doi:10.1038/nature12349

Promoter directionality is controlled by U1 snRNP and polyadenylation signals

Albert E. Almada[1,2]*, Xuebing Wu[1,3]*, Andrea J. Kriz[2], Christopher B. Burge[2,3] & Phillip A. Sharp[1,2]

Transcription of the mammalian genome is pervasive, but productive transcription outside of protein-coding genes is limited by unknown mechanisms[1]. In particular, although RNA polymerase II (RNAPII) initiates divergently from most active gene promoters, productive elongation occurs primarily in the sense-coding direction[2–4]. Here we show in mouse embryonic stem cells that asymmetric sequence determinants flanking gene transcription start sites control promoter directionality by regulating promoter-proximal cleavage and polyadenylation. We find that upstream antisense RNAs are cleaved and polyadenylated at poly(A) sites (PASs) shortly after initiation. *De novo* motif analysis shows PAS signals and U1 small nuclear ribonucleoprotein (snRNP) recognition sites to be the most depleted and enriched sequences, respectively, in the sense direction relative to the upstream antisense direction. These U1 snRNP sites 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 dramatic increase in promoter-proximal cleavage events in the sense direction with slight increases in the antisense direction. These data suggest that a U1–PAS axis characterized by low U1 snRNP recognition and a high density of PASs 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.

Two potential mechanisms for suppressing transcription elongation in the upstream antisense region of gene transcription start sites (TSSs) 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 the positive transcription elongation factor b (P-TEFb)[5]. A detailed characterization of several upstream antisense RNAs (uaRNAs) in mouse embryonic stem cells (ESCs) suggested that P-TEFb is recruited similarly in both sense and antisense directions[6], and in human cells, elongating RNAPII (phosphorylated at Ser 2 in the carboxy-terminal domain) occupies the proximal upstream transcribed region[7]. These data suggest that the upstream antisense RNAPII complex undergoes the initial phase of elongation but probably terminates early owing to an unknown mechanism.

To test globally whether upstream antisense transcripts undergo early termination (compared to coding messenger RNA) by a canonical PAS-dependent cleavage mechanism, we mapped by deep sequencing the 3′ ends of polyadenylated RNAs in mouse ESCs (N. Spies, C.B.B. & D. P. Bartel, unpublished protocol). 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[8]. We sequenced two complementary DNA (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 TSSs and at least 5 kilobases (kb) away from known gene transcription end sites (TESs). Interestingly, in the upstream antisense region we observed a twofold higher number of cleavage sites compared to the downstream sense sites flanking protein-coding-gene TSSs (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[9] (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 phosphorylated at Ser 5-occupied promoters)[10], 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 the 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 TESs of genes, these upstream antisense cleavage sites are associated with the PAS located at the expected position, about 22 nucleotides upstream of the cleavage site (Supplementary Fig. 5a, b)[11,12]. Moreover, the nucleotide sequence composition flanking the cleavage sites resembles that of TESs of genes (Supplementary Fig. 5c–e), including a downstream U-rich region[13,14]. To determine whether members of the canonical cleavage and polyadenylation machinery bind specifically to uaRNA cleavage sites, we analysed available crosslinking immunoprecipitation (CLIP) sequencing data sets for ten human canonical 3′-end-processing factors, namely CPSF160 (also known as CPSF1), CPSF100 (CPSF2), CPSF73 (CPSF3), CPSF30 (CPSF4), Fip1 (FIP1L1), CstF64 (CSTF2) and its paralogue CstF64τ (CSTF2T), CF $\mathrm{I_m}$25 (NUDT21), CF $\mathrm{I_m}$59 (CPSF7) and CF $\mathrm{I_m}$68 (CPSF6), along with poly(A) 3′-end sequencing data generated in HEK293 cells[15]. We detected specific binding of all ten factors at uaRNA cleavage sites with positional profiles identical or very similar to that of mRNA cleavage sites (Supplementary Fig. 6). These results indicate that the poly(A) tails we analysed were products of PAS-dependent cleavage and polyadenylation, rather than either a priming artefact or a PAS-independent polyadenylation representing a transient signal for RNA degradation[16–18].

[1]David H. Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [2]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [3]Computational and Systems Biology Graduate Program, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.
*These authors contributed equally to this work.

