymerase II divergent transcripts are P-TEFb dependent and substrates for the RNA exosome. *Proc Natl Acad Sci* USA 108(26):10460-10465

* These authors contributed equally to this manuscript

Author contributions:

R.A.F. performed qPCR, shRNA knockdowns, and drug treatments. R.A.F designed research, analyzed data and wrote the manuscript

A.E.A. performed 5' and 3'-end cloning and sequencing (with RACE) and 3'-RACE southern blots in wildtype and exosome-deficient mESCs. A.E.A designed research, analyzed data, and wrote the manuscript.

J.R.Z performed qPCR, designed research, analyzed data, and wrote the manuscript

P.A.S supervised research, analyzed data, and wrote the manuscript

Abstract

Divergent transcription occurs at the majority of RNA polymerase II (RNAPII) promoters in mouse embryonic stem cells (mESCs) and this activity correlates with CpG islands. Here we report the characterization of upstream antisense transcription in regions encoding transcription start site associated RNAs (TSSa-RNAs) at four divergent CpG island promoters: Isg20l1, Tcea1, Txn1, and Sf3b1. We find that upstream antisense RNAs (uaRNAs) have distinct capped 5′ termini and heterogeneous non-polyadenylated 3′ ends. uaRNAs are short-lived with average half-lives of 18 minutes and are present at 1-4 copies per cell, approximately one RNA per DNA template. Exosome depletion stabilizes uaRNAs. These uaRNAs are probably initiation products since their capped termini correlate with peaks of paused RNAPII. The pausing factors NELF and DSIF are associated with these antisense polymerases and their sense partners. Knockdown of either NELF or DSIF results in an increase in the levels of uaRNAs. Consistent with P-TEFb controlling release from pausing, treatment with its inhibitor, flavopiridol, decreases uaRNA and nascent mRNA transcripts with similar kinetics. Finally, Isg20l1 induction reveals equivalent increases in transcriptional activity in sense and antisense directions. Together these data show divergent polymerases are regulated after P-TEFb recruitment with uaRNA levels controlled by the exosome.

Introduction

RNA polymerase II (RNAPII) transcription is a highly regulated process controlling cell type and state. Recruitment of chromatin modifying factors and RNAPII to promoters by DNA binding transcription factors are key regulatory steps (Hochheimer and Tjian, 2003; Kuras and Struhl, 1999; Ptashne and Gann, 1997; Roeder, 2005). However, genome-wide profiling of RNAPII indicates that this polymerase is bound and engaged in the early steps of transcriptional initiation at most active and many inactive genes in human embryonic stem cells suggesting post-initiation modes of regulation may occur more frequently than previously appreciated (Guenther et al., 2007). Moreover, divergent transcription, defined by detection of populations of low abundance small RNAs (19-25 nts) generated by non-overlapping (separated by approximately 250 bps) bidirectional transcription, was detected at the majority of transcriptional start sites (TSSs) in mouse embryonic stem cells (mESCs) (Seila et al., 2008). Polymerases engaged in divergent transcription near promoters were simultaneously described in human lung fibroblasts (Core et al., 2008). Surprisingly, RNAPII only productively elongates in the protein-coding sense direction from these divergent promoters. Related results have been reported for several other eukaryotic systems (Fejes-Toth, 2009; Neil et al., 2009; Preker et al., 2008; Taft et al., 2009; Xu et al., 2009). Altogether, these data suggest that control of RNAPII elongation and RNA stability may be major points of transcriptional regulation and that mechanisms controlling these processes may dictate whether a stable RNA molecule is synthesized.

In recent years it has become clear that RNAPII pausing is a major mode of transcriptional regulation (Core et al., 2008; Rahl et al., 2010). The Negative Elongation Factor (NELF) and DRB-Sensitivity Inducing Factor (DSIF) protein complexes bind and arrest RNAPII 20-30 nts downstream of the TSS (Peterlin and Price, 2006). Recruitment of P-TEFb to a paused

RNAPII complex and subsequent phosphorylation of the RNAPII carboxyl-terminal domain (CTD) at serine 2, DSIF, and NELF, results in the dissociation of NELF from the elongation complex and continuation of transcription (Peterlin and Price, 2006). More recently it was recognized, in mESCs, that c-Myc stimulates transcription of over a third of all cellular promoters by recruitment of P-TEFb (Rahl et al., 2010). Intriguingly in these same cells, NELF and DSIF have bimodal binding profiles at divergent TSSs. This suggests divergent RNAPII complexes might be poised for signals controlling elongation and opens up the possibility that in the antisense direction P-TEF-b recruitment may be regulating release for productive elongation.

Cellular mechanisms for removal of improperly processed, spliced, or aberrantly transcribed products likely account for the instability of transcripts from divergent promoters. In *S. cerevisiae* cryptic unannotated transcripts (CUTs) derived from promoter-proximal regions are stabilized in the absence of the exosome (Houseley et al., 2006; Neil et al., 2009; Xu et al., 2009). The exosome, with 3′ to 5′ exonuclease activity, is a multi-subunit protein complex important for degradation and processing of mRNA, rRNA, snoRNA, and tRNA (Houseley et al., 2006; Mitchell et al., 1997; Schmid and Jensen, 2008). The phosphorylation state of the RNAPII CTD (Gudipati et al., 2008; Vasiljeva et al., 2008) and sequence elements within the RNA can influence targeting of transcripts to the exosome (Anderson et al., 2006). Upon exosome depletion in human cells, promoter upstream transcripts (PROMPTs) are stabilized farther upstream (~1kb) from antisense TSSa-RNAs and are detected in both sense and antisense orientations in the upstream promoter region (Preker et al., 2008). However, it is unclear how various promoter associated RNAs, including PROMPTs, relate to transcription from divergent mammalian promoters.

Although multiple studies have identified distinct RNA species from mammalian promoters, the precise mapping of RNAs produced from divergent CpG island promoters has not been described. In light of these questions, we sought to investigate RNAPII divergent transcription through a detailed biochemical analysis of the antisense transcripts. We have characterized upstream antisense RNAs or uaRNAs from four divergent promoters in mESCs. We show that antisense RNAs are predominantly 5′ capped and have heterogeneous 3′ ends ranging in size from 40-1100 bases in length. Both sense and antisense RNAPII complexes were involved in RNAPII pausing and both depend on PTEF-b recruitment and phosphorylation for subsequent elongation. We further show that low steady-state levels of uaRNAs, at least in part, are due to their targeting by the RNA exosome. Finally, we characterize induction of antisense and sense transcription from the divergent promoter of the Isg20l1 gene to show that PTEF-b activation at both sense and antisense RNAPII complexes occur with similar kinetics.

Results

Divergent RNAPII produces low abundant capped upstream antisense RNAs (uaRNAs) with 3′ heterogeneity

To test whether short antisense TSSa-RNAs previously described (Seila et al., 2008) are derived from longer transcripts and to determine the structure of their 5′ termini, we used Rapid Amplification of 5′ Complementary DNA Ends (5′-RACE) to characterize divergent upstream antisense RNAs from the Isg20l1, Tcea1, Txn1, and Sf3b1 genes in V6.5 mESCs (Figure 1, Figure S2A.). These genes were selected as representatives of divergent TSSs associated with CpG islands and spanning a range of expression levels. The positions of oligonucleotides for specific priming for the 5′-RACE overlapped sequences found in antisense TSSa-RNAs from

each promoter. The dependence of the specific 5′-RACE products on treatment with Tobacco Acid Pyrophosphatase (TAP) indicated the presence of a capped structure (Figure S1, lanes 2 and 3). The sequenced 5′-RACE products for the antisense TSSs revealed RNAs initiated upstream of the previously characterized antisense TSSa-RNAs for all four promoters (Figure 1, Figure S2A, leftward arrows). Two predominant uaRNA TSSs for Isg20l1 were identified with the most upstream site 43 nts from the previously characterized clusters of antisense TSSa-RNAs. The Tcea1 gene consists of two predominant antisense capped species with the most upstream 5′ terminus 37 nts from the antisense TSSa-RNAs. One predominant capped 5′ terminus was mapped for both Txn1 and Sf3b1 that extended 18 and 15 nts upstream from the antisense TSSa-RNA, respectively. The 5′ capped termini of these uaRNAs are likely generated by RNAPII initiation events suggesting that the antisense TSSa-RNAs are products of subsequent reactions during elongation.

The 5′ capped antisense RNA for each promoter align under a peak of RNAPII density near the end of the CpG island (Figure 1). The segments between sense and antisense capped RNAs are on average 200-250 bps, comparable to the length of nucleosome-free regions associated with CpG islands (Ramirez-Carrozzi et al., 2009). To test if transcripts extended beyond the sites of antisense 5′ capped termini for these four genes, Reverse Transcriptase-quantitative PCR (RT-qPCR) with strand-specific RT primers was used to determine the orientation of RNA species in the upstream CpG island promoter (Figure S3). As these transcripts might be of low abundance, cellular RNAs were prepared with two rigorous DNase treatment steps and only signals dependent on RT were analyzed. Detectable antisense transcription at all four genes was confined to the region downstream of the mapped uaRNA 5′ cap site. This provides additional evidence for initiation at these cap sites. Sense transcription

within the CpG island upstream of the antisense cap site was probed for all four regions but only detected at Txn1 and was estimated by PCR cycles to be 80-fold less abundant than the antisense product. The inability to detect significant sense or antisense RNA signal upstream of the mapped uaRNA TSSs argues that the majority products from these regions initiate from the identified antisense TSSs.

We previously have characterized antisense RNAs from divergent TSSs by a selective enrichment protocol followed by Northern blot and observed a family of short RNAs spanning 30 to 200 nts (Seila et al., 2008). To more precisely define such RNAs, a 3′-RACE protocol was used to characterize the 3′ ends of uaRNAs from the four divergent promoters. In this approach, adaptor sequences were ligated onto the free 3′-OH of large fractionated RNAs, followed by cDNA synthesis and PCR amplification using target-specific primers. The amplified products were cloned and sequenced to confirm their origin and define their 3′ termini. Multiple non-polyadenylated 3′ ends were observed for uaRNAs at each TSS and were aligned to their respective promoters (Figure 1). As few as 5 distinct antisense RNA 3′ ends were detected for Txn1 and as many as 8 at Tcea1. The longest transcripts cloned were 703, 546, 415, and 1100 nts for Isg20l1, Tcea1, Txn1, and Sf3b1, respectively. However, it is likely that additional 3′ ends exist since only a fraction of the 3′-RACE products were cloned and sequenced (Figure S4). All 3′-RACE products were dependent on RT for amplification (Figure S4). Transcripts under 100 nts were detected in the large fractionated RNA preparation. This probably reflects imperfect fractionation as similar patterns of transcripts were observed for all four genes. Because of this fractionation step, the relative levels of the various length RNAs cannot be estimated from the 3′-RACE products.

We compared the DNA sequences encompassed by uaRNAs with the location of RNAPII, TATA-binding protein (TBP) and chromatin modifications associated with active transcription determined by ChIP-seq in V6.5 mESCs (Kagey et al., 2010; Marson et al., 2008; Seila et al., 2008) (Figure 1). The shorter uaRNAs fell within the peak of bound RNAPII, however the longest transcripts extended farther downstream. It is likely that the density of RNAPII in these downstream regions is below the threshold signal considered positive in the ChIP-seq analysis. Histone H3 lysine 4 trimethylation (H3K4me3) and TBP mark transcription initiation and H3 lysine 79 dimethylation (H3K79me2) correlates with elongation. The TBP density denoting the pre-initiation complex was detected as a broad peak directly between the divergent RNAPII complexes. The H3K4me3 profile generally extended the full length of the uaRNAs in the antisense direction with the exception of the longest Sf3b1 transcript. In contrast, ChIP-seq signal for H3K79me2 is absent in antisense transcribed regions for the four genes. This lack of signal might be due to limits in the sensitivity of the technique, but the same chromatin modification is clearly present in the sense direction downstream of the TSS for these four genes. This suggests differential activity of elongation complexes in the two directions.

To relate levels of uaRNAs to antisense TSSa-RNAs which were previously measured at 1 copy per 10 mESCs (Seila et al., 2008), RT-qPCR probes noted in Figure S2A ("qPCR amplicon") were used for absolute quantification. Copy number defined by molar equivalents as compared to a standard signal in the form of ssDNA was determined per ES cell equivalent of total RNA. The uaRNAs are present at 4.5, 1.8, 1.1, and 1.7 copies per cell for Isg20l1, Tcea1, Txn1, and Sf3b1, respectively (Figure S2B). These results indicate that uaRNAs are roughly 10-fold more abundant then previously characterized antisense TSSa-RNAs; present at approximately one copy per copy of genome sequences.

uaRNAs are substrates of the exosome

As previous studies have linked the exosome to nuclear surveillance of unannotated or cryptic transcripts, uaRNA stabilization and 3′ termini were assayed upon exosome depletion. Exosc5 was targeted for knockdown with an shRNA-lentiviral delivery construct and depletion was confirmed 48 hours after infection (Figure S5). RT-qPCR was used to determine relative steady-state levels of uaRNAs (upstream antisense probe) and spliced sense mRNA (exon1-exon2 probe) between knockdown (shExosc5) and empty vector control (pLKO.1) for all four genes. Across multiple biological replicates, Exosc5 depletion led to a 2.5-3.5 fold increase in uaRNA levels, while spliced sense mRNAs were also slightly elevated yet below statistical significance (Figure 2A). Next, we assayed for uaRNAs by DNA Southern blot of 3′-RACE products in control and exosome depleted cells. After optimization of minimal PCR cycles and multiple probe validation of signal (Figure S6), the most abundant uaRNA forms for each gene were observed. In Exosc5 knockdown samples, the numbers and abundance of long RNAs increased compared to control virus infected cells further supporting uaRNAs as substrates for the exosome (Figure 2B). These results show that upstream antisense RNAPII elongates to produce heterogeneous RNAs that are substrates for the exosome.

RNAPII pausing factors regulate uaRNA transcription

The RNAPII pausing factors that associate with promoter proximal stalled RNAPII are composed of NELF (NELF-A,B,C/D, and E) and DSIF (Supt4h and Supt5h). In addition, RNAPII acquires phosphorylation at Ser5 on the carboxyl-terminal domain early in transcription and this modification peaks in abundance around the pause site. We first aligned ChIP-seq

profiles of RNAPII-Ser5P, Supt5h and NELF-A determined in V6.5 mESCs to uaRNA transcribed regions. For all four genes, the peaks of RNAPII-Ser5P, Supt5h and NELF-A directly overlap the uaRNA TSS supporting post-initiation regulation by RNAPII pausing in the antisense direction (Figure 3A). To test whether RNAPII upstream complexes are poised for transcription in both directions, we performed shRNA-mediated knockdown of NELF-A, NELF-E and Supt4h, with each providing potent targeted mRNA loss (Figure 3B). Depletion of either NELF subunit resulted in near 2-fold increases in uaRNA and spliced mRNA transcripts for all four genes across six biological replicates (Figure 3C). Supt4h knockdown also resulted in 2-fold increases for both uaRNAs and spliced mRNA transcripts. Together these data argue NELF and DSIF complexes are equivalently active in binding and regulating paused RNAPII complexes in both directions at divergent promoters.

P-TEFb regulates elongation of uaRNAs

P-TEFb promotes RNAPII elongation in the sense direction for most if not all genes, however, its role in antisense transcription at divergent promoters has not been examined. We used flavopiridol, a small molecule drug with high specificity for CDK9 inhibition to test P-TEFb's requirement for RNA synthesis at divergent TSSs. uaRNA, spliced mRNA, and nascent mRNA transcripts (exon1-intron1 probe) were measured at all four genes in mock (DMSO) or 1μM flavopiridol treated mESCs. Treatment with this flavopiridol concentration for 1 hour was previously shown to not affect global RNAPII-Ser5P levels while dramatically reducing RNAPII-Ser2P and Supt5h phosphorylation in these cells, indicating a block of transcriptional elongation but not initiation (Rahl et al., 2010). In flavopiridol-treated cells, the nascent mRNA transcripts for all four genes were reduced to 5-12% of mock-treated controls confirming a block

in elongation (Figure 4A). Interestingly, steady-state uaRNA transcript levels decreased to 19-25% of mock-treated controls using RT-qPCR probes that require transcription of ~150 nts or longer from the uaRNA TSS. Spliced mRNA levels were unchanged suggesting stable transcripts over this time course. We next determined uaRNA decay rates using flavopiridol treatment over a 1 hour time course (Figure 4B). Half-lives of 27, 19, 14, and 13 minutes were estimated for uaRNAs from Isg20l1, Tcea1, Txn1, and Sf3b1, respectively (Figure S7).

The large decrease in uaRNA levels following loss of CDK9 activity supports P-TEFb dependent release from paused polymerase and bidirectional recruitment of P-TEFb at CpG island promoters. To confirm P-TEFb-dependent transcription, RNA produced from divergent TSSs for all four genes was measured following removal of flavopiridol. The uaRNAs (Figure 4C, left panel) and nascent mRNA transcripts (Figure 4C, right panel) recovered with similar kinetics and reached control steady-state levels by 30 minutes after flavopiridol removal. The similar recovery rates at both TSSs further supports P-TEFb acting on both divergent RNA polymerases to promote elongation.

Transcriptional induction of Isg20l1 similarly increases mRNA and uaRNA levels

Interferon-stimulated 20 kDa exonuclease-like 1, Isg20l1, is one of two homologs of an apoptosis-enhancing exonuclease. To determine how divergent paused RNAPII complexes respond to gene activation, doxorubicin, a DNA intercalator and inducer of double stranded breaks (Nitiss, 2009), was used to induce apoptosis in mESCs. Treatment with 1μM doxorubicin for 1.5, 4, and 6 hours was followed by measurement of Isg20l1 uaRNA, nascent mRNA, and spliced mRNA levels. Isg20l1 transcriptional output in either direction did not significantly change with 1.5 hours of treatment. However, both uaRNA and spliced mRNA transcripts had

equivalent induction levels of 8 and 12-fold following 4 and 6 hour treatments, respectively (Figure 5). The nascent mRNA transcripts showed only a 2-fold increase at 6 hours of treatment indicating tightly coordinated pre-mRNA processing. Divergent TSS products for thioreductase 1, Txn1, which are not expected to respond to DNA damage, did not change with treatment and served as an additional control for transcription fidelity during cellular stress. These results support a model for gene activation at divergent CpG island promoters proceeded by stimulation of elongation in both directions.

Discussion

The detection of capped 5′ termini on uaRNAs for all 4 studied promoters strongly supports a distinct and specific initiation event from antisense RNAPII complexes at divergent promoters. Further, RNAPII-Ser5P, NELF, and DSIF profiles at divergent TSSs suggest that these antisense RNAPII complexes are poised for transcription elongation. Correspondingly, we find that depletion of NELF and DSIF, factors known to promote the pausing of RNAPII, modestly increases steady-state uaRNA levels. This is consistent with the model that the two divergent and paused complexes are composed of similar factors controlling initial progression into elongation. In addition, we demonstrate that inhibition of P-TEFb activity with flavopiridol decreases both uaRNA and nascent mRNA transcript levels. P-TEFb phosphorylates the RNAPII-CTD, DSIF and NELF promoting elongation and its inhibition blocks elongation (Peterlin and Price, 2006). Since uaRNAs are short-lived, yet detectable in total RNA, upstream antisense RNAPII must be released from the paused state with kinetics comparable to their half-lives.

That RNAPII complexes are poised or stalled in the sense and antisense direction begs the question of how this process contributes to the local chromatin structure and overall gene activity. In Drosophila cells, roughly two-thirds of all changes in gene expression upon NELF depletion were downregulation likely due to loss of RNAPII pausing that allows nucleosome assembly at the promoter (Gilchrist et al.; Gilchrist et al., 2008). However, one-third of all changes in gene expression were increases in transcript levels upon NELF depletion. We observed an increase in uaRNA and nascent mRNA levels upon NELF depletion. It could be that the mechanism and function of RNAPII pausing at our four divergent promoters are similar to the latter class described above. For example, mammalian promoters are frequently CpG-rich, and this tends to destabilize nucleosomes promoting nucleosome-free regions at the 5'-end of genes. Therefore, the impact of RNAPII pausing mechanisms to occlude nucleosome assembly may not be significant at divergently transcribed CpG-rich promoters.

Activation of the Isg20l1 divergent promoter upon doxorubicin treatment yielded simultaneous induction of transcription with similar levels and kinetics of spliced mRNA and uaRNA. These results were compelling as it illustrates the requirement for additional regulatory steps post P-TEFb recruitment to differentiate the sense and antisense RNAPII complexes for production of stable transcripts. In the sense direction, signals for continued productive elongation could involve the recruitment of P-TEFb-type activities by elongation complexes and/or pre-mRNA processing machinery and maintenance of Ser-2P. For example, P-TEFb has been shown to interact with the SR proteins involved in the recognition of exonic sequences (Lin et al., 2008). These signals may not be present in the antisense direction.

Antisense transcription from divergent promoters produce uaRNAs that range from 1-4 copies/cell with relatively short half-lives. Since the uaRNAs are 10-fold more abundant than the

20-25 nt antisense TSSa-RNAs, the latter are likely derived during the synthesis or processing of longer uaRNA through the endonucleolytic cleavage activities of the Transcription Factor II S (TFIIS) (Nechaev et al.) or the RNA exosome, respectively (Lebreton et al., 2008).

The mapping of uaRNA 3′ ends revealed heterogeneous populations possibly arising from nascent transcripts, or RNAPII termination, processing and/or degradation by the exosome. Analysis using 3'-RACE Southern blots on control and exosome depleted cells revealed that uaRNA 3' termini are distinct and longer in exosome-depleted cells. The 2-4 fold increase in uaRNA levels upon exosome depletion is modest but certainly in line with a previous study (Preker et al., 2008) that reports an average 1.5 fold increase in RNA originating 1kb upstream of known TSS**.** This increase is consistent with a dynamic and rapid turnover of antisense transcripts.

Both the act of divergent transcription and the rapidly cycling promoter associated RNA could have multiple functions. In addition to possibly maintaining chromatin structure at promoter regions, it is possible that the nascent RNA tethered to RNAPII and/or disengaged could participate in regulation of local chromatin structure. For example, nascent RNAs transcriptionally engaged upstream of the cyclin D1 gene are thought to recruit TLS, a RNA binding transcriptional regulatory factor sensitive to DNA damage (Wang et al., 2008). In contrast, disengaged short sense RNAs found at the 5′ ends of Polycomb target genes have been reported to form stem-loop structures, which can bind Suz12 to promote silencing of the gene. (Kanhere et al., 2010). These two examples, among many others, suggest that uaRNAs could be involved in control of gene expression.

Methods

Cell Culture Conditions

V6.5 (C57BL/6-129) mouse embryonic stem cells (mESCs) (Koch Institute Transgenic Facility) were grown under standard ES cell culture conditions (Boyer et al., 2006).

Lentiviral infection and total RNA preparation

The shRNA targeting plasmids for knockdown of mRNA and empty plasmid (control) were ordered from Open Biosystems/Thermo Scientific (Huntsville, AL) (Table 1). For Lentivirus production, 293T cells were plated in 6-well dishes at $6x10^5$ cells/well. Forty-eight hours after co-transfection of viral and shRNA plasmids into 293T cells, lentivirus was harvested and used directly to infect mESCs in a 6-well plate at $2x10^5$ cells/well. The infection media was 1:2, viral media: mESC media with 2mM polybrene. Infected mESCs were then selected for 24 hours with 2μM puromycin. Total RNA was collected using the RiboPure Kit (Ambion/Applied Biosystems; Austin, TX) according to the manufacturer's protocol.

