Supporting Online Material

Table of Contents: **Page number**

Materials and Methods

Isolation of capped RNA species

Preparation of nuclear RNA from *Drosophila* S2 cells
Drosophila S2 culture cells were grown to a density of 4-5x10^6 cells/ml at 26°C in 175cm^3 flasks (30 ml in each flask) in M3+BPYE media supplemented with 10% heat-inactivated fetal bovine serum (as described by the Drosophila Genome Resource Center). For each preparation, 3 flasks' worth of cells were collected, harvested by centrifugation at 1,000g for 3 minutes, washed with ice-cold PBS, resuspended in 8ml of buffer A (10mM Tris-HCl pH 8.0; 300mM Sucrose; 3mM $CaCl_2$; 2mM $MgCl_2$; 0.1% Triton X-100; 0.5mM DTT), transferred into a 15ml glass dounce homogenizer and subjected to 25-30 strokes with a loose pestle. Three milliliters of buffer B (10mM Tris-HCl pH 8.0; 2M Sucrose; 5mM $MgCl_2$; 0.5mM DTT) was added by mixing with a pipette, and the suspension was laid onto a cushion of 15ml buffer B in a 45ml centrifuge tube and then spun at 12,500 rpm in Sorvall SS-34 rotor for 25 minutes at 4°C. Both upper and lower phases of the supernatant were carefully withdrawn and the nuclear pellet at the bottom of the tube was resuspended in 300µl of buffer A. Ten microliters was used to prepare genomic DNA or to perform permanganate probing. From the rest of the nuclear suspension, total RNA was extracted using the Trizol reagent (Invitrogen) according to the manufacturer's protocol. The total yield of RNA from ~80ml of S2 cells was 100-150µg.

Preparation of nuclear RNA from *Drosophila* embryos.
Embryos were collected on grape agar plates from 5 grams of OregonR wild type adults for 16 hours at 25°C, dechorionated in 50% bleach for 90 seconds and rinsed thoroughly with distilled water. Embryos (0.5ml) were transferred to a pre-chilled 2ml Potter-Elvejham tissue grinder and rinsed three times with 5 volumes of ENIB buffer (15mM Hepes pH7.6; 10mM KCl; 3mM $CaCl_2$; 2mM $MgCl_2$; 0.1% Triton X-100; 1mM DTT; 1mM PMSF) by gentle vortexing and allowing embryos to settle by gravity. Embryos were then resuspended in 5 volumes of ENIB-0.3 buffer (ENIB with 300mM sucrose) and disrupted with 10 strokes of the pestle. The lysate was transferred to a 2ml Dounce mortar and homogenized with 10 strokes of the tight-fitting pestle. The lysate was then passed through a 50-micron nylon cell strainer to remove debris. The filtrate was gently layered over a two step cushion consisting of an equal part ENIB-1.7 buffer (ENIB with 1.7M sucrose) at the bottom and an equal part ENIB-0.8 buffer (ENIB with 0.8M sucrose), and centrifuged at 15,000xg for 10 minutes at 4^{o}C on an Eppendorf 5804R. The supernatant was withdrawn, the remaining nuclear pellet was resuspended in 0.5ml of ENIB-0.3 buffer, and the nuclei were purified on a second, scaled down two-step cushion as described above. This second nuclear pellet was resuspended in 100µl of ENIB-0.3 buffer and total RNA was extracted with 1ml of Trizol according to the manufacturer's protocol. From 0.5ml embryos, 300µg nuclear RNA was obtained on average.

Preparation of short capped RNA libraries.
The procedure is based on the protocol from the Illumina small RNA cloning kit, with modifications (outlined in fig. S1). In brief, to increase the yield of target RNAs, we

degraded uncapped RNAs, such as rRNAs, with 5’ monophosphate-dependent Terminator exonuclease. To convert 5’-triphosphates on these RNAs to monophosphates, we treated the RNA preparation with 5’ Polyphosphatase. We avoided the use of Alkaline Phosphatase and T4 polynucleotide kinase before ligating the 3’ adapter because these enzymes also remove phosphates from RNA 3’-ends and allow ligation of RNA degradation products (see fig. S1B). Five to ten microgram of nuclear RNA (10µl) was mixed with an equal volume of formamide loading buffer, heated for 5 minutes at 65°C, rapidly placed on ice and loaded on a 15% urea-TBE 10-well 1mm thick gel (Novex) that had been pre-run for 15-30 minutes at 200V. The gel was run at 200V for one hour and stained with ethidium bromide. A section of the gel containing RNAs sized between 25 and 120 bases, as judged by an RNA ladder, was excised and crushed by spinning for 2 minutes at maximum speed through a 0.5ml microcentrifuge tube that was pierced several times with a 22 gauge needle and inserted into a 2ml collection tube. Four hundred microliters of 300mM NaCl was added to the gel slurry and samples were incubated at room temperature for 4 hours with gentle shaking. The sample was then transferred into a 0.22µm spin filter column supplied with the kit and spun for 2 minutes in a microcentrifuge. One microliter of glycoblue precipitant (Ambion) and 2.5 volumes of ethanol were added to the eluate and RNA was allowed to precipitate at -80°C for at least 30 minutes. After centrifugation for 20 minutes at 4°C in a microcentrifuge, the pellet was washed with 70% ethanol, allowed to air dry, and resuspended in 17µl of water. Two microliters of 10x buffer for RNA 5’ polyphosphatase and 1µl of RNA polyphosphatase (Epibio) were added and reaction was incubated for 30 minutes at 37°C. The reaction was stopped by the addition of 100µl of Phenol/Chloroform/Isoamyl alcohol mix and 80µl of TE pH 7.0, extracted with an equal volume of chloroform, and RNA in the water phase was precipitated by the addition of 1/10 volume of 7.5M ammonium acetate and 2.5 volumes of ethanol as described above. The precipitated RNA was resuspended in 17µl of water, to which 2µl of 10x buffer for 5’ Terminator nuclease and 1µl of Terminator exonuclease (Epibio) were added and the reaction was incubated at 30°C for 1 hour, after which it was phenol/chloroform, chloroform extracted and ethanol precipitated as above and resuspended in 6.4µl of water. To the reaction, 0.6 µl of 3’ RNA adapter, 1µl of 10x buffer for RNA ligase, 1µl of RNAse inhibitor, and 1µl of ssRNA ligase 1 were added and the reaction was incubated for 6 hours at 20°C. For 5’-end sequencing, RNA adapters from the Illumina kit were used. For 3’-end sequencing, reciprocally complementary RNA adapters and the appropriate DNA RT primer were made. After incubation, 10µl of formamide RNA loading buffer was added, the reaction was heated for 5 minutes at 65°C, immediately placed on ice and loaded on a 15% urea-TBE gel as above. A section of the gel corresponding to RNAs 47 to 145 bp (for 5’-end sequencing libraries) and 51 to 150 (for 3’-end sequencing libraries) was excised and RNA was eluted from the gel and precipitated as above. RNA was resuspended in 17µl of water, to which 2µl of 10X buffer for Alkaline phosphatase and 1µl Heat Labile Alkaline phosphatase (Epibio) were added, and incubated for 10 minutes at 37°C. The reaction was extracted with phenol/chloroform and chloroform, ethanol-precipitated as above, and resuspended in 44µl of water. Five microliters of 10x buffer for Tobacco Acid Pyrophosphatase and 1µl (2.5 units) of Tobacco Acid Pyrophosphatase (Epibio) were added, the reaction was incubated for 2 hours at 37°C, phenol-extracted and ethanol precipitated as above and resuspended in 5.7µl of water. 1.3 microliters of the appropriate

5’ RNA adapter was added and the reaction was treated with ssRNA ligase as above and run on a 10% urea-TBE 10-well 1mm thick gel (Novex). A section of the gel corresponding to RNA sizes between 70 and 170bp was excised, RNA was extracted from the gel as above and resuspended in 4.5µl of water. Reverse transcription (with the appropriate RT primer), PCR amplification, and separation on a 6% TBE 1mm thick gel (Novex) were performed as specified in the Illumina small RNA kit. A section of the gel corresponding to DNA sizes between 90 and 190 bp was excised and eluted in 400µl of gel elution buffer. DNA was precipitated with ethanol, resuspended in 100µl of water and purified on Qiagen MinElute column, eluted in 12µl of EB (Qiagen). Concentration of the final library was measured on the Nanodrop spectrophotometer, and quality of the library was confirmed by cloning 1µl of the library into pBlueScript vector and sequencing of at least 10 resultant clones.