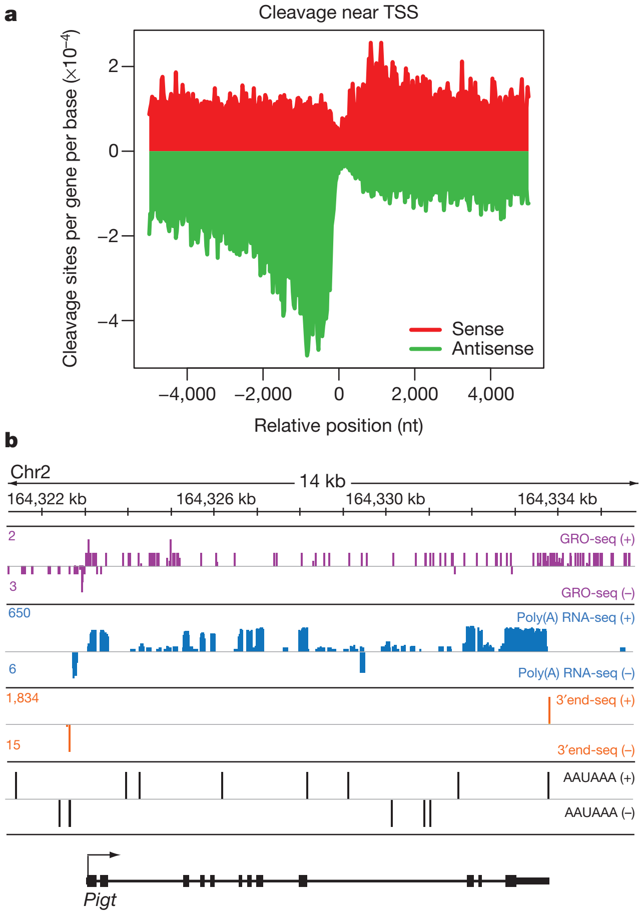

Figure 1 | Promoter-proximal PAS-dependent termination of uaRNA. **a**, Metagene plot of sense (red) or antisense (green) unique cleavage sites flanking coding-gene TSS. The number of unique cleavage sites per gene per base in each 25-base-pair (bp) bin across 5 kb upstream and downstream of the TSS in nucleotides (nt) 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)[28], Poly(A) + RNA-seq (blue)[29], 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 left-hand side represent the maximum read density on each track.

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). As this 33% depletion is unlikely to explain the twofold cleavage bias observed (see simulation results in Supplementary Fig. 8a), we searched for additional discriminative hexamer sequence signals in an unbiased manner. All 4,096 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 TSSs. In addition, we identified 5′-splice-site-related sequences (or sequences recognized by U1 snRNP, 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 base pairs 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).

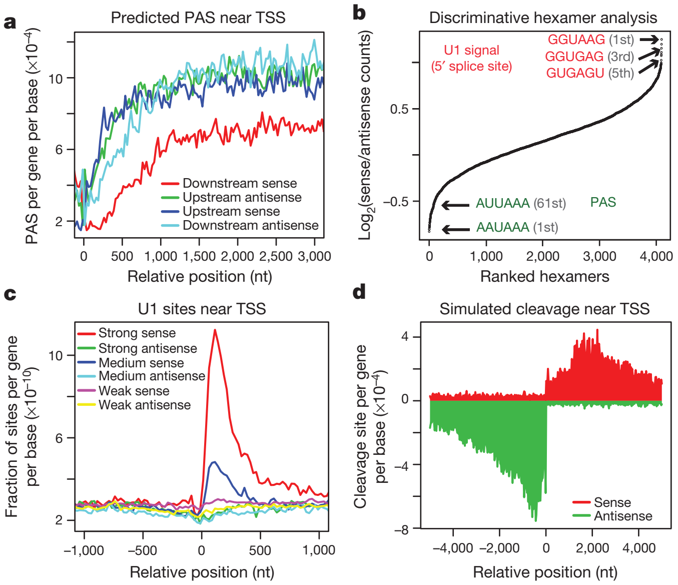

Figure 2 | Asymmetric distribution of PAS and U1 signals flanking coding-gene TSS. **a**, Number of AATAAA sites per gene per base in each 25-bp bin within a 3-kb region flanking the gene TSS on the downstream sense (red), downstream antisense (light blue), upstream antisense (green) and upstream sense (dark blue) strands. **b**, Rank of all 4,096 hexamers by enrichment ($\log_2$ 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.

