Divergent Transcription from Active Promoters

Amy C. Seila,[1]* J. Mauro Calabrese,[1,2]*† Stuart S. Levine,[3] Gene W. Yeo,[4]‡ Peter B. Rahl,[3] Ryan A. Flynn,[1] Richard A. Young,[2,3] Phillip A. Sharp[1,2]§

Transcription initiation by RNA polymerase II (RNAPII) is thought to occur unidirectionally from most genes. Here, we present evidence of widespread divergent transcription at protein-encoding gene promoters. Transcription start site–associated RNAs (TSSa-RNAs) nonrandomly flank active promoters, with peaks of antisense and sense short RNAs at 250 nucleotides upstream and 50 nucleotides downstream of TSSs, respectively. Northern analysis shows that TSSa-RNAs are subsets of an RNA population 20 to 90 nucleotides in length. Promoter-associated RNAPII and H3K4-trimethylated histones, transcription initiation hallmarks, colocalize at sense and antisense TSSa-RNA positions; however, H3K79-dimethylated histones, characteristic of elongating RNAPII, are only present downstream of TSSs. These results suggest that divergent transcription over short distances is common for active promoters and may help promoter regions maintain a state poised for subsequent regulation.

Transcription of DNA by RNAPII is an orchestrated process subject to regulation at numerous levels: binding of RNAPII to the promoter, transcription initiation, and elongation. These phases and their transitions require concerted action by many protein complexes and are accompanied by changes in local chromatin structure (*1*).

When examining short RNA expression in murine embryonic stem (ES) cells, we noted the presence of ~20 nucleotide (nt)–long RNAs located near the transcription start site (TSS) of protein-encoding genes (*2*). To further investigate these low abundance RNAs, 8.4 million sequence reads were analyzed from several murine short RNA cDNA libraries: 7.3 million were derived from ES cells and 1.1 million from differentiated cell types (*3*, *4*). About 42,000 of these reads, referred to as TSSa-RNAs, uniquely mapped within 1.5 kb of protein-encoding gene TSSs (Fig. 1 and table S1). A single TSS frequently had more than one associated TSSa-RNA (Fig. 1B). TSSa-RNAs were associated with more than half of all mouse genes and were detected in all cell types examined (fig. S1). TSSa-RNAs were also found in ES cells lacking Dicer, an RNase III enzyme necessary for microRNA processing, suggesting that they are not Dicer products (fig. S1F). Sequenced TSSa-RNAs were 16 to 30 nt long, with a mean length of 20 nt (fig. S2).

TSSa-RNAs surround promoters in nonrandom, divergent orientations. Sense TSSa-RNAs map downstream of the associated promoter, overlapping genic transcripts and peaking in abundance between +0 and +50 nt downstream of the TSS. Forty percent of TSSa-RNAs map upstream of the TSS and are oriented in the antisense direction relative to their associated genes, peaking between nucleotides –100 and –300 (Fig. 1A). Sense and antisense TSSa-RNAs associated with overlapping sets of 8115 and 6331 gene promoters, respectively (table S2). This distribution is not dependent on mapping to either head-to-head gene pairs or genes with multiple TSSs, nor is it seen in intergenic regions or at gene 3′ ends (figs. S3 and S4).

A majority (67%) of ES cell genes with two or more TSSa-RNAs have both sense and antisense species; thus, individual TSSs produce both RNA subtypes (fig. S3). Based on their direction and position relative to TSSs, we hypothesize that sense and antisense TSSa-RNAs arise from divergent transcription, defined as nonoverlapping transcription initiation events that proceed in opposite directions from the TSS. Divergent transcription is likely a common feature of mammalian TSSs, given the presence of TSSa-RNAs in all cell types examined in this study.

TSSa-RNAs associate with genes expressed at varying levels in ES cells but were biased toward higher levels of gene expression. TSSa-RNAs were found at the majority of highly and moderately expressed genes (Fig. 2 and fig. S5), and 80% associate with promoters having high CpG dinucleotide frequency (CpG islands) (table S2). Additionally, the number of TSSa-RNA observations per gene correlated positively with gene expression levels, with a notable increase in the sense:antisense ratio found at the highest levels of expression (Fig. 2B). This increase suggests that a fraction of these reads from the most active genes arise from mRNA turnover.