Preparation of long capped RNA libraries
RNA obtained from nuclei was subjected to two rounds of purification using Qiagen RNeasy columns to deplete small RNA species. Five microgram of the remaining RNA was treated, sequentially, with 5’ polyphosphatase and 5’ terminator exonuclease, with reaction conditions and phenol/chloroform extraction performed as described above. RNA was resuspended in 75µl of water, placed on ice, and hydrolyzed by the addition of 25µl of 1M NaOH and incubation on ice for 60 minutes. Reaction was stopped with equal volume of Tris-HCl, pH 6.8 and RNA was precipitated with ethanol in the presence of glycoblue coprecipitant as described above. RNA collected by centrifugation was separated on a 15% urea-TBE gel. Section of the gel containing RNA fragments between 70-200nt in size, as judged by low range single-stranded RNA ladder (NEB), was extracted as described above and resuspended in 17µl of water. Two microliters of 10x buffer for T4 DNA ligase with 10mM ATP, and 0.5µl (10 units) of T4 polynucleotide kinase were added and reaction was incubated at 37°C for 30 minutes. Reaction was passed through a Bio-Rad P-30 spin column to remove free ATP, then phenol-extracted and ethanol-precipitated, and resuspended in 17µl of water. RNA was then treated, sequentially, with 5’ terminator exonuclease, alkaline phosphatase, and tobacco acid pyrophosphatase, with phenol-extraction and ethanol-precipitation after each enzyme treatment. The resultant RNA was ligated to the 5’ small RNA linker (Illumina) and resolved on a 15% urea-TBE gel. Section of the gel corresponding to RNA sizes between 90 and 220nt was excised from the gel and RNA was extracted as described above. RNA was then ligated to the 3’ small RNA linker (Illumina), separated on a 10% urea-TBE gel and RNA species between 110 and 240nt were extracted from the gel. Reverse transcription, PCR-amplification, and separation on a 6% TBE gel were carried out as specified in the Illumina small RNA kit. DNA species between 130 and 250 bp were extracted from the gel and the library was treated as described above.

Permanganate probing of nuclei

One hundred microliters of ice-cold solution containing 10mM $KMnO_4$ in 1xPBS was added to 10µl of nuclei suspension. After 60s of incubation on ice, reactions were stopped by the addition of an equal volume of stop solution (10mM Tris-HCl, pH 7.5; 10mM NaCl; 20mM EDTA; 0.5% SDS; 0.4M 2-mercaptoethanol). Samples were

incubated with 10µl of RNAse cocktail (Ambion) for 15 minutes at 37°C, after which Proteinase K was added to 0.2mg/ml and reactions were incubated at 56°C overnight. DNA was extracted once with a mixture of phenol/chloroform/isoamyl alcohol, 1/10 volume of 3M sodium acetate was added and DNA was precipitated with an equal volume of 2-propanol. The pellet was washed with 70% ethanol and resuspended in 150µl of TE buffer (pH 8.0). One microgram of permanganate-treated DNA was incubated in 100µl of 10% (v/v) piperidine for 15 minutes at 90°C. Reactions were extracted twice with an equal volume of chloroform, and DNA was ethanol-precipitated as above and resuspended in 100µl of water. Ten microliters (0.1 µg) of piperidine-treated DNA was used for each ligation-mediated PCR footprinting reaction, which were performed essentially as described (*1*). Sequences of primers for probing each of the genes listed are available upon request.

Preparation of Pol II ChIP material for ChIP-seq

ChIP material was prepared and immunoprecipitations carried out with the anti- Pol II (Rpb3 subunit) antibody as described in (*2*). The immunoprecipitated material was purified on a Qiagen column using the Qiaquick PCR purification kit and ChIP-seq libraries were prepared using the Illumina ChIP-seq kit according to the manufacturer's instructions. The resulting library was sequenced using two lanes on a Solexa Genome Analyzer to achieve appropriate sequence coverage (Table S1).

Data analysis

Sequencing and mapping of RNAs
For each condition, libraries were made and sequenced using the Illumina (Solexa) Genome Analyzer II (G2) system with standard sequencing protocols in the laboratory of Dr. Yuan Gao at Virgina Commonwealth University. Raw sequences were trimmed to 26 nucleotides from the 3'-end and aligned against the *Drosophila melanogaster* genome dm3 downloaded from UCSC (http://hgdownload.cse.ucsc.edu/goldenPath/dm3/bigZips/). 26-mers were aligned using the short read mapping software MOM: Maximum Oligonucleotide Mapping tool (http://go.vcu.edu/mom) with a maximum allowed mismatch of 2, otherwise under default conditions (*3*). The yield of uniquely mappable reads for each set of samples is listed (Table S1). The genomic location and strand of mapped reads were compiled using custom scripts and visually examined using the UCSC genome browser in .bedgraph format. RNA sequences came from discrete locations and were not binned for further analyses, allowing the hits to be examined at nucleotide resolution.

Mapping RNAs to genes
To determine the genes from which uniquely mapped sequences were derived, we generated a list of genome elements from the *D. melanogaster* build r5.17 (April 2009) .gff genome file downloaded from Flybase (ftp://ftp.flybase.net/genomes/Drosophila_melanogaster/dmel_r5.17_FB2009_04/gff/). This list included all Pol II-derived transcripts (mRNA, snoRNA, snRNA, ncRNA) except for miRNA, as the location of miRNAs in the genome build denotes the location

of the mature miRNA product rather than the transcription start site, resulting in a total of 22,202 elements. To ensure that we had a non-redundant search space, we compressed multiple isoforms of the same gene that share a TSS, or have start sites within 50 bp of each other, maintaining the most upstream TSS (longest gene) for further analyses. Additionally, we removed genes on the same strand that overlapped in the window +/- 50 bp around the TSS to avoid counting any sequence read in two such windows. We then created a search space around each of these 17,109 unique TSSs of +/-50 bp and counted the number of strand reads from each 5'-library that fell within this 101 nt region and was derived from transcription in the sense direction. We note that we detected extremely low levels of RNAs mapping in the antisense direction around annotated TSSs (~100-fold lower levels than in the sense direction, when mapping within ~500 bp), indicating that the bidirectional transcripts observed recently in mammalian cells (*4*, *5*) are either less abundant in *Drosophila*, or are not capped and stable enough for isolation using our protocol.

To determine the number of reads that was significant from each data set, we used the Fisher's exact test implemented in R by comparing the 17,109 TSS regions to 10,844 101 nt intergenic regions extracted from intergenic gaps of at least 1,000 nt. Within these gaps, individual control regions were at least 100 nt from the start or end of an annotated gene. The p-values from the Fisher's exact tests were corrected using the Benjamini-Hochberg method to control for false discovery, and a p-value <0.005 was deemed significant. For short RNAs derived from S2 cells, the TSS considered to be significantly represented in the two biological replicates were very similar, so both libraries were combined and this process was repeated. This yielded 7,428 genes that produced significant RNAs in the combined 5'-end libraries. A similar analysis performed on the combined 5'-end libraries derived from long (non size-selected) RNAs isolated from S2 cells indicated that 6778 TSSs were significantly represented.

When the same analysis was performed on the 5'-end short RNA samples derived from 0-16 hour Drosophila embryos, 8,693 genes were found to generate significant levels of short RNAs; including 1565 genes that were not deemed significant in either the 5'-end or 3'-end libraries from S2 cells (see fig. S10).

The process for evaluating 3'-end reads was identical, except that the search space was from the TSS to +100 at each of the 15,629 genes that do not overlap with other genes within the search space from -50 to +100 from the TSS. Using this method, a total of 7,816 genes were deemed significant after Benjamini-Hochberg FDR correction. Genes identified as producing significant short RNAs from the 5'-end and 3'-end libraries are very similar (fig. S3). Additionally, the number of reads from the 5'-end and 3'-end libraries for a given gene was highly correlated (Spearmans ρ=0.76).

Definition of observed TSS and metagene analyses of 3' reads

For the 7,428 genes with a significant number of 5'-end reads, we selected the location with the most sense strand hits and defined that as the experimentally observed TSS (Table S2). If two or more locations had the same number of reads we used the position closest to the Flybase annotated TSS.

Comparison of the quality of match to the initiator element around the annotated vs. observed TSS using WebLogo (weblogo.berkeley.edu) confirmed that the observed TSS was a better indicator of where transcription initiates in our *Drosophila* S2 cells.

To confirm that the TSSs observed from short RNA species also represents the TSSs of full-length mRNA, we sequenced the 5'-ends of long (non size-selected) RNA species as well. The long RNAs mapped to a gene set that was largely overlapping with the gene set derived from short RNAs (e.g. 86% of genes with significant short RNAs also produced significant long nuclear RNAs). The TSSs defined by short and long RNAs agreed extremely well, indicating that transcription starts from the same location whether or not the Pol II undergoes efficient elongation (fig. S7). Thus, the observed TSS was used for subsequent metagene analyses.
For the analyses involving 3'-end libraries or the first 100 nt of the transcribed region (such as determination of melting temperature), we used only the 15,629 genes that did not overlap with other genes in the region between -50 and +100 from the TSS. This included 6,496 of the 7,428 genes with a significant number of 5'-end reads.

