



1-26-2012

Global Analysis of RNA Secondary Structure in Two Metazoans

Fan Li

University of Pennsylvania, fanli.gcb@gmail.com

Qi Zheng

University of Pennsylvania, zhengqi@mail.med.upenn.edu

Paul Ryvkin

University of Pennsylvania, paulnik@gmail.com

Isabelle Dragomir


University of Pennsylvania

Yaanik Desai

University of Pennsylvania

See next page for additional authors

Follow this and additional works at: http://repository.upenn.edu/biology_papers

 Part of the [Amino Acids, Peptides, and Proteins Commons](#), [Biology Commons](#), [Genetics and Genomics Commons](#), and the [Nucleic Acids, Nucleotides, and Nucleosides Commons](#)

Recommended Citation

Li, F., Zheng, Q., Ryvkin, P., Dragomir, I., Desai, Y., Aiyer, S., Valladares, O., Yang, J., Bambina, S., Sabin, L., Murray, J. I., Lamitina, T., Rai, A., Cherry, S., Wang, L., & Gregory, B. D. (2012). Global Analysis of RNA Secondary Structure in Two Metazoans. *Cell Reports*, 1 (1), 69-82. <http://dx.doi.org/10.1016/j.celrep.2011.10.002>

This paper is posted at Scholarly Commons. http://repository.upenn.edu/biology_papers/27

For more information, please contact repository@pobox.upenn.edu.

Global Analysis of RNA Secondary Structure in Two Metazoans

Abstract

The secondary structure of RNA is necessary for its maturation, regulation, processing, and function. However, the global influence of RNA folding in eukaryotes is still unclear. Here, we use a high-throughput, sequencing-based, structure-mapping approach to identify the paired (double-stranded RNA [dsRNA]) and unpaired (single-stranded RNA [ssRNA]) components of the *Drosophila melanogaster* and *Caenorhabditis elegans* transcriptomes, which allows us to identify conserved features of RNA secondary structure in metazoans. From this analysis, we find that ssRNAs and dsRNAs are significantly correlated with specific epigenetic modifications. Additionally, we find key structural patterns across protein-coding transcripts that indicate that RNA folding demarcates regions of protein translation and likely affects microRNA-mediated regulation of mRNAs in animals. Finally, we identify and characterize 546 mRNAs whose folding pattern is significantly correlated between these metazoans, suggesting that their structure has some function. Overall, our findings provide a global assessment of RNA folding in animals.

Disciplines

Amino Acids, Peptides, and Proteins | Biology | Genetics and Genomics | Nucleic Acids, Nucleotides, and Nucleosides

Author(s)

Fan Li, Qi Zheng, Paul Ryvkin, Isabelle Dragomir, Yaanik Desai, Subhadra Aiyer, Otto Valladares, Jamie Yang, Shelley Bambina, Leah R Sabin, John I. Murray, Todd Lamitina, Arjun Rai, Sara Cherry, Li-San Wang, and Brian D. Gregory

Global Analysis of RNA Secondary Structure in Two Metazoans

Fan Li,^{1,2,3,11} Qi Zheng,^{1,2,11} Paul Ryvkin,³ Isabelle Dragomir,¹ Yaanik Desai,⁴ Subhadra Aiyer,⁴ Otto Valladares,⁵ Jamie Yang,^{1,2} Shelly Bambina,⁶ Leah R. Sabin,⁶ John I. Murray,^{2,7} Todd Lamitina,^{2,8} Arjun Raj,^{2,4} Sara Cherry,^{2,6} Li-San Wang,^{2,3,5,9,10,*} and Brian D. Gregory^{1,2,3,*}

¹Department of Biology

²PENN Genome Frontiers Institute

³Genomics and Computational Biology Graduate Program

⁴Department of Bioengineering

University of Pennsylvania, Philadelphia, PA 19104, USA

⁵Department of Pathology and Laboratory Medicine

⁶Department of Microbiology

⁷Department of Genetics

⁸Department of Physiology

⁹Institute on Aging

¹⁰PENN Center for Bioinformatics

School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

¹¹These authors contributed equally to this work

*Correspondence: lswang@mail.med.upenn.edu (L.-S.W.), bdgregor@sas.upenn.edu (B.D.G.)

DOI 10.1016/j.celrep.2011.10.002

SUMMARY

The secondary structure of RNA is necessary for its maturation, regulation, processing, and function. However, the global influence of RNA folding in eukaryotes is still unclear. Here, we use a high-throughput, sequencing-based, structure-mapping approach to identify the paired (double-stranded RNA [dsRNA]) and unpaired (single-stranded RNA [ssRNA]) components of the *Drosophila melanogaster* and *Caenorhabditis elegans* transcriptomes, which allows us to identify conserved features of RNA secondary structure in metazoans. From this analysis, we find that ssRNAs and dsRNAs are significantly correlated with specific epigenetic modifications. Additionally, we find key structural patterns across protein-coding transcripts that indicate that RNA folding demarcates regions of protein translation and likely affects microRNA-mediated regulation of mRNAs in animals. Finally, we identify and characterize 546 mRNAs whose folding pattern is significantly correlated between these metazoans, suggesting that their structure has some function. Overall, our findings provide a global assessment of RNA folding in animals.