Whereas typical RNAPII transcripts have a bias toward G at their 5′ ends, TSSa-RNAs show a nearly random 5′-nucleotide distribution (*4*, *5*) (table S3). This significant distribution difference suggests that the 5′-most base of the TSSa-RNAs does not represent the initial nucleotide transcribed by RNAPII.

Based on sequencing frequency, ~20 nt TSSa-RNAs are estimated to be present at ~1 molecule per 10 cells (*4*). Therefore, an enrichment procedure was developed to determine the nature of the short RNAs surrounding TSSa-RNA–associated genes. Sequenced 21-nt sense and antisense TSSa-RNAs associated with Ring finger protein 12

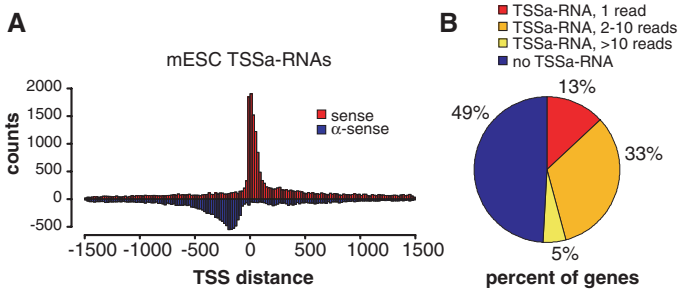

Fig. 1. The distribution of TSSa-RNAs around TSSs shows divergent transcription. (**A**) Histogram of the distance from each TSSa-RNA to all associated gene TSSs (*4*). Counts of TSSa-RNA 5′ positions relative to gene TSSs are binned in 20-nt windows. Red and blue bars represent bins of sense and antisense TSSa-RNAs, respectively. (**B**) Percentage of annotated mouse genes with indicated number of associating TSSa-RNAs.

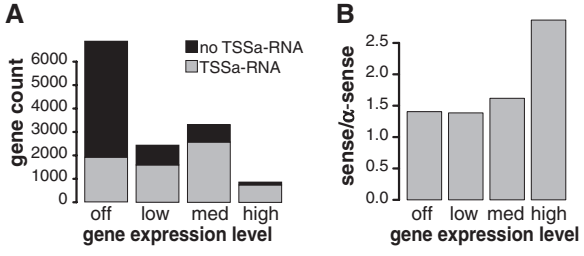

Fig. 2. In ES cells, TSSa-RNA associated genes are primarily expressed. (**A**) ES cell expression data separated into four bins based on Log2 signal intensity. Off, 1 to 4; low, 5 to 8; med, 6 to 12; and high ≥ 13 (*13*). Gene counts for each expression bin are shown. (**B**) Ratio of sense to antisense reads in each expression bin.

[1]Koch Institute, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [2]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [3]Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA 02142, USA. [4]Salk Institute, Crick-Jacobs Center for Theoretical and Computational Biology, 10010 North Torrey Pines Road, La Jolla, CA 92037, USA.

*These authors contributed equally to this work.
†Present address: Department of Genetics and the Carolina Center for Genome Sciences, University of North Carolina, Chapel Hill, NC 27599, USA.
‡Present address: Department of Cellular and Molecular Medicine, University of California, San Diego, CA 92037, USA.
§To whom correspondence should be addressed. E-mail: sharppa@mit.edu

(Rnf12) or Coiled-coil domain containing 52 (Ccdc52) transcripts, respectively, were not detected as unique species in ES cells. Instead, species between 20 and 90 nt were detected at levels estimated to be greater than 10 molecules per cell (*4*) (Fig. 3, B and D). Similar sized fragments were not found in HeLa cell RNA samples using the same sequence probes, demonstrating specificity of the procedure (Fig. 3, B and D). Northern analysis for two other TSSa-RNA–associated genes showed similar results (figs. S6 and S7). We suggest that 20 to 90 nt transcripts are the dominant short RNA species from these promoters and that sequenced TSSa-RNAs represent no more than 10% of promoter-associated transcripts.