RT-qPCR of RNA transcripts

To assess mRNA knockdown and for other PCR analysis, complementary DNA (cDNA) was generated using the QuantiTect Reverse Transcription Kit (Qiagen; Velencia, CA) according to the manufacturer's protocol with the following modification: a 5 minute genomic DNA (gDNA) elimination step. TaqMan gene expression assays (Applied Biosystems; Carlsbad, CA) were used to determine levels of mRNA transcripts after shRNA knockdown (Figure S8, Table 1). The ABI 7500 Real-Time PCR System was used with the accompanying software to analyze quantitative PCR (qPCR) data.

All other qPCR experiments were performed with *Power* SYBR Green PCR Master Mix (Applied Biosystems; Carlsbad, CA). Primers to detect uaRNAs, promoter RNAs, spliced mRNA (exon 1-2 junction, probe), and nascent mRNA (exon1-intron1, probe) are shown in Figure S9, Table 2A. Analysis of relative transcript levels was calculated using the delta-delta Ct method (Livak and Schmittgen, 2001). Once internal controls of β-Actin, GAPDH, and 28S rRNA were shown to be comparable standards, β-Actin was chosen as the internal control for all experiments. Error between biological replicates was calculated using a Standard Error of the Mean. Statistical significance was determined using a two-tailed, paired T-Test. P-values of < 0.05 were reported for all RT-qPCR analysis

For absolute quantitation, ssDNA Ultramer Oligonucleotides (IDT) were designed (Figure S10, Table 3) to contain the 5'-end of uaRNA transcripts. Standard curves were generated using their respective uaRNA qPCR primers. mESCs were collected, counted using a Coulter Counter (Millipore), and total RNA was prepared to determine average RNA concentration per mESC. A quantified number of cells were then subjected to qPCR using the uaRNA primer pairs and resulting qPCR signal was converted to copy number based on the ssDNA molar equivalents.

Rapid Amplification of 5′ complementary DNA ends (5′-RACE)

Total RNA was prepared using the standard Qiazol (Qiagen) protocol. Total RNA (10µg) was DNase treated with the Turbo DNA-Free kit (Ambion). Figure S9, Table 2B contains primers used for 5′-RACE with the FirstChoice RLM-RACE Kit (Ambion) with the following modifications. First, T4 RNA Ligase 1 (ssRNA Ligase) was heat-inactivated for 15 minutes at

65°C. Second, SuperScript III (SSIII, Invitrogen) was used for cDNA synthesis according to the manufacturer's protocol. Reverse transcription was performed with a target specific primer, gsp1, at a 0.25μM final concentration. In addition, two subsequent nested PCR reactions using HotStarTaq (Qiagen) was performed with two forward primers, P_o and P_i, 5′-RACE adaptor-specific primers (Ambion), and two reverse target-specific primers, gsp-2 and gsp-3, at a 0.4μM final concentration (Figure S9, Table 2B). PCR 1 and PCR 2 were amplified for 20 and 25 to 30 cycles, respectively. PCR reaction products were run on a 2% agarose gel, extracted using a QIAquick Gel Extraction Kit (Qiagen), and sequenced using Sanger methods.

Rapid Amplification of 3′-complementary DNA ends (3′-RACE)

Large (> 200 nts) RNA was prepared using the mirVana miRNA Isolation Kit (Ambion). Large fractionated RNA (5μg) was DNase-treated with the Turbo DNA-Free Kit (Ambion). The ligation reaction was performed with a 3′-RACE adaptor, synthesized with a 5′-phosphate and a 3′-dideoxy-C (IDT) and used at a final 50μM concentration for 3′ end ligation. All primers for 3′-RACE are listed in Figure S9, Table 2C. Reverse transcription was modified in the following manner: the SSIII protocol for GC-Rich templates was used in place of the standard and reverse transcription was performed with an adaptor-specific primer, 3′-RACE P_o, at 0.25μM final concentration. Two subsequent PCR reactions were performed similarly as 5′-RACE, but instead with two forward target-specific primers, gsp-1 and gsp-2, and two reverse adaptor-specific primers 3′-RACE P_o and P_i.

Flavopiridol and doxorubicin treatment of mECSs

Flavopiridol (Sigma) and doxorubicin (Sigma) were resuspended to a final concentration of 10mM in DMSO and stored at -80°C until used. Flavopiridol was added directly to the mESC culture media to a final concentration of 1μM and allowed to incubate for 1, 5, 10, 15, 30, and 60 minutes in a tissue culture incubator at 37°C. For the wash-off experiments, flavopiridol treated mESCs were washed once with 37°C 1x Phosphate Buffered Saline. Fresh mESC media was added to the flavopiridol treated mESCs and they were placed at 37°C for 30, 60, or 90 minutes before total RNA was isolated as described above. Doxorubicin was added directly to the mESC culture media to a final concentration of 1μM and incubated for 1.5, 4 or 6 hours at 37°C. After the specified time, RNA was isolated as previously described and RT-qPCR was preformed to assay transcript levels.

Acknowledgments

We would like to thank Grace Zheng, Mohini Jangi, and Allan Gurtan for discussion and critical review of this manuscript. We would also like to thank Pete Rahl and Richard Young for generously sharing ChIP-seq data. This work was supported by United States Public Health Service grants RO1-GM34277 from the National Institutes of Health to PAS and partially by Cancer Center Support (core) grant P30-CA14051 from the National Cancer Institute. A.E.A. was supported by NIH/NIGMS Pre-Doctoral Training in Biological Sciences GM007287.

Figure 1

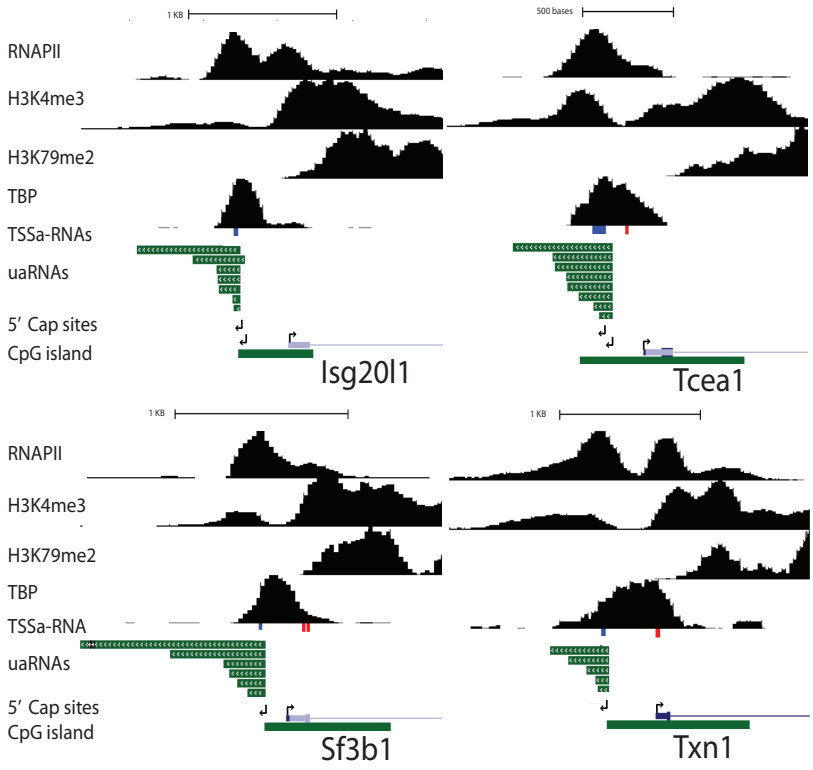


Figure 1. Capped antisense RNA from divergent transcription initiate upstream of antisense TSSa-RNAs and display 3′ heterogeneity.

UCSC Genome browser view showing the location of detected 5′ and 3′ ends using Rapid Amplification of cDNA Ends (RACE) at four selected CpG island divergent promoter genes: Isg20l1, Tcea1, Txn1, and Sf3b1. 5′-RACE analysis was performed on upstream regions containing more than one overlapping antisense TSSa-RNA (Seila et al., 2008). Promoter regions are shown with the TSSs marked (arrows pointing in the direction of transcription). Arrows depicting antisense transcription are pointing to the left while sense TSSs are marked

with arrows pointing to the right. Sense TSSs (right arrows) were labeled according to UCSC genome browser's known genes from UniProt, RefSeq, and GenBank. These were independently confirmed by 5′-RACE for Isg20l1, Tcea1, and Txn1. 3′-RACE analysis yielded various uaRNA transcripts (green) for each divergent promoter. These range in length from approximately 50 nts to 1100 nts. ChIP-seq binding profiles of RNAPII, H3K4me3, H3K79me2, TATA-binding protein (TBP), TSSa-RNA reads (antisense = blue; sense = red), and CpG island regions (green) are shown. ChIP-Seq data was obtained from the following published reports: RNAPII-8W16 (6), H3K4me3 and H3K79me2 (23), and TBP (22). We note the absense of a sense RNAPII ChIP-seq peak at the annotated TSS of Tcea1, likely explained in part to difficulties in mapping reads to this region since it contains high similarity (99%) with a location on chromosome 15. Scale bars are displayed at the top of each promoter region.

Figure 2

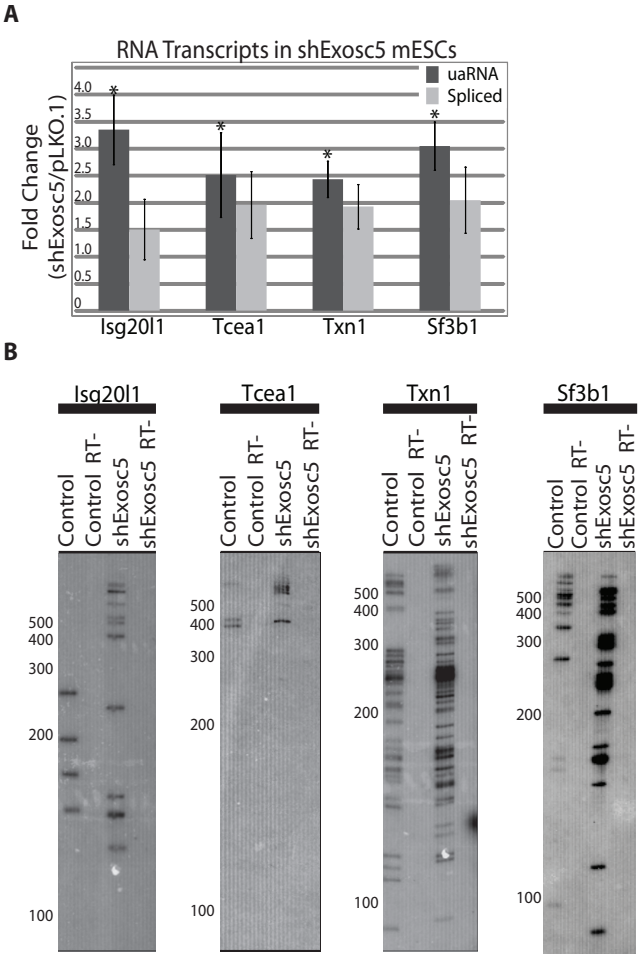


Figure 2. uaRNA transcripts are substrates for the exosome.

A. Relative levels of uaRNA (grey amplicon in Figure S2A, dark gray) and spliced mRNA (exon1-2 probe, light gray) transcripts in mESCs infected with virus containing an shRNA targeting Exosc5 (shExosc5) and assayed by RT-qPCR (probes shown in Figure S9, Table 2A). Transcript levels were normalized to virus-infected cells without shRNA, empty vector

(pLKO.1), and normalized to β-actin levels. Values represent four biological replicates and error represents the respective SEM. Asterisks represent significance of $p < 0.05$ in two-sided T-test. B. 3′-RACE followed by Southern blot analysis of control and shExosc5 treated mESC RNA. The Southern blots were probed with probe 1 shown in Figure S6. Minus RT lanes refer to 3′-RACE experiments with no reverse transcriptase added in the RT step of the procedure. Migration of 100bp molecular weight ladder (NEB) is marked on the left.

Figure 3

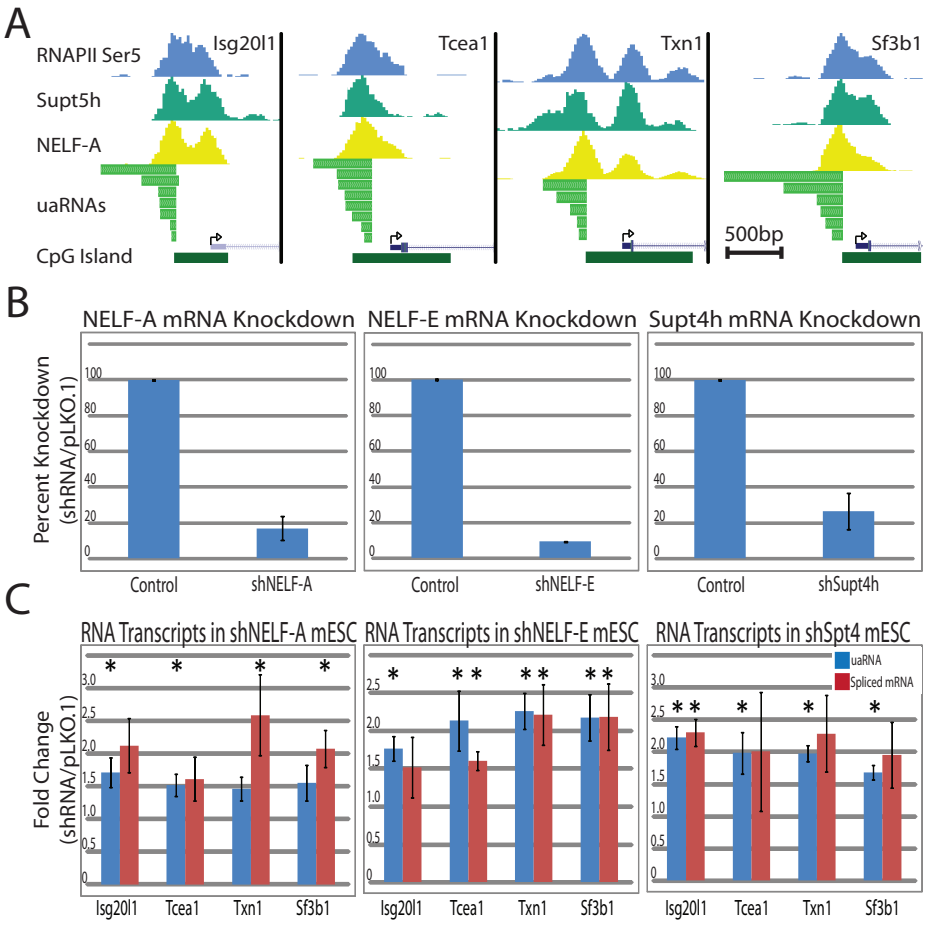


Figure 3. Pausing factors at the antisense RNAPII complex regulate uaRNA levels.

A. UCSC Genome browser views of the four divergent promoter regions displaying ChIP-seq binding profiles (13) of RNAPII-Ser5P, Supt5h, NELF-A, and the full length antisense RNA transcripts. Each region diagrammed spans 2kb and scale bars represent 500 bp. B. Relative gene expression of Nelf-A, Nelf-E, and Supt4h in control and shRNA knockdown mESC after 48 hours of selection measured by Taqman RT-qPCR assay. Values represent six biological replicates and error represents the respective SEM. C. Transcript changes in shNELF-A, shNELF-E, and shSupt4h mESC lines as determined by RT-qPCR. uaRNA and spliced mRNA

levels are represented by blue and red bars, respectively (probes shown in Figure S9, Table 2A). Values represent six biological replicates and error shows the respective SEM. Asterisks represent significance of $p < 0.05$ in two-tailed T-test.

Figure 4

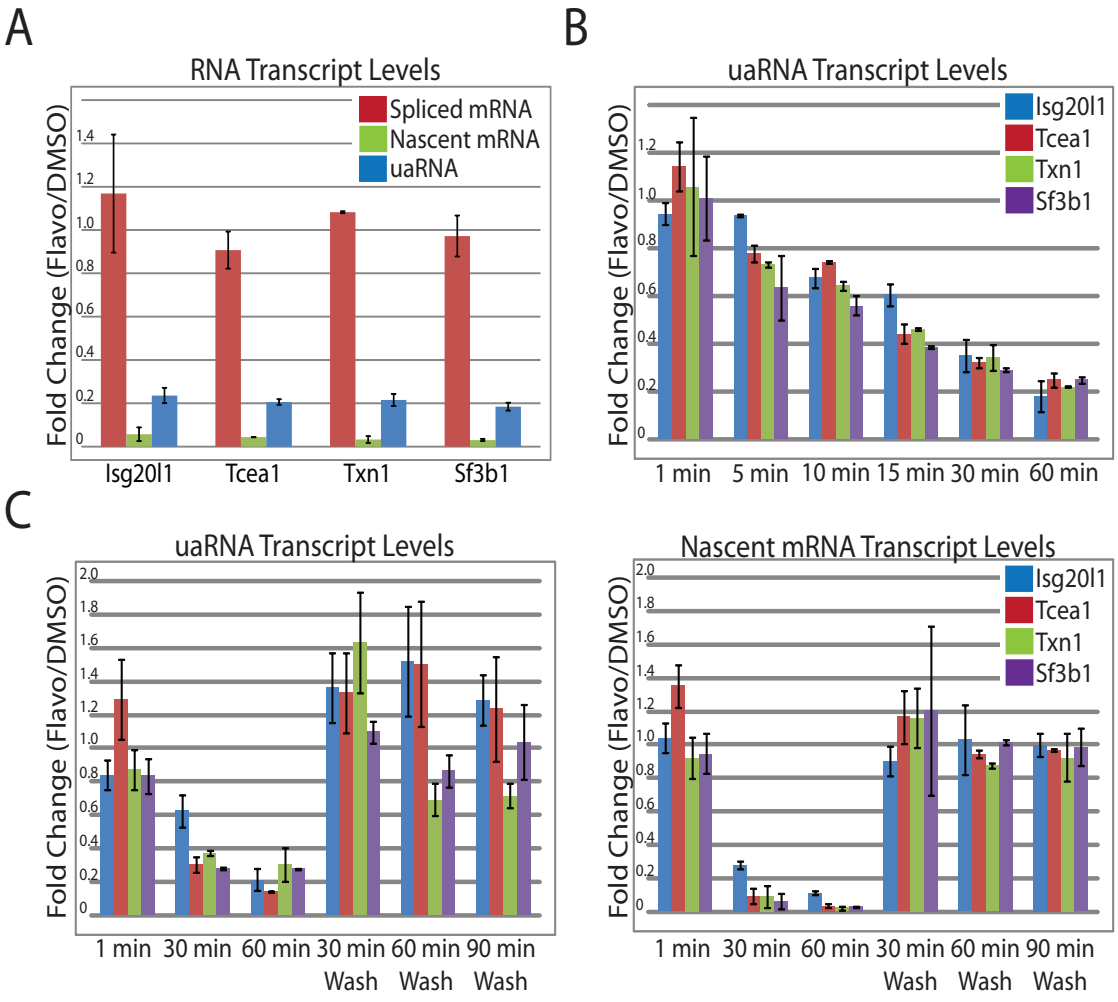


Figure 4. uaRNAs are P-TEFb-dependant transcripts and have short half-lives.

A. Relative levels of spliced mRNA (red), nascent RNA (exon1-intron1 probe, green), and uaRNA (blue) transcripts, as measured by RT-qPCR, after a 1μM flavopiridol treatment for 1 hour. B. uaRNA transcript levels assayed by RT-qPCR from amplicon shown in Figure S2A over a 1 hour time course with 1μM flavopiridol. C. uaRNA (left panel) and nascent sense RNA (right panel) levels assayed by RT-qPCR over an hour 1μM flavopiridol treatment followed by a Phosphate Buffered Saline (PBS) wash off of flavopiridol at the indicated times. Isg20l1, Tcea1, Txn1, and Sf3b1 RNA transcripts are shown in blue, red, green, and purple, respectively. All values are relative to mock (DMSO) treated cells and normalized to β-Actin. Values represent two biological replicates with the error representing the respective SEMs. All probe sequences are shown in Figure S9, Table 2A.

Figure 5

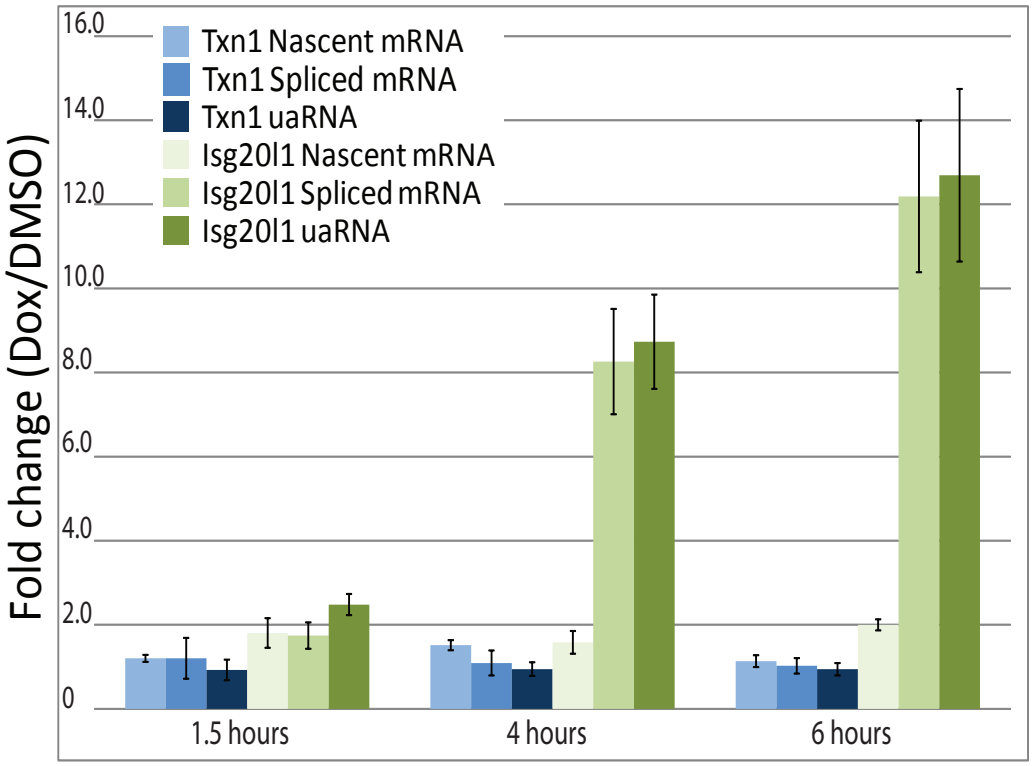


Figure 5. Divergent sense and antisense transcription induced with similar kinetics.

mESCs were treated with 1μM doxorubicin for 1.5, 4, and 6 hours after which total RNA was collected. RT-qPCR analysis to determine the relative fold change of nascent, spliced mRNA, and uaRNA transcription followed. Changes in transcript levels for two genes, Isg20l1 and Txn1, green and blue bars, respectively are shown. Values represent biological triplicates and error of the respective SEM. All probes are shown in Figure S9, Table 2

Table S1

Table 1. Taqman primers and shRNAs used for mRNA Knockdown.

