Published in final edited form as:
Science. 2008 December 19; 322(5909): 1849–1851. doi:10.1126/science.1162253.

Divergent transcription from active promoters

Amy C. Seila[1,*], **J. Mauro Calabrese**[1,2,5,*], **Stuart S. Levine**[3], **Gene W. Yeo**[4,6], **Peter B. Rahl**[3], **Ryan A. Flynn**[1], **Richard A. Young**[2,3], and **Phillip A. Sharp**[1,2,†]

1 *Koch Institute, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*

2 *Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*

3 *Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, Massachusetts 02142, USA*

4 *Salk Institute, Crick-Jacobs Center for Theoretical and Computational Biology., 10010 North Torrey Pines Road, La Jolla, CA 92037, USA*

Abstract

Initiation by RNA polymerase II (RNAPII) is thought to occur unidirectionally from most genes. We have identified a class of short RNAs in mammalian cells that challenges this convention, suggesting the existence of widespread divergent transcription. Transcription start site-associated RNAs (TSSa-RNAs) are transcribed divergently from active protein-coding gene promoters, with anti-sense and sense TSSa-RNA peaks located upstream and downstream of the TSS, respectively. Transcription initiation marks such as bound RNAPII and H3K4-trimethylated histones are located at both sense and anti-sense TSSa-RNA positions; however, H3K79-dimethylated histones, a signature of elongating RNAPII, are present unidirectionally downstream of TSSs. These results suggest that divergent transcription over short distances is common of active promoters, and may help promoter regions maintain a state poised for subsequent regulation.

RNA polymerase II (RNAPII) transcription of DNA is an orchestrated process subject to regulation at numerous levels. When this process begins, RNAPII must bind to promoter DNA, initiate transcription, and transition to an elongation-state compatible with passage through nucleosomes. These transitions require concerted action by many protein complexes and are accompanied by changes in local chromatin structure, including covalent modification and ATP-dependent remodeling (1).

We have previously noted the presence of short RNAs in embryonic stem (ES) cell lines that were located near TSSs of protein-coding genes and not associated with known non-coding RNAs (2). To further investigate these low abundance RNAs, 8.4 million murine short RNA reads were analyzed (3); 7.3 million were derived from ES cells, and 1.1 million from differentiated cell types (4). ~42,000 of these reads, defined here as TSS-associated RNAs (TSSa-RNAs), uniquely mapped within 1.5 kb of protein-coding gene TSSs (Figure 1, Table S1). Multiple RNAs frequently associated with single TSSs (Figure 1B). TSSa-RNAs were found associated with over half of all mouse genes and were detected in all cell types examined (Figure S1). TSSa-RNAs were also found in Dicer$^{-/-}$ ES cells suggesting they are not Dicer products (Figure S1F). Sequenced TSSa-RNAs are most frequently 17 nucleotides (nt) long, with a mean length of 20 nt (Fig S2).

† To whom correspondence should be addressed: E-mail: sharppa@mit.edu.
[5]Present address: Department of Genetics and the Carolina Center for Genome Sciences, University of North Carolina, Chapel Hill, NC 27599, USA
[6]Present address: Department of Cellular and Molecular Medicine, University of California, San Diego, CA 92037, USA
*Co-first authors.

TSSa-RNAs surround promoters in divergent orientations. Sense TSSa-RNAs map downstream of the associated gene TSS, overlapping genic transcripts and peaking in abundance between +0 and +50 nucleotides downstream of the TSS. Surprisingly, 40% of TSSa-RNAs map upstream of the TSS and are oriented in the anti-sense direction relative to their associated genes, peaking between nucleotides -100 and -300. (Figure 1A). Sense and anti-sense TSSa-RNAs were found associated with overlapping sets of 8,115 and 6,331 gene promoters, respectively (Table S2). This distribution is not dependent on either head-to-head gene pairs or genes with multiple TSSs, nor is it seen in intergenic regions or at gene 3′ ends (Figure S3, S4).