To further classify promoters that produce TSSa-RNAs, we examined their local chromatin environment using chromatin immunoprecipitation coupled with DNA sequencing (ChIP-seq) (*3*, *4*). TSSa-RNA–associated promoters are enriched in bound RNAPII and histone H3 lysine 4 trimethylated (H3K4me3) chromatin in ES cells (Fig. 4A). About 90% of TSSa-RNA–associated genes show H3K4me3-modified nucleosomes at their promoters, as compared to ~60% for all mouse genes (Fig. 4A). TSSa-RNA–associated genes also show a ~3-fold enrichment in promoter proximal RNAPII over all genes (Fig. 4A). In contrast, TSSa-RNA–associated genes are depleted of the Polycomb component Suz12, a known transcriptional repressor thought to help maintain pluripotency by repressing developmental regulators (Fig. 4A) (*6*, *7*).

Composite profiles of ChIP-seq data were used to determine RNAPII and histone modification positions relative to TSS, revealing a correlation with sense and antisense TSSa-RNA peaks. In such analyses, the midpoint between the forward and reverse ChIP-seq read maxima defines the average DNA binding site for a factor (Fig. 4B) (*3*). At TSSa-RNA–associated genes, two distinct peaks for RNAPII are detectable with a spacing of several hundred base pairs (Fig. 4, C and D). A sharp RNAPII peak just downstream of the TSS lies directly over the sense TSSa-RNA peak (Fig. 4D). A second RNAPII peak, upstream of the first, is more diffuse but again lies directly over the antisense TSSa-RNA peak (Fig. 4D). The co-occurrence with antisense TSSa-RNAs strongly suggests that the upstream peak of RNAPII is indicative of divergent transcription rather than sense initiation upstream of the TSS, as has been proposed (*8*).

