Divergent transcription from active promoters

Amy C. Seila, J. Mauro Calabrese, Stuart S. Levine, Gene W. Yeo, Peter B. Rahl, Ryan A. Flynn, Richard A. Young, Phillip A. Sharp.

Index:

Supplemental Tables.

Supplemental References.

Supplemental File Index.

Table S5 – Excel file of the full list of genes associated with TSSa-RNAs and enriched chromatin regions.

The following files contain data formatted for upload into the UCSC genome browser (Kent et al., 2002). To upload the files, first copy the files onto a computer with internet access, then use a web browser to go to http://genome.ucsc.edu/cgi-bin/hgCustom?hgsid=105256378 for mouse and http://genome.ucsc.edu/cgi-bin/hgCustom?hgsid=104842340 for human. In the "Paste URLs or Data" section, select "Browse…" on the right of the screen. Use the pop-up window to select the copied files and then select "Submit". The upload process may take some time.

mouse_TSSaRNA_track.mm8.bed – Map of sense and anti TSSa-RNAs in mouse.

mES_chrom_ChIPseq.mm8.WIG.gz – ChIP-seq data for H3K4me3, H3K79me2, RNAPII and Suz12 in V6.5 mES cells. Top track for each data set illustrates the normalized number of reads assigned to each 25bp bin. Bars in the second track identify regions of the genome enriched at $p < 10\text{-}9$.

human_TSSa-RNA_track.hg17.bed – Map of sense and anti-sense TSSa-

RNAs in human.

Murine short RNA data sets.

All short RNA datasets are from (*S1*). Below, a brief description of the cell lines, short RNA cloning, and sequencing is included here for the convenience of the reader. The cloning and sequencing methods are followed by a detailed description of the short RNA read processing and identification and characterization of the TSSa-RNAs.

Growth conditions and quality control for mouse ES cells, neural precursors, and embryonic fibroblasts.

V6.5 (C57BL/6-129) murine ES cells were grown under typical ES cell culture conditions on irradiated mouse embryonic fibroblasts (MEFs) as previously described (*S2*). Briefly, cells were grown on gelatinized tissue culture plates in Dulbecco's modified Eagle medium supplemented with 15% fetal bovine serum (Hyclone), 1000 U/ml leukemia inhibitory factor (Chemicon; ESGRO ESG1106), non-essential amino acids, L-glutamine, Penicillin/Streptomycin and ß-mercaptoethanol. Immunostaining was used to confirm expression of pluripotency markers, SSEA 1 (Developmental Studies Hybridoma Bank) and Oct4 (Santa Cruz, SC-5279). For location analysis, cells were grown for one passage off of MEFs, on gelatinized tissue-culture plates. For RNA analysis by northern, cells were passaged off MEFs for more than one passage.

Embryonic stem cells harboring a doxycycline-repressible Oct4 allele (ZHBT-c4 cells) (*S3*), a gift from A. Smith, were cultured under standard ES cell conditions, described above, on gelatin. Cultures were treated with 2µg/ml doxycycline (SIGMA, D-9891) for 12 hrs or 24 hrs.

To generate neural precursor cells, V6.5 ES cells were differentiated through embryoid body formation for 4 days and selection in ITSFn media for 5–7 days, and maintained in FGF2 and EGF2 (R&D Systems) (*S4*).

Mouse embryonic fibroblasts were prepared from DR-4 strain mice as previously described (*S5*). Cells were cultured in Dulbecco's modified Eagle medium supplemented with 10% cosmic calf serum, β-mercaptoethanol, non-essential amino acids, L-glutamine and penicillin/streptomycin.

Short RNA cloning.

A method of cloning the 18-30nt transcripts previously described (*S6*) was modified to allow for Solexa (Illumina) sequencing (*S1*). Single-stranded cDNA libraries of short transcripts were generated using size selected RNA. RNA extraction was performed using Trizol, followed by RNeasy purification (Qiagen). 5µg of RNA was size selected and gel purified. The 3´ Adaptor (pTCGTATGCCGTCTTCTGTTG [IDT]) was ligated to RNA with T4 RNA ligase and also, separately with RNA ligase (Rnl2(1-249)k->Q). Ligation products were gel purified and mixed. The 5´ adaptor (GUUCAGAGUUCUACAGUCCGACGAUC) was ligated with T4 RNA ligase. RT-PCR (Superscript II, Invitrogen) was performed with 5´ primer (CAAGCAGAAGACGGCATA). Splicing of overlapping ends (SOE PCR) (*S7*) was performed (Phusion, NEB) with 5´ primer and 3´ PCR primer

(AATGATACGGCGACCACCGACAGGTTCAGAGTTCTACAGTCCGA), generating cDNA with extended 3′ adaptor sequence. The PCR product (40 µl) was denatured (85°C, 10 min, formamide loading dye), and the differently sized strands were purified on a 90% formamide, 8% acrylamide gel, yielding single-stranded DNA suitable Solexa sequencing. The single-stranded DNA samples were resuspended in 10 mM Tris (EB buffer)/0.1% Tween and then used as indicated in the standard Solexa sequencing protocol (Illumina). Each library was run on one lane of the Solexa sequencer.

Solexa sequencing of short RNAs

Polony generation on Solexa Flow-Cells.

The DNA library (2-4 pM) was applied to the flow-cell (8 samples per flow-cell) using the Cluster Station device from Illumina. The concentration of library applied to the flow-cell was calibrated such that polonies generated in the bridge amplification step originate from single strands of DNA. Multiple rounds of amplification reagents were flowed across the cell in the bridge amplification step to generate polonies of approximately 1,000 strands in 1 µm diameter spots. Double stranded polonies were visually checked for density and morphology by staining with a 1:5000 dilution of SYBR Green I (Invitrogen) and visualizing with a microscope under fluorescent illumination. Validated flow-cells were stored at 4^{o} C until sequencing.

Sequencing and Analysis.

Flow-cells were removed from storage and subjected to linearization and annealing of sequencing primer on the Cluster Station. Primed flow-cells were loaded into the Illumina Genome Analyzer 1G. After the first base was incorporated in the Sequencing-by-Synthesis reaction the process was paused for a key quality control checkpoint. A small section of each lane was imaged and the average intensity value for all four bases was compared to minimum thresholds. Flow-cells with low first base intensities were re-primed and if signal was not recovered the flow-cell was aborted. Flow-cells with signal intensities meeting the minimum thresholds were resumed and sequenced for 36 cycles. Images acquired from the Illumina/Solexa sequencer were processed through the bundled Solexa image extraction pipeline, which identified polony positions, performed base-calling, and generated QC statistics.

Short RNA read processing.

Bioinformatic extraction of individual short RNA clones from sequence reads was performed on each of the 6 individual data sets as described (*S8*) with minor modifications due to the Solexa sequencing procedure. Short RNAs sequenced by the Solexa method have no 5′ linker in their sequence, therefore the first 6 nucleotides of the 3′ linker was identified and the sequence before the match was extracted as the short RNA sequence “read”. These sequences were collapsed into a list of non-redundant sequences for further analysis. The non-redundant sequence list was compared to the *M. musculus* genome NCBI Build 35 (UCSC mm7) using the bundled ELAND software, and

those sequences with perfect (no gaps or mismatches across their entire length), unique (only a single match to the genome) were retained.

Genomic coordinates from overlapping sequences were collapsed to generate a non-redundant list of "genomic locations" for use in all subsequent genomic analysis. The 5´ coordinate of the genomic location was defined by the 5´ coordinate of the most 5´ sequence while the 3´ coordinate of the genomic location was defined as the most 3´ coordinate of the most 3´ sequence. The number of reads for each genomic location is equal to the sum of the reads that overlap at that genomic location.

Concomitant with Eland analysis, each sequence in the non-redundant list was compared to multiple non-coding RNA libraries; 1) annotated *M. musculus* microRNA hairpins from miRBase 8.2 (*S9*), 2) novel microRNAs found in our previous cloning efforts (*S8*), 3) annotated tRNA sequences (*S10, S11*), 4) the non-code RNA database (*S12*), and 5) the 45S rRNA (*S13*). All sequences with a perfect match to the mouse genome that did not match to any non-coding RNA in the 5 databases above were defined as "novel" and subject to further analysis.

Identification of the TSSa-RNAs.

All novel short RNA genomic locations were compared with all mouse genes annotated in the UCSC Known Gene database as well as those RefSeq annotations that are not found in the UCSC known gene database (*S14-16*). Sequences with start coordinates within ±1.5 kb of an annotated transcription start site (TSS) and with fewer than 100 reads were defined as TSSa-RNAs. The overwhelming majority of short RNA sequences within ±1.5 kb of TSSs were represented by very few reads in the Illumina/Solexa libraries from (*S1*); however, 32 isolated sequences were represented by hundreds of reads, an abundance more typical of tRNA, snoRNA, or rRNA fragments, or of lowly expressed, unannotated miRNAs. Given that such a large majority of sequences falling near TSSs were low in abundance, these high abundant sequences mapping near TSSs likely represent fragments of unannotated ncRNAs rather than TSSa-RNAs, and therefore a 100 read cutoff was applied to TSSa-RNA definition to avoid mis-classification. Exclusion of these 32 high abundance sequences had no significant effect on the TSSa-RNA metagene profile and did not change the number of TSSa-RNA associated genes.

The distance to the TSS is defined as the distance from the TSS to the end of the short RNA closest to the TSS. All histograms are of TSSa-RNA genomic locations with respect to the TSS and were plotted using R (*S17*). The count for all TSSa-RNAs that are on the same strand as the associated transcript were plotted as a histogram above the X-axis while the count for all TSSa-RNAs on the opposite strand as the associated transcript were plotted as a histogram below the X-axis, where each distinct genomic location is counted as 1 in the distribution. The map of all positioned short RNA genomic locations can be visualized using the UCSC browser by uploading supplemental file **mouse_TSSa-RNA_track.mm8.bed.**

Human short RNA data sets.

Growth conditions and quality control for human ES cells, neural precursor cells, and neurons.

Human embryonic stem cell line hues6 was cultured on plates coated with GFR matrigel (BD Biosciences) in MEF-conditioned medium and FGF-2 (20 ng/mL), as previously described (*S18*). Neuroepithelial precursor (NP) cells were generated and cultured as described (*S18*). NP cells were further differentiated into neuronally-enriched cells by culturing in DMEM/F12 medium supplemented with N2 and B27 (all from Invitrogen) in the presence of 1 mM dibutyryl cyclic AMP (Sigma), 20 ng/ml BDNF, 20 ng/ml GDNF (both from Peprotech) and 200 nM ascorbic acid (Sigma). Cells were allowed to differentiate for 4 weeks.

Short RNA cloning.