Comparison of short RNA reads to mRNA expression analyses
GCRMA normalized expression data from our previous work (*2*) were used in these analyses. Of the 7,428 genes with significant numbers of 5'-end reads, 5,515 had available mRNA expression data (4,985 are active, with Log_2 expression levels>2.5; 530 are inactive, with Log_2 values<=2.5), allowing for comparison of the number of short RNAs to steady state levels of mature transcripts as shown in fig. S5. The 5,515 genes were divided into quartiles of expression and we conducted metagene analyses of RNA 5'-ends for each quartile as shown in fig. S13A. For 5'-end analyses, the quartiles each contained 1,379 genes, except for the top quartile that had 1,378 genes. Of the 5,515 genes with significant short RNAs and expression data, 506 overlapped with other transcription units within the range used for 3'-end metagene analyses (extending to +100, as described above) and were therefore excluded when calculating the 3'-end distribution. This resulted in 1,297 genes in the bottom quartile, 1,232 in the second quartile, 1,253 in the third, and 1,227 in the top expression quartile, as shown in fig. S13B.

Approximation of the stability of the RNA-DNA hybrid
To assess the stability of the 9 nt RNA-DNA hybrid at various positions within the initially transcribed region (from +1 to +100), we calculated the Tm of all 9 nt DNA-DNA hybrids throughout this interval as a proxy for the RNA-DNA hybrid. We extracted genomic sequences from the annotated TSS to position +100 for the 15,629 non-overlapping genes detailed above. We then calculated the Tm using EMBOSS dan (*6*) for each possible 9 nt hybrid, plotting this as the 3'-end moves from +9 to +100 at 1 nt intervals (1-9, 2-10, …, 92-100) (Fig. 3B, and fig. S9).
To determine the Tm around the position where the 3'-end is located, we selected 434 genes that both had at least one hundred 3'-end library reads from S2 cell libraries where at least 20% of those reads were derived from a single location. The sequences surrounding these 3'-ends were extracted (+/- 35 nt) and the melting temperatures were determined as above (Fig. 3C).
Analysis of the sequences surrounding the 3'-end of RNAs derived from 0-16h *Drosophila* embryos was also performed to determine the Tm profile around this independent group of genes. For this analysis, we selected 338 genes that had at least

fifty reads from embryo 3'-end libraries where at least 20% of those reads were derived from a single location (fig. S10B).
Given that the 3'-end of the RNA falls within a region of high hybrid stability, surrounded by sequences with lower Tm, we evaluated the nucleotide distribution surrounding the defined 3'-ends. A position weight matrix generated using WebLogo (weblogo.berkeley.edu) revealed that the region directly surrounding the 3'-end has a higher GC-content than regions either upstream or downstream (fig. S9C).

Determination of RNA lengths in IIS-depleted vs. mock-treated samples
The positions of all 3'-ends were determined for the 6,496 genes that generate significant RNA reads and have a unique search space from the TSS to +100 (same as used for all other 3'-end analyses) in IIS-depleted and mock-treated samples (depletion performed as in (*7*). The Mann Whitney U test (a.k.a. Wilcoxon rank sum test) was used to compare the positions of 3'-ends between the two samples, revealing that IIS-depletion lead to a statistically significant lengthening in the RNAs ($p=2.2x10^{-16}$). The average length of the short RNAs for mock-treated and IIS-depleted samples was defined as the sum of the lengths of all 3'-end reads mapping 20 to 65 nucleotides downstream of the observed TSS divided by the number of such reads. The length of a read in these analyses was defined as the distance from the start of the read (which indicates the 3'-end of RNA) to the observed TSS.

Comparison of short RNA data to previously defined stalled genes
Of the 5,403 genes defined as bound by Pol II in our previous study (*2*), 4,788 genes are retained in the updated genome build and have unique search spaces surrounding their TSSs (+/- 50 bp). This includes 852 stalled genes and 3,936 "not stalled" genes that were analyzed for Fig. 1C. Statistical evaluation of the number of short RNAs derived from these two groups of genes using the Mann-Whitney U test demonstrated that they are significantly different ($P<0.001$). Of the 852 stalled genes, 793 (~93%) produced a statistically significant number of short RNAs within +/- 50 bp from the annotated TSS (median of 1477 reads/TSS); of the remaining 58 genes that do not produce a significant number of RNAs that map within 50 bp of the annotated TSS, 35 have statistically significant number of short RNAs on the sense strand that originate within 150 bp of the annotated TSS. Of the 3,936 genes previously defined as bound by Pol II but not stalled, 3,350 (85%) generate significant levels of short RNA (median of 178 reads/TSS).
The metagene analysis shown for 5'-end reads shown in fig. S4A include these 793 stalled and 3,350 not stalled genes with significant short RNAs. After further filtering for genes that are unique within the region from -50 bp to +100 used for 3'-end metagene analysis, 754 stalled and 2,967 not stalled genes are shown in fig. S4B. Finally, of the above genes that have a significant number of 3'-end short RNAs within the region +20 to +65 (>10 reads), 745 stalled genes and 2,904 genes that did not appear to contain stalled Pol II were maintained for analysis in fig. S12.

Pol II ChIP-seq alignment and calculation of Stalling index
ChIP-seq reads were aligned using the same parameters as for the short RNAs. The yield of uniquely mappable reads for each of the technical replicates is listed (Table S1). ChIP-seq hit locations were centered by moving the forward and reverse stand hit location 65

nt in the appropriate direction because the average fragment size was ~130 bp (as per Illumina protocol). Because the hit distribution was less discrete than the short RNAs reads, ChIP-seq data were placed in 25 nt bins for visualization using the UCSC genome browser.

To calculate the Stalling Index, we determined the number of ChIP-seq reads that mapped within a window from +/- 150 bp around the annotated TSS for each gene. This analysis was only performed on genes for which these windows did not overlap with other genes, yielding 13,705 unique genes for which we could calculate the Pol II ChIP-seq signal near the promoter (shown in fig. S4C, left panel). Then, for those genes that had significant Pol II ChIP signal near the promoter (p-value <0.005; determined using FDR corrected Fisher's exact test as described above), were longer than 800 nt, and wherein the region from +500 to +1,500 (or the gene end) did not overlap with other gene windows (from - 150 bp to +1,500 from the annotated TSS), we determined the number of reads that mapped in the window from +500 to +1,500 bp downstream of each gene (or from +500 to the gene end for genes <1,500 bp long). We then calculated the read density in each "promoter" and "downstream gene" window by dividing the number of reads by the window size, and used these numbers to generate a Stalling Index for the 4,046 genes shown in fig. S4C, right panel, and fig. S9B.