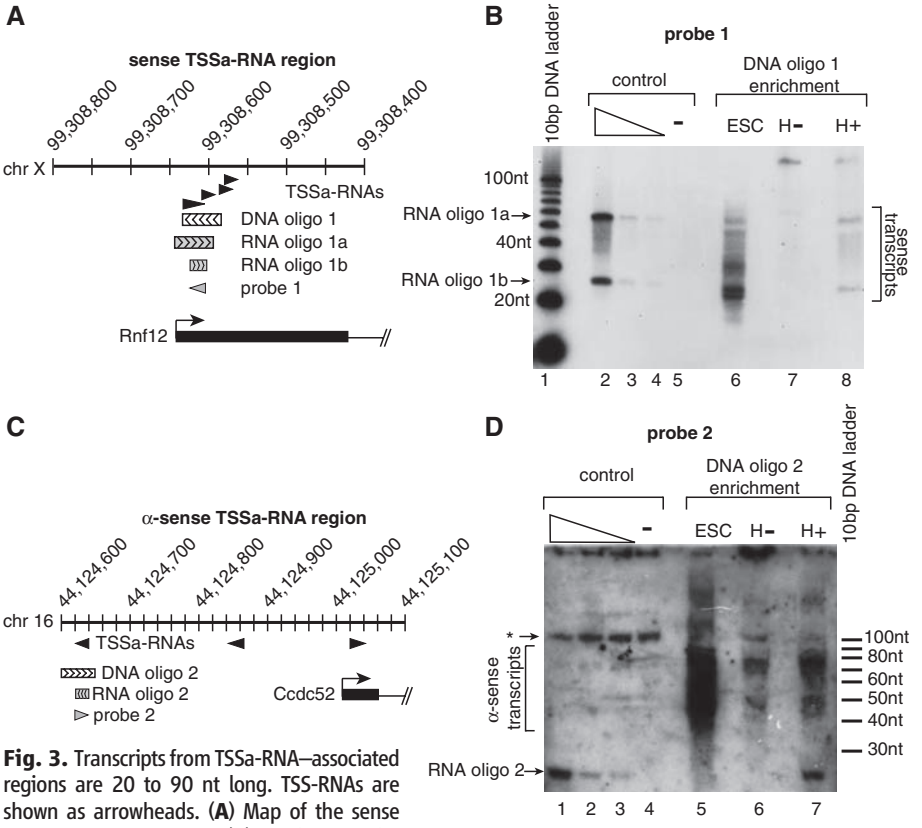

Fig. 3. Transcripts from TSSa-RNA–associated regions are 20 to 90 nt long. TSS-RNAs are shown as arrowheads. (**A**) Map of the sense TSSa-RNA Rnf12 region. (**B**) Northern analysis for Rnf12 sense TSSa-RNA using probe 1 in (A). Lane 1, 10–base pair ladder. Lanes 2 to 5, detection controls with 15, and 1.5, 0.75, and 0 fMol, respectively, of RNA oligo 1a+1b in (A). Lanes 6 to 8, material recovered from ES RNA (ESC), HeLa RNA (H-), and HeLa RNA + 15 fMol RNA oligo 1a+1b (H+), respectively, using DNA oligo 1 in (A). (**C**) Map of the antisense TSSa-RNA Ccdc52 region. (**D**) Northern analysis for the Ccdc52 antisense TSSa-RNA using probe 2 in C. Lanes 1 to 7 are as lanes 2 to 8 in (B), except using RNA oligo 2 for controls and DNA oligo 2 in (C) for enrichment. Bracket marks ESC-specific transcripts; * marks background band.

H3K4me3-modified nucleosome alignment with respect to the TSS shows peaks flanking the TSSa-RNA and RNAPII maxima, consistent with H3K4 methylation at the nucleosomes immediately upstream and downstream of TSSs (Fig. 4, C and D). These flanking peaks suggest that divergently paused RNAPII complexes may recruit H3K4 methyltransferase activity to mark active promoter boundaries. In contrast to the dual peaks of RNAPII and H3K4me3 surrounding TSSs, H3K79me2, a chromatin mark found over RNAPII elongation regions, is solely enriched in the direction of productive transcription (Fig. 4D). These observations suggest that although divergent transcription initiation is widespread, productive elongation by RNAPII occurs primarily unidirectionally, downstream of TSSs.

Sense and antisense TSSa-RNAs with bound RNAPII are found at a large number of mammalian promoters, suggesting that divergent initiation by RNAPII at TSSs is a general feature of transcriptional processes. Supporting this hypothesis, genome-wide nuclear run-on assays by Core *et al*. show that divergent transcripts arise from transcriptionally engaged RNAPII at many genes in human fibroblasts (*9*).

Because TSSa-RNAs do not represent the 5′ end of transcripts, they likely mark regions of RNAPII pausing rather than initiation. Pausing of RNAPII 20 to 50 nt downstream of the TSS has been observed at many genes, most notably *Drosophila* Hsp^{70}, and is thought to maintain a chromatin structure permissive to transcription initiation (*10*, *11*). The results presented here suggest the presence of antisense paused RNAPII upstream of many TSSs. The position of paused, antisense RNAPII centers around 250 nt upstream of the TSS, as inferred by the presence of bound RNAPII and antisense short RNAs colocalizing at this location. Considering that chromatin marks associated with elongating RNAPII are only found downstream of TSSs, it appears that antisense RNAPII frequently does not elongate after TSSa-RNA production (Fig. 4D) (*12*–*14*). This suggests the existence of an undefined mechanism that discriminates between the sense and antisense polymerase for productive elongation.

RNAPII initiation complex polarity at promoters is thought to be established by TFIID/TBP complex binding together with TFIIB (*15*). RNAPII/TFIIF binding and DNA unwinding by the TFIIH helicase then gives rise to the open preinitiation complex (*10*). The prevalence of divergently oriented RNAPII at most promoters suggests a more complex situation. We hypothesize that transcription factors first nucleate a sense-oriented preinitiation complex at the TSS. Transcription by this complex generates at least two signals that could subsequently promote upstream antisense paused polymerase. First, the RNAPII carboxy-terminal domain and other initiation complex components can activate transcription when tethered to DNA, suggesting