Small RNAs were purified and cloned using the Illumina/Solexa Small RNA Sample Prep Protocol (version 1.4B). Briefly, small RNAs of 18-30 nucleotides (nt) in length were isolated from 10 ug total RNA by electrophoresis on a 15% acrylamide/urea gel stained with SybrGold. Synthetic RNA oligonucleotides run alongside served as markers. Eluted RNAs were successively ligated to 5´ and 3´ adaptors with RNA ligase (Promega) and products purified on polyacrylamide/urea gels after each ligation step. RNAs were reverse transcribed with Superscript II (Invitrogen) and the resultant cDNA was amplified by PCR with Phusion high-fidelity DNA polymerase (NEB) for 15-17 cycles. PCR products were purified on a non-denaturing acrylamide gel, assayed for DNA concentration by PicoGreen assay (Invitrogen), and subjected to massively parallel sequencing on the Solexa/Illumina 1G sequencing platform.

Solexa sequencing of short RNAs.

Sequencing for the human cDNA libraries was performed at Solexa/Illumina using their in-house procedure.

Short RNA read processing.

After trimming adaptor sequences from Illumina/Solexa 1G analyzer sequenced reads; removing sequences with contiguous repeats of more than 9 ‘A’, ‘C’, ‘G’,’ T’ and more than 2 ‘Ns’ with a perl script, the trimmed sequences were mapped to the human genome (hg17) with MEGABLAST on a 32-node quad-core linux cluster retaining hits with 100% identify and hits that hit at most 5 locations in the genome. Genome sequences of human (hg17) were obtained from the University of California Santa Cruz (UCSC), as were the whole-genome multiz alignments. Sequences and annotation for known RNAs were obtained as follows: non-coding RNA from Rfam; microRNAs from Mirbase (Release 11.0); CD and H/ACA Box snoRNA and microRNA annotation from the Weber and Griffiths-Jones RNA table from the UCSC genome browser (UCSCgb); non-coding RNA gene annotation from the rnaGene table from the UCSCgb; 28S rRNA sequences were obtained from NCBI; repetitive sequence annotation were obtained from

Repeatmasker tracks from UCSCgb; piRNA and literature-curated non-coding RNA sequences from RNAdb; other non-coding RNA from ncRNA database. Protein-coding gene annotations were obtained from Refseq gene annotation obtained from the UCSCgb. Sequences that did not map to known non-coding RNAs and predicted microRNAs from MIResque (Yeo *et al.* submitted) were considered novel RNAs.

Identification of the TSSa-RNAs.

Human TSSa-RNAs were identified and histograms were plotted as described for mouse TSSa-RNAs above. The map of all positioned short RNA reads can be visualized using the UCSC browser by uploading supplemental file **human_TSSa-RNA_track.hg17.bed.**

Analysis of TSSa-RNA associated genes.

All analysis of TSSa-RNA associated genes was performed on the 4 mouse ESC data sets, ZHBTc4-0hDox (Oct4-0h), ZHBTc4-12hDox (Oct4-12h), ZHBTc4-24hDox (Oct4-24h) and mESC v6.5 combined, unless otherwise indicated (Table S1). For all TSSa-RNA associated gene analyses, gene symbols were used for transcripts where the gene symbol was defined. If no gene symbol was associated with the gene annotation, transcripts that were overlapping and in the sense orientation with respect to each other were defined by a representative gene annotation. All data analysis and plotting was performed using the statistical software package R unless otherwise indicated (*S17*).

Identification and analysis of multiple start site genes.

A large number of genes in the genome have multiple TSSs. This complicates the analysis of the TSSa-RNA distribution with respect to the TSS since the TSSa-RNA reads that map to multiple TSSs for the same gene will be counted more than once in the distribution. Therefore, it was important to determine if removing the TSSa-RNAs that map to multiple TSSs for the same gene would greatly affect the distribution. For each gene annotation that has an ES cell TSSa-RNA mapping to it, we determined if the gene annotation was unique or if it overlaps other gene annotations that are in the same orientation. All genes that are unique have a single known TSS and are called “single transcript genes”. Gene annotations that do overlap another gene annotation in the same orientation have multiple TSSs and are called “multi-transcript genes”. 52% of TSSa-RNA genes are single transcript genes while the other 48% are multi-transcript genes (Figure S3A). The distribution of TSSa-RNAs around the TSS of all single transcript genes (Figure S3B, C) shows a similar distribution relative to TSSs as the entire set of TSSa-RNAs (Figure 1A). Therefore, all analysis described in this manuscript included both TSSa-RNA subpopulations.

Identification and analysis of head-to-head gene pairs.

Approximately 10% of genes in the genome are “head-to-head” gene pairs, defined as two genes with TSSs located within 1 kb of each other and transcribed from opposite strands of the DNA (*S19*). TSSa-RNAs associated with head-to-head gene pairs are both sense and anti-sense to each gene (Table S1, sense and anti-sense read column, and Figure S3A). Therefore, it was important to determine how many reads mapped to

"head-to-head" gene pairs. Only 14% of TSSa-RNA associated genes are head-to-head genes (Figure S3D). Comparing the distribution of all TSSa-RNAs to those associated with all non-head-to-head genes reveals that the two distributions of TSSa-RNAs around the TSS are similar (Figure S3E). This suggests that the anti-sense TSSa-RNAs do not map to head-to-head genes alone and therefore the assignment of TSSa-RNAs as anti-sense is not due solely to head-to-head gene pair mapping.

Analysis of TSSa-RNAs clusters.

Analysis of TSSa-RNA associated genes shows that most genes have more than one short RNA mapping to their TSS (Figure 1B) and that these TSSa-RNAs are produced from multiple genomic locations. 28% of TSSa-RNA associated genes have a single associated TSSa-RNA while 78% have multiple TSSa-RNAs clustered at the TSS (Figure S3F). We then subdivided the multiple gene type into two subtypes, unidirectional clusters and divergent clusters (Figure S3F). If all the TSSa-RNAs associated with a particular gene are in the same orientation with respect to each other, then they are defined as a unidirectional cluster. If at least one of the TSSa-RNAs associated with a gene is in the opposite orientation with respect to the other associated TSSa-RNAs, then they are defined as a divergent cluster. Divergent cluster genes make up 59% of the multiple gene type. The TSSa-RNA distribution around the TSS for divergent cluster genes is remarkably similar to the distribution for all TSSa-RNAs (Figure S3G).

ES cell expression data and Gene Ontology analysis.

Expression data for murine ES cells was from Mikkelsen *et al*. (*S20*). The TSSa-RNA gene set was joined to each Affymetrix data set using Gene Symbol and Log2 of the MAS5 expression data was calculated. In order to determine TSSa-RNA gene expression levels with respect to the total population expression levels, the Log2 signal intensity data was divided into 4 equal bins (off = 1-4, low = 5-8, med = 9-12, and high $\geq$ 13) A stacked barplot showing the contributions of the no TSSa-RNA genes and the TSSa-RNA associated genes was generated. The subset of TSSa-RNA associated genes that are included in the expression arrays shows the same patterns of sense and anti-sense short RNAs as the total gene population, suggesting that this gene subset is an unbiased sample of the total population (Figure S3H).

Gene ontology analysis was performed using GOstat (*S21*). In order to determine which GO categories were enriched or depleted for the TSSa-RNA associated genes, the list of TSSa-RNA associated genes was analyzed against the list of genes used to identify the TSSa-RNAs (see Identification of the TSSa-RNAs, above). Based on GOstat results, only categories which have significant differences from the expectation value and p-value $\leq 10^{-20}$ were analyzed. The proportion of genes with GO annotations that are associated with TSSa-RNAs defines the expectation value used to determine GO category enrichment or depletion. The number of TSSa-RNA associated genes with GO annotations is 8,194 from 15,995 total genes with GO annotations, meaning that 51% of genes (8,194/15,995) randomly selected from a list of GO genes would be expected to associate with TSSa-RNAs. The fraction of TSSa-RNA associated genes in a specific GO category is calculated by dividing the number of TSSa-RNA associated genes in the GO category by the total gene number in that category. The difference from the expected

value for a subset of GO categories was plotted using Kaleidagraph software (Figure S5A). The expectation value (51%) is centered on the graph. Bars to the right of center indicate enrichment, while bars to the left of center indicate depletion.

Production of TSSa-RNAs might vary with the expression level of the associated gene. Therefore, level of gene expression was compared to number of TSSa-RNAs associating per gene. The TSSa-RNA associated genes were first divided into 3 bins based on the number of observed TSSa-RNAs (0, 1-3, 3-100) and the Log2 expression values for each bin were compared by boxplot analysis (Figure S5B). Statistical significance between the bins was calculated with a pairwise T-test using the Bonferroni method.

Analysis of CpG island overlap.

Mammalian promoters can be classified into two different categories in terms of their frequency of CpG dinucleotides: those with windows of high CpG content and those without such windows. CpG dinucleotides are generally depleted throughout the mammalian genome, presumably due to C → T conversion of methylated cytosine. CpG islands are regions that retain the expected random level of CpG dinucleotides (approximately 10-fold higher G+C content than the genome average). These regions are strongly associated with transcription initiation and tend to be devoid of CpG methylation. To determine the association of TSSa-RNAs with CpG islands, CpG island coordinates were downloaded from UCSC genome database (*S14, 22*). For each TSSa-RNA associated gene it was determined if there was a midpoint of a CpG island within 1 kb of the TSS (Table S2). To determine the enrichment for CpG islands at all genes, this analysis was also performed on the non-redundant set of all genes in the UCSC known gene and RefSeq databases combined. Chi-Sq test comparing the TSSa-RNA associated gene CpG overlap and the all genes gives a p-value = 0.0002, suggesting that the difference in the overlap is statistically significant.

Estimation of TSSa-RNA per cell abundance.

To approximate TSSa-RNA per cell abundance, we analyzed two previously published short cDNA libraries derived from J1 ES cells, referred to as “J1” and “J1aza” (*S8*). The J1 and J1aza libraries contain ~105,000 and ~115,000 reads, respectively, or approximately one single ES cell’s equivalent of short RNAs. Therefore, the overlap of TSSa-RNAs between these two libraries is a rough estimate of the number of TSSa-RNAs common to two single ES cells. TSSa-RNAs are produced from 40 common locations out of the 484 and 691 TSSa-RNA locations in the J1 and J1aza libraries, respectively. Using the J1 library as a reference, a single TSSa-RNA sequence is present in about 1 in 10 (40/484) ES cells. The average TSSa-RNA sequencing frequency in these two libraries was 1.22 reads per sequence. Together, these data suggest that individual TSSa-RNAs are present at 1 copy per every 10 ES cells.

Enrichment and Northern analysis.

In order to visualize transcripts arising from genomic regions with TSSa-RNAs, we developed an assay that included an enrichment procedure followed by northern analysis based on (*S23, 24*). This assay was used to assess the length of transcripts arising from 2 genic regions that produce sense TSSa-RNAs as well as regions upstream of 2

TSSs that produce anti-sense TSSa-RNAs (Figure 3, S6 and S7). Oligos were synthesized by either Dharmacon or IDT and are found in Table S4. For each genomic region, a synthetic 50 nt biotinylated DNA oligo spanning the TSSa-RNA region and extending 20 nts upstream and 10 nts downstream of the region was designed. As positive and negative controls for the analysis, we utilized Hela cell RNA with or without the synthetic short RNA added, respectively. Total RNA was prepped from approximately $8x10^8$ J1 ES cells or suspension Hela cells using Trizol (Invitrogen), following the standard protocol. Varying milligram concentrations of RNA were used in each enrichment procedure. The milligram amounts of RNA were 9 mg, 17 mg, 16 mg, and 8 mg for Rnf12, Ccdc52, Cops8, and Isg20l1, respectively. We have found that we can see signal with this enrichment procedure using as little as 2.5mg of RNA.