Gsym, Complex	Accession ID	Open Biosystems shRNA ID	Applied Biosystems Expression Assay
Supt4h, DSIF	NM_009296	RMM4534-NM_009296	Mm02527263_g1
Supt5h, DSIF	NM_013676	RMM4534-NM_013676	Mm01170629_m1
NELF-E, NELF	NM_138580	RMM4534-NM_138580	Mm01134804_m1
NELF-A, NELF	NM_011914	RMM4534-NM_011914	Mm01217228_m1
Exosc5, Exosome	NM_138586	RMM4534-NM_138586	Mm00506830_m1
GAPDH	NM_008084.2	N/A	4352932E

Table S2

Table 2A. Primers used in RT-qPCR for relative transcript level analysis.

Gene – Location	Forward Primer (5′ -> 3′)	Reverse Primer (5′ -> 3′)
Isg20l1-Upstream Antisense	CCTAGTGAACCGCAACCTTC	GACTTCTCACAACCCCAAGC
Isg20l1-Exon1-Exon2	GGGTTGGTTTGCAACTAGGC	GCTCACAGGTTGGGGTAAGA
Isg20l1-Exon1-Intron1	GGGTTGGTTTGCAACTAGGC	CCCAAAAGCTTACAGACCA
Isg20l1-Promoter	TGATCCTGCTCCTCCTCAGT	AGATCGGGATGTGCTCTTTG
Tcea1-Upstream Antisense	CTATCCGGACTCGCGTTG	CTTTAAGCCCTCGGCAATG
Tcea1-Exon1-Exon2	GATGGACAAAATGGTGCAGA	TTCATCTGTGCTCTGCTTGC
Tcea1-Exon1-Intron1	GTTCGCATTGCCAAGAAGAT	GCAGCACGGACCTGAAAG
Tcea1-Promoter	GATCGCAGGAGACTGGAAAG	GGGTTTCGATGGAACTCGTA
Txn1-Upstream Antisense	GCCTCAAGGGCACTTTAACA	GGTCTAGTTTGGGGCATGG
Txn1-Exon1-Exon2	GCCAAAATGGTGAAGCTGAT	TGATCATTTTGCAAGGTCCA
Txn1-Exon1-Intron1	TGGATCCATTTCCATCTGG	CCGAGAGTGTCCTCTTCAGC
Txn1-Promoter	GCTGCCGAACAAGAACCTTA	TTGGCTCTTAGGGGTAGCTG
Sf3b1-Upstream Antisense	GCGGAAGAGGATGGCTACT	GTCTGTACAGCCCTGGCTTC
Sf3b1-Exon1-Exon2	GTGGACAAAATGGCGAAGAT	TGCCTTCTTGCCTTGAATTT
Sf3b1-Exon1-Intron1	GTGGACAAAATGGCGAAGAT	CTCGGTCGAGACCAGAGATG
Sf3b1-Promoter	TCCTTAAAAAGCCAGCGAAA	GACAGGCTACAGCCCTCTTG
β-Actin mRNA	GACGAGGCCCAGAGCAAGAGAGG	GGTGTTGAAGGTCTCAAACATG
28S rRNA	AGCAGCCGACTTAGAACTGG	TAGGGACAGTGGGAATCTCG
GAPDH	GTGTTCCTACCCCCAATGTGT	AATGTGATACCAGGAAATGAGCTT

Table 2B. 5′-RACE Primers.

	Gsp-1 (5'-3')	Gsp-2 (5'-3')
Isg20l1	GCTCTTTGAGGCTTTAAGTCTTTGAAGG	CTTTGAGGCTTTAAGTCTTTGAAGGTTGC
Tcea1	TAGCAACCTACCGCCTACTGCC	CTACCGCCTACTGCCTGATCC
Txn1	GTGCCCTTGAGGCAGCTGGAAGTTGG	CCTTGAGGCAGCTGGAAGTTGGCTC
Sf3b1	GGACAGGCTACAGCCCTCTTGG	CTACAGCCCTCTTGGGAAGTAGC
	Gsp-3 (5'-3')	
Isg20l1	GAGGCTTTAAGTCTTTGAAGGTTGCGG	
Tcea1	CTCTGGCGTGAAAGCCGGACTCC	
Txn1	GAGGCAGCTGGAAGTTGGCTCTTAGG	
Sf3b1	CAGCCCTCTTGGGAAGTAGCCATCC	

Table 2C. 3′-RACE Primers and Adaptor Sequence.

	Gsp-1 (5'-3')	Gsp-2 (5'-3')
Isg20l1	CGACGCCTAGTGAACCGCAACC	ACCGCAACCTTCAAAGACTTAAAGCC
Tcea1	ACGGAGTCCGGCTTTCACGCCAGAG	GGAGTCCGGCTTTCACGCCAGAG
Txn1	GGCAGCTACCCCTAAGAGCC	CCTAAGAGCCAACTTCCAGCTGCC
Sf3b1	CGCGGAAGAGGATGGCTACTTCC	CCAAGAGGGCTGTAGCCTGTCC
Po	CGACTACCGCTACTTACTTGTGAC	
Pi	CTTGTGACGCAAACGACCGAAACTAC	
Adaptor Sequence	5′-/5Phos/rArArArGrUrArGrUrUrUrCrGrGrUrCrGrUrUrUrGrCrGrUrCrArCrArArGrUrArArGrUrArGrCrGrGr UrArGrUrCrG/3ddC/ -3′	

Table S3

Table 3. ssDNA Ultramer Oligonucleotides for absolute copy number determination.

	Sequence (5'-3')
Isg20l1	TTATTTCACGGACTTCTCACAACCCCAAGCCTGGAGGGGCTGCAGTTCCCCGAGGCAGATC GGGATGTGCTCTTTGAGGCTTTAAGTCTTTGAAGGTTGCGGTTCACTAGGCGTCGGGTC
Tcea1	TGCTCATGCGCTTTAAGCCCTCGGCAATGCCTGTCCTGCGTCCCAGAGAACGCTCTGCCGG AGGGGTTTCGATGGAACTCGTAGCAACCTACCGCCTACTGCCTGATCCCTCTGGCGTGAAA GCCGGACTCCGTCCAACTCCAGCTCGCCAGCAACGCGAGTCCGGATAGGGCCGGAAGT
Txn1	ATCTGACTTAGGTCTAGTTTGGGGCATGGGCAGTGTGATTACAGAAGGACTCTACGGTGTG AGAGAGGACCGTGATCTACCCCGGCGCTGTTCGCTGTTAAAGTGCCCTTGAGGCAGCTGG AAGT
Sf3b1	GACAGGCTTTGTCTGTACAGCCCTGGCTTCGGGAACTCTCTTTGTAGACCAGGCTGGCCTC GAACTGCCTCTTCTCTTCCGAGTGCTTGAATTAACGGCACGTTACCCACCACTGGCCGGAC AGGCTACAGCCCTCTTGGGAAGTAGCCATCCTCTTCCGCGTTTT

Figure S1

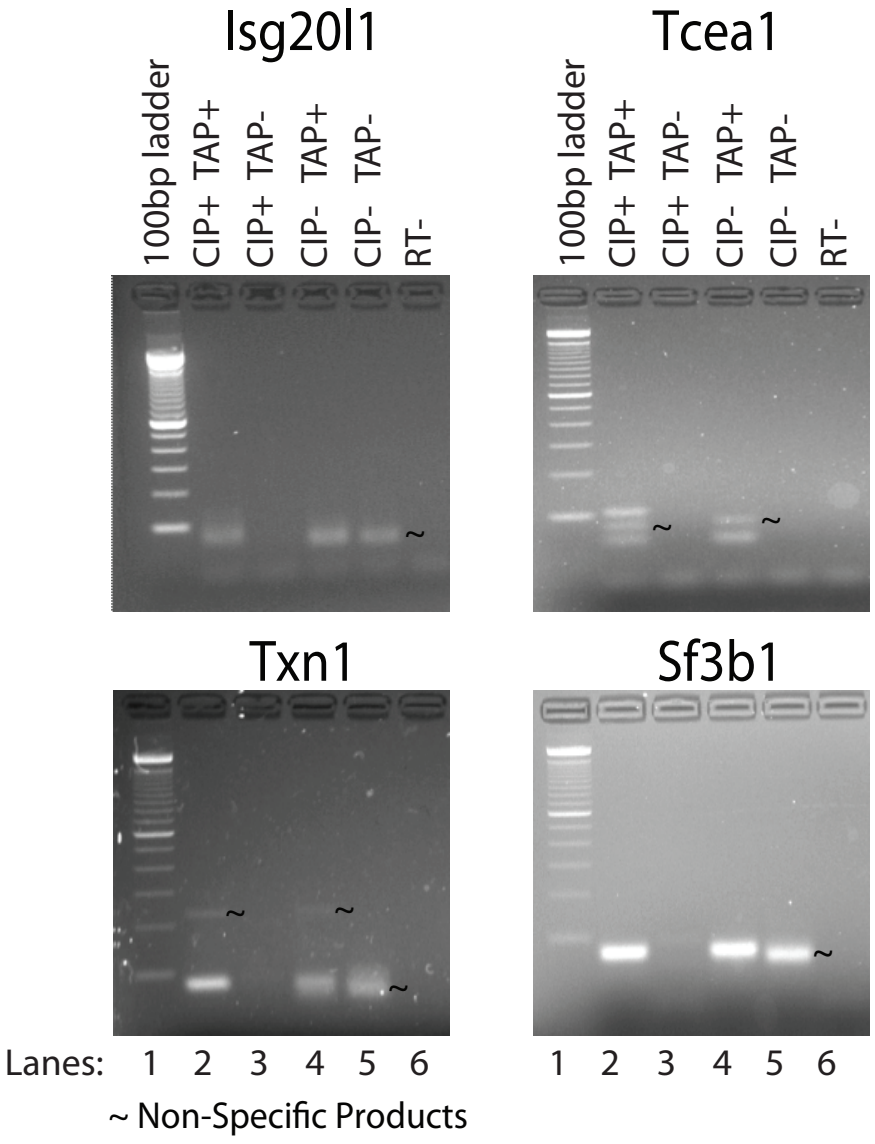


Figure S1. uaRNAs contain a cap structure at the 5′ termini.

All PCR products were collected from the gel, cloned into a TOPO vector, and transformed into competent cells. Five colonies for each band were selected for sequencing. From these sequencing results, specific products from the upstream antisense regions were only detected in

the CIP+TAP+ lane. Despite appearances of similarly sized bands in the CIP+TAP+ and CIP-TAP- lanes, these products did not contain sequences from the promoter regions. For Tcea1, the addition of CIP prior to TAP treatment allowed for the detection of an additional antisense 5′ end cap site likely explained by the increase in PCR efficiency upon removal of uncapped background RNA (lanes 2 and 4). Size markers are shown to the left of each gel.

Figure S2

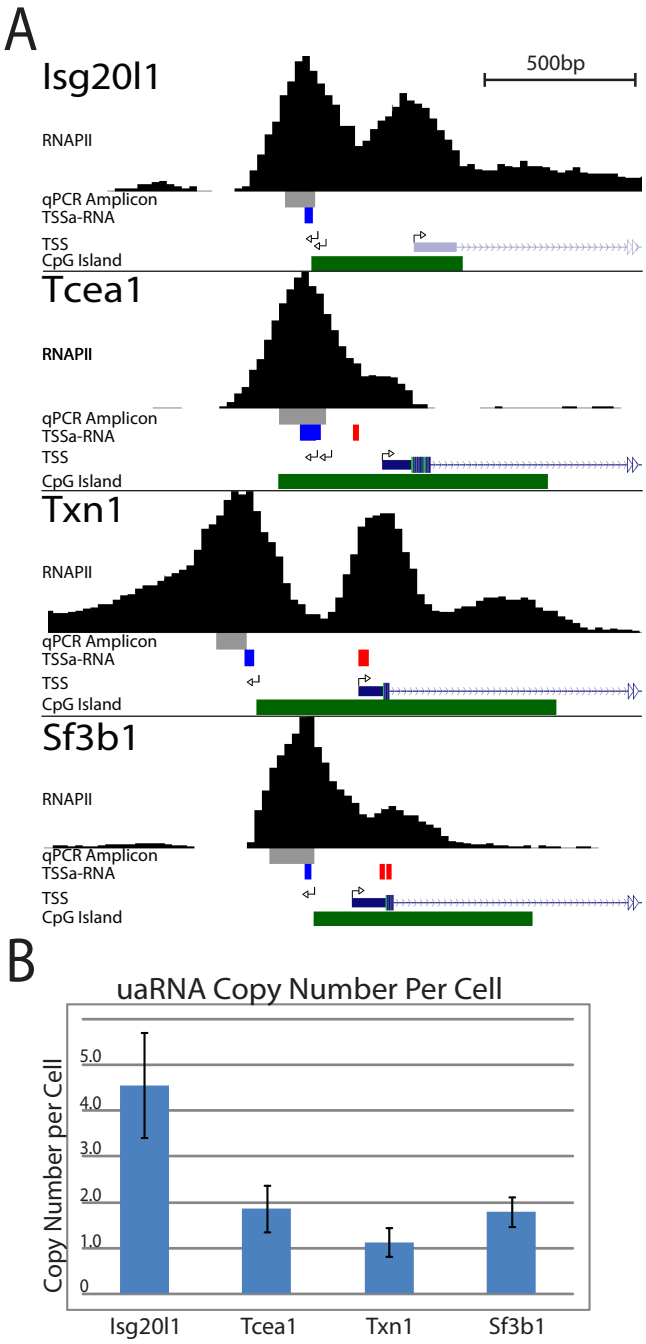


Figure S2. Capped antisense RNA from divergent transcription initiate upstream of antisense TSSa-RNAs.

A. UCSC Genome browser view showing the location of detected antisense TSSs using Rapid Amplification of 5′ cDNA Ends (5′-RACE) at four selected CpG island promoter genes: Isg20l1, Tcea1, Txn1, and Sf3b1 (Figure 1, Figure S1). Arrows depicting antisense transcription are pointing to the left while sense TSSs are marked with arrows pointing to the right. Tracks: RNAPII ChIP-seq profiles in mESCs (black), qPCR amplicon (gray), antisense (blue) and sense (red) TSSa-RNAs, and CpG island (green). Each genomic region displayed spans 2kb and the scale bar represents 500 bp. B. Absolute quantification of the upstream antisense RNAs (uaRNAs) determined by qPCR using a ssDNA oligonucleotide standard. The values represent biological triplicates and error bars are standard error of the mean (SEM).

Figure S3

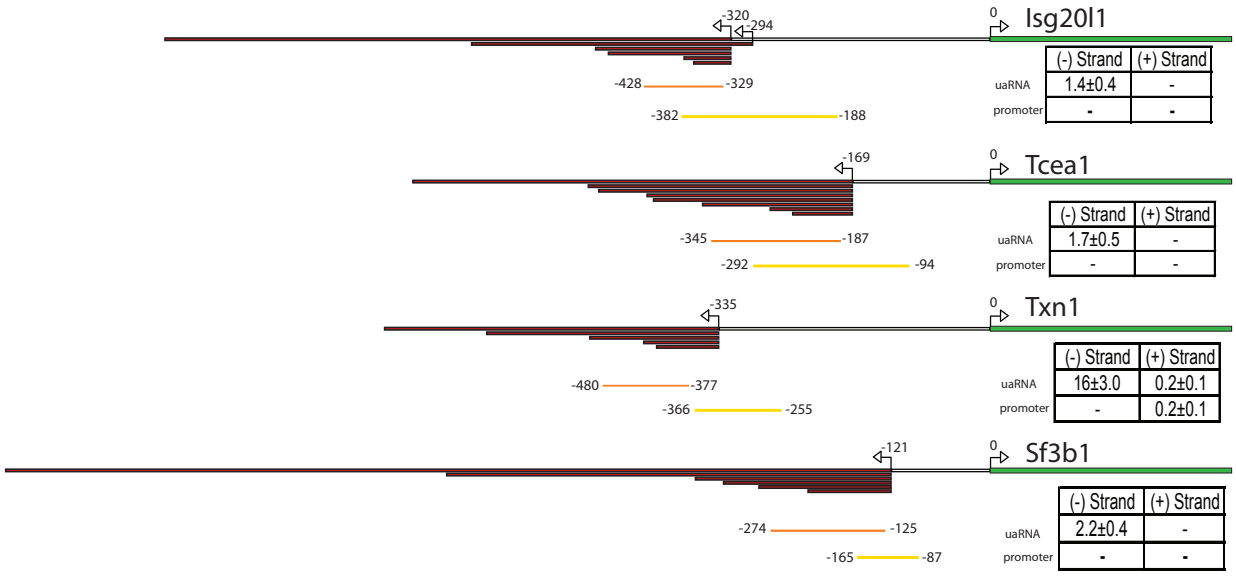


Figure S3. Strand-specific RT-qPCR on RNAs that originate in the upstream antisense or promoter regions.

Promoter regions are diagrammed for Isg20l1, Tcea1, Txn1, and Sf3b1, respectively. Sense TSSs are noted with a right facing arrow and set to position zero. Upstream antisense TSSs are noted with a left facing arrow. All uaRNA identified are represented by red bars and the first 300 nts of each sense transcript is shown as green bars. Amplicons for the "upstream" and "promoter" qPCR primers (Figure S9, Table 2) are shown for each gene as orange and yellow bars, respectively. Absolute quantitation was used to calculate copy number from strand-specific RT-qPCR reactions and copy numbers appear in the table below each sense TSS. The (-) and (+) strand columns represent transcripts amplified from either the "-" or "+" strand using a strand-specific RT primer, while the rows "upstream" and "promoter" correspond to the primer pair used in the experiment.

Figure S4

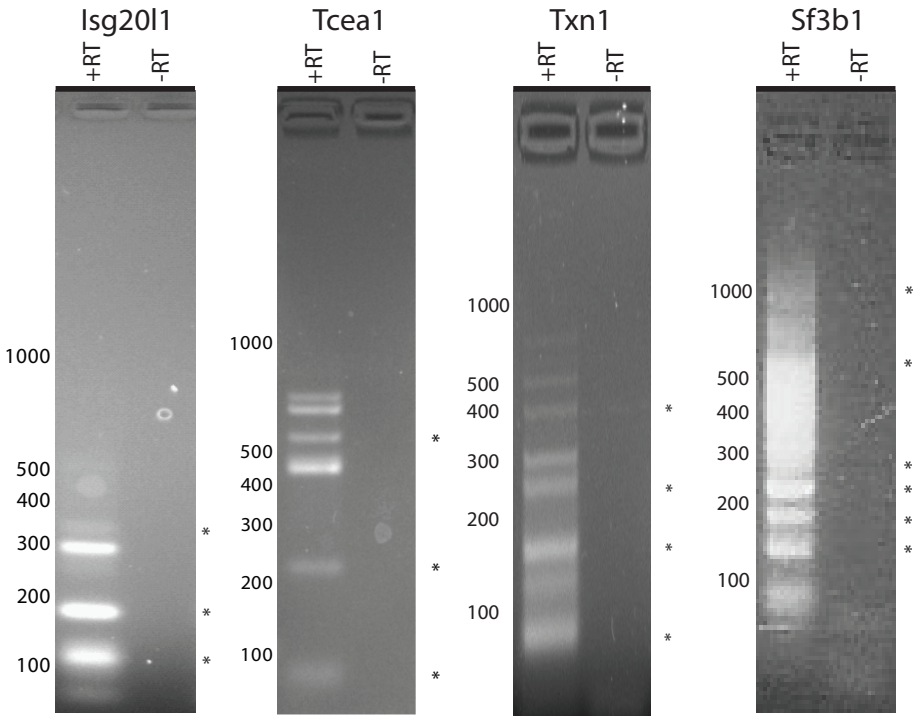


Figure S4. 3′-RACE products are reverse-transcriptase dependant.

A. 3′-RACE products were separated on a 2% agarose gel and the specific bands that were cloned and sequenced are marked with an asterisk (*) and shown in the UCSC Genome Browser in Figure 1. The above gels are examples that yielded uaRNA sequences shown in Figure 1. Samples displaying –RT did not receive were reverse transcriptase during cDNA synthesis. Size markers are shown to the left of each gel.

Figure S5

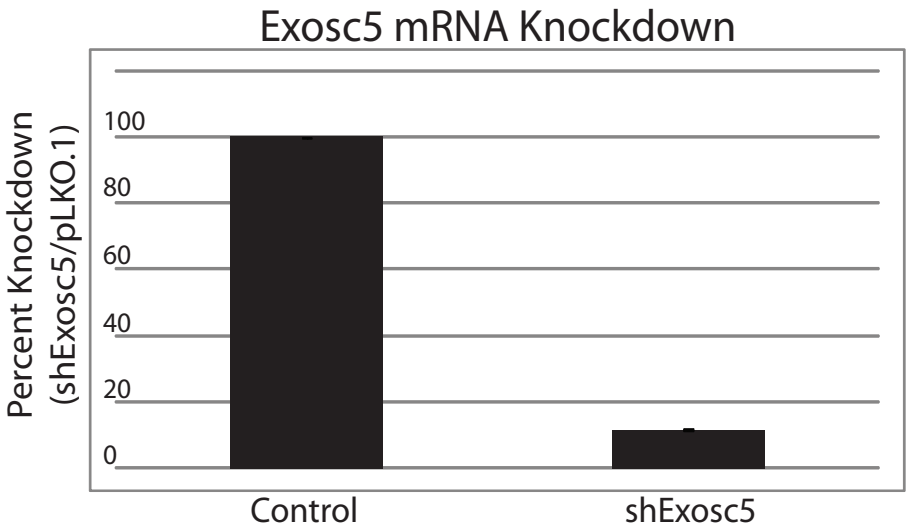


Figure 5. shRNA-mediated Exosc5 knockdown in V6.5 ES cells.

Relative gene expression of Exosc5, component of the exosome, as determined by RT-qPCR using Taqman probes (Figure S8, Table 1). Percent mRNA levels are shown compared to mock knockdown RNA samples. Values represent biological triplicates and error bars represent SEM of the biological replicates.