Supplementary figures S1-S13

Figure S1

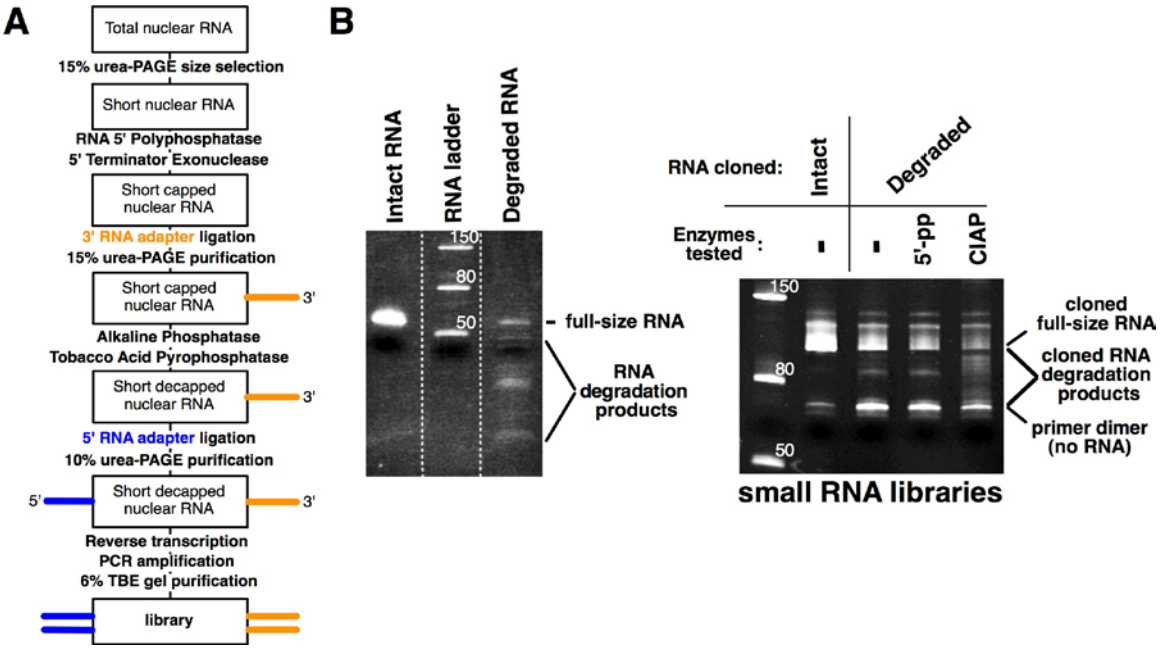

Figure S1. Description of protocol for generating RNA libraries. (A) General scheme for the preparation of short RNA libraries. Experimental steps as described in the Supplementary Methods are shown in bold and significant products of each step are enclosed in squares. The 5' and 3' RNA adapters are highlighted by color. (B) Our strategy favors cloning of intact RNA transcripts. Unlike intact RNA transcripts, which contain a hydroxyl group at their 3'-ends, products of RNAse or chemical RNA degradation contain a 3'-phosphate and will be cloned only if this phosphate is removed. *Left panel:* a 49-nt RNA fragment of *Arabidopsis thaliana* cab gene was synthesized *in vitro* using T7 megascript kit (Ambion), capped at its 5'-end using ScriptCap m7G capping kit (Epibio) and partially degraded using RNAse A+T1 cocktail (Ambion). *Right panel*: intact or RNAse-treated RNA fragments were used to prepare short RNA libraries as in A, except that instead of RNA 5' polyphosphatase at the step indicated in A the RNA was treated with no enzyme (-), RNA 5' polyphosphatase (5'-pp), or alkaline phosphatase+T4 polynucleotide kinase (CIAP). Alkaline phosphatase + T4 polynucleotide kinase remove the 3' phosphate and, if used prior to ligating the 3' RNA adapter, allow cloning of RNA degradation products. In contrast, 5' polyphosphatase does not remove the 3' phosphate and therefore favors cloning of intact RNA transcripts.

Figure S2

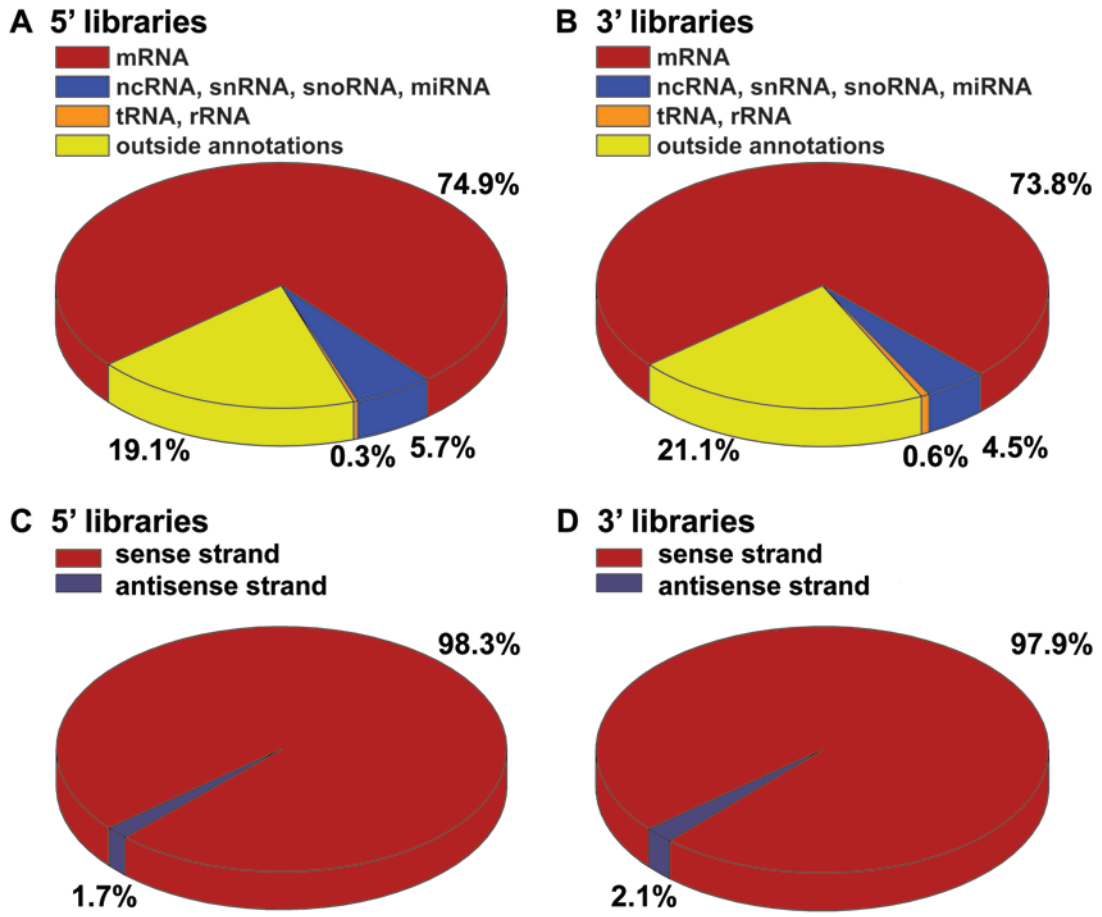

Figure S2. Solexa sequencing of short RNA libraries reveals consistent enrichment near gene promoters. Short capped RNAs were isolated using two independent biological replicates from *Drosophila* S2 cell nuclei and prepared for 5'- and 3'-end sequencing (see Table S1). The percentages of reads sequenced from the 5'-end (A) and 3'-end (B) libraries that map within 200 bp of the annotated transcription start sites for features in the *Drosophila* Genome (build 5.17) are shown. Approximately three quarters of the mappable reads fall within 200 bp of the annotated TSS for mRNA species. The percentage of reads that fall near annotated TSS (+/- 500 bp) that map to the sense vs. the antisense strand are shown for 5'-end (C) and 3'-end (D) reads. This analysis was performed on those promoters that do not overlap with other transcription units within the window +/- 500 bp, and indicates that the levels of short capped transcripts that arise from divergent or antisense transcription are low in *Drosophila* S2 cells.

Figure S3

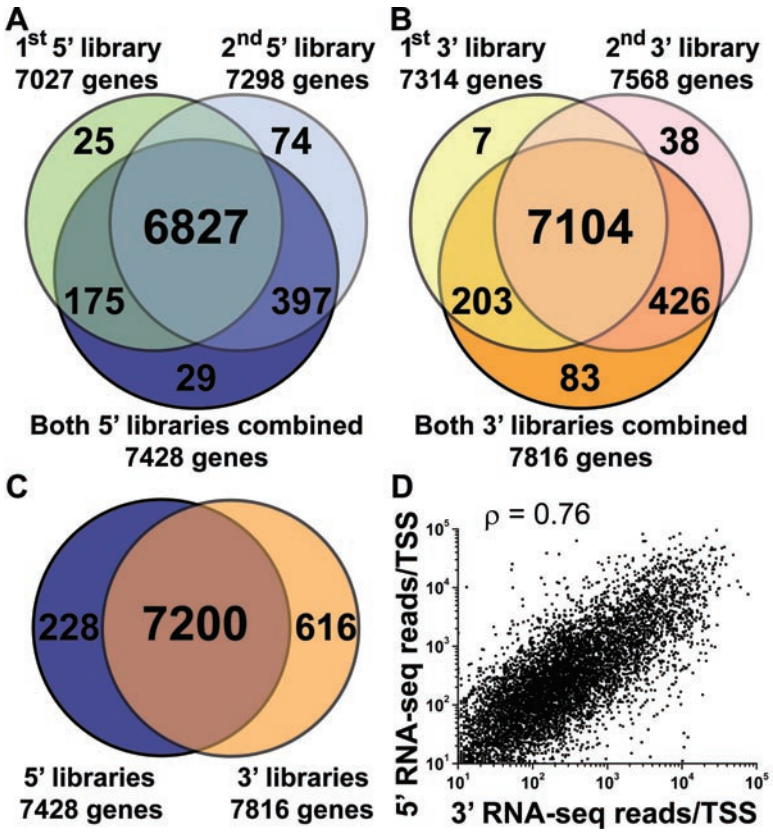

Figure S3. Short RNAs are detected from the same set of genes consistently. Genes that produce a significant number of short RNAs that were present in the 5'-end (A) or 3'-end (B) libraries are identified using the Fisher's exact test as described in Materials and Methods. Genes identified in this way are shown for the two biological replicates. Given the excellent correspondence between the two samples for each type of library, we combined the reads and repeated this analysis to generate a final list of 7,428 genes that have significant 5'-end reads and 7,816 genes with significant 3'-end reads. The overlap between these gene lists was high (C) and the overall correlation between the number of 5'- and 3'-reads arising from each gene was significant (D; Spearman's ρ=0.76).