For enrichment, the DNA oligo was incubated with cellular RNA at 4º C in 10 mM Tris-Mes pH 6.0, 1 mM EDTA, 0.5 M NaCl with rotation. After 30 minutes, Dynal streptavidin beads (Invitrogen) were added and the solution was incubated for an additional 30 minutes at 4º C. Beads were captured with a magnetic particle concentrator (Invitrogen) and washed 4 times in low salt buffer (10mM Tris-Mes pH 6.0, 1mM EDTA, 10mM NaCl) for 2 minutes at 4º C. The bound RNA was then released from the DNA oligo by incubating the beads in low salt buffer at 95º C for 5 minutes. The RNA solution was concentrated 10-fold on a speed vac concentrator, mixed with 1x volume of formamide loading buffer, incubated as 95º C for 5 minutes, and loaded onto a 15% denaturing 10 mM MOPS pH 7.0 polyacrylamide gel (Sequagel, National Diagonistics) after pre-running the gel for 30 minutes. The samples were electrophoresed for 1 hour at 35W. Afterwards, the RNA was transferred to Hybond-NX nylon membrane (Amersham Biosciences) in water at 18 V in a semi-dry transfer apparatus. Transfer occurred at 4º C for 1.5 hours. The RNA was cross-linked to the membrane using EDC by incubating for overnight at 55º C (*S25*). After chemical crosslinking, the membrane was washed with H_2O and incubated with Ultrahyb oligo (Ambion) for 1 hour, after which 5´ end labeled LNA/DNA hybrid oligo, in which every 3rd nucleotide starting at position 2 was an LNA, was added to the Ultrahyb oligo solution (*S26*). Hybridization took place overnight and in the morning the membrane was washed 2 times for 1 minute with 2xSSC, 0.1% SDS at room temperature followed by 2, 30 minute washes with 2x SSC, 0.1% SDS at 37º C. The membrane was wrapped in Saran Wrap and exposed to film with an intensifying screen at -70º C. The length of exposure was 10 days for Ccdc52 associated enrichments (Figure 3 and S6), and 4 days for all other enrichments (Figure 3, S6 and S7). Prior to hybridizing with a different probe, membranes were stripped by incubating the membrane in boiling 0.1% SDS for 30 minutes and loss of signal was confirmed prior to rehybridization.

All northern films were scanned in and quantified using the gel analyzer in Imagej (*S27*). An estimate for the recovery can be determined by dividing the signal for the synthetic RNA oligo in the positive control (H+) sample by the 15 fMol or 1.5 fMol standards. We estimated the recovery to be 72% by taking the average between these two estimates. An estimate of the number of molecules/cell for the 20-90nt long TSSa-RNA transcript can be determined by dividing the area under the curve for the ES cell RNA sample by the area under the curve for the 15 fMol or 1.5 fMol standard (10 molecules/cell and 1 molecule/cell, respectively, based on an average of 20 pg RNA/ES cell). The average of the estimate with 15 fMol and 1.5 fMol standards was used to give

an approximate value for the molecules/cell for the short RNA transcripts. The quantification for each of the ES cell samples are 100 molecules/cell for the anti-sense transcripts upstream of Ccdc52, 76 molecules/cell for the anti-sense transcripts upstream of Isg20l1, 429 molecules/cell for the short sense transcripts overlapping the Cops8 gene, and 83 molecules/cell for the Rnf12 gene.

In order to gain more information about the longer TSSa-RNA species detected by northern analysis, each blot was stripped and re-probed using DNA/LNA hybrid oligos complementary to the regions surrounding the cloned TSSa-RNA (Figure S6 and S7). In general, a large amount of signal with a length distribution similar to that seen with the probe complementary to the cloned TSSa-RNA is seen with the probe just downstream of the cloned TSSa-RNA (Figure S6E and H and Figure S7E and H). In contrast, the probe just upstream of the cloned TSSa-RNA gives very little signal. This could be due to the placement of the DNA oligo used for the enrichment since the overlap with the region for the downstream probe is longer than for the upstream probe, or it could be that the longer TSSa-RNA species do not extend much further upstream past the 5´end of the cloned TSSa-RNA.

Comparison with ChIP-Seq data.

The ChIP-seq data sets used in this manuscript for H3K4me3, H3K79me2 and Suz12 are those in (*S1*). The ChIP-seq method for these data sets is described in full below for the convenience of the reader. Also included is a full detailed description of the ChIP-seq method for RNAPII as well as the method for high-resolution mapping of RNAPII, H3K4me3, and H3K79me2.

Antibodies.

RNA polymerase II (RNAPII) bound genomic DNA was isolated from whole cell lysate using 8WG16, a mouse monoclonal antibody (*S28*). This antibody preferentially binds a form of RNA polymerase II that lacks phosphorylation at the C-terminal domain of the largest subunit of polymerase (*S29-31*) although this preference can be subject to experimental conditions. A full list of genes associated with RNAPII enriched regions can be found in Table S5. RNAPII ChIP-seq data can be visualized on the UCSC browser by uploading supplemental file mES_chom_ChIPseq.mm8.WIG.gz

H3K4me3-modified nucleosomes were enriched from whole cell lysate using an epitope-specific rabbit polyclonal antibody purchased from Abcam (AB8580) (*S32, 33*). Samples were analyzed using ChIP-seq. Comparison of this data with ChIP-seq published previously (*S20*) showed near identity in profile and bound regions. A full list of genes associated with H3K4me3-modified nucleosomes can be found in Table S5 and has been previously published (*S1*). H3K4me3 ChIP-seq data can be visualized on the UCSC browser by uploading supplemental file mES_chom_ChIPseq.mm8.WIG.gz

H3K79me2-modified nucleosomes were isolated from mES whole cell lysate using Abcam antibody AB3594 (*S32*). Samples were analyzed using ChIP-seq. A full list of genes associated with H3K79me2 enriched regions can be found in Table S5 and has been previously published (*S1*). H3K79me2 ChIP-seq data can be visualized on the UCSC browser by uploading supplemental file mES_chom_ChIPseq.mm8.WIG.gz

Suz12-bound genomic DNA was enriched from whole cell lysate using an affinity purified rabbit polyclonal antibody purchased from Abcam (AB12073) and compared to a

reference whole cell extract (*S34*). A full list of genes associated with Suz12 enriched regions can be found in Table S5. Suz12 ChIP-seq data can be visualized on the UCSC browser by uploading supplemental file mES_chom_ChIPseq.mm8.WIG.gz

Chromatin Immunoprecipitation.

Protocols describing all materials and methods have been previously described (*S35*) and can be downloaded from http://web.wi.mit.edu/young/hES_PRC.

Briefly, we performed independent immunoprecipitations for each analysis. Embryonic stem cells were grown to a final count of $5x10^7$ – $1x10^8$ cells for each location analysis experiment. Cells were chemically crosslinked by the addition of one-tenth volume of fresh 11% formaldehyde solution for 15 minutes at room temperature. Cells were rinsed twice with 1xPBS and harvested using a silicon scraper and flash frozen in liquid nitrogen. Cells were stored at –80˚ C prior to use.

Cells were resuspended, lysed in lysis buffers and sonicated to solubilize and shear crosslinked DNA. Sonication conditions vary depending on cells, culture conditions, crosslinking and equipment. We used a Misonix Sonicator 3000 and sonicated at approximately 28 watts for 10 x 30 second pulses (90 second pause between pulses). For ChIP of Suz12 in murine ES cells, SDS was added to lysate after sonication to a final concentration of 0.1%. Samples were kept on ice at all times.

The resulting whole cell extract was incubated overnight at 4° C with 100 µl of Dynal Protein G magnetic beads that had been preincubated with approximately 10 µg of the appropriate antibody. Beads were washed 4-5 times with RIPA buffer and 1 time with TE containing 50 mM NaCl. For ChIP of Suz12 in murine ES cells, the following 4 washes for 4 minutes each were used instead of RIPA buffer: 1X low salt (20 mM Tris pH 8.1, 150 mM NaCl, 2 mM EDTA, 1% Triton X-100, 0.1% SDS), 1X high salt (20 mM Tris pH 8.1, 500 mM NaCl, 2 mM EDTA, 1% Triton X-100, 0.1% SDS), 1X LiCl (10 mM Tris pH 8.1, 250 mM LiCl, 1 mM EDTA, 1% deoxycholate, 1% NP-40), and 1X TE+ 50mM NaCl. Bound complexes were eluted from the beads by heating at 65° C with occasional vortexing and crosslinking was reversed by overnight incubation at 65° C. Whole cell extract DNA (reserved from the sonication step) was also treated for crosslink reversal.

ChIP-Seq Sample Preparation and Analysis.

All protocols for Illumina/Solexa sequence preparation, sequencing and quality control are provided by Illumina (http://www.illumina.com/pages.ilmn?ID=203). A brief summary of the technique and minor protocol modifications are described below.

Sample Preparation.

Purified immunoprecipitated (ChIP) DNA was prepared for sequencing according to a modified version of the Illumina/Solexa Genomic DNA protocol. Fragmented DNA was prepared for ligation of Solexa linkers by repairing the ends and adding a single adenine nucleotide overhang to allow for directional ligation. A 1:100 dilution of the Adaptor Oligo Mix (Illumina) was used in the ligation step. A subsequent PCR step with limited (18) amplification cycles added additional linker sequence to the fragments to prepare them for annealing to the Genome Analyzer flow-cell. After amplification, a narrow range of fragment sizes was selected by separation on a 2% agarose gel and

excision of a band between 150-300 bp (representing shear fragments between 50 and 200 nt in length and ~100 bp of primer sequence). The DNA was purified from the agarose and diluted to 10 nM for loading on the flow cell.

Polony generation on Solexa Flow-Cells.

The DNA library (2-4 pM) was applied to the flow-cell (8 samples per flow-cell) using the Cluster Station device from Illumina. The concentration of library applied to the flow-cell was calibrated such that polonies generated in the bridge amplification step originate from single strands of DNA. Multiple rounds of amplification reagents were flowed across the cell in the bridge amplification step to generate polonies of approximately 1,000 strands in 1μm diameter spots. Double stranded polonies were visually checked for density and morphology by staining with a 1:5000 dilution of SYBR Green I (Invitrogen) and visualizing with a microscope under fluorescent illumination. Validated flow-cells were stored at 4^{o}C until sequencing.

Sequencing.