Figure S6

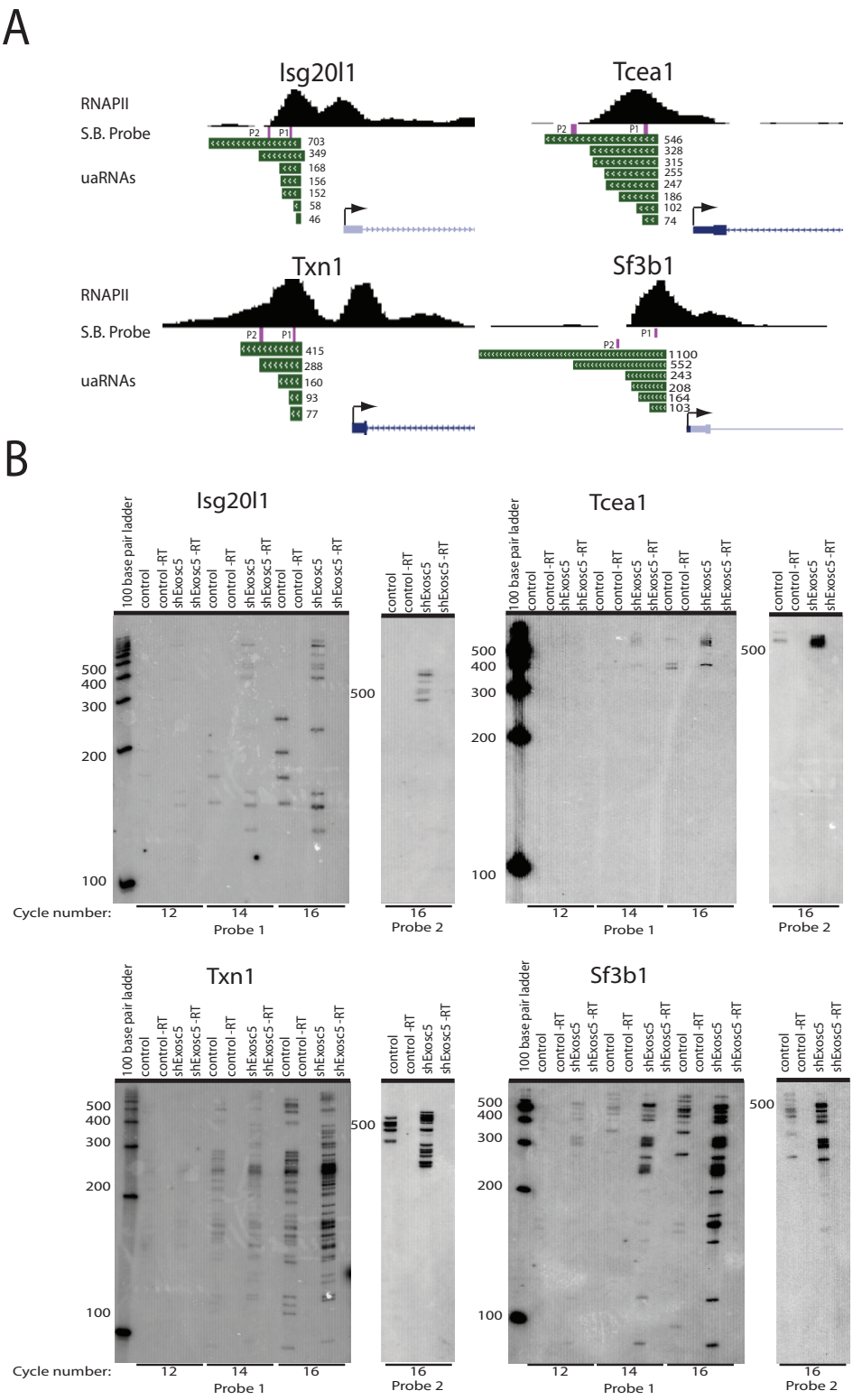

A
Isg20l1
Tcea1
RNAPII
S.B. Probe
uaRNAs
P2
P1
703
349
168
156
152
58
46
546
328
315
255
247
186
102
74
Txn1
Sf3b1
415
288
160
93
77
1100
552
243
208
164
103
B
100 base pair ladder
control
control -RT
shExosc5
shExosc5 -RT
500
400
300
200
100
Cycle number:
12
14
16
Probe 1
Probe 2

Figure S6. uaRNAs have 3′ heterogeneity and their lengths are altered upon exosome depletion.

A. Genome browser views of the four divergent promoter regions displaying ChIP-Seq signal for RNAPII, the mapped uaRNA transcripts, and the southern blot probes used in Figure 2B and Figure S6B. For each region, two probes were designed to be either proximal (probe 1) or distal (probe 2) to the antisense TSS. B. 3′-RACE followed by Southern blot analysis from either control (pLKO.1) or knockdown (shExosc5) mESCs. For each uaRNA region, both probes described above were used to visualize the RNA species. A range of PCR cycles was used during the 3′-RACE protocol to assay the most abundant transcripts as determined by the detection of the initial products. The number of cycles used for control and knockdown samples are indicated below each blot.

Figure S7

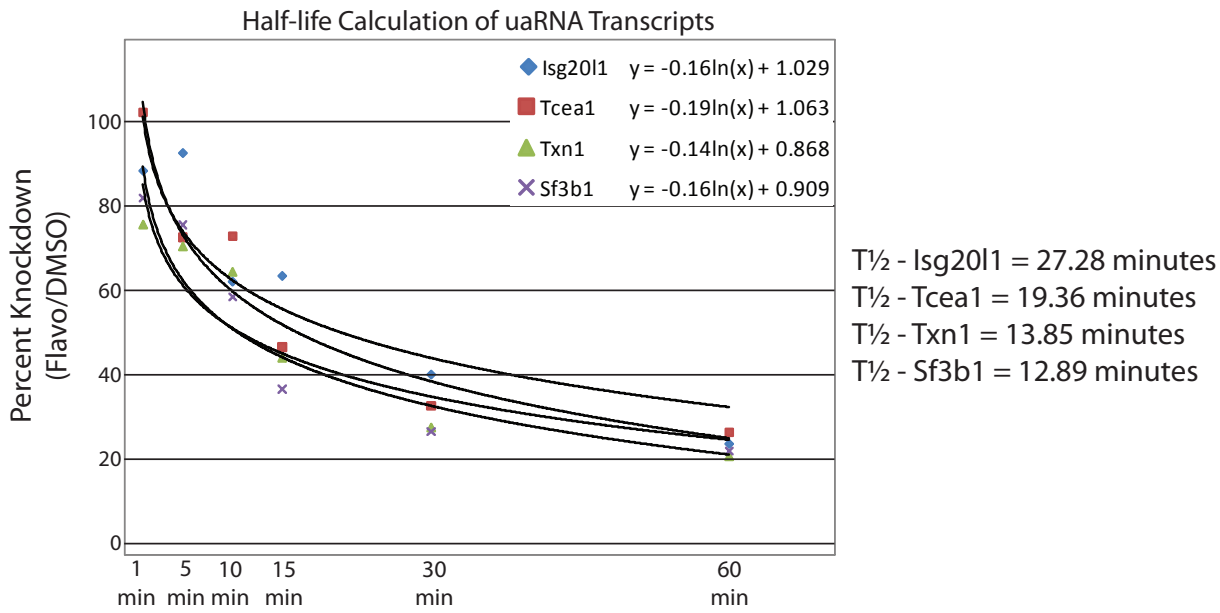


Figure S7. uaRNA half-life calculations.

Scatter plot of the relative uaRNA transcript abundances over a one hour flavopiridol treatment. For each gene, a best fit logarithmic curve was determined and equations corresponding to each gene are shown in the upper right-hand corner. Using the equations found in part A, half-lives were calculated and shown for each gene to the right.

References

Anderson, J.R., Mukherjee, D., Muthukumaraswamy, K., Moraes, K.C., Wilusz, C.J., and Wilusz, J. (2006). Sequence-specific RNA binding mediated by the RNase PH domain of components of the exosome. RNA *12*, 1810-1816.

Boyer, L.A., Plath, K., Zeitlinger, J., Brambrink, T., Medeiros, L.A., Lee, T.I., Levine, S.S., Wernig, M., Tajonar, A., Ray, M.K.*, et al.* (2006). Polycomb complexes repress developmental regulators in murine embryonic stem cells. Nature *441*, 349-353.

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.

Fejes-Toth, K. (2009). Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs. Nature *457*, 1028-1032.

Gilchrist, D.A., Dos Santos, G., Fargo, D.C., Xie, B., Gao, Y., Li, L., and Adelman, K. Pausing of RNA polymerase II disrupts DNA-specified nucleosome organization to enable precise gene regulation. Cell *143*, 540-551.

Gilchrist, D.A., Nechaev, S., Lee, C., Ghosh, S.K., Collins, J.B., Li, L., Gilmour, D.S., and Adelman, K. (2008). NELF-mediated stalling of Pol II can enhance gene expression by blocking promoter-proximal nucleosome assembly. Genes Dev *22*, 1921-1933.

Gudipati, R.K., Villa, T., Boulay, J., and Libri, D. (2008). Phosphorylation of the RNA polymerase II C-terminal domain dictates transcription termination choice. Nat Struct Mol Biol *15*, 786-794.

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.

Hochheimer, A., and Tjian, R. (2003). Diversified transcription initiation complexes expand promoter selectivity and tissue-specific gene expression. Genes Dev *17*, 1309-1320.

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

Kagey, M.H., Newman, J.J., Bilodeau, S., Zhan, Y., Orlando, D.A., van Berkum, N.L., Ebmeier, C.C., Goossens, J., Rahl, P.B., Levine, S.S.*, et al.* (2010). Mediator and cohesin connect gene expression and chromatin architecture. Nature *467*, 430-435.

Kanhere, A., Viiri, K., Araujo, C.C., Rasaiyaah, J., Bouwman, R.D., Whyte, W.A., Pereira, C.F., Brookes, E., Walker, K., Bell, G.W.*, et al.* (2010). Short RNAs are transcribed from repressed polycomb target genes and interact with polycomb repressive complex-2. Mol Cell *38*, 675-688.

Kuras, L., and Struhl, K. (1999). Binding of TBP to promoters in vivo is stimulated by activators and requires Pol II holoenzyme. Nature *399*, 609-613.

Lebreton, A., Tomecki, R., Dziembowski, A., and Seraphin, B. (2008). Endonucleolytic RNA cleavage by a eukaryotic exosome. Nature *456*, 993-996.

Lin, S., Coutinho-Mansfield, G., Wang, D., Pandit, S., and Fu, X.D. (2008). The splicing factor SC35 has an active role in transcriptional elongation. Nat Struct Mol Biol *15*, 819-826.

Livak, K.J., and Schmittgen, T.D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods *25*, 402-408.

Marson, A., Levine, S.S., Cole, M.F., Frampton, G.M., Brambrink, T., Johnstone, S., Guenther, M.G., Johnston, W.K., Wernig, M., Newman, J.*, et al.* (2008). Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. Cell *134*, 521-533.

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.

Nechaev, S., Fargo, D.C., dos Santos, G., Liu, L., Gao, Y., and Adelman, K. Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila. Science *327*, 335-338.

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.

Nitiss, J.L. (2009). Targeting DNA topoisomerase II in cancer chemotherapy. Nat Rev Cancer *9*, 338-350.

Peterlin, B.M., and Price, D.H. (2006). Controlling the elongation phase of transcription with P-TEFb. Mol Cell *23*, 297-305.

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.

Ptashne, M., and Gann, A. (1997). Transcriptional activation by recruitment. Nature *386*, 569-577.

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.

Roeder, R.G. (2005). Transcriptional regulation and the role of diverse coactivators in animal cells. FEBS Lett *579*, 909-915.

Schmid, M., and Jensen, T.H. (2008). The exosome: a multipurpose RNA-decay machine. Trends Biochem Sci *33*, 501-510.

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.

Taft, R.J., Glazov, E.A., Cloonan, N., Simons, C., Stephen, S., Faulkner, G.J., Lassmann, T., Forrest, A.R., Grimmond, S.M., Schroder, K.*, et al.* (2009). Tiny RNAs associated with transcription start sites in animals. Nat Genet *41*, 572-578.

Vasiljeva, L., Kim, M., Mutschler, H., Buratowski, S., and Meinhart, A. (2008). The Nrd1-Nab3-Sen1 termination complex interacts with the Ser5-phosphorylated RNA polymerase II C-terminal domain. Nat Struct Mol Biol *15*, 795-804.

Wang, X., Arai, S., Song, X., Reichart, D., Du, K., Pascual, G., Tempst, P., Rosenfeld, M.G., Glass, C.K., and Kurokawa, R. (2008). Induced ncRNAs allosterically modify RNA-binding proteins in cis to inhibit transcription. Nature *454*, 126-130.

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.

CHAPTER 3

Promoter directionality is controlled by U1 snRNP and polyadenylation signals

The material was adapted with permission from the following publication:

Almada AE*, Wu X*, Kriz AJ, Burge CB, Sharp PA. (2013) Promoter directionality is controlled by U1 snRNP and polyadenylation signals. *Nature* USA ahead of print (June 23)

* These authors contributed equally to this manuscript

Author contributions:

A.E.A performed poly (A) 3'-end sequencing and U1 inhibition experiments. A.E.A designed research, analyzed data, and wrote the manuscript

X.W. performed computational analysis on generated datasets and quantified U1 and PAS sites in the genome of mouse and various other metazoans. X.W designed research, analyzed data, and wrote the manuscript.

A.J.K. aided X.W. in computational analysis.

C.B.B analyzed data and wrote the manuscript

P.A.S. supervised research, analyzed data, and wrote the manuscript

Abstract

Transcription of the mammalian genome is pervasive but productive transcription outside protein-coding genes is limited by unknown mechanisms (Djebali et al., 2012). In particular, although RNA polymerase II (RNAPII) initiates divergently from most active gene promoters, productive elongation occurs primarily in the sense coding direction (Core et al., 2008; Preker et al., 2008; Seila et al., 2008). Here we show that asymmetric sequence determinants flanking gene transcription start sites (TSS) control promoter directionality by regulating promoter-proximal cleavage and polyadenylation. We find that upstream antisense RNAs (uaRNAs) are cleaved and polyadenylated at poly (A) sites (PAS) shortly after their initiation. De novo motif analysis reveals PAS signals and U1 snRNP (U1) recognition sites as the most depleted and enriched sequences, respectively, in the sense direction relative to the upstream antisense direction. These U1 and PAS sites are progressively gained and lost, respectively, at the 5' end of coding genes during vertebrate evolution. Functional disruption of U1 snRNP activity results in a significant increase in promoter-proximal cleavage events in the sense direction with slight increases in the antisense direction. These data suggests that a U1-PAS axis characterized by low U1 recognition and high density of PAS in the upstream antisense region reinforces promoter directionality by promoting early termination in upstream antisense regions whereas proximal sense PAS signals are suppressed by U1 snRNP. We propose that the U1-PAS axis limits pervasive transcription throughout the genome.

Introduction

Two potential mechanisms for suppressing transcription elongation in the upstream antisense region of gene TSS include inefficient release of paused RNAPII and / or early termination of transcription. RNAPII pauses shortly after initiation downstream of the gene TSS and the paused state is released by the recruitment and activity of p-TEFb (Adelman and Lis, 2012). A detailed characterization of several uaRNAs in mouse embryonic stem cells (mESCs) suggested that p-TEFb is recruited similarly in both sense and antisense directions (Flynn et al., 2011), and in human cells, elongating RNAPII (phosphorylated at serine 2 in the C-terminal domain) occupies the proximal upstream transcribed region (Preker et al., 2011). These data argue that the upstream antisense RNAPII complex undergoes the initial phase of elongation but likely terminates early due to an unknown mechanism.

Results

To test globally whether upstream antisense transcripts undergo early termination (compared to coding mRNA) by a canonical PAS-dependent cleavage mechanism, we mapped by deep sequencing the 3'-ends of polyadenylated RNAs in mESCs. For most protein-coding genes, transcription termination is triggered by cleavage of the nascent RNA upon recognition of a PAS whose most essential feature is an AAUAAA sequence or a close variant located about 10-30 nucleotides upstream of the cleavage site (Proudfoot, 2011). We sequenced two cDNA libraries and obtained over 230 million reads, of which 114 million mapped uniquely to the genome with at most two mismatches. We developed a computational pipeline to identify 835,942 unique 3'-ends (cleavage sites) whose poly (A) tails are likely to be added post-

transcriptionally and are also associated with the canonical PAS hexamer or its common variants (Supplementary Fig. 1, see Methods).

To investigate whether uaRNAs are terminated by PAS-dependent mechanisms, we focused our analysis on cleavage sites proximal to gene TSS and at least 5 kilobases (kb) away from known gene transcription end sites (TES). Interestingly, in the upstream antisense region we observed a 2-fold higher number of cleavage sites compared to the downstream sense sites flanking protein-coding gene TSS (Fig. 1a). The peak of the upstream antisense cleavage sites is about 700 bases from the coding gene TSS. This observation suggests that upstream antisense transcripts are frequently terminated by PAS-directed cleavage shortly after initiation, a trend we also observe in various tissues of mouse and human (Derti et al., 2012) (Supplementary Fig. 2). Inspection of gene tracks at the *Pigt* locus reveals upstream antisense cleavage shortly after a PAS (AATAAA) less than 400 bases from the *Pigt* TSS, whereas in the sense direction cleavage is confined to the TES (Fig. 1b). Similar patterns were observed for subsets of promoters (promoters without nearby genes, Global Run-On Sequencing (GRO-seq) defined divergent promoters, and Chromatin immunoprecipitation sequencing (ChIP-seq) defined RNAPII-occupied promoters (Rahl et al., 2010), or for high confidence cleavage sites, cleavage reads, and cleavage clusters (Supplementary Fig. 3). Of all divergent promoters, nearly half (48%) produce PAS-dependent upstream antisense cleavage events within 5 kb of coding gene TSS, compared to 33% downstream of the TSS. We validated several of these promoter proximal sense and antisense cleavage sites using Rapid Amplification of 3' cDNA ends (3'-RACE) (Supplementary Fig. 4).

Similar to annotated cleavage sites at TES of genes, these upstream antisense cleavage sites are associated with the PAS located at the expected position, about 22 nucleotides upstream

the cleavage site (Supplementary Fig. 5a-b). Moreover, the nucleotide sequence composition flanking the cleavage sites resembles that of TES of genes (Supplementary Fig. 5c-e) including a downstream U-rich region (Gil and Proudfoot, 1987; MacDonald et al., 1994). To determine whether members of the canonical cleavage and polyadenylation machinery bind specifically to uaRNA cleavage sites, we analyzed available cross-linking immunoprecipitation (CLIP) sequencing datasets for 10 canonical 3' end processing factors, including CPSF-160, CPSF-100, CPSF-73, CPSF-30, Fip1, CstF-64, CstF-64τ, CF I_m25, CF I_m59, and CF I_m68 along with poly (A) 3'-end sequencing data generated in HEK293 cells (Martin et al., 2012). We detect specific binding of all 10 factors at uaRNA cleavage sites with positional profiles identical or very similar to that of mRNA cleavage sites (Supplementary Fig. 6). These results indicate the poly (A) tails that we analyzed are products of PAS-dependent cleavage and polyadenylation, rather than either a priming artifact or PAS-independent polyadenylation representing a transient signal for RNA degradation (LaCava et al., 2005; Vanacova et al., 2005; Wyers et al., 2005).

As a first step to understand the molecular mechanism underlying the cleavage bias, we examined the frequency of PAS in a 6 kb region on the four strands flanking the coding gene TSS. We observed an approximately 33% depletion of the canonical AATAAA PAS hexamer specifically downstream of the TSS on the coding strand of genes as compared to the other regions (Fig. 2a). Since this 33% depletion is unlikely to explain the 2-fold cleavage bias observed (see simulation results in Supplementary Fig. 8a), we searched for additional discriminative 6-mer sequence signals in an unbiased manner. All 4096 hexamers were ranked by enrichment in the first 1 kb of the sense strand of genes relative to the corresponding upstream antisense region (Fig. 2b). Interestingly, we identified the PAS as the most depleted sequence in sense genes relative to the upstream antisense region of gene TSS. In addition, we

identified 5' splice site related sequences (or sequences recognized by U1 referred to as U1 sites) as the most enriched hexamers in sense genes (Fig. 2b) relative to antisense regions. This includes the consensus GGUAAG (first) that is perfectly complementary to the 5' end of the U1 snRNA, as well as GGUGAG (third) and GUGAGU (fifth), which represent common 5' splice site sequences (with the first GU in each motif located at the intron start). Consistent with the hexamer enrichment analysis, a metagene plot displaying an unbiased prediction of strong, medium, and weak U1 sites (see Methods) revealed strong enrichment of U1 signals in the first 500 bps downstream of the TSS, with essentially only background levels observed in all other regions and a small depletion in the upstream antisense direction (Fig. 2c).

The asymmetric distribution of U1 sites and PAS sites flanking the TSS could potentially explain the biased cleavage pattern shown in Fig. 1a if the U1 complex suppresses cleavage and polyadenylation near a U1 site, as has been observed in various species including human and mouse (Andersen et al., 2012; Berg et al., 2012; Kaida et al., 2010). Consistent with this model, we observed a depletion of cleavage sites, especially frequent cleavage sites, downstream of strong U1 sites (Supplementary Fig. 7a). Focusing on the upstream antisense direction, the presence of proximal PAS sites (within 1 kb of coding gene TSS) is significantly associated with shorter uaRNAs (p < 1e-15), whereas the presence of proximal U1 sites is significantly associated with longer uaRNAs but only in the presence of proximal PAS sites (p < 0.0006), consistent with a model where U1 promotes RNA lengthening by suppressing proximal PAS (Supplementary Fig. 7b). To test whether the encoded bias in U1 and PAS signal distribution explains the cleavage bias observed from our 3'-end sequencing analysis, we performed a cleavage site simulation using predicted strong U1 sites and canonical PAS (AATAAA) sequences. Specifically, we defined a protection zone of 1 kb downstream of a strong U1 site and

used the first unprotected PAS as the cleavage site. The metagene plot of simulated cleavage events (Fig. 2d) recapitulate the major features of the observed distribution (Fig. 1a), including an antisense peak around 700 bases upstream and a ~2-fold difference between sense and antisense strands. Similar patterns were robustly observed when varying the size of the protection zone (Supplementary Fig. 8). Thus, we identified a U1-PAS axis flanking gene promoters that may explain why uaRNAs undergo early termination.

To validate the U1-PAS axis model, we functionally inhibited U1 in mESCs. Specifically, we transfected mESCs with either an antisense morpholino oligonucleotide (AMO) complementary to the 5' end of U1 snRNA to block its binding to 5' splice sites (or similar sequences) or a control AMO with scrambled sequences followed by 3'-end RNA sequencing (Berg et al., 2012; Kaida et al., 2010). Interestingly, we observe in two biological replicates a dramatic increase in promoter-proximal cleavage events in coding genes but only a slight increase in upstream antisense regions, which eliminates the asymmetric bias in promoter-proximal cleavage we observed in either the wild-type cells or cells treated with scrambled control AMOs (Fig. 3). These observations confirm that U1 protects sense RNA in protein-coding genes from premature cleavage and polyadenylation in promoter proximal regions, thus, reinforcing transcriptional directionality of genes. However, in the antisense direction, the activity of U1 is much less and there is little enhancement in cleavage sites upon inhibition of U1 recognition.

The conservation of the asymmetric cleavage pattern across human and mouse (Supplementary Fig. 2) led us to examine if there is evolutionary selection on the U1-PAS axis. Previously, mouse protein-coding genes have been assigned to 12 evolutionary branches and dated by analyzing the presence or absence of orthologs in the vertebrate phylogeny (Zhang et

al., 2010). We find strong trends of progressive gain of U1 sites depending on the age of a gene (Fig. 4a) and loss of PAS sites (Fig. 4b) over time at the 5' end (the first 1 kb) of protein-coding genes, suggesting that suppression of promoter-proximal transcription termination is important for maintaining gene function. Interestingly, the same trends, although weaker, are observed in upstream antisense regions, suggesting at least a subset of uaRNAs may be functionally important in that over time they gain U1 sites and lose PAS sites to become more extensively transcribed. In addition to the coding strand of genes (downstream sense region), PAS sites were also progressively lost on the other three strands flanking TSS (Fig. 4b). This observation probably reflects on the increases in CpG-rich sequences within 1 kb of gene TSS and suggests that coding genes acquire CpG islands as they age (Fig. 4c). However, the bias of low PAS site density in the sense direction extends across the total transcription unit (Supplementary Fig. 9) and is distinct from the CpG density near the promoter.