A majority (59%) of ES cell TSSa-RNA associated genes have both sense and anti-sense TSSa-RNAs, indicating that individual TSSs produce both RNA sub-types (Figure S3E-F). Based on their direction and position relative to TSSs, we hypothesize that sense and anti-sense TSSa-RNAs arise from divergent transcription, defined as non-overlapping 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 are biased towards higher levels of gene expression. TSSa-RNAs were found at the majority of highly and moderately expressed genes (Figure 2, S5) and 80% were associated with CpG island promoters (Table S2). Additionally, the number of TSSa-RNA observations per gene correlates positively with gene expression levels, with a notable increase in the sense:anti-sense ratio found at the highest levels of expression (Figure 2B). This increase suggests that a fraction of these reads from the most active genes arise from mRNA turnover.

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

Based on cloning levels, a TSSa-RNA sequence is estimated to be present at ~1 molecule per 10 cells (4). Therefore, an enrichment procedure was developed to determine the nature of the short RNA species surrounding TSSa-RNA associated genes (4). Sequenced 21 nt sense and anti-sense TSSa-RNAs associated with Rnf12 or Ccdc52, respectively, were not detected as unique species in ES cells. Instead, species between 20 and 90 nucleotides were detected at levels estimated to be greater than 10 molecules per cell (4) (Figure 3B, D). Similar sized fragments were not found in Hela cell RNA samples using the same sequence probes, demonstrating specificity of the procedure (Figure 3B, D). Northern analysis for 2 other TSSa-RNA associated genes showed similar results (Figure S6, S7). We suggest that 20-90 nt transcripts are the dominant short RNA species from these two promoters and the TSSa-RNAs likely represent no more than 10% of the total associated transcripts.

To further classify promoters that produce TSSa-RNAs, and by inference, promoters that show evidence of divergent transcription, we examined their local chromatin environment using chromatin immunoprecipitation coupled with DNA sequencing (ChIP-seq) (3). TSSa-RNA associated promoters are enriched in bound RNAPII and H3K4me3 modified chromatin in ES cells (Figure 4A). ~90% of TSSa-RNA associated genes show H3K4me3-modified nucleosomes at their promoters, as compared to ~60% for all mouse genes (Figure 4A). TSSa-RNA associated genes also show a ~3-fold enrichment in promoter proximal RNAPII over all genes (Figure 4A). In contrast, TSSa-RNA associated genes are depleted of the Polycomb component Suz12 (Figure 4A).

Composite profiles of ChIP-seq data were used to determine RNAPII and histone modification positions relative to TSS. These profiles revealed a striking correlation with sense and anti-sense 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 (Figure 4B) (3). At TSSa-RNA associated genes, two distinct peaks for RNAPII are detectable with a spacing of several hundred base pairs (Figure 4C,D). A sharp RNAPII peak just downstream of the TSS lies directly over the sense TSSa-RNA peak (Figure 4D). A second RNAPII peak, upstream of the first, is more diffuse but again lies directly over the anti-sense TSSa-RNA peak (Figure 4D). The co-occurrence with anti-sense 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 (6).

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 (Figure 4D). These flanking peaks suggest that divergently paused RNAP II 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 (Figure 4D). These observations suggest that although divergent transcription initiation is widespread, productive elongation by RNAPII occurs primarily unidirectionally, downstream of TSSs.

Sense and anti-sense TSSa-RNAs with bound RNAPII are found at a surprisingly 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 divergent transcripts arise from transcriptionally engaged RNAPII at many genes in human fibroblasts.

Because TSSa-RNAs do not represent the 5′ end of transcripts, they likely mark regions of RNAPII pausing rather than initiation. Pausing has been observed at many genes, most notably *Drosophila* Hsp70, where it maintains RNAPII in a state poised for activation upon heat shock (7). RNAPII has been shown to pause 20-50 nt downstream of the TSS (7). The results presented here now suggest the presence of anti-sense paused RNAPII upstream of many TSSs. The position of paused, anti-sense RNAPII centers around 250 nt upstream of the TSS as inferred by the presence of bound RNAPII and anti-sense short RNAs co-localizing at this location. Considering that chromatin marks associated with elongating RNAPII are only found downstream of TSSs (Figure 4E) (8-10), it appears that anti-sense RNAPII frequently does not elongate after TSSa-RNA production. This suggests the existence of an undefined mechanism that discriminates between the sense and anti-sense polymerase for productive elongation.