Flow-cells were removed from storage and subjected to linearization and annealing of sequencing primer on the Cluster Station. Primed flow-cells were loaded into the Illumina Genome Analyzer 1G. After the first base was incorporated in the Sequencing-by-Synthesis reaction the process was paused for a key quality control checkpoint. A small section of each lane was imaged and the average intensity value for all four bases was compared to minimum thresholds. Flow-cells with low first base intensities were re-primed and if signal was not recovered the flow-cell was aborted. Flow-cells with signal intensities meeting the minimum thresholds were resumed and sequenced for 26 cycles.

Solexa Data Analysis.

Images acquired from the Illumina/Solexa sequencer were processed through the bundled Solexa image extraction pipeline that identified polony positions, performed base-calling and generated QC statistics. Sequences were aligned using the bundled ELAND software using murine genome NCBI Build 36 (UCSC mm8) as the reference genome. Only sequences perfectly and uniquely mapping to the genome were used.

The analysis methods used were derived from previously published methods (*S20*, *36*). Sequences from all lanes for each chromatin IP were combined, extended 200 bp (maximum fragment length accounting for ~100 bp of primer sequence), and allocated into 25 bp bins. Genomic bins containing statistically significant ChIP-seq enrichment were identified by comparison to a Poissonian background model, using a p-value threshold of 10^{-9}. Additionally, we used an empirical background model obtained from identical Solexa sequencing of DNA from whole cell extract (WCE) from matched cell samples (> 5x normalized enrichment across the entire region, see below). Reads per million are calculated by dividing the number of counts in each 25bp bin by the number of millions of reads (Figure 4C). A summary of the bound regions and their relation to gene targets can be found in Table S5.

The p-value threshold was selected to minimize the expected false-positive rate. Assuming background reads are spread randomly throughout the genome, the probability of observing a given number of counts can be modeled as a Poisson process where the

expectation can be calculated as the number of mapped reads times the number of bins per read (8) divided by the total number of bins available (we assumed 50% as a very conservative estimate).

The Poissonian background model assumes a random distribution of background reads, however we have observed significant deviations from this expectation in ChIP-seq datasets. These non-random events can be detected as sites of enrichment using control IPs and create a significant number of false positive events for actual ChIP-seq experiments. To remove these regions, we compared genomic bins and regions that meet the statistical threshold for enrichment to an empirical distribution of reads obtained from Solexa sequencing of DNA from whole cell extract (WCE) from matched cell samples. We required that enriched regions have five-fold greater ChIP-seq density in the specific IP sample as compared with the non-specific WCE sample, normalized for the total number of reads. This served to filter out genomic regions that are biased to having a greater than expected background density of ChIP-seq reads. We observed that ~200-500 regions in the genome showed non-specific enrichment in these experiments.

Analysis of enrichment for H3K4me3, RNAPII, Suz12, and H3K4+H3K27 at TSSa-RNA associated genes.

In order to determine the number of TSSa-RNA associated genes that also show a specific ChIP-seq signal, the TSSa-RNA associated gene list was joined with the ChIP-seq data sets using gene symbol (Table S5). ChIP-seq data for H3K4me3, RNAPII, and Suz12 are from the ChIP-seq data described above while the data for H3K4+H3K27 are from Mikkelsen *et al* (*S20*). The percent of all genes or TSSa-RNA associated genes that are associated with a specific ChIP-seq signal was determined by calculating the ratio of genes with a specific chromatin mark to all genes in the data set. The ChIP signal vs. the percent associated genes was plotted using Kaleidagraph software. The p-value for the difference between all genes and TSSa-RNA associated genes was calculated using a one-sample t-test where mu, the true value of the mean, was equal to the mean of the all gene data set.

In order to assess the status of chromatin marks with respect to gene expression we analyzed the expression data and ChIP data for all TSSa-RNA associated genes from Mikkelsen *et al*. (*S20*). The Log2 signal intensity data was divided into 4 equal bins (off = 1-4, low = 5-8, med = 9-12, and high $\geq$ 13). Within each bin, the proportion of genes with each chromatin mark was determined. A stacked barplot showing the contributions of each chromatin mark within each expression bin for all TSSa-RNA associated genes was generated using R (Figure S5C). Analysis of the H3K4me3 status for genes in each of the four expression bins shows that over 50% of genes in each bin show the H3K4me3 modification (Figure S5C).

High-resolution Composite Profile Analysis of ChIP-seq data.

Because of the extreme proximity of the expected divergent promoters, we modified our standard technique, generally used for ChIP-chip analysis, to generate composite profiles to take advantage of the high resolution provided by ChIP-seq data (*S37*). All reads that mapped to within 10 kb of an annotated start site were collected and sorted based on their strand relative to the start site. The strands were separated in order to correct for the variable fragment length expected from the processing of ChIP DNA for

the ChIP-seq procedure. This fragment distribution is expected to create a separation of forward and reverse reads of ~100-150 bp. Therefore the peaks of distributions of the reads should be displaced by about half that amount from the center of average binding.

Forward and reverse reads were binned based on the 5´ most base sequenced into 10 bp bins and summed across all promoters containing TSSa-RNAs. Similar analyses performed using transcription factors show very simple distributions of forward and reverse data with peaks flanking the predicted motif (*S1*). Analysis with RNAPII, H3K4me3 and H3K79me2 enriched DNA show distinct peaks in the forward and reverse direction. Because the distribution is centered on the transcription start site in the forward direction, the forward peak and downstream events, such as positioned nucleosomes, are easily visualized. The site of initiation in the reverse direction is less fixed and therefore represents an average of a relatively large spectrum of positions. Similar results were seen when examining all promoters and non-divergent promoters in addition to the TSSa-RNA associated promoters shown in Figure 4 (data not shown).

Supplementary Figures and Legends.

Legends

Figure S1- The distribution of short RNAs around TSSs of known genes in human and mouse cells types as well as Dicer$^{-/-}$ ES cells. (A-F) Distribution of the distance from each TSSa-RNA to each associated gene annotation for mouse neural precursor cells, NPC (A), human ES cell, hESC (B), mouse embryonic fibroblasts, mEFs (C), human neural precursor cells, NPC (D), human neurons (E) and mouse Dicer$^{-/-}$ ES cells (F). 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.

Figure S2- Length distribution for all TSSa-RNAs from the 6 murine data sets. The mean length for this distribution is 20nts while the median is 19nts.

Figure S3- Analysis of varying TSSa-RNA sub-populations does not change the overall nature of the TSSa-RNA distribution around the TSS. (A) Genomic DNA segments that show different classes of TSSa-RNA associated genes; Red triangles represent sense TSSa-RNAs, blue triangles represent anti-sense TSSa-RNAs, and red and blue striped triangles represent TSSa-RNAs sense to one transcript and anti-sense to another. (i) A gene with a single TSS that shows divergent TSSa-RNAs. (ii) A gene with two TSSs that shows divergent TSSa-RNAs. (iii) TSSa-RNAs mapping to a genomic locus with a head-to-head gene pair. (B-C) Analysis of TSSa-RNAs that map to a single transcript gene annotation compared to multi-transcript gene annotations. Single transcript genes are those that are defined by a single UCSC known gene or Refseq gene annotation. Multi-transcript genes have multiple UCSC known gene or Refseq gene annotations. In order to be considered a multi-transcript gene, transcript annotations have to overlap, arise from the same DNA strand, and have different TSSs. (B) Table of the

number of single transcript genes vs. multi-transcript genes by both genomic location as well as the number of reads. Reads that overlap in the genome are collapsed into a single genomic location. (C) Distribution of the distance from each TSSa-RNA to their associated gene annotations for all mouse ES cell single transcript genes. 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. (D-E) Analysis of TSSa-RNAs that map to "head to head" and "non-head to head" gene pairs. Head to head gene pairs were defined as transcripts for UCSC known gene or Refseq annotations that are in the opposite orientation with respect to each other and are associated with the same TSSa-RNA. (D) Table of the number of "head-to-head" and "non-head-to-head" genes as in B. (E) Distribution of the distance from each TSSa-RNA to their associated gene annotations for all mouse ES cell non-head-to head-genes. The histogram is as in C. (F-G) Analysis of the number of genes with a single or multiple associated TSSa-RNAs. Genes that have multiple associated TSSa-RNAs are subdivided into two classes based on the orientation of the TSSa-RNAs with respect to each other. TSSa-RNAs associated with a particular gene in the same orientation with respect to each other, are defined as a unidirectional cluster. If at least one of the TSSa-RNAs associated with a gene is in the opposite orientation with respect to the other associated TSSa-RNAs, then they are defined as a divergent cluster. (F) Table of the number of genes and % total in the single and multiple TSSa-RNA classes. (G) Distribution of the distance from each TSSa-RNA to their associated gene annotations for mouse ES cell genes in the multiple TSSa-RNA, divergent cluster class. The histogram is as in C. (H) Distribution of the distance from each TSSa-RNA to their associated gene annotations for all mouse ES cell genes for which there is Affymetrix expression array data, as in B (*S20*).

Figure S4- The distribution of short RNAs around TSSs of known genes, 3´ends of genes, and random points in intergenic regions. (A) Distribution of the distance from each TSSa-RNA to each associated gene annotation for mouse ES cells. Histogram of the distance from each TSSa-RNA to all associated gene TSSs. 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) Distribution of the distance from all RNAs to the 3´end to each associated gene annotation for mouse ES cells. Histogram of the distance from each RNA to all associated gene 3´ends. Counts of TSSa-RNA 5´ positions relative to gene 3´ends 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. (C) Histogram of the distance from all cloned RNAs to randomly selected locations in non-repetitive intergenic. Counts of RNA 5´ positions relative to the associated location are binned in 20 nucleotide windows. Red and blue bars represent bins of RNAs in the sense and anti-sense orientation with respect to a randomly selected orientation.

Figure S5- Analysis of TSSa-RNA associated gene populations. (A) Gene ontology (GO) analysis for TSSa-RNA associated genes. Shown is the difference from the expected value for GO categories that are significantly, p-value $\leq 10^{-20}$,

enriched/depleted for the TSSa-RNA associated genes. The expected value (51%) is centered on the graph. Bars to the right of center indicate enrichment, while bars to the left of center indicate depletion. (B) Boxplot of $\log_2$ (ES cell expression data) vs. TSSa-RNA associated genes with 0, 1-3 reads, and >3 associated TSSa-RNAs. (C) The percentage of genes with H3K27me3 (K27, white), H3K4me3 (K4, light grey), the bivalent mark H3K4me3+K3K27me3 (K4+K27, dark grey), and no chromatin marks (None, black) are shown for each gene expression bin. The expression bins are as in B.