We also propose that some long noncoding RNAs (lncRNAs) generated from bidirectional promoters might represent an evolutionary intermediate between uaRNAs and protein-coding genes. Consistent with this, annotated head-to-head mRNA-lncRNA pairs as a whole showed a bias (in terms of promoter-proximal cleavage site, U1 site, and PAS site distributions flanking coding gene TSS) weaker than head-to-head mRNA-uaRNA pairs but stronger than mRNA-mRNA pairs (Supplementary Fig. 10). This is also consistent with recent results suggesting that *de novo* protein-coding genes originate from lncRNAs at bidirectional promoters (Xie et al., 2012).

The U1-PAS axis likely has a broader role in limiting pervasive transcription throughout the genome. The enrichment of U1 sites and depletion of PAS sites are confined to the sense strand within the gene body, whereas intergenic and antisense regions show relatively high PAS

but low U1 density (Supplementary Fig. 9), indicating the U1-PAS axis may serve as a mechanism for terminating transcription in both antisense and intergenic regions.

Discussion

Together, we propose that a U1-PAS axis is important in defining the directionality for transcription elongation at divergent promoters (Supplementary Fig. 11). Although the U1-PAS axis may explain the observed cleavage bias at promoters surprisingly well, it seems likely that additional cis-elements may influence PAS usage (Hu et al., 2005) and will need to be integrated into this model. There may also be other PAS-independent mechanisms that contribute to termination of transcription in upstream antisense regions and across the genome (Arigo et al., 2006; Connelly and Manley, 1989; Zhang et al., 2013). However, evidence for the U1-PAS axis is found in several different tissues of mouse and human, indicating its wide utilization as a general mechanism to regulate transcription elongation in mammals. Like protein-coding transcripts, lncRNAs must also contend with the U1-PAS axis. These RNAs and short non-coding RNAs from divergent transcription of gene promoters may be considered part of a continuum that varies in the degree of the activity of the U1-PAS axis.

Acknowledgements

The authors wish to dedicate this paper to the memory of Officer Sean Collier, for his caring service to the MIT community and for his sacrifice. We would like to thank Noah Spies for generously sharing his optimized 3'-end sequencing protocol, Charles Lin for providing computational assistance, Mary Lindstrom for assistance on constructing Supplementary Figure 11, and Sidi Chen, Anthony Chiu, Mohini Jangi, Qifang Liu, Jeremy Wilusz, and Jesse Zamudio for critical reading of the manuscript. We also thank the Core Facility in the Swanson Biotechnology Center at the David H. Koch Institute for Integrative Cancer Research at M.I.T for their assistance with high-throughput sequencing. This work was supported by United States Public Health Service grants RO1-GM34277 and R01-CA133404 from the National Institutes of Health (P.A.S.), partially by Cancer Center Support (core) grant P30-CA14051 from the National Cancer Institute, and by Public Health Service research grant (GM-085319) from the National Institute of General Medical Sciences (C.B.B.). X.W. is a Howard Hughes Medical Institute International Student Research fellow.

Methods

Cell Cuture. V6.5 (C57BL/6-129) mouse embryonic stem cells (mESCs) (Koch Institute Transgenic Facility) were grown under standard ES cell culture conditions (Seila et al., 2008).

Poly(A) 3'-End sequencing. Total RNA was extracted from V6.5 mESCs using Ambion's Ribopure kit (AM1924M). Poly (A) selected RNA was fragmented using RNase T1 (AM2283). Reverse transcription was performed with an RT oligo (Table S1) at 0.25 uM final concentration using Invitrogen's Superscript III Reverse Transcriptase (18080-44) according to the manufacturer's protocol. The resulting cDNA was run on a 6% TBE-Urea polyacrylamide gel (National Diagnostics) and the 100-300 size range of products were gel extracted and eluted overnight. The gel-purified cDNA products were circularized using CircLigase II (CL9025K) according to the manufacturer's protocol. Circularized cDNA was PCR-amplified using the Phusion High-Fidelity DNA Polymerase (MO530L) for 15-18 cycles using the primers described in Table S1. Amplified products were run on a 1.5 % agarose gel and the 200-400 size range was extracted using Qiagen's MinElute Gel Extraction Kit (28604). The 3'-end library was then submitted for Illumina sequencing on the HI-Seq 2000 platform.

U1 inhibition with antisense morpholino oligonucleotides (AMO). V6.5 mESCs were transfected using the Amaxa Nucleofector II with program A-23 (mESC-specific) according to the manufacturers protocol. Specifically, 2.5 million V6.5 mESCs were transfected with 7.5 uM of U1-targeting or a scrambled AMO for 8 hrs, (Berg et al., 2012; Kaida et al., 2010) prior to RNA sequencing analysis.

3'-RACE. Total RNA was extracted using Ambion's Ribopure kit and DNase-treated using Ambion's DNA Free-Turbo. 3'-RACE was performed using Ambion's Gene Racer Kit according to the manufacturer's instructions. 3'-end PCR products were run on a 1.5% agarose gel, gel extracted using Qiagen's gel extraction kit, and Sanger sequenced. All primers are described in Table S1.

Reads mapping. Raw reads were processed with the program *cutadapt* (Martin et al., 2011) to trim the adaptor sequence (TGGAATTCTCGGGTGCCAAGGAACTCCAGTCACATCAC) from the 3' end. Reads longer than 15 nts after adaptor trimming were mapped to the mouse genome (mm9) with *bowtie*[31] requiring unique mapping with at most two mismatches (options: -n 2 -m 1 --best --strata). Mapped reads were collapsed by unique 3' end positions.

Internal priming filter. To remove reads whose A-tail is encoded in the genome rather than added post-transcriptionally, we filtered reads that have 1) more than 10 As in the first 20 nt window or 2) more than 6 As in the first 10 nt window downstream from the detected cleavage site of the 3'-end. The threshold used is based on the bimodal distribution of the number of As downstream of annotated TES.

PAS filter. In addition to a set of 12 hexamers identified previously in mouse and human EST analysis (Beaudoing et al., 2000; Tian et al., 2005), we analyzed the annotated TES in the mouse genome to identify additional potential PAS variants. All hexamers with at most two mismatches to the canonical AATAAA motif were used to search in the sequence up to 100 nts upstream of annotated TES. The distribution of the position of each hexamer relative to the TES (a

histogram) is compared to that of AATAAA. Hexamers with a position profile similar to AATAAA will have a peak around position 20-24. We quantified the similarity by Pearson correlation coefficient and used a cut-off of 0.5 after manual inspection. In total, 24 new hexamers were identified as potential PAS and a hierarchy was assigned for the 36 hexamers (PAS36): first, the 12 known variants are ranked by their frequency of usage in the mouse genome, and then the newly identified PAS ranked by their correlation with AATAAA in terms of the positional profile defined above. To define a window where most PAS or variants are located, we searched for each of the 36 PAS variants within 100 nts of annotated gene 3' ends and chose the best one according to the designated hierarchy. We summarized the distance of the best PAS to the annotated TES and defined a window of (0-41) around the position 22 peak such that 80% of the annotated TES have their best-matched PAS within that window. Using these criteria, we searched for PAS36 variants within the 0-41 window upstream of our experimentally sequenced 3'-ends. If there were multiple PAS hexamers identified within this window for a given 3'-end, we chose the best one defined by the hierarchy described above. Reads without any of the 36 PAS variants within the 0-41 window were discarded.

Remove potential false positive cleavage sites. Due to sequencing error, abundant transcripts such as ribosomal gene mRNAs can produce error-containing 3' end reads that mapped to other locations in the genome, leading to false positive cleavage sites. To remove such potential false positive sites, we defined a set of 71674 (7.5%) abundant cleavage sites that are supported with more than 100 reads from the pooled library. A bowtie reference index was built using sequences within 50 nts upstream of those abundant sites. Non-abundant sites within these 50 nts reference regions were not used to search for false positives. Reads initially mapped to sites outside these

reference regions were re-mapped against the new index allowing up to two mismatches. Reads mapped to any of the reference regions in this analysis were treated as potential false positive reads. Cleavage sites containing only potential false positive reads are defined as potential false positive sites and were removed from subsequent analysis. In total, 7.2% (389185) of initially mapped reads are outside the reference regions. 0.34% of all mapped reads were classified as potential false positive reads and 9.1% (86425) of all cleavage sites were identified as potential false positive sites.

Remove B2 SINE RNA associated cleavage sites. We further removed cleavage sites associated with B2_Mm1a and B2_Mm1t SINE RNAs. These B2 SINE RNAs are transcribed by RNA Pol III but contain AAUAAA sequences near the 3' end. In total, 3.5% (33696) of all cleavage sites passing the internal priming filter and the PAS filter were mapped within B2 regions or within 100 nts downstream of B2 3' end. These sites were removed.

Prediction of U1 sites / putative 5' splice sites. A nucleotide frequency matrix of 5' splice sites (3 nt in exon and 6 nt in intron) was compiled using all annotated constitutive 5' splice sites in the mouse genome. The motif was then used by FIMO (Grant et al., 2011)to search significant matches ($p<0.05$) on both strands of the genome. Matches were then scored by a Maximum Entropy model (Yeo and Burge, 2004). Maximum entropy scores for all annotated 5' splice sites were also calculated to define thresholds used to classify the predicted sites into strong, medium and weak. Sites with scores larger than the median of annotated 5' splice sites (8.77) were classified as 'strong'. Sites with scores lower than 8.77 but higher than the threshold dividing the first and second quarter of annotated 5' splice sites (7.39) were classified as 'medium', and the

rest of the predicted sites with scores higher than 4 were classified as 'weak'. Sites with scores lower than 4 were discarded.

Define a set of divergent promoters. GRO-seq data from mESCs (Min et al., 2011) were used to define a set of active and divergent promoters. Active promoters were defined as promoters with GRO-seq signal detected within the first 1 kb downstream sense strand. A promoter was considered divergent if it contained GRO-Seq signal in the first 1 kb downstream the sense strand and within the first 2 kb of the upstream antisense strand. A minimum number of two reads within the defined window (downstream 1 kb or upstream 2 kb) were used as a cut-off for background signals.

Define RNAPII Ser5P bound TSS. ChIP-seq data for ser5p RNA Pol II and corresponding input was downloaded from GEO database (accession number GSE20530 (Rahl et al., 2010)) and peaks called using MACS (Zhang et al., 2008) with default settings. TSS less than 500bps away from a peak summit are defined as bound.

Discriminative hexamer analysis. An unbiased exhaustive enumeration of all 4096 hexamers was performed to find hexamers that are discriminative of downstream sense and upstream antisense strands of protein-coding gene promoters. Specifically, the first 1000 nucleotides downstream sense and upstream antisense of all protein-coding gene TSS were extracted from repeat masked genome (from UCSC genome browser, non-masked genome sequence gave similar results). For each hexamer, the total number of occurrences on each side was counted and

then the log2 ratio of the occurrences on sense versus antisense strand was calculated as a measure of enrichment on the sense but depletion on the antisense strand.

Cleavage site simulation. Protein-coding genes and 10 kb upstream antisense regions were scanned for strong U1 sites and PAS sites (AATAAA). Starting from protein coding gene TSS, the first unprotected PAS was predicted to be the cleavage site. A PAS is protected only if it is within a designated protection window (in nucleotides) downstream (+) of a strong U1 site.

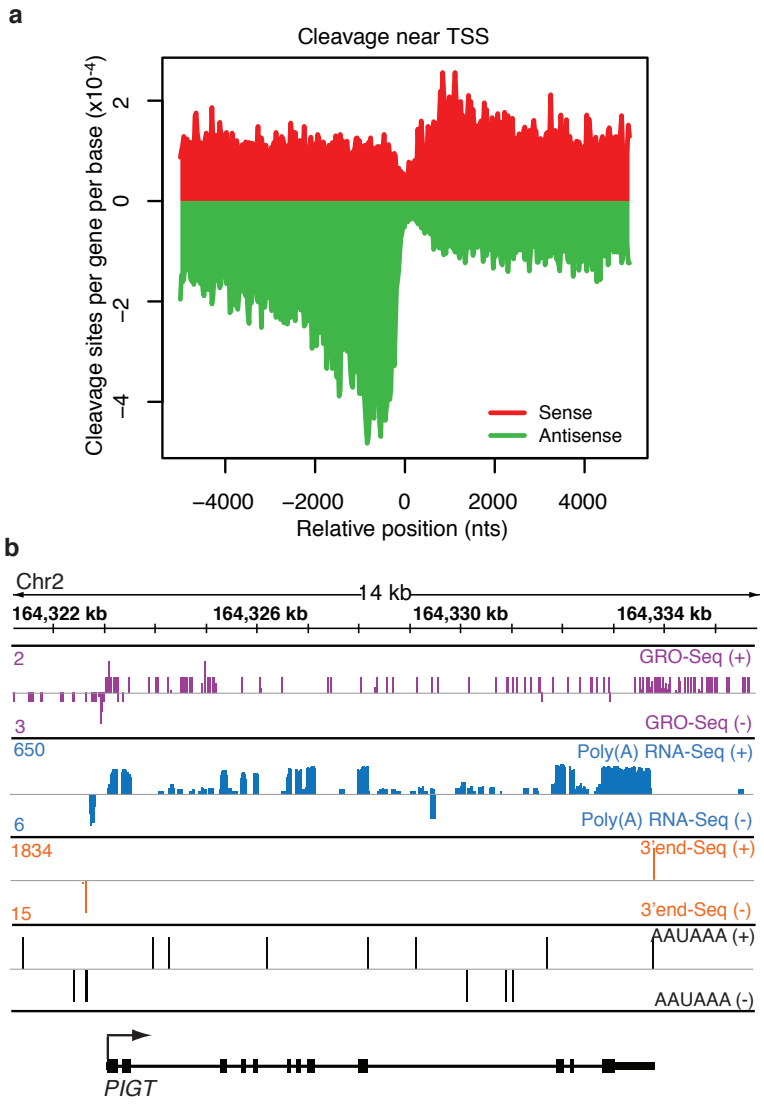
Binding of 3' end processing factors in uaRNA regions. RNA 3' end cleavage and polyadenylation sites and CLIP-seq read density of ten 3' end processing factors in wild type HEK293 cells were downloaded from Gene Expression Omnibus (GEO) dataset GSE37401. A cleavage site is defined as a uaRNA cleavage site if it is outside any protein-coding gene but locates within 5 kb upstream antisense of a protein-coding gene. mRNA cleavage sites are defined as cleavage sites within 100 bases of annotated protein-coding gene ends. For each 3' end-processing factor, CLIP read density within 200 bases of all cleavage sites are added up every 5bp bin and then normalized such that the max value is 1.

Evolutionary analysis of U1 sites, PAS sites, and CpG islands. Mouse protein-coding gene branch/age assignment was obtained from a previous analysis (Zhang et al., 2010). The number of strong U1 sites, PAS (AATAAA) sites, and CpG islands (UCSC mm9 annotations) in the first 1 kb region flanking TSS on each strand were calculated, and the average number of sites in each branch/age group was plotted against gene age. Pearson correlation coefficient and linear regression fitting were done using R. Significance of the correlation was assessed by comparing

to a null distribution of correlation coefficients calculated by shuffling gene branch/age assignments 1000 times.

Bidirectional promoter analysis. For each annotated TSS the closest upstream antisense TSS was identified and those TSS pairs within 1 kb were defined as head-to-head pairs. LncRNAs were defined as noncoding RNAs longer than 200 bps. UCSC mm9 gene annotations were used in this analysis.

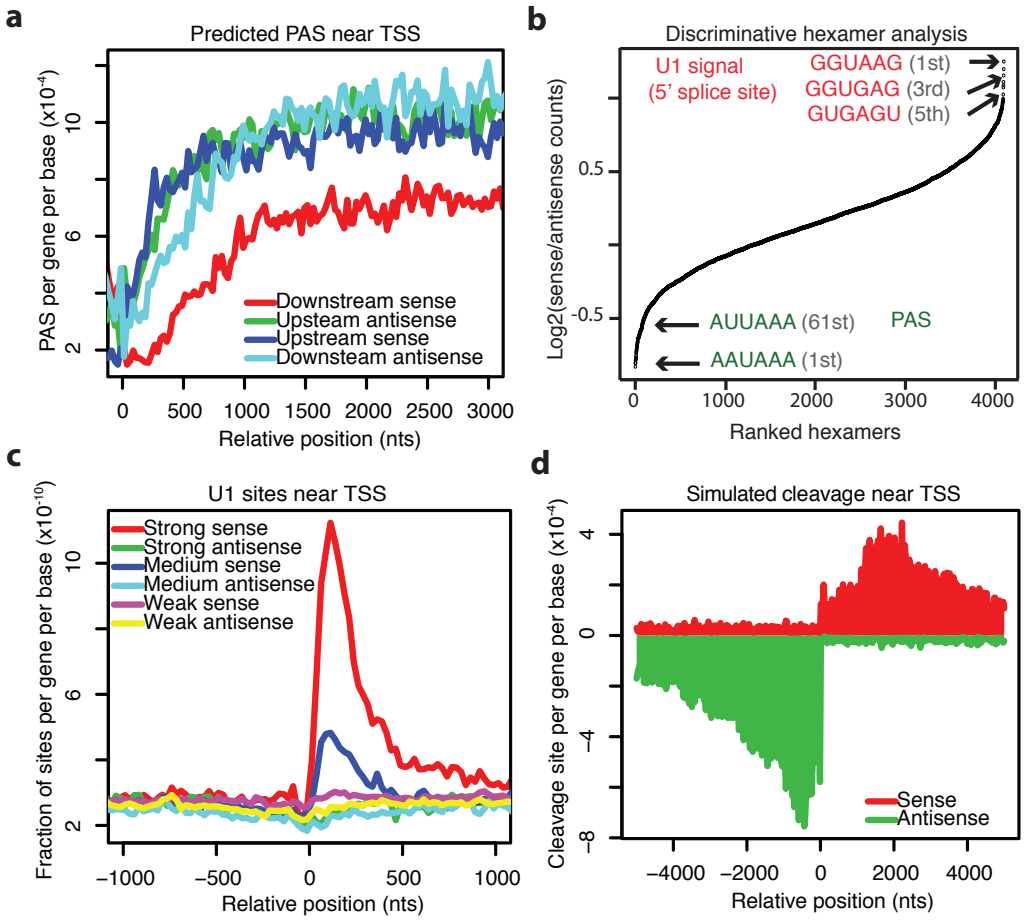
Figure 1



Promoter-proximal PAS-dependent termination of uaRNA. **a**, Metagene plot of sense (red) or antisense (green) unique cleavage sites flanking coding gene TSS. The number of unique cleavage sites per gene per base in each 25 bp bin across 5 kb upstream and downstream of the TSS is plotted. Mean cleavage density of first 2 kb: sense/antisense = 1.45/3.10. **b**, Genome browser view from the *PIGT* locus (shown in black on the + strand) displaying the following

tracks with + strand (top) and – strand (bottom) represented: GRO-Seq (purple)(Min et al., 2011), Poly (A)+ RNA-Seq (blue) (Sigova et al., 2013), 3'end RNA-Seq (orange), and PAS (AAUAAA, black). For each gene track, the x-axis represents the linear sequence of genomic DNA. The numbers on the top left corner represent the maximum read density on each track.

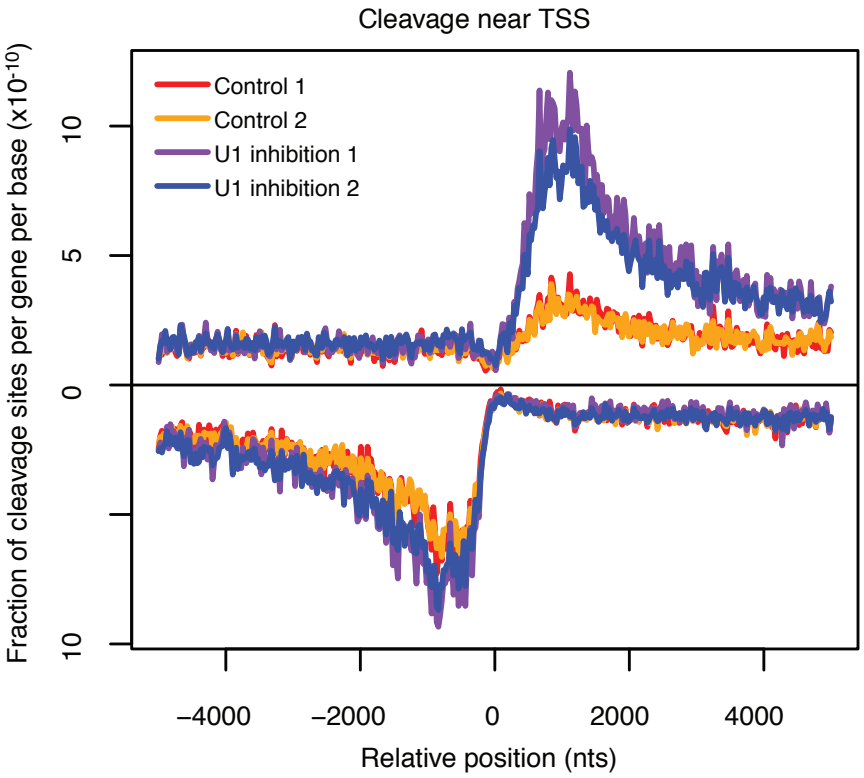
Figure 2



Asymmetric distribution of PAS and U1 signals flanking coding gene TSS. **a**. Number of AATAAA sites per gene per base in each 25 bp bin within a 3 kb region flanking gene TSS on

the downstream sense (red), downstream antisense (light blue), upstream antisense (green), and upstream sense (dark blue) strands. **b**, Rank of all 4096 hexamers by enrichment (log2 ratio) in the first 1 kb of all coding genes in the sense direction relative to 1 kb in the upstream antisense direction of the TSS. **c**, Density of predicted 5' splice sites within a 1 kb region flanking gene TSS. Strong, medium, and weak 5' splice sites are defined in Methods. **d**, Metagene plot of simulated cleavage sites around gene TSS. The first unprotected PAS (AAUAAA) that is not within 1 kb downstream of a strong U1 site for all coding genes is plotted. Mean cleavage density of first 2 kb: sense/antisense = 2.08/4.99.