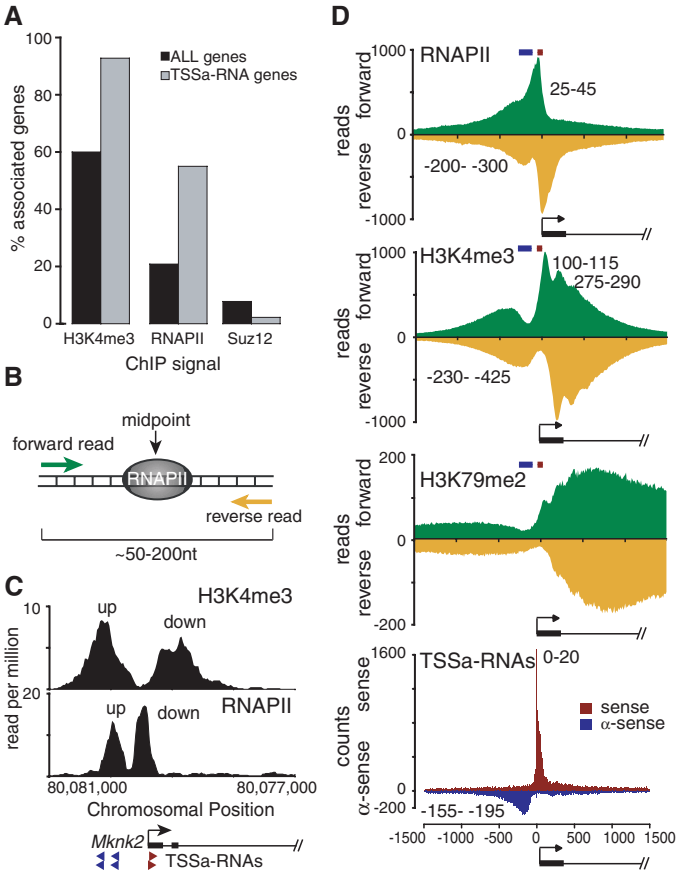

Fig. 4. Relationship between TSSa-RNAs, chromatin structure, and RNAPII. (**A**) Percentage of genes associated with H3K4me3, RNAPII and Suz12. ALL, all mouse genes. *t* test gives $P < 2.2 \times 10^{-16}$ for all marks. (**B**) Schematic of factor binding site mapping using Chip-seq reads. (**C**) Chromosomal position versus enrichment ratio for H3K4me3-modified nucleosomes and RNAPII for representative gene Mknk2. TSSa-RNAs are shown as arrowheads. (**D**) Metagene profiles for forward (green) and reverse (yellow) reads for ChIP-seq data (first three panels) and TSSa-RNAs from the sense (red) and antisense (blue) strand (bottom panel). Panels are aligned at the TSS. The TSS is denoted by the arrow. Black numbers on the profiles define the midpoint between forward and reverse peaks. Red and blue bars above the ChIP-seq profiles represent sense and antisense TSSa-RNA peak maxima.

that the sense complex may promote antisense preinitiation complex formation in the upstream region (*16*). Second, as RNAPII elongates the sense transcript, negative supercoiling of the DNA will occur upstream, perhaps promoting the antisense initiation process (*17*). This divergent transcription could structure chromatin and nascent RNA at the TSS for subsequent regulation.

References and Notes

1. G. Orphanides, D. Reinberg, *Cell* **108**, 439 (2002).
2. J. M. Calabrese, A. C. Seila, G. W. Yeo, P. A. Sharp, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 18097 (2007).
3. A. Marson *et al.*, *Cell* **134**, 521 (2008).
4. Materials and methods are available as supporting material on *Science* Online.
5. P. Carninci *et al.*, *Science* **309**, 1559 (2005).
6. L. A. Boyer *et al.*, *Nature* **441**, 349 (2006).
7. T. I. Lee *et al.*, *Cell* **125**, 301 (2006).
8. M. Sultan *et al.*, *Science* **321**, 956 (2008).
9. L. J. Core, J. J. Waterfall, J. T. Lis, *Science* **322**, 1845 (2008); published online 4 December 2008 (10.1126/science.1162228).
10. A. Saunders, L. J. Core, J. T. Lis, *Nat. Rev. Mol. Cell Biol.* **7**, 557 (2006).
11. D. A. Gilchrist *et al.*, *Genes Dev.* **22**, 1921 (2008).
12. M. G. Guenther, S. S. Levine, L. A. Boyer, R. Jaenisch, R. A. Young, *Cell* **130**, 77 (2007).
13. A. Barski *et al.*, *Cell* **129**, 823 (2007).
14. T. S. Mikkelsen *et al.*, *Nature* **448**, 553 (2007).
15. A. R. Kays, A. Schepartz, *Chem. Biol.* **7**, 601 (2000).
16. H. Xiao, J. T. Lis, H. Xiao, J. Greenblatt, J. D. Friesen, *Nucleic Acids Res.* **22**, 1966 (1994).
17. L. F. Liu, J. C. Wang, *Proc. Natl. Acad. Sci. U.S.A.* **84**, 7024 (1987).