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 snRNP complex suppresses cleavage and polyadenylation near a U1 site, as has been observed in various species including human and mouse[19–21]. 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 the coding-gene TSS) is significantly associated with shorter uaRNAs ($P < 10^{-15}$), 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 in which the U1 snRNP promotes RNA lengthening by suppressing proximal PASs (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) recapitulated the main features of the observed distribution (Fig. 1a), including an antisense peak around 700 bases upstream and a ∼twofold 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 the U1 snRNP in mouse ESCs. Specifically, we transfected ESCs with either an antisense morpholino oligonucleotide (AMO) complementary to the 5′ end of the 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[19,20]. Interestingly, we observed in two biological replicates a marked 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 the cells treated with scrambled control AMOs (Fig. 3). These observations confirm that U1 snRNP 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 the U1 snRNP is much lower and there is little enhancement in cleavage sites upon U1 snRNP inhibition.

The conservation of the asymmetric cleavage pattern across human and mouse (Supplementary Fig. 2) led us to examine whether there is evolutionary selection on the U1–PAS axis. Previously, mouse protein-coding genes have been assigned to 12 evolutionary branches and dated by analysing the presence or absence of orthologues in the vertebrate phylogeny[22]. 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 that 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 the increases in CpG-rich sequences within 1 kb of gene TSSs 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 the 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[23].

The U1–PAS axis probably 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 that the U1–PAS axis may serve as a mechanism for terminating transcription in both antisense and intergenic regions.

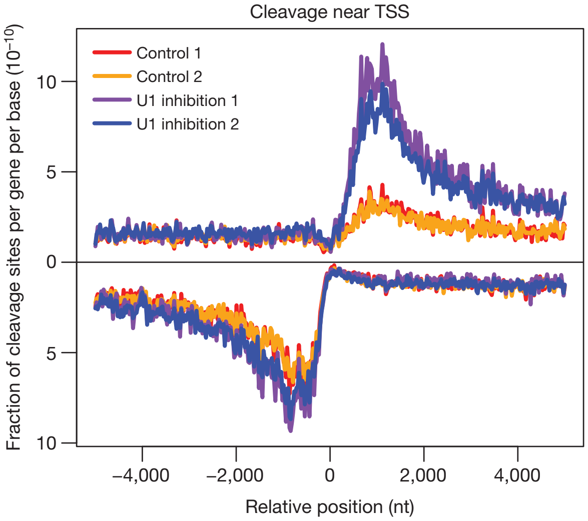

Figure 3 | Promoter-proximal cleavage sites are altered upon functional U1 inhibition. *y* axis represents the number of cleavage sites per gene per base divided by the total number of cleavage sites identified in each 3′-end sequencing library in a 5-kb region flanking the coding-gene TSS. Signal for the antisense strand is set as negative. U1 inhibition 1 (purple) and U1 inhibition 2 (blue) represent 3′-end sequencing libraries generated from mouse ESCs treated with a U1-targeting AMO. Control 1 (red) and control 2 (orange) represent 3′-end sequencing libraries generated from mouse ESCs 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).