Figure S4

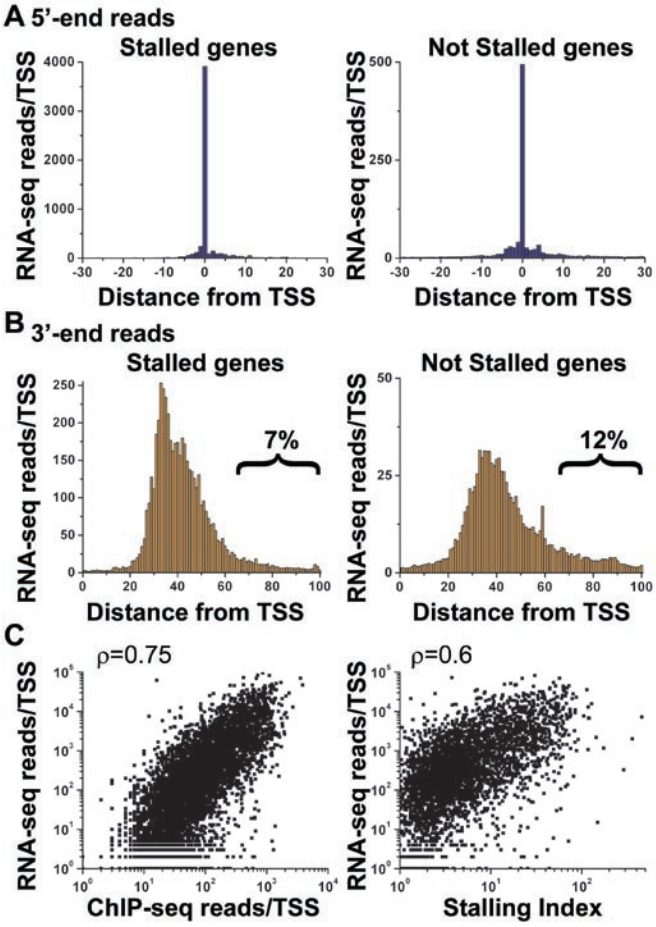

Figure S4. Short RNAs are produced by both stalled and non-stalled genes. Distribution of 5'-end reads (A) and 3'-end reads (B) around genes that were previously defined as bound by polymerase that was stalled (left panels) or not stalled (right panels) (*2*). We note that while the number of 5'-end reads is significantly greater at the stalled genes, the distribution of reads around the TSSs is similar. Likewise, there are more 3'-end reads/TSS at the stalled genes than the non-stalled genes. Interestingly, the non-stalled genes show a significantly greater percentage of reads mapping downstream of +65, indicating that these RNA species are being elongated. (C) Scatter plots showing the relationship between the number of 5'-end short RNA reads/TSS and the Pol II ChIP-seq signal (left panel), or Stalling Index (right panel). Pol II ChIP-seq signal was calculated as the number of reads that mapped within 150 bp of the annotated TSS for the 13,705 promoters for which this region did not overlap with neighboring genes. Stalling Index was calculated as described in Supplementary Methods, and reflects the density of Pol II reads at a gene promoter divided by the read density in the downstream region. Previous work has shown that a high Stalling Index is a good indicator of Pol II stalling (*2*). This analysis is shown for the 4,046 genes that were unique (non-overlapping) within the region -150 to +1,500 from the TSS. The correlation between these values, calculated as Spearman's ρ, is shown.

Figure S5

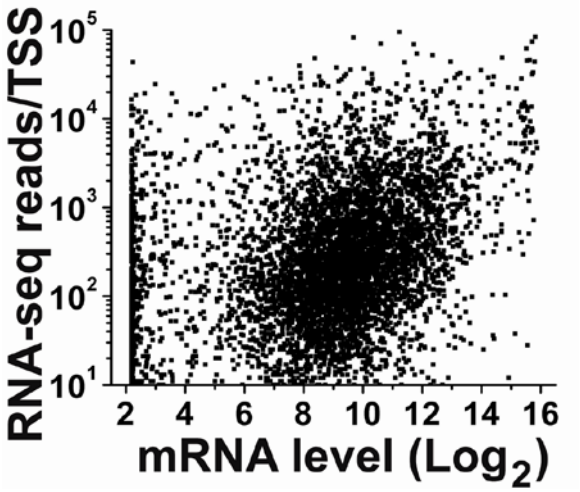

Figure S5. The relationship between the number of short RNA reads and gene expression levels. Of the 7,428 TSSs that produce significant numbers of short RNAs, we have expression data for 5,515 of these. Shown is the number of short RNAs detected per TSS plotted against the GCRMA normalized expression level of that gene, as determined by our previous microarray expression analysis (*2*). There is only a weak correlation between these values (Spearman's ρ=0.35).

Figure S6

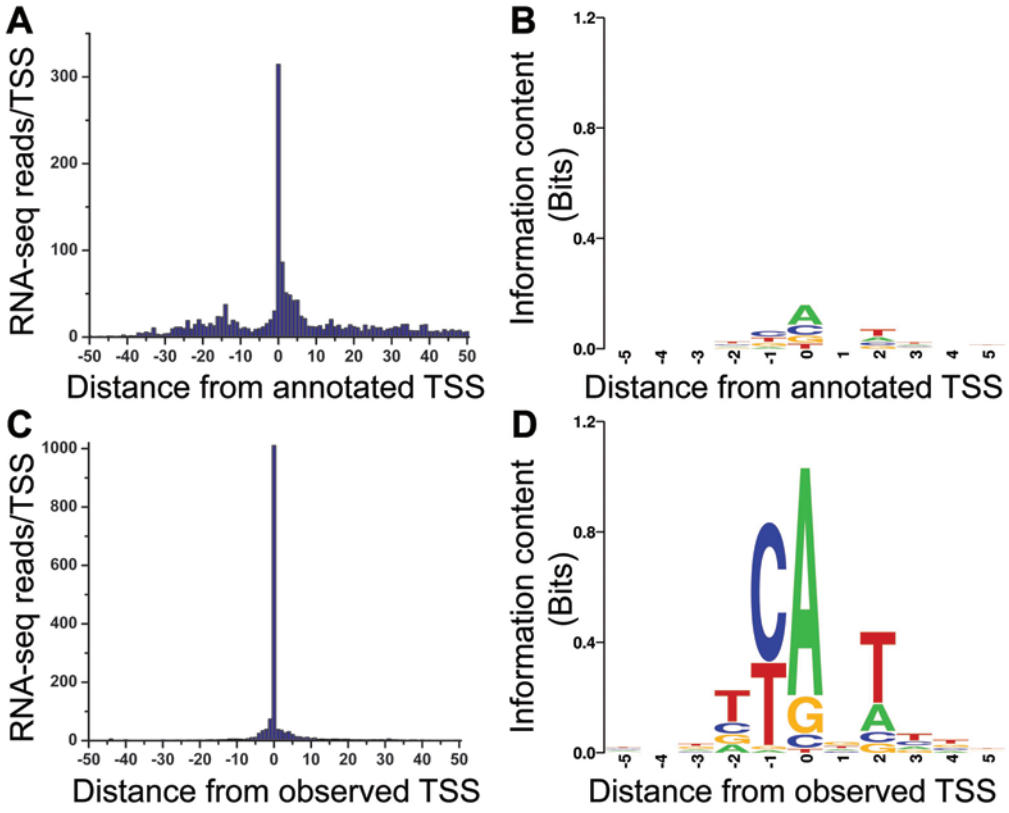

Figure S6. The 5'-ends of short RNAs accurately reflect start sites of transcription, which are highly focused around the Initiator sequence. Sequences from the combined 5'-end libraries were aligned with respect to annotated TSSs. Metagene analysis was performed on the 7,428 genes with a significant number of 5'-end reads, shown in (A) as the number of reads per gene originating at each nucleotide position from +/- 50 relative to the TSS. This distribution reveals that many reads do not map precisely to the annotated TSS, indicating dispersed initiation or differences between annotated TSS and sites of transcription initiation in S2 cells. **(B)** Evaluation of the sequence surrounding these 7,428 annotated promoters found a very weak position-specific similarity to the highly conserved *Drosophila* Initiator element, which has a consensus sequence TCAKTY where A represents the TSS (*8*). We then defined the observed TSS for each of these genes as the position within the interval +/- 50 bp from the annotated TSS from which the most sense strand reads arise. The metagene analysis was then repeated (C), aligning the 5'-end reads with respect to the observed TSSs for these 7,428 genes. This analysis reveals a significant enrichment in the signal at the +1 position, and much less evidence of dispersed initiation, consistent with the fact that *Drosophila* shows largely focused transcription initiation events (*8*). Using the observed TSSs also significantly increases the match to the Initiator element (D).