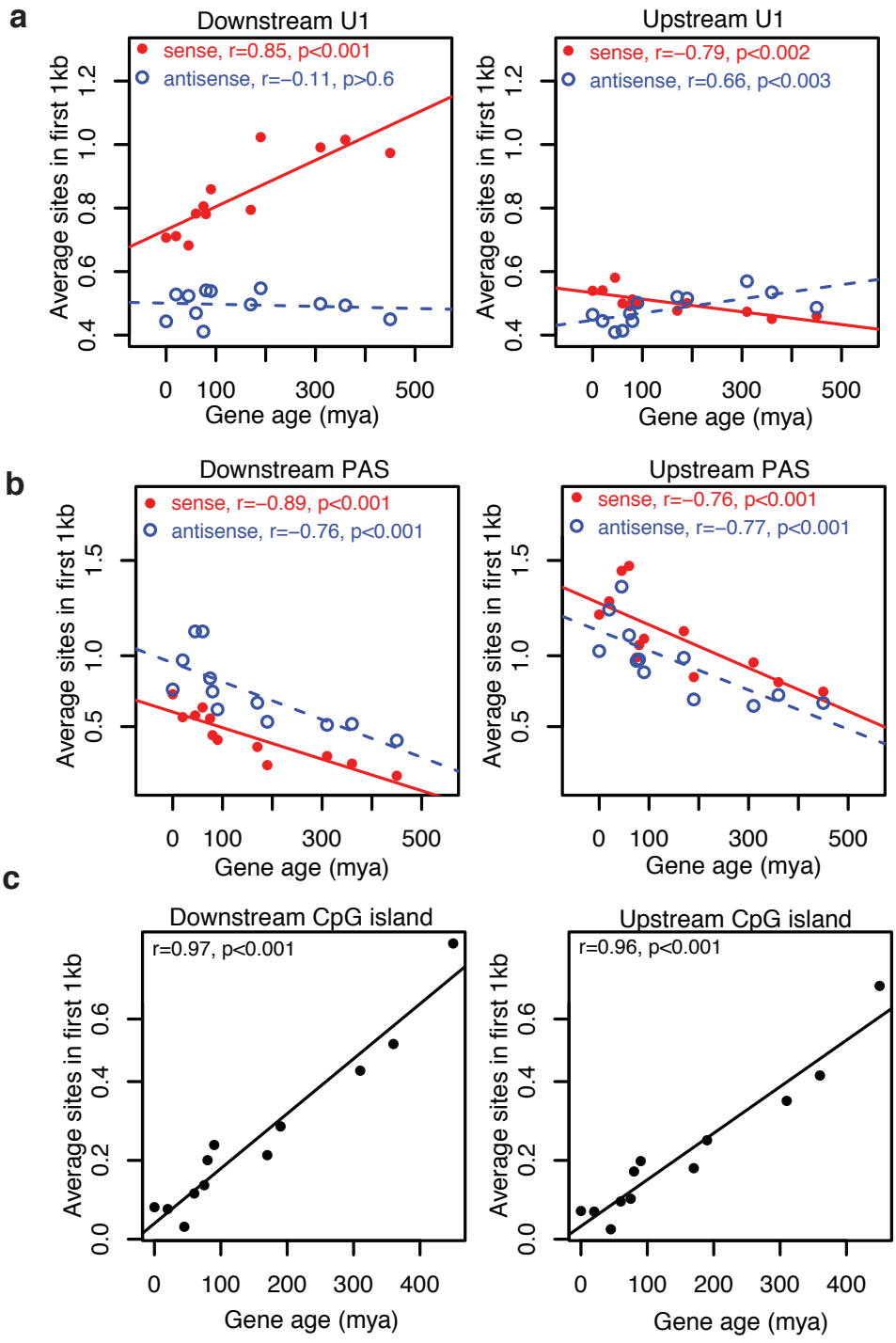
Figure 3



Promoter-proximal cleavage sites are altered upon functional U1 inhibition.

Y-axis is the number of cleavage sites per gene per base divided by the total number of cleavage sites identified in each 3' end-sequencing library in a 5 kb region flanking coding gene TSS. Signal for the antisense strand is set as negative. U1 inhibition 1 (purple) and U1 inhibition 2 (blue) represent 3'-end sequencing libraries generated from mESCs treated with a U1-targeting AMO. Control 1 (red) and Control 2 (orange) represent 3'-end sequencing libraries generated from mESCs treated with a scrambled control AMO. Mean cleavage density of first 2 kb: sense/antisense = 2.5/4.4 (Control 1), 2.4/4.3 (Control 2), 7.0/5.8 (U1 inhibition 1), 5.9/5.5 (U1 inhibition 2).

Figure 4


a
Downstream U1
sense, r=0.85, p<0.001
antisense, r=−0.11, p>0.6
Upstream U1
sense, r=−0.79, p<0.002
antisense, r=0.66, p<0.003
Average sites in first 1kb
Gene age (mya)
b
Downstream PAS
sense, r=−0.89, p<0.001
antisense, r=−0.76, p<0.001
Upstream PAS
sense, r=−0.76, p<0.001
antisense, r=−0.77, p<0.001
c
Downstream CpG island
r=0.97, p<0.001
Upstream CpG island
r=0.96, p<0.001

Evolutionary gain and loss of U1 and PAS sites. **a**, Average number of strong U1 sites in the first 1 kb of protein-coding genes and upstream regions. **b**, Average number of PAS sites in the first 1 kb downstream and upstream of coding gene TSS, respectively. **c,** Average number of CpG islands overlapping the first 1 kb of protein-coding genes and upstream regions. Genes are divided into 12 ordered groups by gene age. X-axis indicates the age (myr, million years) of gene groups. The number of genes in each group (from old to young): 11934, 1239, 914, 597, 876, 1195, 279, 175, 198, 315, 926, and 1143. Solid red dots and blue circles indicate sites on the sense and antisense strands, respectively.

Table S1

3’-End Sequencing

Name	Sequence (5'-3')
RT oligo	/5phos/TGGAATTCTCGGGTGCCAAGGAACTCCA GTCAC/iSp18/CTTTCCCTACACGACGCTCTTCCGAT CTTTTTTTTTTTTTTTTTTTVN
Illumina_PCR_F	AATGATACGGCGACCACCGAGATCTACACTCTTT CCCTACACGACGCTCTTCCGATCT
Illumina_PCR_R	CAAGCAGAAGACGGCATACGAGATCGTGATGTG ACTGGAGTTCCTTGGCACCCGAGAATTCCA

3’-RACE

Name	Sequence (5'-3')
GeneRacer oligo-dt (RT)	GCTGTCAACGATACGCTACGTAACGGCATGACA GTG(T)24VN
GeneRacer 3′ Primer	GCTGTCAACGATACGCTACGTAACG
GeneRacer 3′ Nested Primer	CGCTACGTAACGGCATGACAGTG
Pgm2_Gsp-1	GGTCTATGGGAGAAGAGAGAACAGG
Pgm2_Gsp-2	GATCAAGAACTTCAGAAAACCCTGACC
Ccm2_Gsp-1	GGA CAT GGT TCA AGA GGC TCC TTC G
Ccm2_Gsp-2	GTT CAA GAG GCT CCT TCG TTT CTG TAG
Zcchc2_Gsp-1	CTG CCG GTC AAG AGT TCT TGG C
Zcchc2_Gsp-2	GAG TTC TTG GCA TGC TGT TTG TCA GTC C
Mapk4_Gsp-1	GGA TTG CTG CCA ATG CCT AGT AAC CTG
Mapk4_Gsp-2	GCC TAG TAA CCT GTA TTT GAT AGC CAG G

Oligonucleotides for 3’-end sequencing and 3’-RACE. (T)24 denotes the 24 ‘T’ nucleotides at the 3’-end of the RT oligonucleotide. V denotes bases (A,G,C) while N denotes bases (A,T,G,C).

Table S2

	WT.pooled	WT.1	WT.2
raw reads	-	105878435	151249729
>15nt after adaptor trimming	230652917	90198305	140454612
uniquely mapped reads	113951376	45503131	68448245
reads mapped to mulitple locations	98227112	38930372	59296740
reads failed to map	18474429	5764802	12709627

	Scr.1	Scr.2	U1.1	U1.2
raw reads	76743312	162067250	45149070	160547799
>15nt after adaptor trimming	74861868	148304257	45138339	148855654
uniquely mapped reads	34759636	75570463	20695978	75100565
reads mapped to mulitple locations	32979897	64108442	19607666	63659018
reads failed to map	7122335	8625352	4834695	10096071

Summary statistics for adaptor trimming and genome mapping

Tables S3

WT Pooled

	unique cleavage sites	mapped reads
starting reads/3' ends	1598504	113951376
after internal priming filter	1173847	102627131
after PAS filter	953864	99043279
potential false positive	86425	389185
moving potential false positi	867439	98654094
B2 repeat associated sites	33696	357080
emoving B2 repeat associate	835942	98411432

WT 1

starting reads/3' ends	unique cleavage sites	mapped reads
after internal priming filter	1263848	45503131
after PAS filter	929350	40427355
potential false positive	744611	38569397

WT 2

starting reads/3' ends	unique cleavage sites	mapped reads
after internal priming filter	752851	68448245
after PAS filter	566827	62199776
potential false positive	500472	60473882

Scr 1

starting reads/3' ends	unique cleavage sites	mapped reads
after internal priming filter	1588278	34759636
after PAS filter	1004222	28408510
potential false positive	804486	24500514

Scr 2

starting reads/3' ends	unique cleavage sites	mapped reads
after internal priming filter	1964575	75570463
after PAS filter	1366425	65708494
potential false positive	1039972	57343146

U1 1

starting reads/3' ends	unique cleavage sites	mapped reads
after internal priming filter	1349625	20695978
after PAS filter	844187	16704681
potential false positive	683584	14292669

U1 2

starting reads/3' ends	unique cleavage sites	mapped reads
after internal priming filter	2443769	75100565
after PAS filter	1584859	61525590
potential false positive	1193338	53109971

Summary statistics for cleavage sites identification and filtering

Tables S4

a

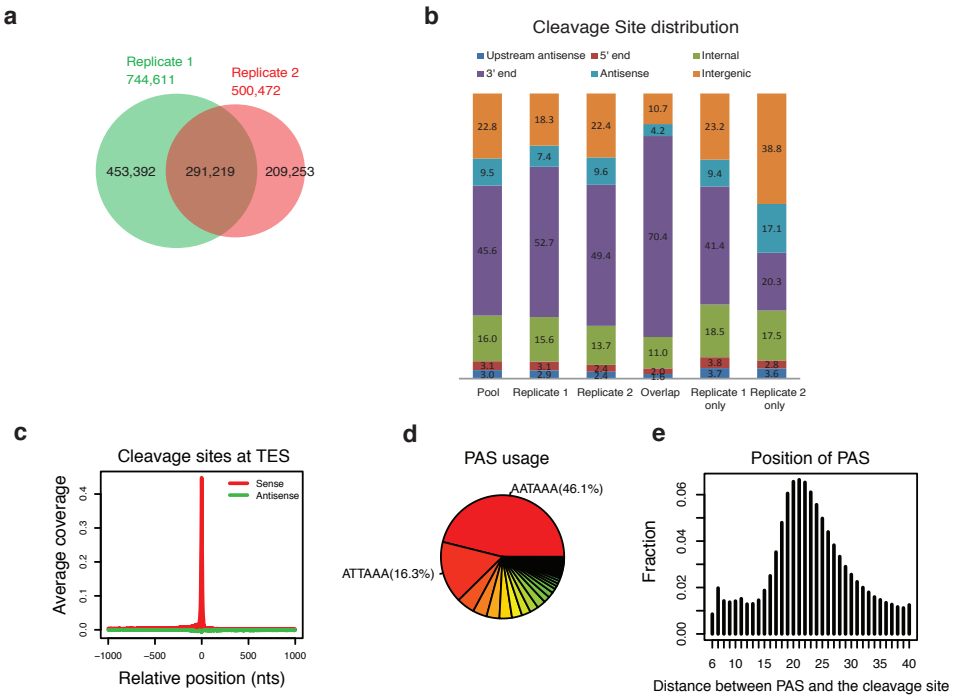
PAS hexmaer	Percentage
AATAAA	46.1
ATTAAA	16.3
TATAAA	4.7
AGTAAA	3.7
AAAAAA	3.4
TAAAAA	3.3
AAGAAA	2.7
AATATA	2.3
CATAAA	2.3
AATACA	2.2
GATAAA	1.7
AATGAA	1.5
ACTAAA	1.2
AAATAA	0.8
AATAGA	0.8
AAAACA	0.7
AAAATA	0.7
AAAAGA	0.5
AATAAT	0.5
AATTAA	0.5
TATTAA	0.4
AACAAT	0.4
AAAAAT	0.4
AACAAA	0.3
AAGAAT	0.3
AAACAA	0.3
AAAGAA	0.3
CATTAA	0.3
AAAAAG	0.2
CATAAT	0.2
ACTAAT	0.2
AAGTAA	0.2
AAAAAC	0.2
GATTAA	0.2
ATAAAA	0.1

b

PAS hexmaer	Percentage
AATAAA	50.2
ATTAAA	16.2
AAAAAA	4.5
TAAAAA	3.7
AGTAAA	3.6
TATAAA	3.4
AATACA	2.1
CATAAA	2.1
AAGAAA	2
AATATA	1.7
GATAAA	1.4
ACTAAA	1.1
AATGAA	1.1
AATAGA	0.8
AAAACA	0.6
AAATAA	0.6
AAAATA	0.5
AATTAA	0.4
AATAAT	0.4
AAAAGA	0.3
AAACAA	0.3
AAGTAA	0.3
AAAAAT	0.3
AAAGAA	0.3
AACAAT	0.2
AAAAAC	0.2
AACAAA	0.2
AAGAAT	0.2
CATAAT	0.2
TATTAA	0.2
AAAAAG	0.2
ACTAAT	0.2
CATTAA	0.1
ATAAAA	0.1
GATTAA	0.1

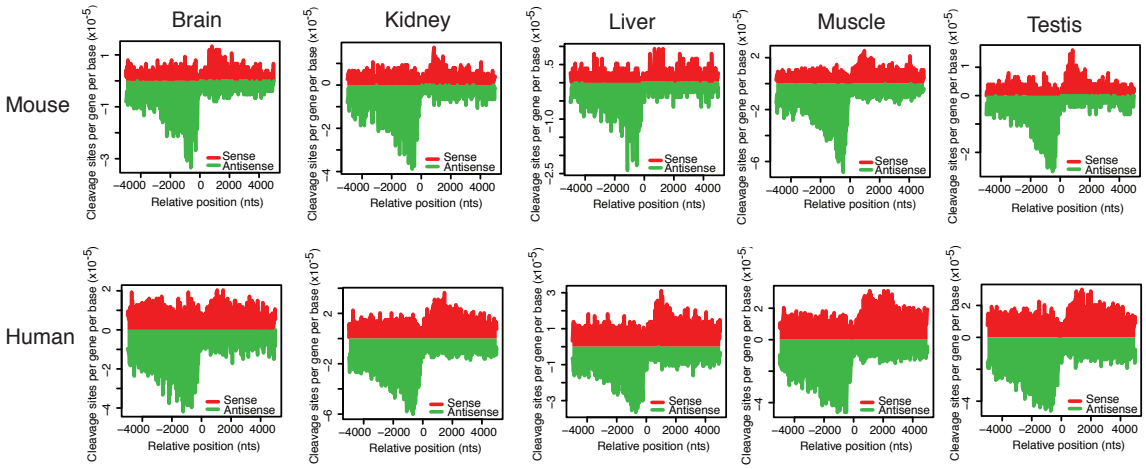
Distribution of PAS hexamers at experimentally detected 3'-ends. PAS hexamers used at all cleavage sites (Supplementary Figure 1d) or from cleavage sites restricted to the upstream antisense region (Supplementary Figure 5a). A red box is placed around the two most highly used PAS hexamers (AAUAAA and AUUAAA). All numbers are rounded to the nearest 10th decimal place. The PAS usage distribution at all cleavage sites and from upstream antisense regions are similar.

Figure S1



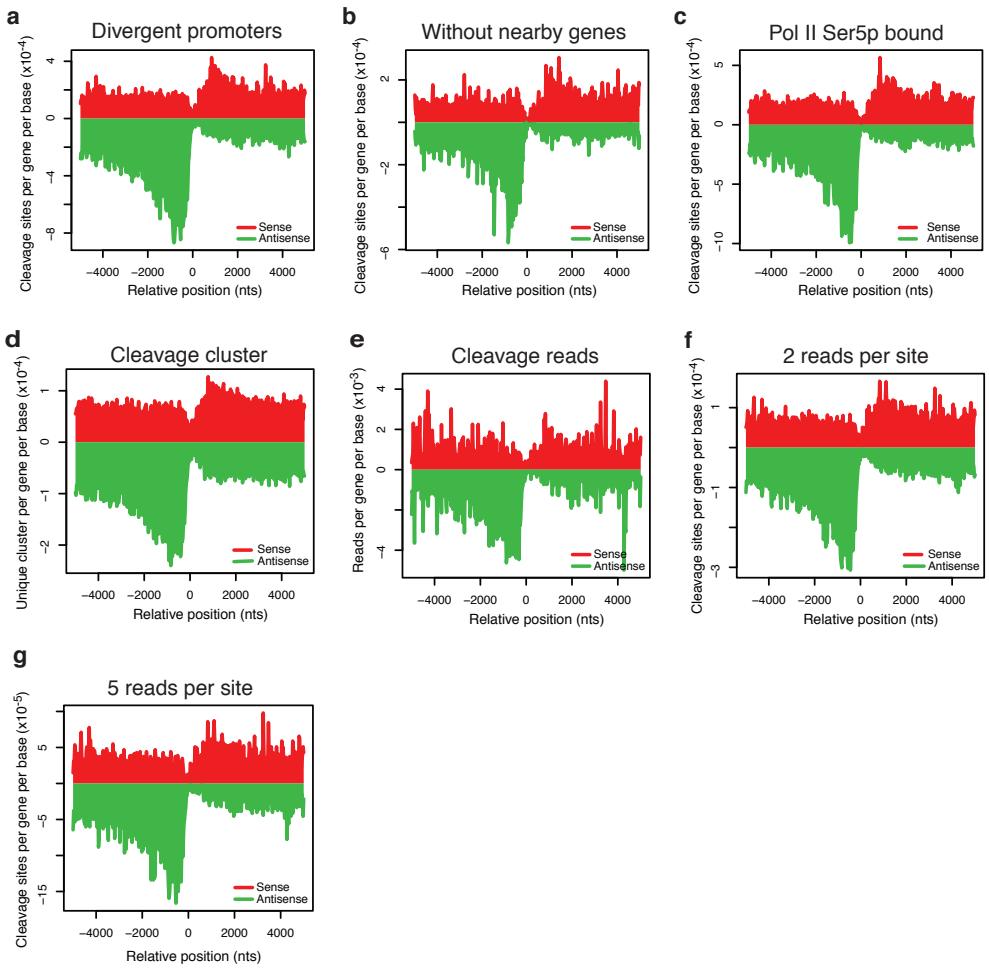
Supplementary Figure 1. Mapping the 3' ends of polyadenylated RNAs by deep sequencing in mESCs. (a) Venn diagram depicts the overlap of unique cleavage sites between two 3'-end libraries that were constructed and denoted as library replicate 1 and replicate 2. (b) The fraction of cleavage sites in six non-overlapping categories including: 2 kb flanking 3' end of the gene (3' end), 5 kb downstream the TSS in the gene (5' end), internal of the gene (Internal, not 5' end or 3' end), upstream antisense of the TSS within 5 kb (Upstream antisense), antisense to the gene (Antisense), and other intergenic regions (Intergenic) in pool (combining replicate 1 and 2), replicate 1, replicate 2, overlap (only common to replicate 1 and 2), and sites unique to replicate 1 or replicate 2. (c) Density of unique cleavage sites at annotated 3' ends of genes with sense and antisense sites shown in red and green, respectively. Position zero denotes the annotated TES. Average coverage equals the number of unique cleavage sites per nucleotide per gene. (d) Pie chart displaying the usage of each PAS (all percentages shown in Supplementary Table 4a) among all unique cleavage sites. (e) Histogram showing the distance of the PAS (all 36 hexamers) 5' end relative to the cleavage site (indicated as position zero on the x-axis) and fraction of all cleavage sites that have a PAS at each position is shown on the y-axis.

Figure S2



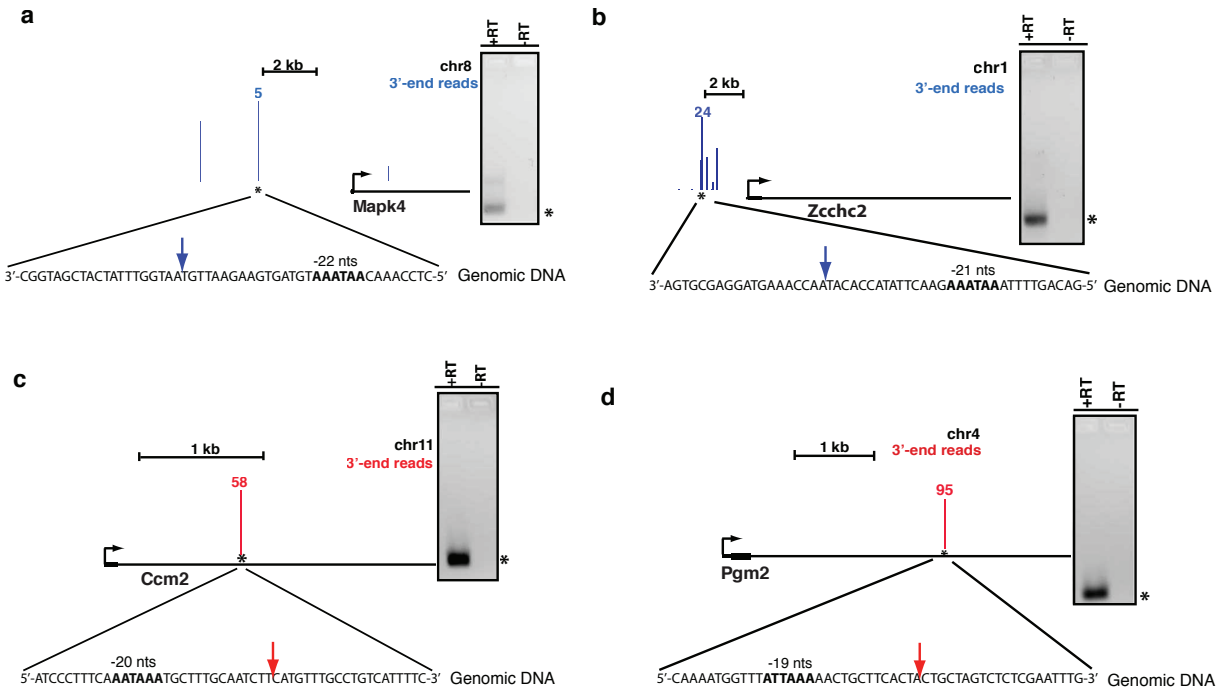
Supplementary Figure 2. The cleavage bias near gene TSS is conserved in various tissues in mouse and human. To determine if the bias found in mouse ES cells can be observed in other mouse tissues or another mammalian species, we examined published 3'-end sequencing data. Panels display metagene plots of sense (red) or antisense (green) unique cleavage sites flanking coding gene TSS. The number of unique cleavage sites in each 25 bp bin across 5 kb upstream and downstream of the TSS is plotted and unique cleavage sites within 5 kb of annotated 3'-ends were removed. In all tissues of human and mouse, we observed more upstream antisense cleavage and a promoter proximal antisense peak. Despite different sets of genes being expressed across various tissues and analyzing 3'-end sequencing data generated from another mammalian species, the pattern is consistent with the biased distribution of PAS and U1 sites that is generally encoded in gene sequences.