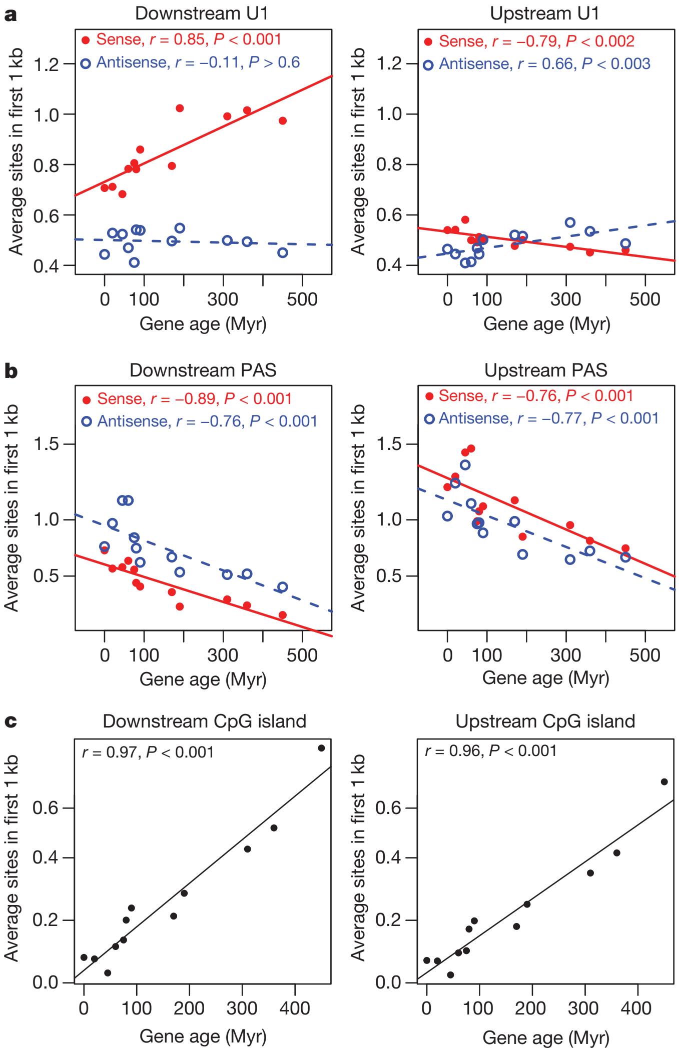

Figure 4 | Evolutionary gain and loss of U1 and PAS sites. a, Average number of strong U1 sites in the first 1 kb of protein-coding genes and upstream regions. **b**, Average number of PAS sites in the first 1 kb downstream and upstream of coding-gene TSS, respectively. **c**, Average number of CpG islands 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. Number of genes in each group (from old to young): 11,934, 1,239, 914, 597, 876, 1,195, 279, 175, 198, 315, 926 and 1,143. Solid red dots and blue circles indicate sites on the sense and antisense strands, respectively. Pearson correlation coefficient (*r*) between gene age and average site counts is calculated and a line is fitted by linear regression. *P* values are estimated from 1,000 randomizations of gene age assignments.

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 probable that additional *cis*-elements may influence PAS usage[24] and will need to be integrated into this model. There may also be other PAS-independent mechanisms that contribute to the termination of transcription in upstream antisense regions and across the genome[25–27]. However, evidence for the U1–PAS axis is found in several different tissues of mouse and human, indicating its wide use 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 to be part of a continuum that varies in the degree of U1–PAS axis activity.

METHODS SUMMARY

Total RNA was extracted from V6.5 mouse ESCs that were grown under standard ESC culture conditions[2]. Poly(A) RNA was selected, fragmented using a limited RNase T1 digestion and reverse transcribed using an oligo-dT-containing primer, and the resulting cDNA was circularized and PCR amplified using Illumina-specific primers. U1 inhibition experiments were performed as previously described[19,20].

Full Methods and any associated references are available in the online version of the paper.

Received 4 February; accepted 4 June 2013.
Published online 23 June 2013.

1. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489,** 101–108 (2012).
2. Seila, A. C. *et al.* Divergent transcription from active promoters. *Science* **322,** 1849–1851 (2008).
3. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322,** 1845–1848 (2008).
4. Preker, P. *et al.* RNA exosome depletion reveals transcription upstream of active human promoters. *Science* **322,** 1851–1854 (2008).
5. Adelman, K. & Lis, J. T. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nature Rev. Genet.* **13,** 720–731 (2012).
6. Flynn, R. A., Almada, A. E., Zamudio, J. R. & Sharp, P. A. Antisense RNA polymerase II divergent transcripts are P-TEFb dependent and substrates for the RNA exosome. *Proc. Natl Acad. Sci. USA* **108,** 10460–10465 (2011).
7. Preker, P. *et al.* 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 (2011).
8. Proudfoot, N. J. Ending the message: poly(A) signals then and now. *Genes Dev.* **25,** 1770–1782 (2011).
9. Derti, A. *et al.* A quantitative atlas of polyadenylation in five mammals. *Genome Res.* **22,** 1173–1183 (2012).
10. Rahl, P. B. *et al.* c-Myc regulates transcriptional pause release. *Cell* **141,** 432–445 (2010).
11. Beaudoing, E., Freier, S., Wyatt, J. R., Claverie, J. M. & Gautheret, D. Patterns of variant polyadenylation signal usage in human genes. *Genome Res.* **10,** 1001–1010 (2000).
12. Tian, B., Hu, J., Zhang, H. & Lutz, C. S. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res.* **33,** 201–212 (2005).
13. Gil, A. & Proudfoot, N. J. Position-dependent sequence elements downstream of AAUAAA are required for efficient rabbit β-globin mRNA 3′ end formation. *Cell* **49,** 399–406 (1987).
14. MacDonald, C. C., Wilusz, J. & Shenk, T. 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 (1994).
15. Martin, G., Gruber, A. R., Keller, W. & Zavolan, M. 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 (2012).
16. LaCava, J. *et al.* RNA degradation by the exosome is promoted by a nuclear polyadenylation complex. *Cell* **121,** 713–724 (2005).
17. Wyers, F. *et al.* Cryptic pol II transcripts are degraded by a nuclear quality control pathway involving a new poly(A) polymerase. *Cell* **121,** 725–737 (2005).
18. Vaňáčová, S. *et al.* A new yeast poly(A) polymerase complex involved in RNA quality control. *PLoS Biol.* **3,** e189 (2005).
19. Berg, M. G. *et al.* U1 snRNP determines mRNA length and regulates isoform expression. *Cell* **150,** 53–64 (2012).
20. Kaida, D. *et al.* U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature* **468,** 664–668 (2010).
21. Andersen, P. K., Lykke-Andersen, S. & Jensen, T. H. Promoter-proximal polyadenylation sites reduce transcription activity. *Genes Dev.* **26,** 2169–2179 (2012).
22. Zhang, Y. E., Vibranovski, M. D., Landback, P., Marais, G. A. & Long, M. Chromosomal redistribution of male-biased genes in mammalian evolution with two bursts of gene gain on the X chromosome. *PLoS Biol.* **8,** e1000494 (2010).
23. Xie, C. *et al.* Hominoid-specific *de novo* protein-coding genes originating from long non-coding RNAs. *PLoS Genet.* **8,** e1002942 (2012).
24. Hu, J., Lutz, C. S., Wilusz, J. & Tian, B. Bioinformatic identification of candidate *cis*-regulatory elements involved in human mRNA polyadenylation. *RNA* **11,** 1485–1493 (2005).
25. Connelly, S. & Manley, J. L. A CCAAT box sequence in the adenovirus major late promoter functions as part of an RNA polymerase II termination signal. *Cell* **57,** 561–571 (1989).
26. Arigo, J. T., Eyler, D. E., Carroll, K. L. & Corden, J. L. Termination of cryptic unstable transcripts is directed by yeast RNA-binding proteins Nrd1 and Nab3. *Mol. Cell* **23,** 841–851 (2006).
27. Zhang, L., Ding, Q., Wang, P. & Wang, Z. An upstream promoter element blocks the reverse transcription of the mouse insulin-degrading enzyme gene. *Biochem. Biophys. Res. Commun.* **430,** 26–31 (2013).
28. Min, I. M. *et al.* Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells. *Genes Dev.* **25,** 742–754 (2011).
29. Sigova, A. A. *et al.* Divergent transcription of lncRNA/mRNA gene pairs in embryonic stem cells. *Proc. Natl Acad. Sci. USA* **110,** 2876–2881 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements The authors acknowledge the service to the MIT community of the late Sean Collier. We would like to thank N. Spies for generously sharing his optimized 3′-end sequencing protocol, C. Lin for providing computational assistance, M. Lindstrom for assistance on constructing Supplementary Figure 11, and S. Chen, A. Chiu, M. Jangi, Q. Liu, J. Wilusz and J. Zamudio for 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 MIT for their assistance with high-throughput sequencing. This work was supported by United States Public Health Service grants R01-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 a 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.

Author Contributions A.E.A., X.W. and P.A.S. conceived and designed the research. A.E.A. performed experiments. X.W. and A.J.K. performed computational analysis. A.E.A., X.W., C.B.B. and P.A.S. analysed the data and wrote the manuscript.

Author Information 3′-end sequencing data is deposited in the Gene Expression Omnibus under accession number GSE46433. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to P.A.S. (sharppa@mit.edu).