RNAPII initiation complex polarity at promoters is thought to be established by TFIID/TBP complex binding together with TFIIB (11). RNAPII/TFIIF binding and DNA unwinding by the TFIIH helicase then gives rise to the open pre-initiation complex (7). The prevalence of divergently oriented RNAPII at most promoters suggests a more complex situation. We hypothesize that transcription factors first nucleate a sense oriented pre-initiation complex at the TSS. Transcription by this complex generates at least two signals that could subsequently promote upstream anti-sense paused polymerase. First, the RNAPII carboxy-terminal domain and other initiation complex components can activate transcription when tethered to DNA, suggesting that the sense complex may promote anti-sense pre-initiation complex formation in the upstream region (12). Secondly, as RNAPII elongates the sense transcript, negative supercoiling of the DNA will occur upstream, perhaps promoting the anti-sense initiation process (13). This divergent transcription would structure chromatin and nascent RNA at the TSS for subsequent regulation.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Many thanks to Grace Zheng, Charlie Whittaker, Sebastian Hoersch and Andrew Seila. ACS was supported by NIH postdoctoral fellowship 5-F32-HD051190 and GWY 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.

References and Notes

1. Orphanides G, Reinberg D. Cell Feb 22;2002 108:439. [PubMed: 11909516]
2. Calabrese JM, Seila AC, Yeo GW, Sharp PA. Proc Natl Acad Sci U S A Nov 13;2007 104:18097. [PubMed: 17989215]
3. Marson A, et al. Cell Aug 8;2008 134:521. [PubMed: 18692474]
4. Supplementary text.
5. Carninci P, et al. Science Sep 2;2005 309:1559. [PubMed: 16141072]
6. Sultan M, et al. Science Aug 15;2008 321:956. [PubMed: 18599741]
7. Saunders A, Core LJ, Lis JT. Nat Rev Mol Cell Biol Aug;2006 7:557. [PubMed: 16936696]
8. Guenther MG, Levine SS, Boyer LA, Jaenisch R, Young RA. Cell Jul 13;2007 130:77. [PubMed: 17632057]
9. Barski A, et al. Cell May 18;2007 129:823. [PubMed: 17512414]
10. Mikkelsen TS, et al. Nature Aug 2;2007 448:553. [PubMed: 17603471]
11. Kays AR, Schepartz A. Chem Biol Aug;2000 7:601. [PubMed: 11048951]
12. Ptashne M, Gann A. Nature Apr 10;1997 386:569. [PubMed: 9121580]
13. Liu LF, Wang JC. Proc Natl Acad Sci U S A Oct;1987 84:7024. [PubMed: 2823250]

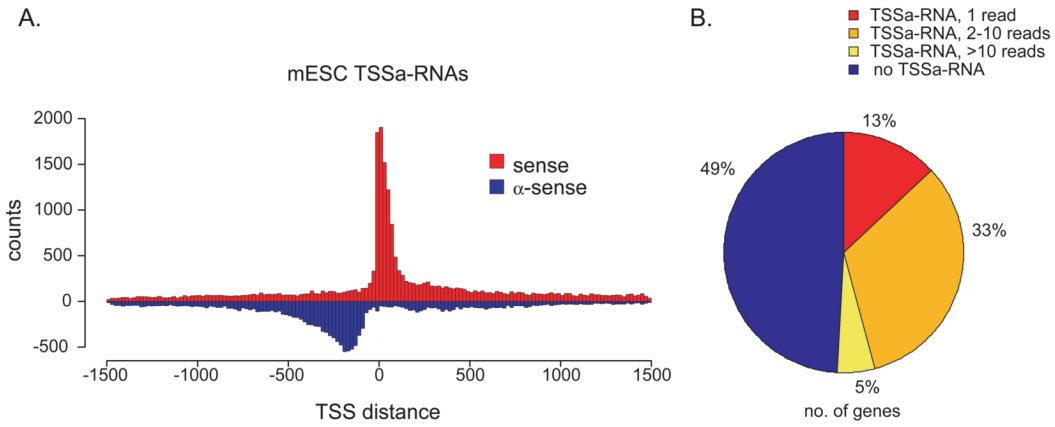

Figure 1.