Figure S6- Transcripts from the TSSa-RNA associated regions genes Cops8 and Isg20l1 are primarily 20-90 nts long. (A) Map of the sense TSSa-RNA region at the Cops8 gene. (B) Map of the anti-sense TSSa-RNA region at the Isg20l1 gene. (C) The membrane shown in D was stripped and re-probed with probe 1 in A. Lanes 6-8 of the membrane in B are shown. (D) Northern analysis for the Cops8 sense TSSa-RNA using probe 2 in A. Lane 1 is a 10 bp DNA ladder. Lanes 2-5 are detection controls where 15, and 1.5, 0.75, and 0 fMol of synthetic RNAs (RNA oligo 1 in A) were loaded directly on the gel along with 2ug of Hela RNA as carrier. Lanes 6-8 are material recovered from the enrichment procedure using DNA oligo 1 in A. Lane 6 is the material recovered from 16 mg J1 ES cell RNA, lane 7 is material recovered from 16 mg Hela RNA (H-), and lane 8 is material recovered from 18 mg Hela RNA and 15 fMol synthetic RNA oligos 1 in A (H+). (E) The membrane shown in D was stripped and re-probed with probe 3 in A. Lanes 6-8 of the membrane in D are shown. (F) The membrane shown in G was stripped and re-probed with probe 4 in B. Lanes 5-7 of the membrane in G are shown. (G) Northern analysis for the Isg20l1 anti-sense TSSa-RNA using probe 5 in B. Lane 1 is a 10 bp ladder. Lanes 2-5 are detection controls where 15, and 1.5, 0.75, and 0 fMol of synthetic RNAs (RNA oligo 2 in B) were loaded directly on the gel along with 2 ug of Hela RNA as carrier. Lanes 6-8 are material recovered from the enrichment procedure using the DNA oligo. Lane 6 is the material recovered from 8 mg J1 ES cell RNA, lane 7 is material recovered from 10mg Hela RNA (H-), and lane 8 is material recovered from 10 mg Hela RNA and 15 fMol synthetic RNA oligo 2 in B (H+). A bracket marks ES cell specific anti-sense transcripts. (H) The membrane shown in G was stripped and re-probed with probe 6 in B. Lanes 6-8 of the membrane in G are shown.

Figure S7- Transcripts from the TSSa-RNA associated regions genes Rnf12 and Ccdc52 are primarily 20-90 nts long. (A) Map of the sense TSSa-RNA region for the Rnf12 gene. (B) Map of the anti-sense TSSa-RNA region for the Ccdc52 gene. (C) The membrane shown in D was stripped and re-probed with probe 1 in A. Lanes 6-8 of the membrane in D are shown. (D) Northern analysis for the Rnf12 sense TSSa-RNA using probe 2 in A. Lane 1 is a 10 bp DNA ladder. Lanes 2-5 are detection controls where 15, and 1.5, 0.75, and 0 fMol of synthetic RNAs (RNA oligo 1a and 1b in A) were loaded directly on the gel along with 2ug of Hela RNA as carrier. Lanes 6-8 are material recovered from the enrichment procedure using DNA oligo 1 in A. Lane 6 is the material recovered from 9 mg J1 ES cell RNA, lane 7 is material recovered from 9 mg Hela RNA (H-), and lane 8 is material recovered from 9 mg Hela RNA and 15 fMol synthetic RNA oligos 1a and 1b in A (H+). (E) The membrane shown in D was stripped and re-probed with probe 3 in A. Lanes 6-8 of the membrane in D are shown. (F) The membrane shown in G was stripped and re-probed with probe 4 in B. Lanes 5-7 of the membrane in G are

shown. (G) Northern analysis for the Ccdc52 anti-sense TSSa-RNA using probe 5 in B. Lanes 1-4 are detection controls where 15, and 1.5, 0.75, and 0 fMol of synthetic RNAs (RNA oligo 2 in B) were loaded directly on the gel along with 2ug of Hela RNA as carrier. Lanes 5-7 are material recovered from the enrichment procedure using the DNA oligo. Lane 5 is the material recovered from 17 mg J1 ES cell RNA, lane 6 is material recovered from 17mg Hela RNA (H-), and lane 7 is material recovered from 17 mg Hela RNA and 15 fMol RNA oligo 2 in B (H+). A bracket marks ES cell specific anti-sense transcripts; * marks a background band. (H) The membrane shown in G was stripped and re-probed with probe 6 in B. Lanes 5-7 of the membrane in G are shown along with the 10 bp DNA ladder in lane 8.

FIgure S1

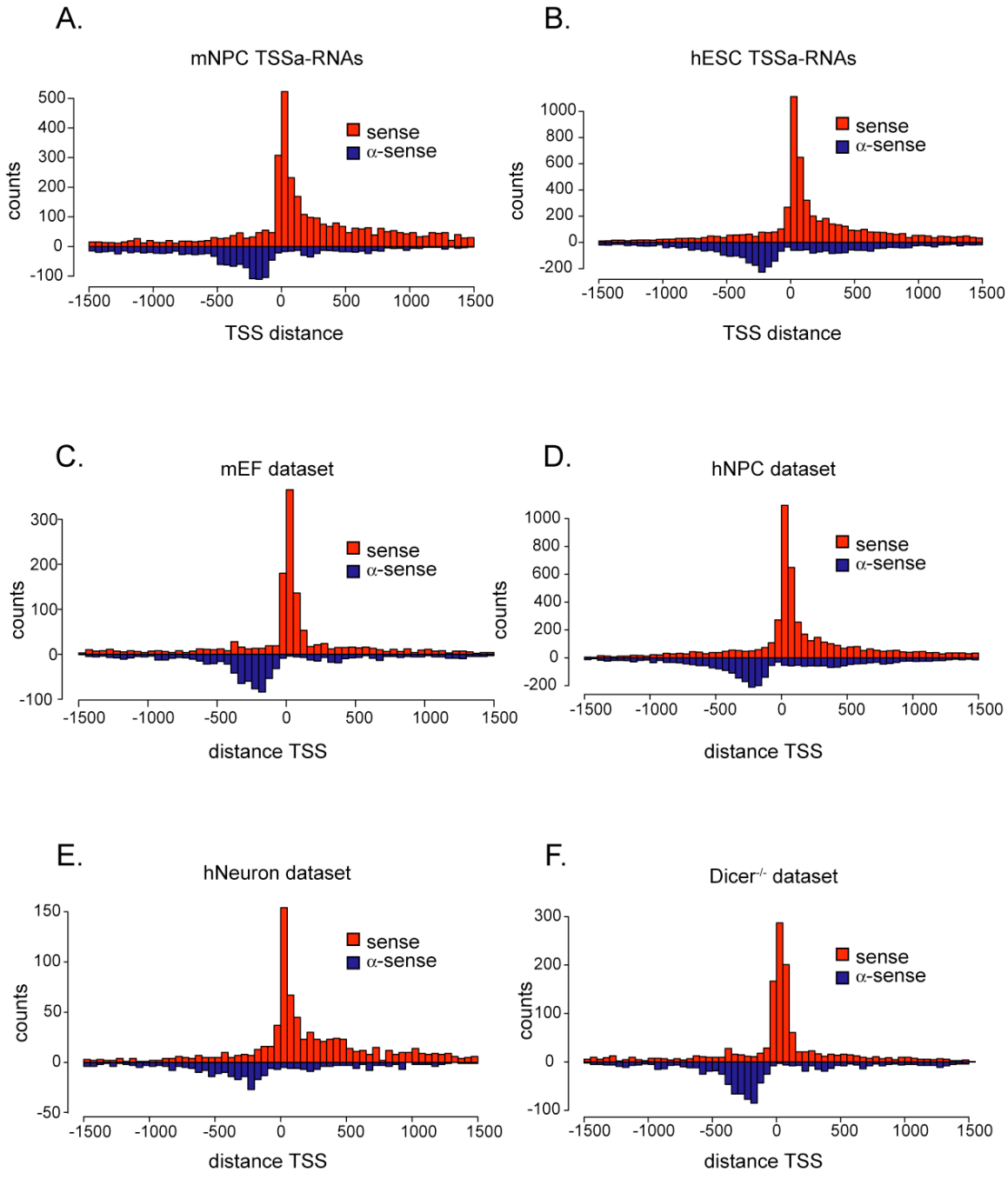

A.
mNPC TSSa-RNAs
sense
α-sense
counts
TSS distance
B.
hESC TSSa-RNAs
sense
α-sense
counts
TSS distance
C.
mEF dataset
sense
α-sense
counts
distance TSS
D.
hNPC dataset
sense
α-sense
counts
distance TSS
E.
hNeuron dataset
sense
α-sense
counts
distance TSS
F.
Dicer-/- dataset
sense
α-sense
counts
distance TSS

Figure S2

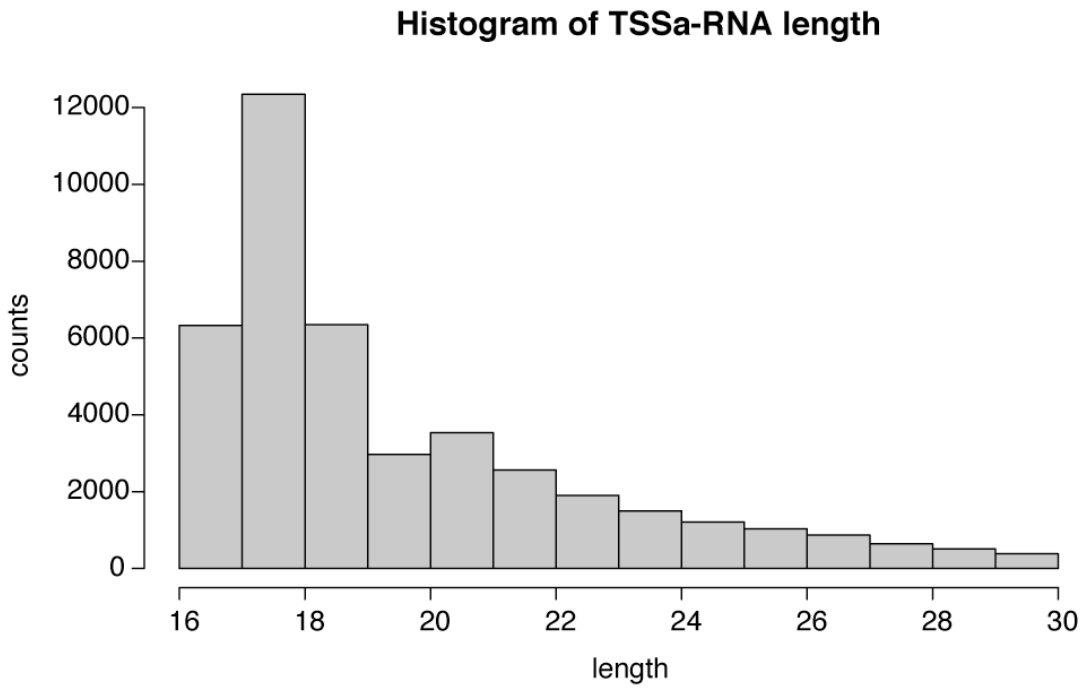
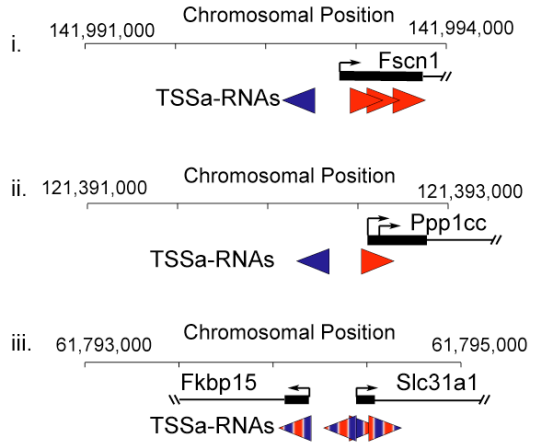

Histogram of TSSa-RNA length
counts
12000
10000
8000
6000
4000
2000
0
16
18
20
22
24
26
28
30
length

Figure S3

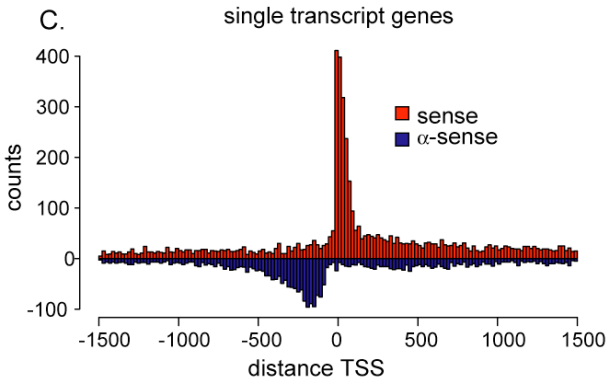
A.