Figure S3



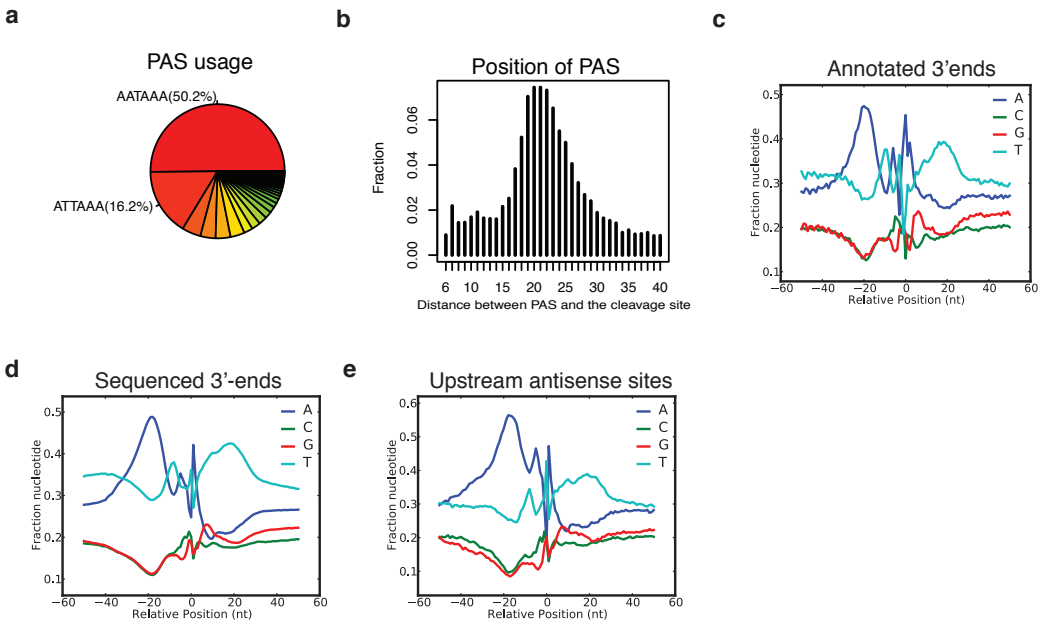
Supplementary Figure 3. Metagene analysis of cleavage sites near gene TSS. Displayed metagene plots (a-g) were generated in the same way as Figure 1a with the specified modifications. (a) Plot focusing on divergent promoters (details in methods), (b) or a subset of promoters where the gene is at least 6 kb in size and there are no other TSS or TES within the 10 kb window. Unlike Figure 1a, sites within 5 kb of TES were not removed. (c) A plot displaying a subset of promoters that showed significant Ser5 phosphorylated Pol II peaks in mESCs. For metagene plots a-c, only unique cleavage sites are being plotted. (d) Plotting the density of unique cleavage clusters (cleavage sites within 24 bps were clustered together and the most 5' sites are used as a reference site of the cluster). (e) Plotting read density instead of unique cleavage sites. Sites with more than 500 supporting reads were removed from the plot since they could be unannotated gene ends. Metagene plots (f-g) were generated in the same way as Figure 1a except taking a subset of unique cleavage sites with at least two (f) or five (g) supporting reads.

Figure S4



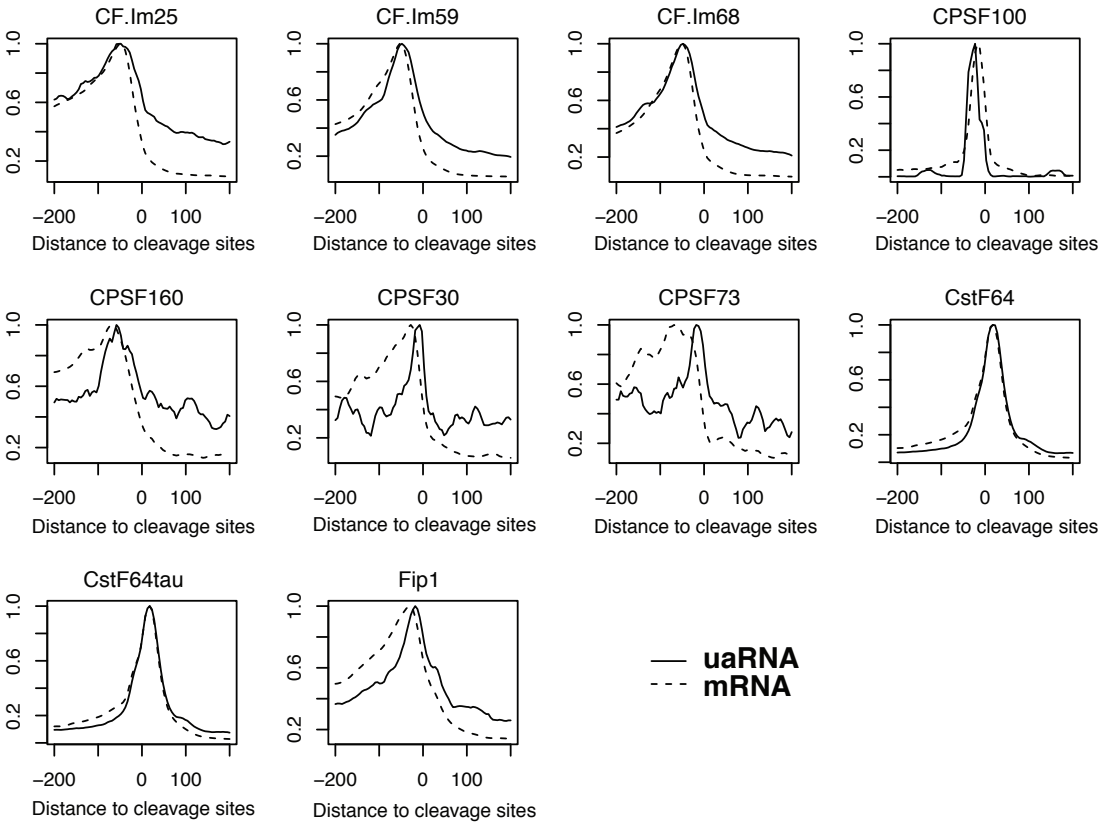
Supplementary Figure 4. Validation of promoter proximal antisense (a-b) and sense (c-d) cleavage sites using 3'-RACE. Each panel displays a genome browser view of the promoter proximal region at four coding genes: Mapk4 (a), Zcchc2 (b), Ccm2 (c), Pgm2 (d) with the gene TSS denoted with a black arrow pointing towards the right. Promoter proximal 3'-end cleavage reads for uaRNA (blue) and mRNA (red) are displayed above each gene schematic shown in black. The assayed cleavage site is denoted with an asterisk and the number of reads supporting each site is displayed above each site. We validated the most prominent cleavage site (supported by the most number of reads) for each uaRNA loci. Agarose gels of 3'-RACE PCR products are displayed to the right and each assayed cleavage site (asterisk) was cloned and sequenced using Sanger sequencing methods. Scale bars are represented in black above genes. The encoded genome sequence is displayed including the sequence of the PAS (bold) and the distance between the cleavage site (blue and red arrow for uaRNA and mRNA, respectively) and the 5'-end nucleotide of the PAS is noted above.

Figure S5



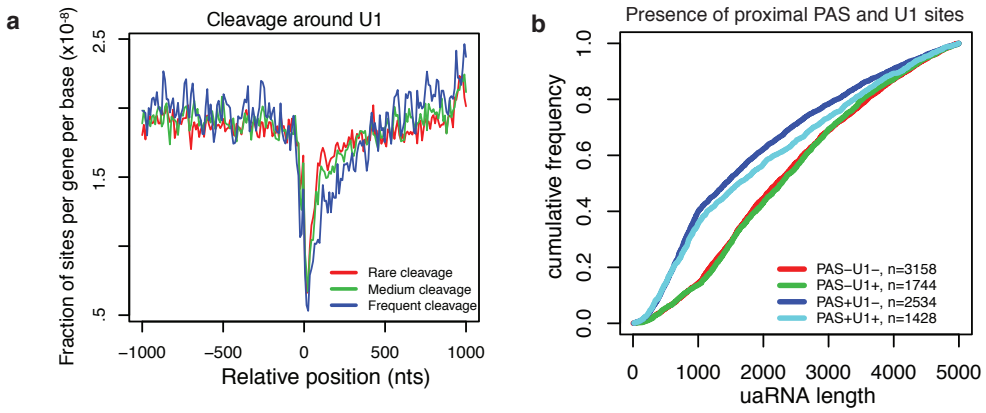
Supplementary Figure 5. Upstream antisense cleavage sites resemble annotated gene TES. (a) Pie chart displaying the usage of each PAS among unique cleavage sites in the upstream antisense region. (b) Histogram showing the distance of the PAS 5' end relative to the cleavage site indicated as position zero. For (a-b), figures include all 36 PAS hexamers with the percentage of all PAS hexamers in (a) described in Supplementary Table 4b. (c-e) The nucleotide frequency flanking cleavage sites (position 0): annotated end of genes (c), cleavage sites detected from our 3' end sequencing -- sites within 2 kb of annotated gene ends (d), and upstream antisense sites (e).

Figure S6



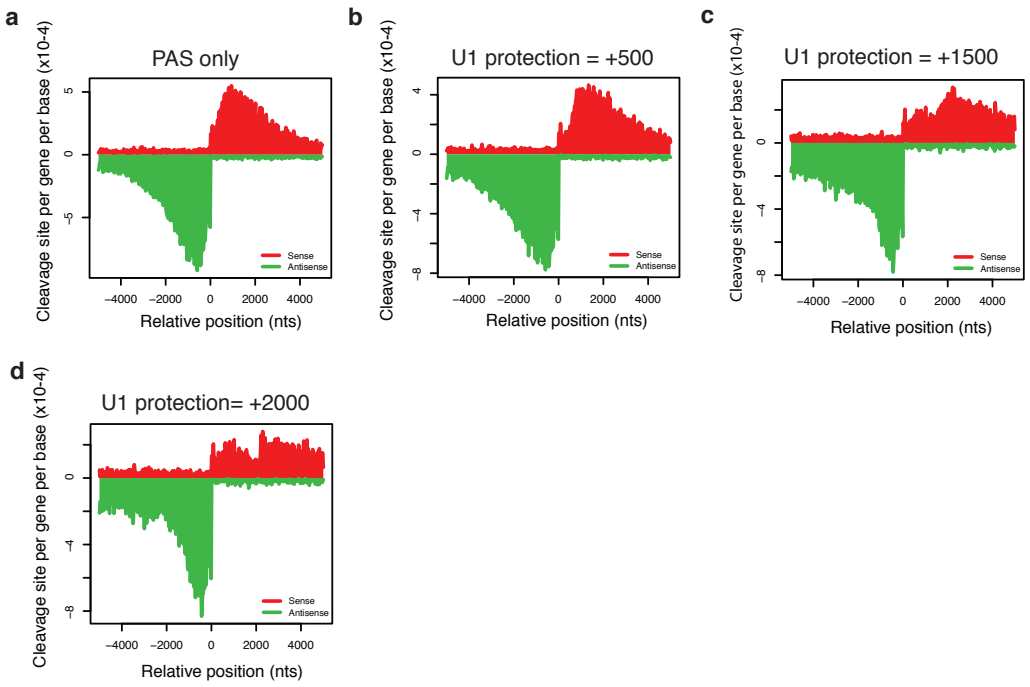
Supplementary Figure 6. Binding profiles of ten 3' end processing factors around cleavage sites in uaRNA regions and mRNA ends. A cleavage site is defined as a uaRNA cleavage site if it is outside any protein coding gene but locates within 5 kb upstream antisense of a protein-coding gene TSS. mRNA cleavage sites are defined as cleavage sites within 100 bases of annotated protein-coding gene ends. For each 3' end processing factor, CLIP read density within 200 bases of all cleavage sites are summed up in every 5 bp bin and subsequently normalized such that the max value is 1.

Figure S7



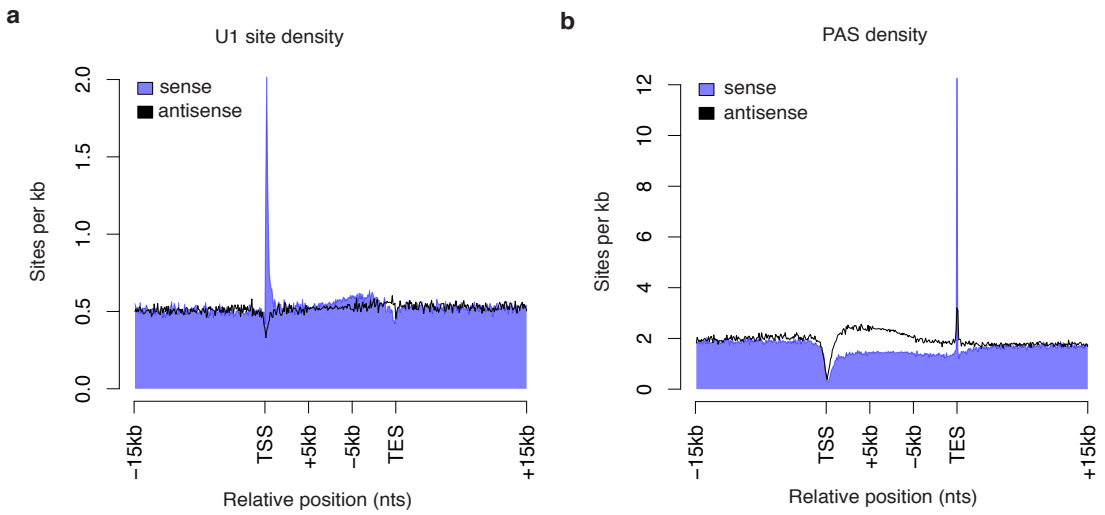
Supplementary Figure 7. Proximal U1 sites are associated with uaRNA length. (a) Distribution of cleavage sites flanking strong U1 sites (position 0). Cleavage sites are classified as rare, medium, and frequent sites based on the number of reads supporting each cleavage site (rare: 1 read, medium: 2-9 reads, frequent: >9 reads). Y-axis is shown as the fraction of sites per gene per base. (b) CDF plot comparing the length of uaRNAs grouped by the presence/absence of promoter proximal PAS and U1 sites. PAS+/- (U1+/-) indicates the presence/absence of PAS or U1 sites in the first 1 kb of uaRNA region. The length of uaRNAs is estimated using the distance from cleavage sites to coding gene TSS.

Figure S8



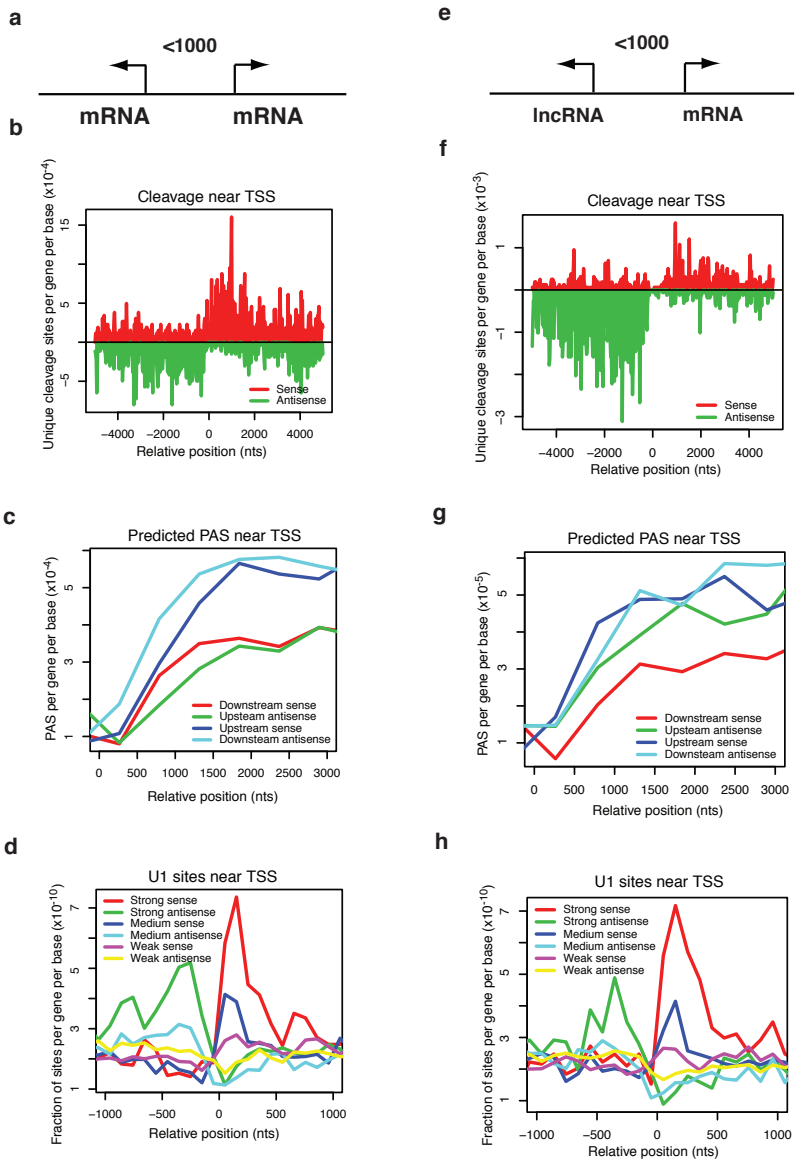
Supplementary Figure 8. Cleavage site simulation near coding gene TSS. Plots were generated in the same way as Figure 2d with unique simulated cleavage sites being plotted. Above each simulation plot, U1 protection refers to the zone of protection in nucleotides downstream (+) conferred by a strong U1 site. Metagene plot of simulated cleavage events considering the PAS (AATAAA) alone (a), or parameters where a PAS is protected if it contains a strong U1 site at least 500 (b), 1500 (c), or 2000 (d) nts upstream. These data demonstrate that the cleavage bias from the simulation is robust when considering protection zones of various sizes.

Figure S9



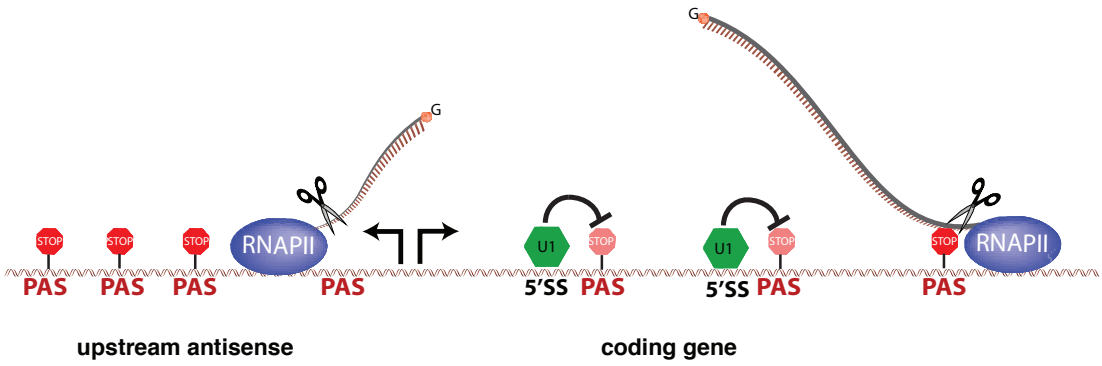
Supplementary Figure 9. **Density of U1 and PAS signals at coding genes and intergenic regions.** The density of strong U1 sites (a) and AAUAAA polyadenylation signals (b) in sites per kb for protein-coding genes longer than 15 kb and flanking 15 kb of intergenic sequences. U1 or PAS signals located on sense or antisense regions are depicted in purple and black, respectively. In addition to the strong U1 enrichment in the proximal sense direction of the gene, we observe a modest increase in the frequency of strong U1 signals internal to genes. We also observed a strong strand bias of PAS in coding transcription units, both exon and intron sequences, as compared to intergenic regions. Specifically, PAS are depleted on the sense strand when compared to the antisense strand throughout coding genes prior to the TES. In absolute terms, the genome background has a relatively high density of PAS (~ 2 sites per kb on average) but lower density of strong U1 sites (~0.5 sites per kb on average). Together, the observed distributional patterns support a general model of a U1-PAS axis favoring elongation to produce long transcripts such as precursors to mRNA but limiting transcription from antisense and intergenic regions.

Figure S10



Supplementary Figure 10. U1-PAS axis at mRNA:mRNA and lncRNA:mRNA gene pairs. 1047 and 629 mRNA:mRNA and lncRNA:mRNA gene pairs, respectively, were analyzed similarly as in Fig 1a, Fig. 2a, and Fig. 2c, except that larger bins were used (500 bps bin for PAS and 100 bps bin for U1) to smooth the curve due to the low number of genes used to make the plot. For mRNA:mRNA gene pairs position zero represents the TSS of all genes on the + strand. For lncRNA:mRNA gene pairs position zero represents the TSS of the coding gene.

Figure S11



Supplementary Figure 11. Illustration of the U1-PAS axis for divergent non-coding RNA control. At divergent promoters, RNAPII (depicted as a purple oval) transcribes in both downstream sense and upstream antisense directions, yet upstream antisense RNAs are frequently terminated shortly after initiation due to the high density of PAS (red stop sign) and a lack of strong U1 signals to suppress these sites. In contrast, PAS signals are low in the downstream sense direction and are generally protected by the binding of U1 snRNP (green hexagon) to a nearby 5' splice site denoted as 5'SS in black. A pink stop sign denotes a protected PAS. The U1-PAS axis may function to promote continued elongation throughout the gene and to ensure transcription is suppressed outside protein-coding genes.

References

Adelman, K., and Lis, J.T. (2012). Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. Nat Rev Genet *13*, 720-731.

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

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.

Beaudoing, E., Freier, S., Wyatt, J.R., Claverie, J.M., and Gautheret, D. (2000). Patterns of variant polyadenylation signal usage in human genes. Genome Res *10*, 1001-1010.

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.

Connelly, S., and Manley, J.L. (1989). A CCAAT box sequence in the adenovirus major late promoter functions as part of an RNA polymerase II termination signal. Cell *57*, 561-571.

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.

Derti, A., Garrett-Engele, P., Macisaac, K.D., Stevens, R.C., Sriram, S., Chen, R., Rohl, C.A., Johnson, J.M., and Babak, T. (2012). A quantitative atlas of polyadenylation in five mammals. Genome Res *22*, 1173-1183.

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.

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.

Gil, A., and Proudfoot, N.J. (1987). Position-dependent sequence elements downstream of AAUAAA are required for efficient rabbit beta-globin mRNA 3' end formation. Cell *49*, 399-406.

Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: scanning for occurrences of a given motif. Bioinformatics *27*, 1017-1018.

Hu, J., Lutz, C.S., Wilusz, J., and Tian, B. (2005). Bioinformatic identification of candidate cis-regulatory elements involved in human mRNA polyadenylation. RNA *11*, 1485-1493.

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.

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.

MacDonald, C.C., Wilusz, J., and Shenk, T. (1994). The 64-kilodalton subunit of the CstF polyadenylation factor binds to pre-mRNAs downstream of the cleavage site and influences cleavage site location. Mol Cell Biol *14*, 6647-6654.