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 nucleotide windows. Red and blue bars represent bins of TSSa-RNAs in the sense and anti-sense orientation with respect to gene transcription, respectively. (B) Percentage of annotated mouse genes with indicated number of associating TSSa-RNAs.

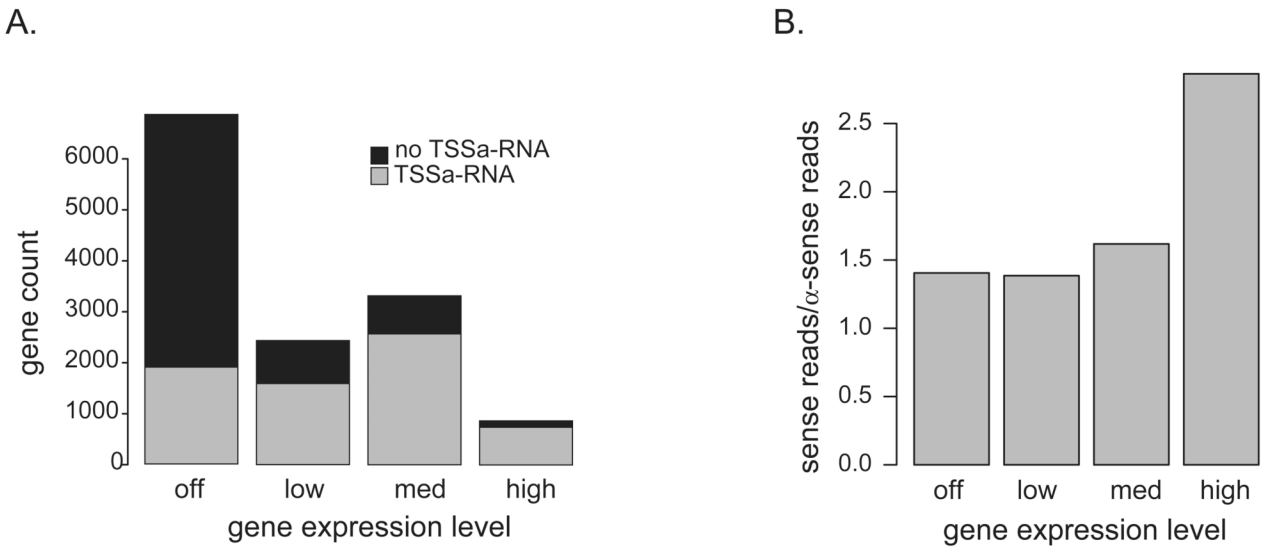

Figure 2.

In ES cells, TSSa-RNA associated genes are primarily expressed. (A) ES cell expression data was separated into 4 bins based on Log2 signal intensity levels; off = 1-4, low = 5-8, med = 6-12, and high ≥ 13 (9). Gene counts for each gene expression level bin are shown. (B) The ratio of sense to anti-sense reads in each expression bin.