Figure S7

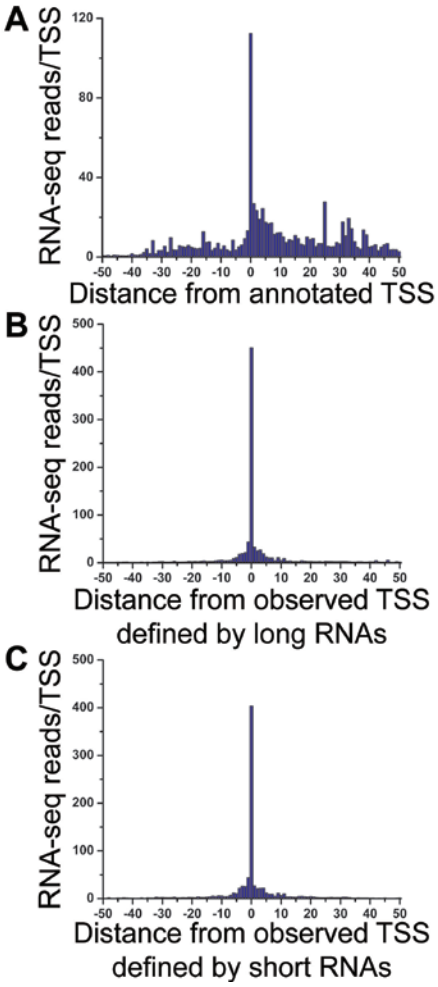

Figure S7. The 5'-ends of long RNAs agree well with the observed TSSs defined using short RNAs, but not with annotated TSSs. Sequences from the combined 5'-end libraries derived from long (>100nt) capped RNA were aligned with respect to annotated TSSs. Metagene analysis was performed on the 6,778 genes with a significant number of 5'-end reads, shown in (A) as the number of reads per gene originating at each nucleotide position. This distribution reveals that many reads do not map to the annotated TSS, consistent with our finding for short RNAs. We then defined the observed TSS for each of these genes as the position within the interval +/- 50 bp from the annotated TSS from which the most sense strand reads arise. The metagene analysis was then repeated (B), aligning the 5'-end reads with respect to the observed TSSs. This analysis reveals a significant enrichment in the signal at the +1 position, and much less evidence of dispersed initiation, again consistent with the data from short RNAs. (C) To determine whether the same TSSs were used to generate short and long RNAs, we mapped the long RNAs against the observed TSSs derived from our short RNA reads (Table S2). This analysis demonstrates that short and long RNAs initiate from the same TSS. Consistent with this finding, the observed TSSs defined by long RNAs agreed precisely (same nt position) with those derived from analysis of short RNAs >50% of the time, while long RNA species mapped to the annotated TSS at only 12% of genes.

Figure S8

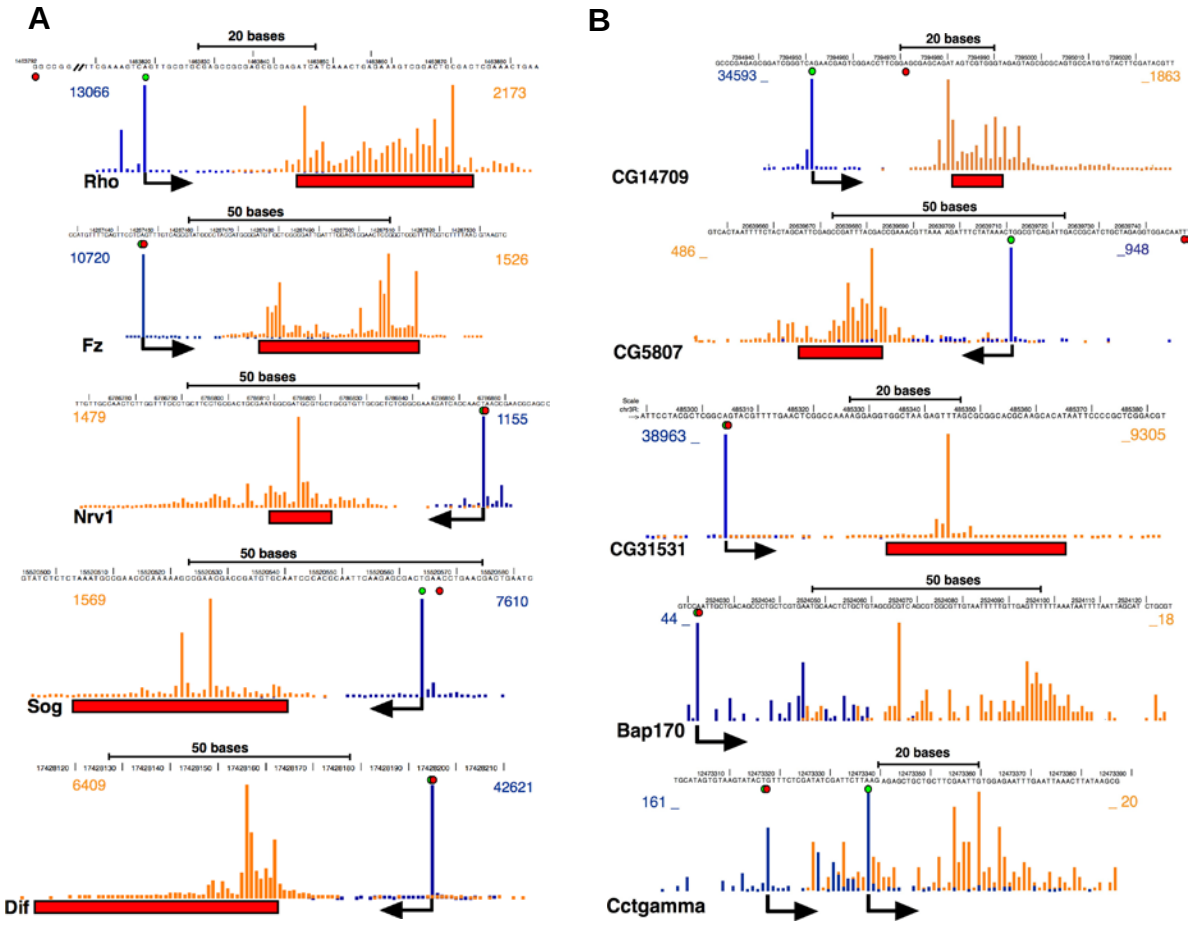

Figure S8. Location of 3'-end reads corresponds well to previously published permanganate footprints. (A) Distribution of RNA reads for the five genes that were analyzed by permanganate probing in (S2). Their direction of transcription is shown with arrows and genomic locations are indicated above the graphs. 5'-end reads are shown in blue and 3'-end reads in orange. The number of reads at the peak location in each library is shown with the corresponding color. Regions of permanganate reactivity noted in previous work are indicated with red rectangles. (B) Distribution of RNA reads for five genes that were analyzed by permanganate probing in (*9*). Permanganate reactivity had been detected in the top three genes – these genes show higher numbers of RNA reads. Permanganate reactivity had not been noted in the bottom two genes, which show much smaller - yet detectable - numbers of RNA reads. Red and green circles indicate UCSC gene bank annotated TSSs and our observed TSSs, respectively. We note that our observed TSSs agree remarkably well with those independently identified for the genes shown on the left in (*9*).