B.

Gene Type	genomic locations	No. of Reads
single transcript	10,376	18,903
multi-transcript	9,630	17,970

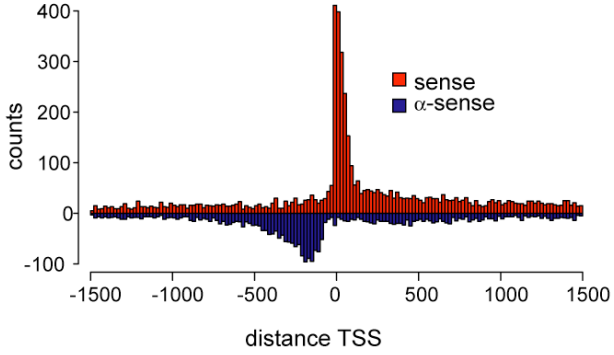
C.



D.

Gene Type	genomic locations	No. of Reads
head-to-head	2,844	5,522
non-head-to-head	17,162	31,351

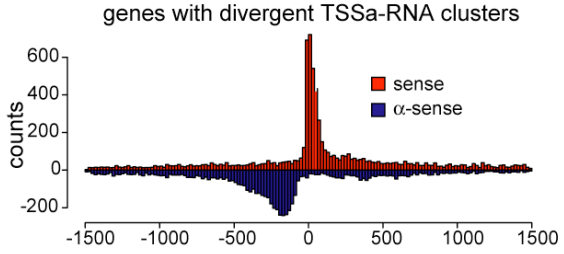
E.



F.

TSSa-RNA/gene		No. of genes		% total genes	
Single		2,281		28%	
Multiple	unidirectional clusters	5,875	2,430	72%	41%
	divergent clusters		3,445		59%

G.



H.

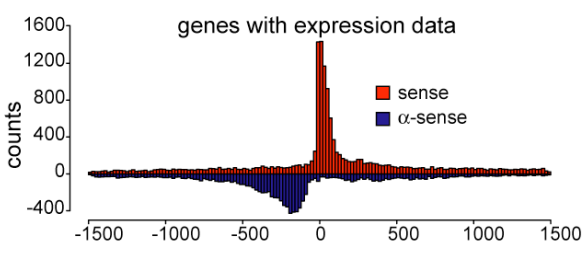


Figure S4

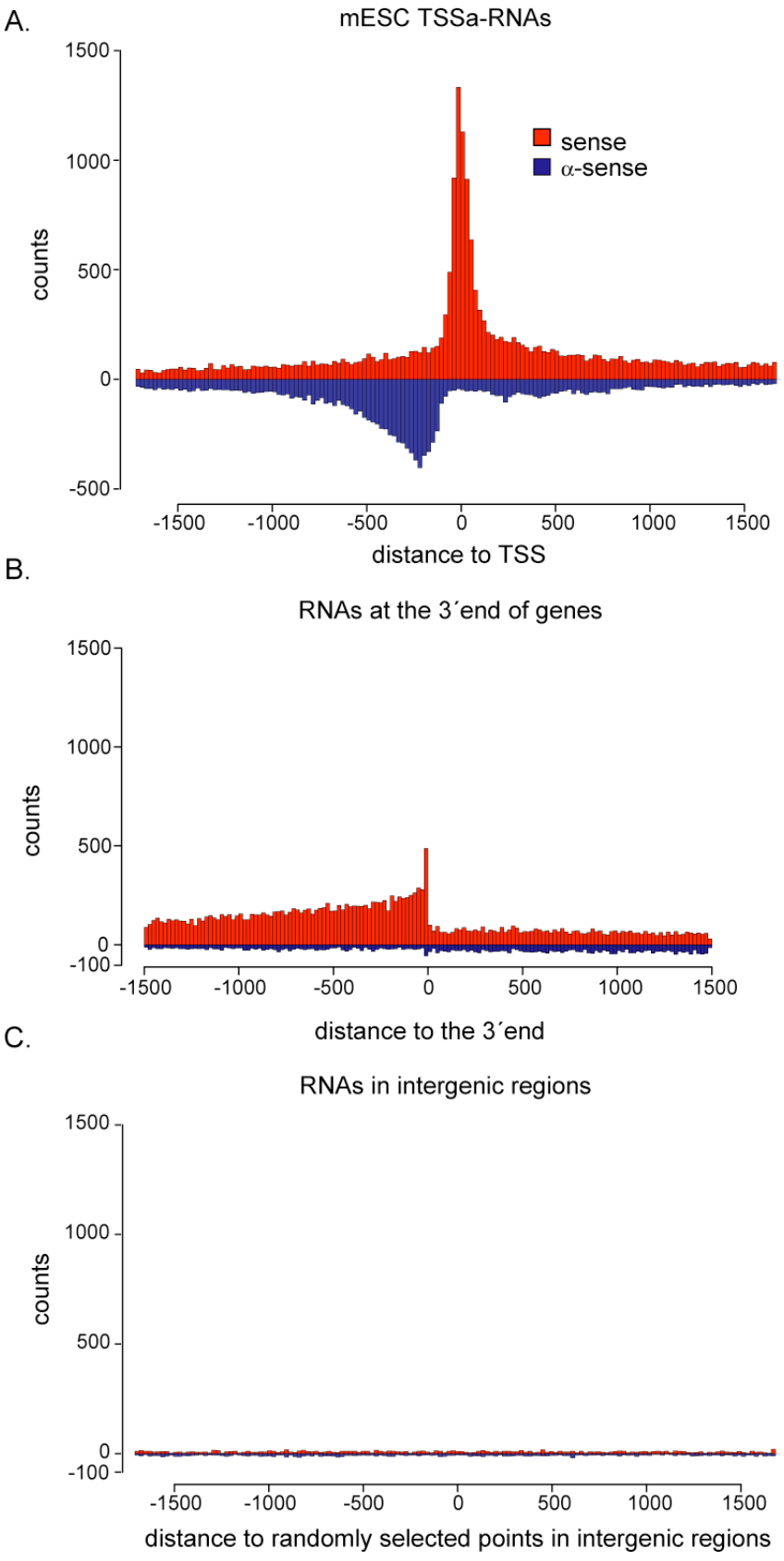

A.
mESC TSSa-RNAs
1500
1000
500
0
-500
counts
sense
α-sense
-1500
-1000
-500
0
500
1000
1500
distance to TSS
B.
RNAs at the 3´end of genes
1500
1000
500
0
-100
counts
-1500
-1000
-500
0
500
1000
1500
distance to the 3´end
C.
RNAs in intergenic regions
1500
1000
500
0
-100
counts
-1500
-1000
-500
0
500
1000
1500
distance to randomly selected points in intergenic regions