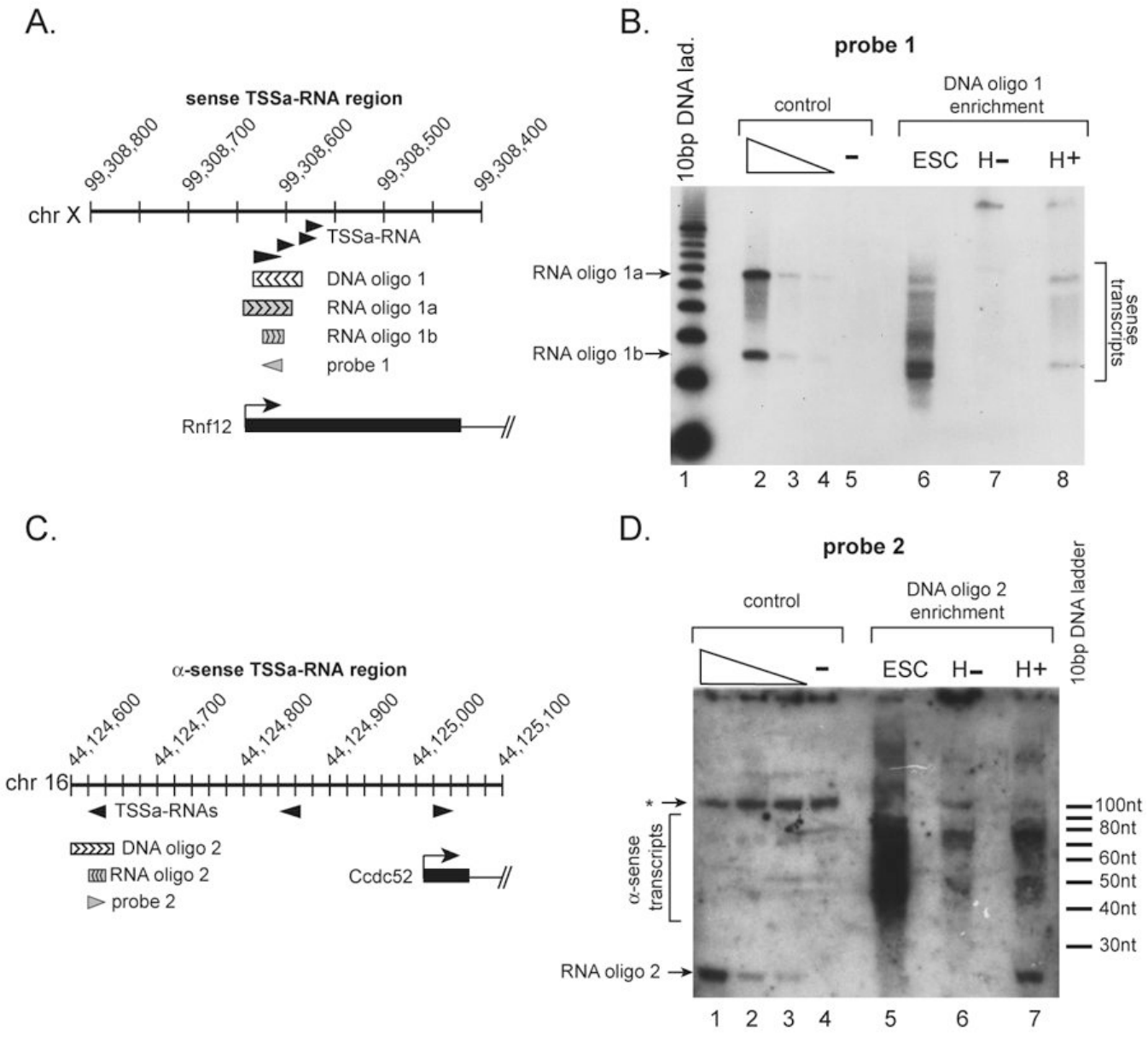

Figure 3.

Transcripts from TSSa-RNA associated regions are primarily 20-90 nts long. (A) Map of the sense TSSa-RNA Rnf12 region. (B) Northern analysis for the Rnf12 sense TSSa-RNA using probe 1 in A. Lane 1 is a 10 bp ladder. Lanes 2-5 are detection controls with 15, and 1.5, 0.75, and 0 fMol of synthetic RNAs (RNA oligo 1a/b in A). Lanes 6-8 are material recovered from the enrichment procedure using DNA oligo 1 in A. Lane 6 is J1 ES cell enriched material (ESC), lane 7 is Hela enriched material (H-), and lane 8 is Hela + 15 fMol synthetic RNA oligos 1a/b in A enriched material (H+). (C) Map of the anti-sense TSSa-RNA Ccdc52 region. (D) Northern analysis for the Ccdc52 anti-sense TSSa-RNA using probe 1 in A. Lanes 1-4 are as lanes 2-5 in B, except using RNA oligo 2 in C. Lanes 5-7 are material recovered from the enrichment procedure using DNA oligo 2. Lane 5 is J1 ES cell enriched material (ESC), lane 6 is Hela enriched material (H-), and lane 7 is Hela + 15 fMol synthetic TSSa-RNA-2 in C enriched material (H+). Bracket marks ESC specific transcripts; * marks background band.