Figure S9

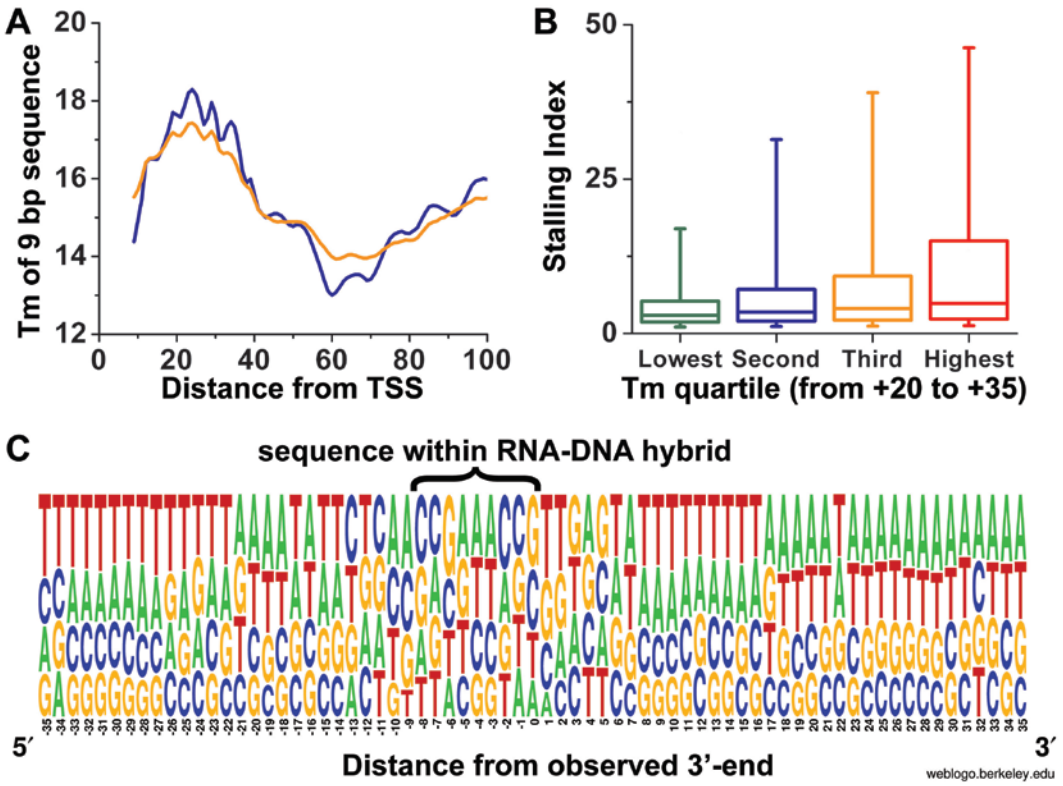

Figure S9. The 3’-ends of short RNAs map to a region of locally high hybrid stability surrounded by AT-rich sequences. Genes that generate significant levels of short RNAs have a distinct Tm profile, with a peak between +20 and +35 with respect to the TSS, followed by a decline (Fig. 3B). We note that the Pause Button was detected at ~1/3 of these genes (2,008 of 6,496 genes; P<0.002) in agreement with prior reports; however, no novel motifs were observed in our analyses (*10*). (A) While the characteristic Tm profile is most striking when genes are aligned with respect to the observed TSSs defined by short RNAs (shown in blue), it is also apparent when aligning sequences with respect to the annotated TSSs (shown in orange and in Fig. 3B). (B) Plotting the relationship between Stalling Index and average Tm of 9-bp sequences between +20 and +35 demonstrates that genes with a higher Tm in the promoter-proximal region are more likely to possess stalled Pol II (P<0.001; Kruskal-Wallis test for each pairwise comparison except second vs. third quartile P<0.05). The average Tm was calculated for all 4,046 genes shown in fig. S4C (right panel) and this gene list divided into quartiles. Box plots depict the 25th, 50th and 75th percentiles and the whiskers show the range between 5th and 95th percent. (C) Weblogo (weblogo.berkeley.edu) was used to create a position weight matrix for the sequences analyzed in Fig. 3C in the ~70 nt region centered around the observed 3’-end. The size of each letter denotes its frequency at each location with respect to the 3’-end of the RNA. The 9-nt of sequence that would form the RNA-DNA hybrid within the polymerase is shown by a bracket. Consistent with the high Tm of the sequence encompassed by the bracket (Fig. 3C), this region is significantly more GC rich than the surrounding sequences.

Figure S10

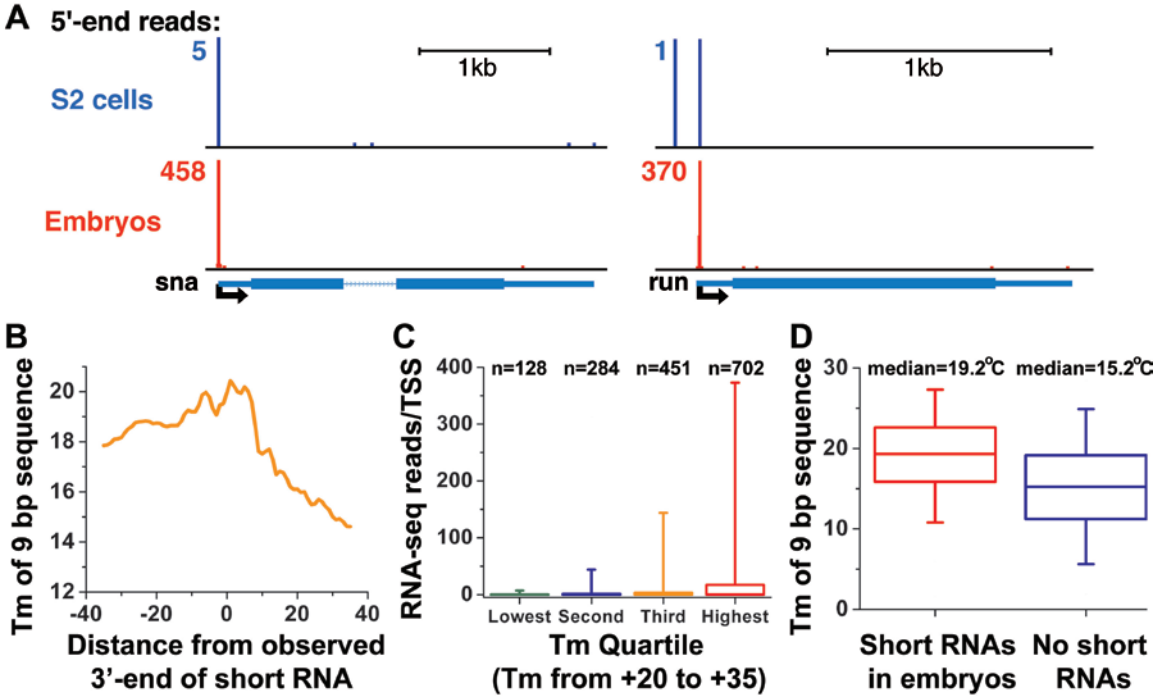

Figure S10. 1,565 genes that produce short RNAs in embryos, but not S2 cells, show a characteristic Tm profile within their promoter-proximal region. (A) Shown are two examples of genes that lack significant levels of short RNAs in libraries derived from S2 cells (shown in blue), but that display high levels of short RNAs in 0-16h *Drosophila* embryos (shown in red). Snail (sna) is CG3956 and Runt (run) is CG1849. The peak number of reads from RNA-seq experiments is denoted for each panel in the corresponding color. (B) Tm analysis of 9-bp sequences surrounding the primary 3'-end location for 338 genes with focused 3'-end positions in libraries derived from embryos. (C) 8,492 genes that do not produce significant short RNAs in S2 cells were divided into quartiles based on the average Tm of the region from +20 to +35. The number of genes in each quartile that generate significant levels of short RNAs in embryos (shown above each box) increase with Tm, with nearly a third of genes in the highest quartile producing short RNAs in embryos (702 of 2,123 genes in this quartile). Moreover, box plots showing the number of RNA reads detected in embryos from each of these quartiles reveal that the number of reads observed per TSS also increases significantly with increasing Tm ($P<0.0001$; Kruskal-Wallis test for each pairwise comparison). (D) Of the 8,492 genes shown in (C) those that produce significant short RNAs in embryos have a higher average Tm in the promoter-proximal region (average Tm from +20 to +35) than genes that do not produce short RNAs ($P<0.0001$; Mann-Whitney U-test). Box plots depict the 25th, 50th and 75th percentiles and the whiskers show the range between 5th and 95th percent.

Figure S11

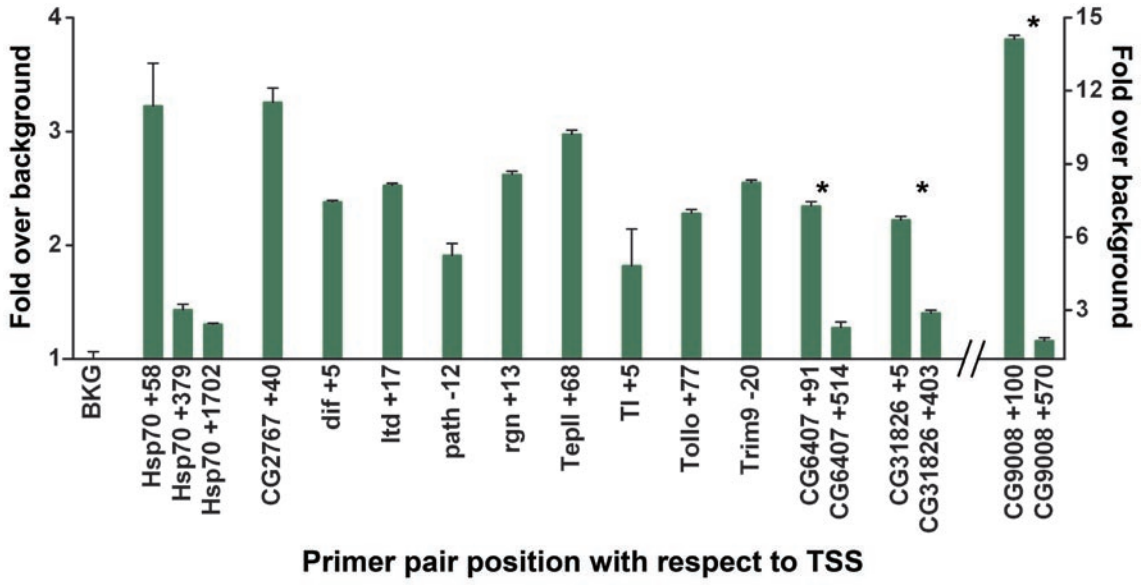

Figure S11. ChIP using an antibody against IIS in untreated S2 cells shows that IIS localizes to the promoter-proximal region of genes that produce short RNAs. IIS binding was determined by qPCR as in (*7*) and is expressed as fold enrichment over background (the inactive Ubx gene, where IIS binding is absent). The positions given for each gene are the center point of primer pairs with respect to the TSS of each gene. Genes shown to the left of the hash marks on the X axis are scaled to the left-hand Y-axis whereas CG9008 is scaled using the right-hand Y-axis. Genes shown in Fig. 4 are denoted by an asterisk.