Figure S5

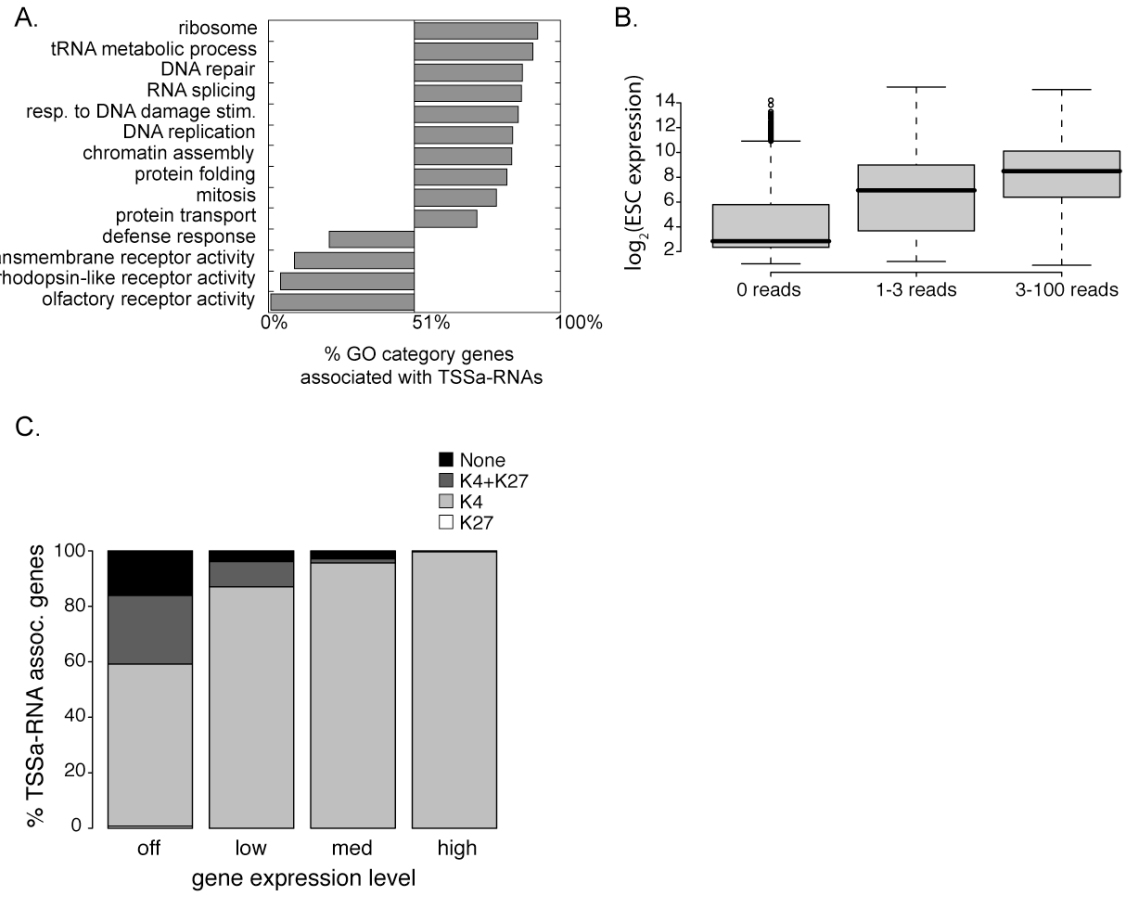
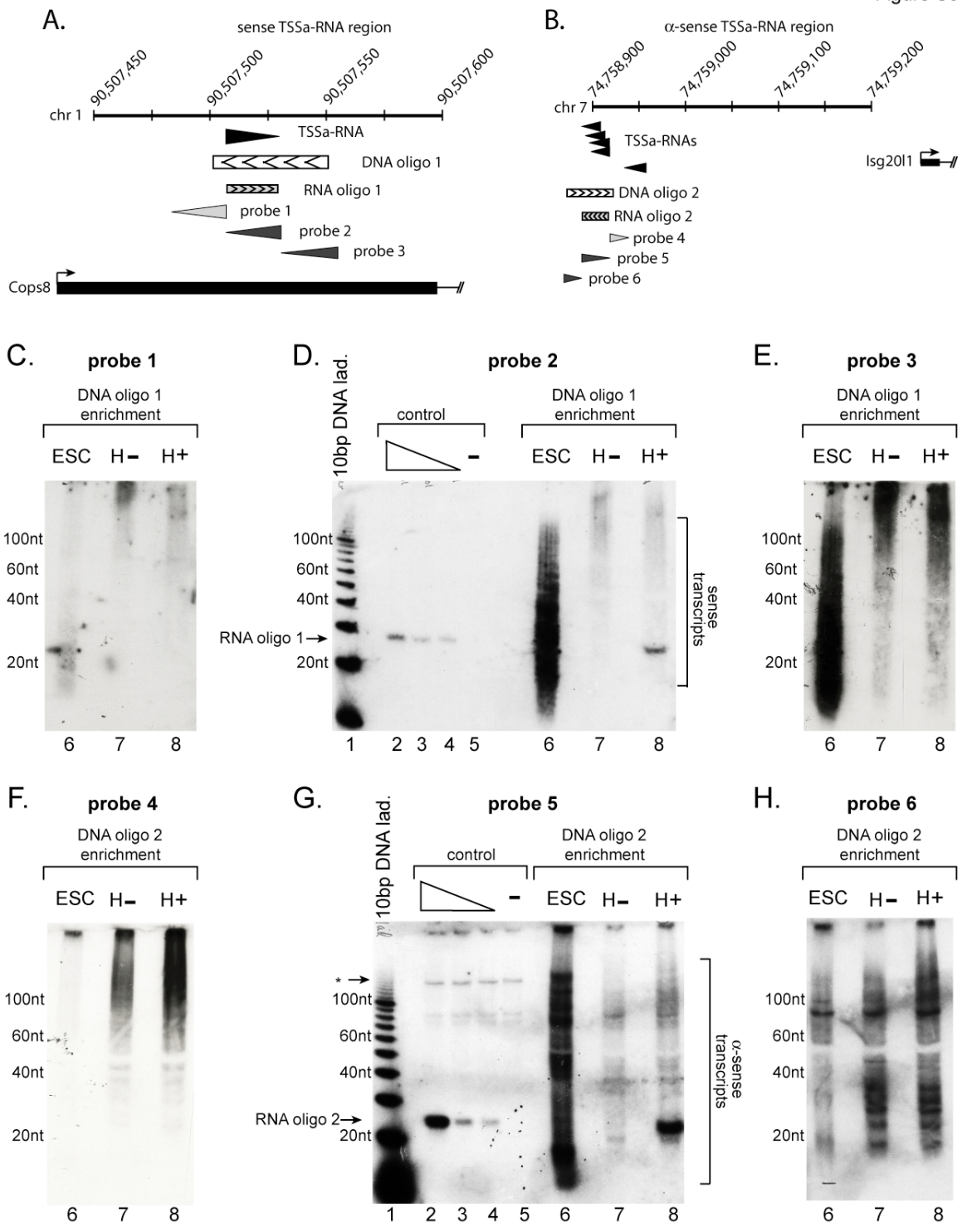
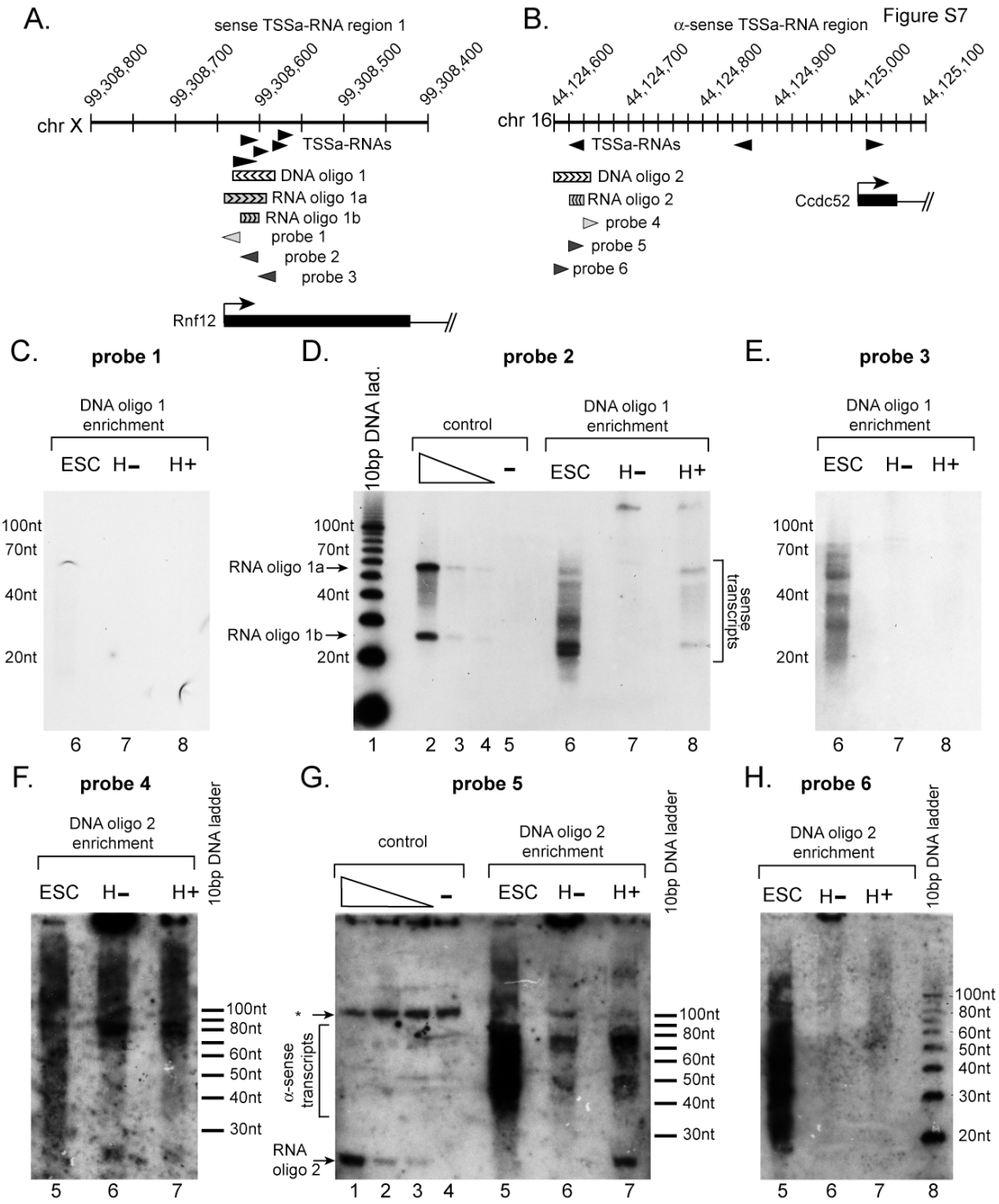

A.
ribosome
tRNA metabolic process
DNA repair
RNA splicing
resp. to DNA damage stim.
DNA replication
chromatin assembly
protein folding
mitosis
protein transport
defense response
transmembrane receptor activity
rhodopsin-like receptor activity
olfactory receptor activity
0%
51%
100%
% GO category genes
associated with TSSa-RNAs
B.
log2(ESC expression)
14
12
10
8
6
4
2
0 reads
1-3 reads
3-100 reads
C.
None
K4+K27
K4
K27
% TSSa-RNA assoc. genes
100
80
60
40
20
0
off
low
med
high
gene expression level

Figure S6


A.
sense TSSa-RNA region
90,507,450
90,507,500
90,507,550
90,507,600
chr 1
TSSa-RNA
DNA oligo 1
RNA oligo 1
probe 1
probe 2
probe 3
Cops8
B.
α-sense TSSa-RNA region
74,758,900
74,759,000
74,759,100
74,759,200
chr 7
TSSa-RNAs
Isg20l1
DNA oligo 2
RNA oligo 2
probe 4
probe 5
probe 6
C.
probe 1
DNA oligo 1 enrichment
ESC H– H+
100nt
60nt
40nt
20nt
6 7 8
D.
probe 2
10bp DNA lad.
control
DNA oligo 1 enrichment
ESC H– H+
100nt
60nt
40nt
RNA oligo 1
20nt
sense transcripts
1 2 3 4 5 6 7 8
E.
probe 3
DNA oligo 1 enrichment
ESC H– H+
100nt
60nt
40nt
20nt
6 7 8
F.
probe 4
DNA oligo 2 enrichment
ESC H– H+
100nt
60nt
40nt
20nt
6 7 8
G.
probe 5
10bp DNA lad.
control
DNA oligo 2 enrichment
ESC H– H+
*
100nt
60nt
40nt
RNA oligo 2
20nt
α-sense transcripts
1 2 3 4 5 6 7 8
H.
probe 6
DNA oligo 2 enrichment
ESC H– H+
100nt
60nt
40nt
20nt
6 7 8