METHODS

Cell culture. V6.5 (C57BL/6-129) mouse ESCs (Koch Institute Transgenic Facility) were grown under standard ESC culture conditions[2].

Poly(A) 3′-end sequencing. Total RNA was extracted from V6.5 mouse ESCs using Ambion's Ribopure kit. Poly(A)-selected RNA was fragmented using Invitrogen's RNase T1 (biochemistry grade). Reverse transcription was performed with an RT oligo (Supplementary Table 1) at 0.25 μM final concentration using Invitrogen's Superscript III Reverse Transcriptase 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 Epicentre's CircLigase II according to the manufacturer's protocol. Circularized cDNA was PCR amplified using the New England Biolab's Phusion High-Fidelity DNA Polymerase for 15–18 cycles using the primers described in Supplementary Table 1. 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. The 3′-end library was then submitted for Illumina sequencing on the HI-Seq 2000 platform.

U1 inhibition with AMO. V6.5 mouse ESCs were transfected using the Amaxa Nucleofector II with program A-23 (mouse ESC specific) according to the manufacturer's protocol. Specifically, 2.5 million V6.5 ESCs were transfected with 7.5 μM of U1-targeting or a scrambled AMO for 8 h[19,20], 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 Supplementary Table 1.

Reads mapping. Raw reads were processed with the program cutadapt[30] to trim the adaptor sequence (TGGAATTCTCGGGTGCCAAGGAACTCCAGTCACATCAC) from the 3′ end. Reads longer than 15 nt 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 had: (1) more than ten As in the first 20-nt window or (2) more than six 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 the annotated TES.

PAS filter. In addition to a set of 12 hexamers identified previously in mouse and human expressed sequence tag analysis[11,12], we analysed 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 nt upstream of the 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 using the Pearson correlation coefficient and used a cutoff of 0.5 after manual inspection. In total, 24 new hexamers were identified as potential PASs and a hierarchy was assigned for the 36 hexamers (PAS36): first, the 12 known variants were 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 in which most PASs or variants are located, we searched for each of the 36 PAS variants within 100 nt 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. Owing 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 71,674 (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 nt upstream of those abundant sites. Nonabundant sites within these 50-nt reference regions were not used to search for false positives. Reads initially mapped to sites outside of 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 were defined as potential false-positive sites and were removed from subsequent analysis. In total, 7.2% (389,185) of initially mapped reads were outside of the reference regions. 0.34% of all mapped reads were classified as potential false-positive reads and 9.1% (86,425) of all cleavage sites were identified as potential false-positive sites.

Remove mouse B2 short interspersed elements (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% (33,696) of all cleavage sites passing the internal priming filter and the PAS filter were mapped within B2 regions or within 100 nt 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[32] to search significant matches ($P < 0.05$) on both strands of the genome. Matches were then scored by a maximum entropy model[33]. 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 mouse ESCs[28] 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 of the sense strand. A promoter was considered divergent if it contained GRO-seq signal in the first 1 kb downstream of 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) was used as a cutoff for background signals.

Define Ser 5-phosphorylated (Ser5P) RNAPII-bound TSS. ChIP-seq data for Ser5P RNA Pol II and corresponding input was downloaded from the Gene Expression Omnibus (GEO) database (accession number GSE20530 (ref. 10)) and peaks called using MACS[34] with default settings. TSSs less than 500 bp away from a peak summit were defined as bound.

Discriminative hexamer analysis. An unbiased exhaustive enumeration of all 4,096 hexamers was performed to find hexamers that are discriminative of downstream sense and upstream antisense strands of protein-coding gene promoters. Specifically, the first 1,000 nucleotides downstream sense and upstream antisense of all protein-coding gene TSSs 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 the $\log_2$ 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 TSSs, 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 GEO data set 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 5-bp 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[22]. The number of strong U1 sites, PAS (AATAAA) sites and CpG islands (UCSC mm9 annotations) in the first 1-kb region flanking the TSS on each strand were calculated, and the average number of sites in each branch/age group was plotted against gene age. The 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 1,000 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 bp. UCSC mm9 gene annotations were used in this analysis.

30. Martin, M. *Cutadapt* removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17,** 10–12 (2011).

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

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

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

34. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* **9,** R137 (2008).