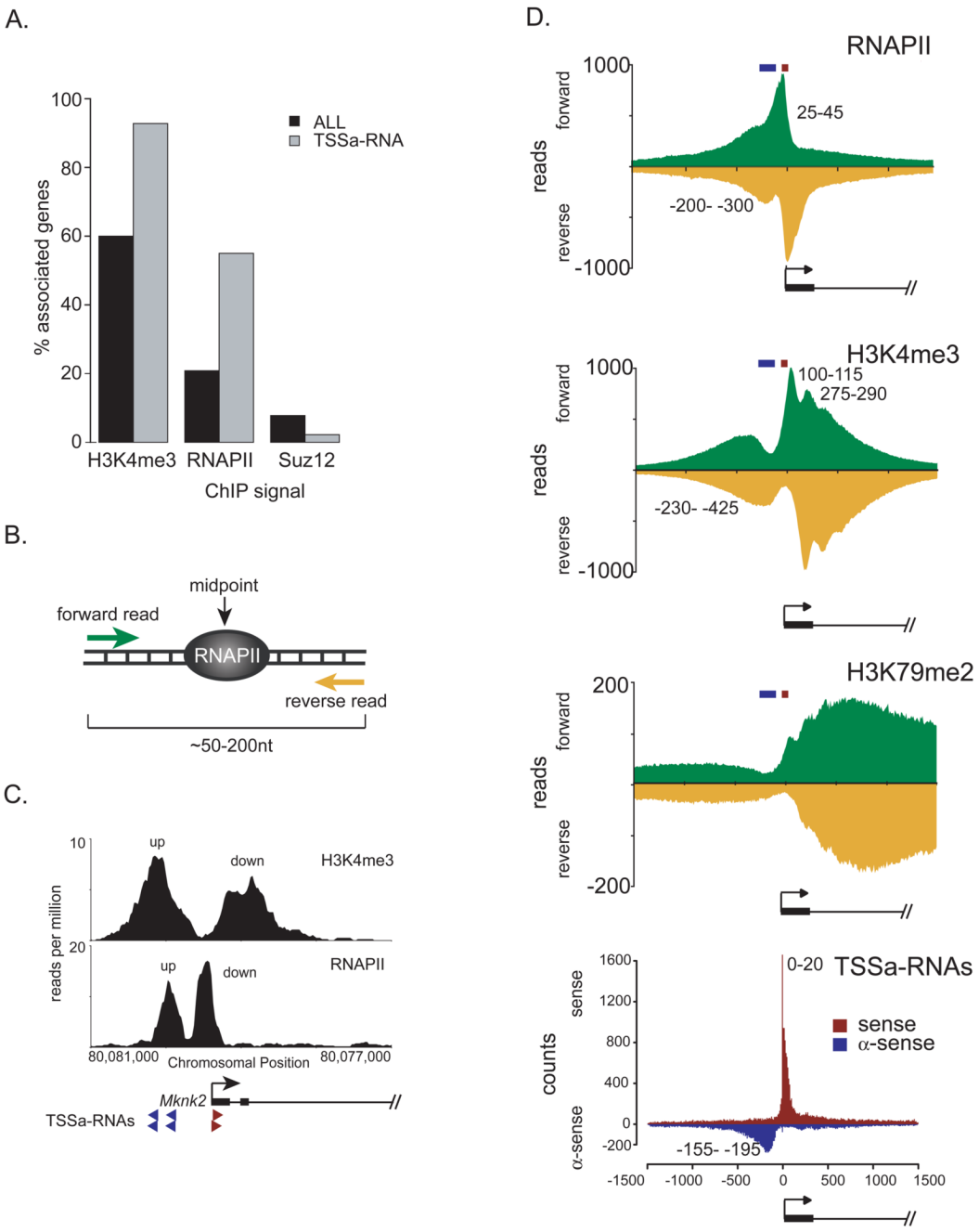

Figure 4.

Relationship between TSSa-RNAs and chromatin structure. (A) Percentage of genes associated with indicated chromatin marks. T-test gives p-values < $2.2e^{-16}$ for all marks. (B) Schematic of factor binding site mapping using forward and reverse Chip-seq reads. The midpoint between the forward and reverse reads defines the bound factor location. (C) Chromosomal position vs. enrichment ratio for H3K4me3-modified nucleosomes and RNAPII for a representative gene Mknk2. (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 anti-sense (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 anti-sense TSSa-RNA peak maxima.