18. We thank G. Zheng, C. Whittaker, S. Hoersch, and A. F. Seila. A.C.S. was supported by NIH postdoctoral fellowship 5-F32-HD051190 and G.W.Y. by the Crick-Jacobs Center for Computational Biology. This work was supported by NIH grants RO1-GM34277 and HG002668, NCI grant PO1-CA42063, and the NCI Cancer Center Support (core) grant P30-CA14051. The data discussed in this publication have been deposited in National Center for Biotechnology Information's Gene Expression Omnibus under accession numbers GSE13483 and GSE12680.

Supporting Online Material

www.sciencemag.org/cgi/content/full/1162253/DC1
Materials and Methods
Figs. S1 to S7
Tables S1 to S5
Data Files
References

24 June 2008; accepted 12 November 2008
Published online 4 December 2008;
10.1126/science.1162253
Include this information when citing this paper.

RNA Exosome Depletion Reveals Transcription Upstream of Active Human Promoters

Pascal Preker,[1] Jesper Nielsen,[2] Susanne Kammler,[1]* Søren Lykke-Andersen,[1] Marianne S. Christensen,[1] Christophe K. Mapendano,[1] Mikkel H. Schierup,[2] Torben Heick Jensen[1]†

Studies have shown that the bulk of eukaryotic genomes is transcribed. Transcriptome maps are frequently updated, but low-abundant transcripts have probably gone unnoticed. To eliminate RNA degradation, we depleted the exonucleolytic RNA exosome from human cells and then subjected the RNA to tiling microarray analysis. This revealed a class of short, polyadenylated and highly unstable RNAs. These promoter upstream transcripts (PROMPTs) are produced ~0.5 to 2.5 kilobases upstream of active transcription start sites. PROMPT transcription occurs in both sense and antisense directions with respect to the downstream gene. In addition, it requires the presence of the gene promoter and is positively correlated with gene activity. We propose that PROMPT transcription is a common characteristic of RNA polymerase II (RNAPII) transcribed genes with a possible regulatory potential.

Recent high-throughput analyses have revealed that >90% of all human DNA is transcribed (*1*). The vast majority of these transcripts are noncoding, thus challenging the classical definition of what constitutes a gene and, by association, a promoter (*2*–*4*). Furthermore, additional short-lived RNAs might have escaped detection. With the aim of identifying such transcripts, we used RNA interference in HeLa cells to deplete hRrp40, a core component of the human 3′ to 5′ exoribonucleolytic exosome, one of the major RNA degradation complexes (fig. S1A) (*5*). This resulted in a severe processing defect of the known exosome substrate 5.8S ribosomal RNA (fig. S1B), demonstrating diminished exosome function. Oligo dT-primed, double-stranded cDNA from cells that had been treated with either a control [enhanced green fluorescent protein (eGFP)] or hRrp40 small interfering RNA (siRNA) was hybridized to an encyclopedia of DNA elements (ENCODE) tiling array, which covers a representative ~1% of the human genome (*1*). Comparison of array data to public gene annotations revealed overall stabilization of mRNAs (exons in Fig. 1A), as expected. RNA from intronic and intergenic regions were largely unaffected, with the exception of a 1.5-kb region immediately upstream of transcription start sites (TSSs) that was stabilized ~1.5-fold on average (Fig. 1A). The relative stabilization of