Figure S12

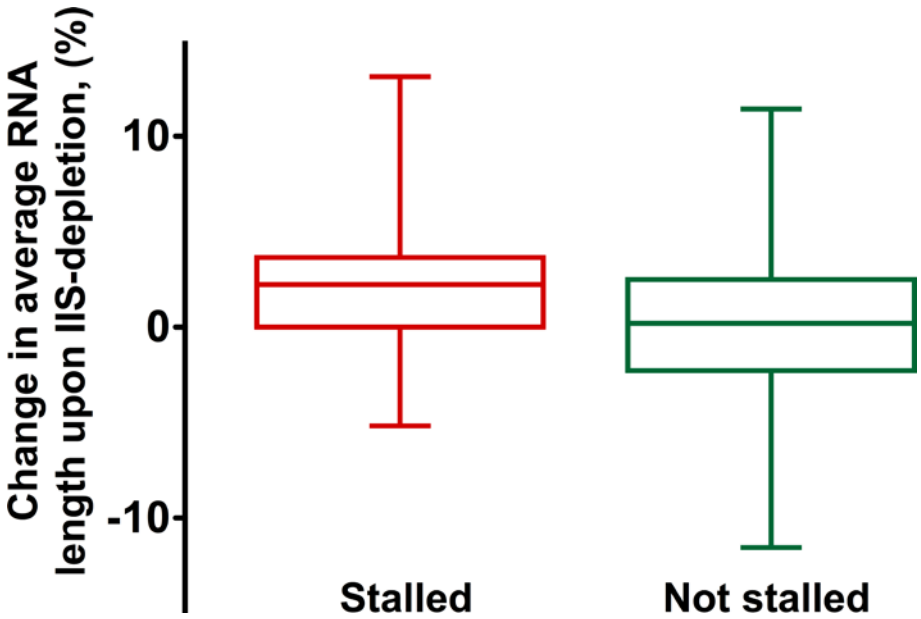

Figure S12. A majority of genes that harbor stalled Pol II have longer promoter-proximal RNAs in cells depleted of IIS, indicating that transcription of these genes involves IIS-mediated RNA cleavage. The median position of 3'-ends was calculated for genes that were defined as stalled, or not stalled, in mock-treated and IIS-depleted samples, as described in Materials and Methods. The change in the average RNA lengths that resulted from depletion of IIS was calculated as a percentage of the total RNA length, and the distribution of these values is shown for genes that were determined to possess stalled Pol II (average shift 2.2% ± 0.2; 745 genes), as compared to those genes that were bound by Pol II, but were not considered stalled (average shift 0.2% ± 0.15; 2,904 genes) (*2*). This difference is significant, as determined by the Mann-Whitney U-test ($P<10^{-4}$). The box plot depicts the 25^{th}, 50^{th} and 75^{th} percentiles and the whiskers show the range between 5^{th} and 95^{th} percent.

Figure S13

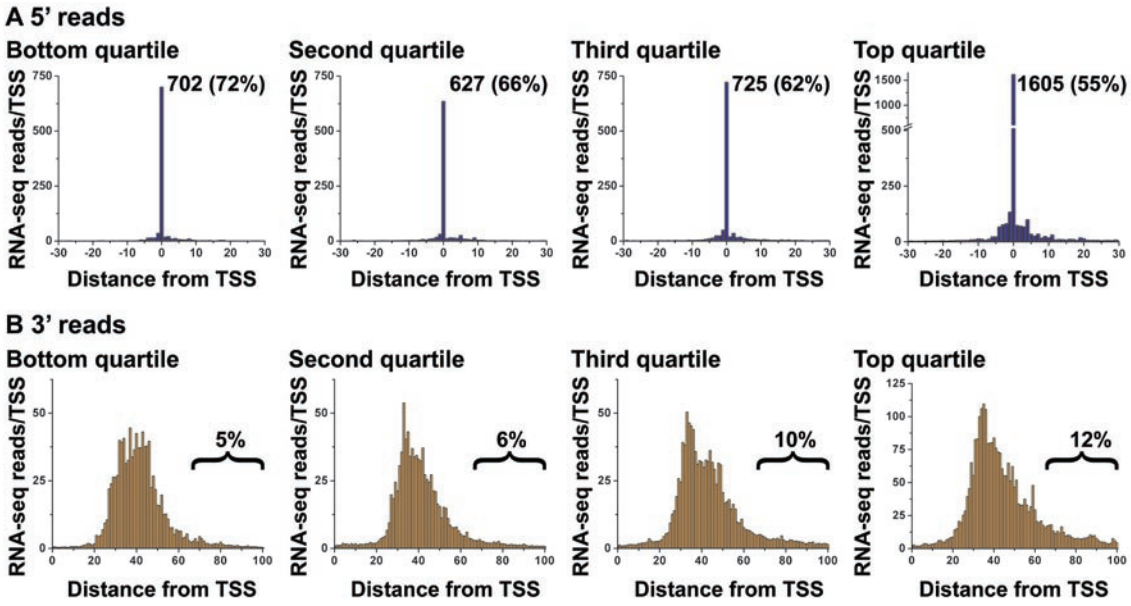

Figure S13. Correlation between the number of short RNAs and expression level of a given gene. Genes with significant numbers of short RNAs were separated into quartiles based on their expression levels (as determined in (*2*). Metagene analyses were performed for genes in each quartile: the 5'-end (A) and 3'-end (B) libraries ends were aligned to the observed TSS for each gene as described in Materials and Methods. The numbers above each graph in (A) represent the average number of reads per gene at the observed TSS and the percentage of total reads in the +/- 50 bp interval that are located at the observed TSS. The brackets in (B) denote the interval from +65 to +100, and the values show the percentage of total reads (from +1 to +100) that fall into this downstream interval. We note that as transcription levels increase, the fraction of reads that map downstream of the major stalling sites increases, indicating that these RNAs are associated with productively elongating Pol II.

Supplementary Table S1

	5'-end short RNAs	3'-end short RNAs	3'-end short RNAs	3'-end short RNAs
Treatment	**Untreated**	**Untreated**	**Mock-treated**	**IIS-depleted**
Sample 1	**8,729,554**	**7,960,788**	**9,237,021**	**6,823,304**
Sample 2	**7,794,030**	**7,918,307**	**7,712,926**	**8,997,291**
Combined	**16,523,584**	**15,879,095**	**16,949,947**	**15,820,595**

	5'-end long RNAs	Pol II ChIP-seq	5'-end short RNAs	3'-end short RNAs
Treatment	**Untreated**	**Untreated**	**0-16 hour embryos**	**0-16 hour embryos**
Sample 1	**2,704,931**	**3,400,645**	**6,373,904**	**5,849,170**
Sample 2	**5,539,955**	**4,509,932**		
Combined	**8,244,886**	**7,910,577**		

Table S1. The number of mappable sequences obtained for each sample described in this work. Reads from independent biological replicates of short RNA libraries (one lane each) were combined for the majority of the analyses described in this work. Reads from two independent lanes of Pol II ChIP-seq samples were also combined and were derived from technical replicates. Fastq and bed files for the sequences generated during this study are available in GEO under the accession number GSE18643.

Supplementary References

S1. D. S. Gilmour, R. Fan, *Methods* **48**, 368 (Aug, 2009).
S2. G. W. Muse *et al.*, *Nat Genet* **39**, 1507 (Dec, 2007).
S3. H. L. Eaves, Y. Gao, *Bioinformatics* (Feb 19, 2009).
S4. L. J. Core, J. J. Waterfall, J. T. Lis, *Science* **322**, 1845 (Dec 19, 2008).
S5. A. C. Seila *et al.*, *Science* **322**, 1849 (Dec 19, 2008).
S6. P. Rice, I. Longden, A. Bleasby, *Trends Genet* **16**, 276 (Jun, 2000).
S7. K. Adelman *et al.*, *Mol Cell* **17**, 103 (Jan 7, 2005).
S8. T. Juven-Gershon, J. Y. Hsu, J. W. Theisen, J. T. Kadonaga, *Curr Opin Cell Biol* **20**, 253 (Jun, 2008).
S9. C. Lee *et al.*, *Mol Cell Biol* **28**, 3290 (May, 2008).
S10. D. A. Hendrix, J. W. Hong, J. Zeitlinger, D. S. Rokhsar, M. S. Levine, *Proc Natl Acad Sci U S A* **105**, 7762 (Jun 3, 2008).