A.
sense TSSa-RNA region 1
99,308,800
99,308,700
99,308,600
99,308,500
99,308,400
chr X
TSSa-RNAs
DNA oligo 1
RNA oligo 1a
RNA oligo 1b
probe 1
probe 2
probe 3
Rnf12
B.
α-sense TSSa-RNA region
Figure S7
44,124,600
44,124,700
44,124,800
44,124,900
44,125,000
44,125,100
chr 16
TSSa-RNAs
DNA oligo 2
RNA oligo 2
probe 4
probe 5
probe 6
Ccdc52
C.
probe 1
DNA oligo 1 enrichment
ESC H– H+
100nt
70nt
40nt
20nt
6 7 8
D.
probe 2
10bp DNA lad.
control
DNA oligo 1 enrichment
ESC H– H+
100nt
70nt
RNA oligo 1a
40nt
RNA oligo 1b
20nt
sense transcripts
1 2 3 4 5 6 7 8
E.
probe 3
DNA oligo 1 enrichment
ESC H– H+
100nt
70nt
40nt
20nt
6 7 8
F.
probe 4
DNA oligo 2 enrichment
ESC H– H+
10bp DNA ladder
100nt
80nt
60nt
50nt
40nt
30nt
5 6 7
G.
probe 5
control
DNA oligo 2 enrichment
ESC H– H+
10bp DNA ladder
*
α-sense transcripts
RNA oligo 2
100nt
80nt
60nt
50nt
40nt
30nt
1 2 3 4 5 6 7
H.
probe 6
DNA oligo 2 enrichment
ESC H– H+
10bp DNA ladder
100nt
80nt
60nt
50nt
40nt
30nt
20nt
5 6 7 8

Supplementary Tables

Cell Type	genomic locations	no. of Reads	sense reads	anti-sense reads	Sense & anti-sense reads
All Mouse	23105	42139	24151	12522	5466
Oct4-0h	5245	6864	3979	2027	858
Oct4-12h	6606	10300	5638	3303	1359
Oct4-24h	7244	11054	6048	3500	1506
mESC v6.5	4176	8655	5172	2312	1171
mNPC	2804	3520	2337	828	355
mEF	1178	1746	977	552	217
hESC	11030	32477	11631	5672	479
hNPC	9374	14695	9242	5074	379
hNeurons	1429	7678	5437	2077	164
Dicer-/- mESC	1720	1837	918	658	261

Table S1- Total TSSa-RNA reads and genomic locations by dataset. Genomic locations are defined as regions of DNA that have at least one read. Overlapping reads are collapsed into single locations. Number of reads (no. of Reads) is the number of TSSa-RNAs identified in the data set. Sense reads are reads transcribed from the same DNA strand as the associated mRNA transcript. Anti-sense reads are reads transcribed from the opposite DNA strand as the associated mRNA transcript. Sense & anti-sense reads are reads that are transcribed from the same DNA strand as one associated transcript as well as from the opposite DNA strand of another associated transcript. Together, the Oct4-0h, Oct4-12h, Oct4-24h and mESC v6.5 datasets from (*S1*) are defined as the mESC dataset.

Cell Type	no. of genes	no. of genes with sense reads only	no. of genes with anti-sense reads only	no. of genes with both sense and anti-sense reads	no. of genes with CpG island
All Mouse	10159	3828	2044	4287	8019
mESC	9453	2086	3624	3743	7537
mNPC	2545	1633	731	181	2032
mEF	1273	754	492	27	1090

Table S2- TSSa-RNA associated gene characteristics from each cell type. Number of genes (no. of genes) which have associated TSSa-RNAs for each cell type. Genes with sense reads are genes which have TSSa-RNAs transcribed from the same DNA strand as the associated mRNA transcript. Genes with anti-sense reads are genes which have TSSa-RNAs transcribed from the opposite DNA strand as the associated mRNA transcript. Genes with sense and anti-sense reads are genes that have TSSa-RNAs transcribed from both the same DNA strand as well as the opposite DNA strand as the associated mRNA transcript. Genes with a CpG island are genes that have a CpG island within 1kb of their TSS.

Data Set	A	C	G	U
First nt TSSa-RNA reads	0.33	0.19	0.28	0.20
First nt genes with TSSa-RNA	0.25	0.20	0.47	0.08

Table S3- Fraction of reads or genes that have A, C, G, or U at their 5´ end. P-value=0.0002 (Chi-squared).

Oligo name	Sequence
DNA oligo Cops8	/5Biosg/CCG CTC TCG CCG CCA CCA CCT CTG CCA GAC CAA AGC ACA AGC TGG CCC TC
RNA oligo Cops8	CAG CUU GUG CUU UGG UCU GGC AGA
S6 probe1, Cops8	G+CCC+TCG+GGG+CGT+CGC+GCC+CGC+GA
S6 probe 2, Cops8	T+CTG+CCA+GAC+CAA+AGC+ACA+AGC+TG
S6 probe 3, Cops8	C+CGG+CCG+CTC+TCG+CCG+CCA+CCA+CC
DNA oligo Isg20l1	/5Biosg/ATC GGG ATG TGC TCT TTG AGG CTT TAA GTC TTT GAA GGT TGC GGT TCA CT
RNA oligo Isg201l	ACC GCA ACC UUC AAA GAC UUA
S6 probe 4, Isg20l1	C+ACT+AGG+CGT+CGG+GTC+AGA+A
S6 probe 5, Isg20l1	T+AAG+TCT+TTG+AAG+GTT+GCG+GT
S6 probe 6, Isg20l1	G+CAG+ATC+GGG+ATG+TGC+TCT+T
DNA oligo Rnf12	/5Bio/AGA CGT AGC TCA ATC AGC CAT TAT CTT CCC CAT TGT TAC CTA GCA CTG AC
RNA oligo 1a, Rnf12	UGU CAU UUC CUG UCA GUG CUA GGU AAC AAU GGG GAA GAU AAU GGC UGA UUG
RNA oligo 1b, Rnf12	AGG UAA CAA UGG GGA AGA UAA
S7 probe 1, Rnf12	A+GCA+CTG+ACA+GGA+AAT+GAC+TG
S7 probe 2, Rnf12	T+TAT+CTT+CCC+CAT+TGT+TAC+CT
S7 probe 3, Rnf12	G+AGA+CGT+AGC+TCA+ATC+AGC+CA
DNA oligo Ccdc52	/5Bio/AGC ATG TAT ACA GTT TTC ACC TTA TAC ATT TCG GTC GTC AAT AGC TTG CT
RNA oligo Ccdc52	UUG ACG ACC GAA ATG TAT AAG
S7 probe 4 Ccdc52	T+AGC+TTG+CTG+CTA+TGG+GGT+TG
S7 probe 5, Ccdc52	C+TTA+TAC+ATT+TCG+GTC+GTC+AA
S7 probe 6 Ccdc52	T+AGC+ATG+TAT+ACA+GTT+TTC+AC

Table S4- Oligos used for northern analysis. A "+" before a nucleotide represents an LNA base./5Biosg/ represents a 5´biotin.

Supplemental References

S1. A. Marson *et al.*, *Cell* **134**, 521 (Aug 8, 2008).

S2. L. A. Boyer *et al.*, *Nature* **441**, 349 (May 18, 2006).

S3. H. Niwa, J. Miyazaki, A. G. Smith, *Nat Genet* **24**, 372 (Apr, 2000).

S4. S. Okabe, K. Forsberg-Nilsson, A. C. Spiro, M. Segal, R. D. McKay, *Mech Dev* **59**, 89 (Sep, 1996).

S5. K. L. Tucker, Y. Wang, J. Dausman, R. Jaenisch, *Nucleic Acids Res* **25**, 3745 (Sep 15, 1997).

S6. N. C. Lau, L. P. Lim, E. G. Weinstein, D. P. Bartel, *Science* **294**, 858 (Oct 26, 2001).

S7. R. M. Horton, H. D. Hunt, S. N. Ho, J. K. Pullen, L. R. Pease, *Gene* **77**, 61 (Apr 15, 1989).

S8. J. M. Calabrese, A. C. Seila, G. W. Yeo, P. A. Sharp, *Proc Natl Acad Sci U S A* **104**, 18097 (Nov 13, 2007).

S9. S. Griffiths-Jones, *Nucleic Acids Res* **32**, D109 (Jan 1, 2004).

S10. T. M. Lowe, S. R. Eddy, *Nucleic Acids Res* **25**, 955 (Mar 1, 1997).

S11. P. Schattner, A. N. Brooks, T. M. Lowe, *Nucleic Acids Res* **33**, W686 (Jul 1, 2005).

S12. C. Liu *et al.*, *Nucleic Acids Res* **33**, D112 (Jan 1, 2005).

S13. P. Grozdanov, O. Georgiev, L. Karagyozov, *Genomics* **82**, 637 (Dec, 2003).

S14. http://genome.ucsc.edu/.

S15. F. Hsu *et al.*, *Bioinformatics* **22**, 1036 (May 1, 2006).

S16. K. D. Pruitt, T. Tatusova, D. R. Maglott, *Nucleic Acids Res* **33**, D501 (Jan 1, 2005).

S17. http://www.r-project.org/.

S18. G. W. Yeo *et al.*, *PLoS Comput Biol* **3**, 1951 (Oct, 2007).

S19. Y. Y. Li *et al.*, *PLoS Comput Biol* **2**, e74 (Jul 7, 2006).

S20. T. S. Mikkelsen *et al.*, *Nature* **448**, 553 (Aug 2, 2007).

S21. T. Beissbarth, T. P. Speed, *Bioinformatics* **20**, 1464 (Jun 12, 2004).

S22. M. Gardiner-Garden, M. Frommer, *J Mol Biol* **196**, 261 (Jul 20, 1987).

S23. P. Carninci *et al.*, *Science* **309**, 1559 (Sep 2, 2005).

S24. S. Bashiardes *et al.*, *Nat Methods* **2**, 63 (Jan, 2005).

S25. J. P. Noonan *et al.*, *Science* **314**, 1113 (Nov 17, 2006).

S26. G. S. Pall, C. Codony-Servat, J. Byrne, L. Ritchie, A. Hamilton, *Nucleic Acids Res* **35**, e60 (2007).

S27. A. Valoczi *et al.*, *Nucleic Acids Res* **32**, e175 (2004).

S28. W. S. Rasband, *U. S. National Institutes of Health, Bethesda, Maryland, USA, http://rsb.info.nih.gov/ij/, 1997-2008.* (1997-2008).

S29. N. E. Thompson, T. H. Steinberg, D. B. Aronson, R. R. Burgess, *Journal of Biological Chemistry* **264**, 11511 (JUL 5, 1989).

S30. E. J. Cho, M. S. Kobor, M. Kim, J. Greenblatt, S. Buratowski, *Genes Dev* **15**, 3319 (Dec 15, 2001).

S31. M. Patturajan *et al.*, *J Biol Chem* **273**, 4689 (Feb 20, 1998).

S32. J. C. Jones *et al.*, *J Biol Chem* **279**, 24957 (Jun 11, 2004).

S33. M. G. Guenther, S. S. Levine, L. A. Boyer, R. Jaenisch, R. A. Young, *Cell* **130**, 77 (Jul 13, 2007).

S34. H. Santos-Rosa *et al.*, *Nature* **419**, 407 (Sep 26, 2002).

S35. T. I. Lee *et al.*, *Cell* **125**, 301 (Apr 21, 2006).

S36. T. I. Lee, S. E. Johnstone, R. A. Young, *Nat Protoc* **1**, 729 (2006).

S37. D. S. Johnson, A. Mortazavi, R. M. Myers, B. Wold, *Science* **316**, 1497 (Jun 8, 2007).

S38. D. K. Pokholok *et al.*, *Cell* **122**, 517 (Aug 26, 2005).