Martin, G., Gruber, A.R., Keller, W., and Zavolan, M. (2012). Genome-wide analysis of pre-mRNA 3' end processing reveals a decisive role of human cleavage factor I in the regulation of 3' UTR length. Cell Rep *1*, 753-763.

Min, I.M., Waterfall, J.J., Core, L.J., Munroe, R.J., Schimenti, J., and Lis, J.T. (2011). Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells. Genes Dev *25*, 742-754.

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. (2011). Ending the message: poly(A) signals then and now. Genes Dev *25*, 1770-1782.

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.

Sigova, A., Almada, A.E., Sharp, P.A., and Young, R.A. (2013). Divergent transcription of lncRNA/mRNA gene pairs in embryonic stem cells. Proc Natl Acad Sci U S A.

Tian, B., Hu, J., Zhang, H., and Lutz, C.S. (2005). A large-scale analysis of mRNA polyadenylation of human and mouse genes. Nucleic Acids Res *33*, 201-212.

Vanacova, S., Wolf, J., Martin, G., Blank, D., Dettwiler, S., Friedlein, A., Langen, H., Keith, G., and Keller, W. (2005). A new yeast poly(A) polymerase complex involved in RNA quality control. PLoS Biol *3*, e189.

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, C., Zhang, Y.E., Chen, J.Y., Liu, C.J., Zhou, W.Z., Li, Y., Zhang, M., Zhang, R., Wei, L., and Li, C.Y. (2012). Hominoid-specific de novo protein-coding genes originating from long non-coding RNAs. PLoS Genet *8*, e1002942.

Yeo, G., and Burge, C.B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J Comput Biol *11*, 377-394.

Zhang, L., Ding, Q., Wang, P., and Wang, Z. (2013). An upstream promoter element blocks the reverse transcription of the mouse insulin-degrading enzyme gene. Biochem Biophys Res Commun *430*, 26-31.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W.*, et al.* (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol *9*, R137.

Zhang, Y.E., Vibranovski, M.D., Landback, P., Marais, G.A., and Long, M. (2010). Chromosomal redistribution of male-biased genes in mammalian evolution with two bursts of gene gain on the X chromosome. PLoS Biol *8*.

Chapter 4

Conclusions

Although RNAPII initiates transcription divergently at most gene promoters, it has been unclear why a full-length stable mRNA molecule is not produced in the upstream antisense direction of gene promoters similar to a coding sense mRNA. In Chapter 2 of this thesis we focused on the characterization of uaRNAs from a small cohort of divergent promoters in mESCs. This analysis demonstrated that the previously identified TSSa-RNAs are part of longer transcripts (not the 5'-ends) we refer to as uaRNA. uaRNAs are less than 1 kb in size, capped at their 5'-ends, heterogeneous at their 3'-ends, and are expressed at 1-4 copies per cell at the steady state. In addition, uaRNAs are unstable with half-lives of 15-20 minutes. This is due, in part, to uaRNA being targeted for degradation by the RNA exosome since we show their levels and size increase upon inactivation of a core exosome component. Surprisingly, we also find that uaRNA are sensitive to NELF/DSIF (pausing factors) knockdown and flavopiridol treatment (inhibitor of P-TEFb) since their levels increase and decrease, respectively, indicating that the antisense RNAPII complex undergoes pausing and is released from the paused state via recruitment and activity of P-TEFb. That uaRNA overcome the pausing barrier and proceed to the initial stages of productive elongation was unexpected, given the lack of H3K79me^2 and H3K36me^3 in the upstream antisense region of gene promoters. In addition, upon induction of the Isg20l1 gene, we found a comparable rate of change in mRNA and uaRNA over several time points, suggesting uaRNA and mRNA may be transcribed with similar kinetics. Altogether, the data in Chapter 2 suggested that the mechanism regulating promoter directionality likely occurs after P-TEFb recruitment and may involve premature termination followed by degradation by the RNA exosome. We suspect that the uaRNA cloned and sequenced in Chapter 2 likely represent degradation products for the following reasons: uaRNA are exosome substrates and that uaRNA ends are heterogeneous and non-polyadenylated. Because early transcription events were similar

between uaRNA and mRNA, it seemed probable that, like their sense counterparts, uaRNA may undergo PAS-dependent transcription termination.

In Chapter 3, we explored the possibility that promoter directionality, in mammals, is controlled by regulating PAS-dependent transcription termination. To test this, we utilized a high-throughput sequencing technique to capture the 3'-ends of polyadenylated RNAs in mESCs. Interestingly, we detected 2-fold more upstream antisense cleavage sites compared to downstream sense cleavage sites in the promoter-proximal region flanking all coding gene TSSs (Chapter 3, Figure 1), suggesting uaRNA may undergo early termination shortly after being transcribed. Consistent with uaRNA terminating using PAS-dependent mechanisms, we also detect the binding of canonical cleavage and polyadenylation factors near uaRNA cleavage sites (Chapter 3, Supplementary Figure 6).

In addition, we find an asymmetric distribution of U1 and PAS signals flanking gene TSSs (Chapter 3, Figure 2). These observations are consistent with a known role for U1 snRNP binding in suppressing downstream PAS signals (Berg et al., 2012; Kaida et al., 2010). Indeed, functional inhibition of U1 snRNP using morpholinos led to a drastic increase in proximal sense cleavage sites where U1 signals are high and modest increases in cleavage sites in the upstream antisense direction where U1 signals are low. These data reveal that under normal conditions, sense proximal U1 signals function to suppress proximal sense termination signals.

Lastly, we show evidence for evolutionary selection on the U1-PAS axis in vertebrates. First, we demonstrated a selection for high U1 and low PAS signals in the first 1 kb of genes throughout vertebrate evolution, such that, as genes age (or have more time to evolve) they acquire more U1 and less PAS signals at the 5'-end. The selection for U1 signals in the first 1 kb of genes is likely to reflect the pressure to suppress proximal termination (Chapter 3, Figure 4A).

Interestingly, a similar trend, albeit weaker, is also observed in the upstream antisense region of gene TSSs, indicating the selection for some uaRNAs to retain U1 signals. This is consistent with a recent study that indicates that the majority of long-noncoding RNAs (LncRNAs) expressed in mouse and human embryonic stem cells arise from divergent transcription and about a quarter of the divergent lncRNAs are spliced (Sigova et al., 2013). Second, we propose an evolutionary relationship between uaRNA, lncRNA, and mRNA at bidirectional promoters (Chapter 3, Supplementary Figure 10). For example, lncRNA:mRNA gene pairs have a weaker bias (in terms of cleavage pattern, U1 signals, and PAS signal) than uaRNA:mRNA gene pairs but stronger than mRNA:mRNA gene pairs. These findings beg the question whether some uaRNAs over time are selected to become longer non-coding RNA. We also suggest the U1-PAS axis may extend past regions proximal to gene TSSs, as we reveal that the enrichment of U1 signals and depletion of PAS extends throughout the coding gene, whereas intergenic regions are relatively low in U1 and high in PAS signal (Chapter 3, Supplemental Figure 9).

To conclude, in this thesis we have uncovered a mechanism to explain why full-length mRNAs are not produced in the upstream antisense direction of divergent mammalian promoters through a detailed structure and sequence analysis of uaRNA at the gene and genome-level. We demonstrate that promoter directionality is encoded in the DNA sequence as a U1-PAS axis and that this mechanism likely extends throughout the genome. Not only does the U1-PAS axis ensure coding genes are read in the correct direction, but likely serves as a mechanism to suppress transcription outside of coding genes. Although it seems paradoxical to transcribe RNAs that will be destroyed shortly after they are transcribed, perhaps pervasive transcription serves as a playground for natural selection to occur and, that over time, with the right mutations

and selection, any given uaRNA may evolve to become a longer non-coding RNA or even coding mRNA.

Future directions

The mechanism explaining why uaRNAs are targeted for degradation despite acquiring a PAS-dependent adenylated tail remains elusive. Furthermore, it is unclear why elongation chromatin marks, such as H3K79me2 and H3K36me3, are absent in the upstream direction of gene promoters where uaRNAs are transcribed. Therefore, in this section these two questions will be expanded upon and when appropriate models and hypothesis will be suggested.

Poly (A) tail length and RNA decay Despite the addition of a PAS-dependent, adenylated tail to the 3'-end, uaRNAs are very unstable (15-20 minute half-lives) and rapidly targeted for degradation by the RNA exosome (Chapter 2, Figure 2 and Supplementary Figure 7). This is surprising given that a PAS-dependent adenylated tail is thought to function in protecting the 3'-end from nuclear decay machineries (LaCava et al., 2005; Proudfoot, 2011). However, it is possible that differences in the length of the poly(A) tail may impact whether the RNA exosome or its cofactors have access to the 3'-end. For example, in yeast, a non-canonical poly (A) polymerase, either *Trf4* or *Trf5*, is part of the (Trf4/Air2/Mtr4p) TRAMP complex and functions to target cryptic unstable transcripts (CUTS) for degradation by recruiting the exosome (LaCava et al., 2005; Vanacova et al., 2005). Compared to the canonical Poly (A) polymerase (PAP), *Trf4* displays less processivity and catalyzes the addition of a short oligo (A) tail. Thus, it remains possible that uaRNAs acquire short oligo (A) tails that are not large enough to be protected by RNA binding proteins, and thus, are subsequently targeted for degradation by the exosome.

Consistent with this model, human orthologs of *Trf4* and *Trf5* (PAPD5 and PAPD7) have been linked to the degradation of improperly processed ribosomal RNAs (Shcherbik et al., 2010), snoRNA maturation (Berndt et al., 2012), and more recently oligo (A) tails synthesized by PAPD5 were detected at the 3'-end of PROMPTs in human cells (Preker et al., 2011). These antisense PROMPTs have recently been shown to be the human equivalent of uaRNAs (Ntini et al., 2013). On the other hand, the relationship between poly(A) tail length and RNA decay can be complicated given that in fission yeast hyperadenylated poly(A) tails (as long as 1 kb) can also act as a signal to recruit the exosome for target degradation (Chen et al., 2011). Therefore, acquiring the proper tail length may be critical in establishing a stable mRNA and that either shorter and, in some cases, longer adenylated tails may mark the transcript as an improperly processed mRNA that is subsequently destroyed. Further experimentation to define the length of uaRNA poly(A) tails genome-wide will clarify whether they contain a stabilizing 3'-end poly (A) tail of a normal mRNA length.

Poly (A) binding protein nuclear 1 (Pab2) and RNA decay The Pab2, has been established as an important 3'-end processing factor. Specifically, early *in vitro* polyadenylation experiments revealed Pab2 interacted with the growing poly (A) tail and the canonical poly (A) polymerase to stimulate the addition of roughly 200-250 adenines to the 3'-end of mRNAs (Kerwitz et al., 2003; Kuhn et al., 2003). Surprisingly recent studies investigating the function and activity of Pab2 *in vivo* demonstrate that upon knockdown (Apponi et al., 2010) or deletion (Hurschler et al., 2011; Lemay et al., 2010) of Pab2, expression of mRNAs were unaffected globally. A similar observation was made in human cells when Pab2 was knocked down, but more interestingly, a subset of long noncoding RNAs arising from divergent transcription were stabilized nearly 2-4

fold (Beaulieu et al., 2012). Furthermore, they demonstrated that these lncRNAs displayed short half-lives, required the poly(A) tail for degradation, and that the poly (A) binding protein recruited the exosome through a direct interaction with hRrp6 (catalytic nuclear component of the exosome) to promote target degradation. Interestingly, this stabilization was independent of hTrf4 (non-canonical poly (A) polymerase), which has previously been linked to the degradation of CUTS in yeast (described above). Therefore, these data represent a new mechanistic link between cleavage and polyadenylation and RNA decay involving Pab2. We plan to test whether Pab2 functions to selectively recruit the exosome complex to uaRNAs by knocking down Pab2 in mESCs followed by RNA-Seq to determine whether uaRNA levels are stabilized.

Splicing and deposition of active chromatin marks As mentioned earlier in Chapter 1, elongation-specific chromatin marks (H3K79me^2 and H3K36me^3) are largely absent in the upstream direction of divergent promoters. One hypothesis that will be elaborated on in this section is the possibility that a lack of strong cis- splicing elements (and weak recruitment of trans- splicing factors) may result in a failure to recruit chromatin modifiers to the upstream antisense region of divergent promoters. Recently, there has been an accumulating amount of evidence indicating a link between splicing and the deposition of histone H3 chromatin modifications (de Almeida et al., 2011; Kim et al., 2011). For example, genome-wide analysis of histone methylation demonstrates that intron-containing genes are marked with higher levels of H3K36me^3 and splicing inhibition (with spliceostatin) results in less recruitment of Set2d (H3K36me^3 methyltransferase) and reduced H3K36me^3. This data suggests that splicing signals recruit histone modifiers to the chromatin when they are needed during the transcription cycle.

A direct involvement of U1 snRNP in modulating chromatin was revealed in a study that demonstrated that strengthening the base pairing between the U1 snRNP and a 5'SS (by either mutating the 5'SS or U1 snRNA) resulted in chromatin reorganization when assayed with MNase digestions, implicating a role for U1 snRNP in chromatin remodeling (Keren-Shaul et al., 2013). Therefore, it will be interesting to directly test the role that U1 snRNP plays in the deposition of active histone marks by performing genome-wide ChIP-seq on histone H3 upon U1 snRNP inhibition, focusing on modifications, H3K79me2 and H3K36me3, that are largely absent in the upstream antisense region. In parallel, it will be interesting to assess how the insertion of 5' splice site sequences in the upstream antisense region, using genome-integrated reporter constructs, may impact the deposition of active histone modification.

Additional mechanisms that may impact promoter directionality

In addition to the U1-PAS axis, there may be other mechanisms to influence directionality of transcription, which will be the focus of this section. First, it is possible that intrinsic cis-regulatory elements within the promoter sequence may dictate the degree of divergent transcription. For example, two major classes of promoters have been described in mammals: TATA-containing and TATA-less (often CpG rich). The TATA box was the first identified core promoter element and early *in vitro* studies revealed its functional importance in recruiting the transcriptional apparatus to the promoter (Smale and Kadonaga, 2003). However, recent computational analysis would argue that at human genes only 24% contain a TATA-like box element in the promoter, whereas the other 76% are CpG rich (Yang et al., 2007). Interestingly, 77% of bidirectional promoters (instances where two mRNAs share a promoter and are in a head-to-head orientation) contain CpG-rich sequences, compared to only 38% of

unidirectional promoters (Trinklein et al., 2004). Moreover, among divergent promoters (defined by the presence of at least 1 sense and antisense TSSa-RNA), 80% are associated with CpG-rich sequences (Seila et al., 2008). Together, these data indicate a positive connection between CpG island promoters and divergent transcription (bidirectional), and suggest TATA-containing promoters may provide directionality of transcription. Consistent with this, small RNA cloning performed in fruit flies revealed that 95% of promoter-associated reads mapped in the sense direction (compared to 58% at human promoters), indicating a lack of divergent transcription in flies (Nechaev et al., 2010). This may be due to the increase in TATA-like sequences in *Drosophila melanogaster* promoters (compared to human), which were observed to strongly correlate with unidirectional transcription (Core et al., 2012; FitzGerald et al., 2006). Interestingly, TATA containing sequences within CpG island promoters were highly unidirectional, indicating the TATA box is dominant in this orientation (Core et al., 2012). Further investigation of *Drosophila* promoters will likely reveal additional cis-elements that promote unidirectional transcription that may be less prominent at mammalian promoters.

Promoter directionality may also be controlled at the level of transcription initiation. For example, RNAPII and general transcription factors may be more efficiently recruited in the sense direction, leading to more sense transcription. However, several lines of evidence from our work and others would argue against this in mammals. First, analysis of promoter-proximal sense and antisense GRO-Seq reads display a 50% difference in transcriptionally engaged RNAPII (Core et al., 2008), which cannot account for the 10-fold difference between the sense and antisense RNA steady-state levels (Sigova et al., 2013). Furthermore, we observe a coordinated increase in both sense and antisense RNAs upon induction of the Isg20l1 gene and after release from the

transcription inhibitor, flavopiridol (Chapter 2, Fig 4,5). However, its been recently shown that promoter directionality in yeast is controlled at the level of initiation, mainly through Ssu72-dependent gene loops (Tan-Wong et al., 2012). Interestingly, in a Ssu72 and Rrp6 (exosome defective) null strain they observe reduced gene looping and detect a dramatic increase in upstream antisense transcription, indicating that gene loops function to ensure RNAPII and GTFs are efficiently recruited in the sense direction. Although our data would argue against this model, we cannot rule out the possibility that at some divergent promoters gene loops function to enforce sense transcription.

Another potential mechanism to regulate promoter directionality includes differential recruitment of elongation factors in the sense and antisense direction. I focus here on the recruitment of P-TEFb, which has been revealed as the elongation factor that functions to release RNAPII from the paused state. Prior to the work presented in this thesis, it was presumed that uaRNAs were poorly elongated and likely destroyed either prior to or at the level of RNAPII pausing. The mapping of NELF and DISF (pausing factors) genome-wide detected a peak of both factors upstream of gene TSSs, which indicated that the antisense RNAPII complex undergoes RNAPII pausing (Rahl et al., 2010) but likely is never released from the paused state. In our studies, I demonstrate at four divergent promoters that uaRNA levels are sensitive to P-TEFb inhibition, arguing that they are likely released from the paused state via P-TEFb (Chapter 2, Figure 4) and undergo the initial stages of transcription elongation. However, it is possible that at some divergent promoters P-TEFb recruitment may be limiting. For example, an SR protein (SC35) was shown to actively recruit P-TEFb to the 5'-ends of genes (Ji et al., 2013; Lin et al., 2008) and, more recently, was demonstrated to be part of a repressive complex including the

7SK noncoding RNA, Hexim1, and P-TEFb (Ji et al., 2013). Interestingly, they find that 5'-proximal splicing enhancer signals (that bind Sc35) function to release SC35 and P-TEFb from the 7SK-repressive complex to promote pause release and gene activation. Thus, it is likely that, in addition to 5'SSs, splicing enhancer signals that function to recruit SC35 and P-TEFb are weak in the upstream antisense direction of gene promoters and may impact productive elongation.

References

Apponi, L.H., Leung, S.W., Williams, K.R., Valentini, S.R., Corbett, A.H., and Pavlath, G.K. (2010). Loss of nuclear poly(A)-binding protein 1 causes defects in myogenesis and mRNA biogenesis. Hum Mol Genet *19*, 1058-1065.

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.

Berndt, H., Harnisch, C., Rammelt, C., Stohr, N., Zirkel, A., Dohm, J.C., Himmelbauer, H., Tavanez, J.P., Huttelmaier, S., and Wahle, E. (2012). Maturation of mammalian H/ACA box snoRNAs: PAPD5-dependent adenylation and PARN-dependent trimming. RNA *18*, 958-972.

Chen, H.M., Futcher, B., and Leatherwood, J. (2011). The fission yeast RNA binding protein Mmi1 regulates meiotic genes by controlling intron specific splicing and polyadenylation coupled RNA turnover. PLoS One *6*, e26804.

Core, L.J., Waterfall, J.J., Gilchrist, D.A., Fargo, D.C., Kwak, H., Adelman, K., and Lis, J.T. (2012). Defining the status of RNA polymerase at promoters. Cell Rep *2*, 1025-1035.

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 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.

FitzGerald, P.C., Sturgill, D., Shyakhtenko, A., Oliver, B., and Vinson, C. (2006). Comparative genomics of Drosophila and human core promoters. Genome Biol *7*, R53.

Hurschler, B.A., Harris, D.T., and Grosshans, H. (2011). The type II poly(A)-binding protein PABP-2 genetically interacts with the let-7 miRNA and elicits heterochronic phenotypes in Caenorhabditis elegans. Nucleic Acids Res *39*, 5647-5657.

Ji, X., Zhou, Y., Pandit, S., Huang, J., Li, H., Lin, C.Y., Xiao, R., Burge, C.B., and Fu, X.D. (2013). SR Proteins Collaborate with 7SK and Promoter-Associated Nascent RNA to Release Paused Polymerase. Cell *153*, 855-868.

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.

Keren-Shaul, H., Lev-Maor, G., and Ast, G. (2013). Pre-mRNA splicing is a determinant of nucleosome organization. PLoS One *8*, e53506.

Kerwitz, Y., Kuhn, U., Lilie, H., Knoth, A., Scheuermann, T., Friedrich, H., Schwarz, E., and Wahle, E. (2003). Stimulation of poly(A) polymerase through a direct interaction with the nuclear poly(A) binding protein allosterically regulated by RNA. EMBO J *22*, 3705-3714.

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.

Kuhn, U., Nemeth, A., Meyer, S., and Wahle, E. (2003). The RNA binding domains of the nuclear poly(A)-binding protein. J Biol Chem *278*, 16916-16925.

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.

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.

Lin, S., Coutinho-Mansfield, G., Wang, D., Pandit, S., and Fu, X.D. (2008). The splicing factor SC35 has an active role in transcriptional elongation. Nat Struct Mol Biol *15*, 819-826.

Nechaev, S., Fargo, D.C., dos Santos, G., Liu, L., Gao, Y., and Adelman, K. (2010). Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila. Science *327*, 335-338.

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.

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.

Proudfoot, N.J. (2011). Ending the message: poly(A) signals then and now. Genes Dev *25*, 1770-1782.

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.

Shcherbik, N., Wang, M., Lapik, Y.R., Srivastava, L., and Pestov, D.G. (2010). Polyadenylation and degradation of incomplete RNA polymerase I transcripts in mammalian cells. EMBO Rep *11*, 106-111.

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.* (2013b). Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells. Proc Natl Acad Sci U S A *110*, 2876-2881.

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

Tan-Wong, S.M., Zaugg, J.B., Camblong, J., Xu, Z., Zhang, D.W., Mischo, H.E., Ansari, A.Z., Luscombe, N.M., Steinmetz, L.M., and Proudfoot, N.J. (2012). Gene loops enhance transcriptional directionality. Science *338*, 671-675.

Trinklein, N.D., Aldred, S.F., Hartman, S.J., Schroeder, D.I., Otillar, R.P., and Myers, R.M. (2004). An abundance of bidirectional promoters in the human genome. Genome Res *14*, 62-66.

Vanacova, S., Wolf, J., Martin, G., Blank, D., Dettwiler, S., Friedlein, A., Langen, H., Keith, G., and Keller, W. (2005). A new yeast poly(A) polymerase complex involved in RNA quality control. PLoS Biol *3*, e189.

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

Appendix

Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells

Alla A. Sigova, Alan C. Mullen, Benoit Molinie, Sumeet Gupta, David A. Orlando, Mathew Guenther, Albert E. Almada, Charles Lin, Phillip A. Sharp, Cosmas C. Giallourakis, and Richard A. Young

Author contributions: I contributed experimental data, validating the sequence and structure of divergent lncRNAs, to the following work previously published as:

Alla A. Sigova, Alan C. Mullen, Benoit Molinie, Sumeet Gupta, David A. Orlando, Mathew Guenther, Albert E. Almada, Charles Lin, Phillip A. Sharp, Cosmas C. Giallourakis, and Richard A. Young. (2013). Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells. Proc Natl Acad Sci U S A *110,* 2876-2881