INTRODUCTION

Recent findings have revealed unexpectedly pervasive transcription within animal genomes (Birney et al., 2007; Celniker et al., 2009), but only a very small proportion of these RNA transcripts are actually predicted to encode proteins. These

observations provide further emphasis for the growing population of functional RNA molecules, which have been implicated in gene expression regulation (e.g., small RNAs [smRNAs]; functional, long, noncoding RNAs [lncRNAs]; and riboswitches), RNA splicing (e.g., small nuclear RNAs [snRNAs]), RNA editing (e.g., small nucleolar RNAs [snoRNAs]), translation (e.g., ribosomal RNAs [rRNAs] and transfer RNAs [tRNAs]), and catalytic activities (e.g., group I introns and ribozymes). The functionality of these non-protein-coding RNAs is intimately linked to their three-dimensional structure (Brierley et al., 2007; Montange and Batey, 2008), which is determined by specific base-pairing interactions encoded within their primary sequences (Buratti et al., 2004; Cooper et al., 2009; Cruz and Westhof, 2009; Sharp, 2009). These interactions can either be within (intramolecular) or between (intermolecular [heteroduplex]) RNA molecules. Furthermore, there is increasing evidence suggesting that mRNA maturation processes (e.g., splicing, polyadenylation) require that pre-mRNA molecules be folded into a precise secondary structure in eukaryotic organisms (Buratti et al., 2004; Cooper et al., 2009; Cruz and Westhof, 2009; Sharp, 2009). In fact, specific secondary structures within the pre-mRNA transcript can either repress or aid splicing by masking or organizing splice sites, respectively (Raker et al., 2009; Warf and Berglund, 2010) and can also modulate polyadenylation (Klasens et al., 1998; Zarudnaya et al., 2003). Thus, the secondary structure of all RNA classes is abundantly important for the functionality, maturation, and regulation of these molecules.

The discovery of RNA interference (RNAi) pathways has brought our attention to a vast, evolutionarily conserved, post-transcriptional, regulatory network dependent on self or foreign double-stranded RNAs (dsRNAs) (Bartel, 2004; Carthew and Sontheimer, 2009). In animals, production of intra- and intermolecularly base-paired RNAs gives rise to 20–30 nt smRNAs

through the activity of DICER RNase III-type ribonucleases (Bartel, 2004; Carthew and Sontheimer, 2009). These smRNAs are the sequence-specific effectors of RNAi pathways that direct transcriptional or posttranscriptional regulation of genes, repetitive sequences, viruses, and transposable elements by pairing with complementary RNAs from these sources (Almeida and Allshire, 2005).

In most animals, smRNAs are composed of microRNAs (miRNAs), several classes of endogenous small interfering RNAs (esiRNAs), and Piwi-interacting RNAs (piRNAs), with the first two classes being somatically dominant because piRNAs are found only in the germline (Aravin and Hannon, 2008; Kim et al., 2009). Upon incorporation into an RNA-induced silencing complex (RISC), animal miRNAs typically inhibit translation of their target mRNAs, but can also induce degradation of these transcripts (Chekulaeva and Filipowicz, 2009). The regulatory interaction of miRNA-bound RISC (miRISC) and a target mRNA mostly involves complementary base pairing only between nucleotides 2–8 of a miRNA (counted from its 5' end) and a binding site in that transcript (seed region). There is usually limited or no interaction between the miRNA and its target transcript outside of this seed region (nucleotides 9–22 of the miRNA). Interestingly, esiRNA-incorporated RISC complexes (esiRISC) normally cleave their target RNAs (Okamura et al., 2008) or direct dimethylation and trimethylation of histone H3 lysine 9 (H3K9me2 and H3K9me3, respectively), the latter leading to heterochromatin formation and transcriptional silencing of the target loci (Fagegaltier et al., 2009). The regulatory interaction of esiRISC and a target mRNA involves extensive complementary base pairing between the entire esiRNA and a binding site in that transcript (Girard and Hannon, 2008; Watanabe et al., 2008). The difference in regulatory outcomes directed by miRNAs and esiRNAs is thought to be a consequence of the difference in complementary base-pairing interactions between these smRNAs and their targets. In total, base-paired RNAs are required for both the biogenesis and function of all animal small silencing RNAs, further emphasizing the importance of RNA secondary structure in regulating gene expression.

Recently, we and others have developed and employed high-throughput, sequencing-based, structure-mapping approaches to interrogate RNA secondary structure on a genome-wide scale (Kertesz et al., 2010; Underwood et al., 2010; Zheng et al., 2010). These first studies focused on determination of secondary structure for all *Arabidopsis thaliana* RNAs (Zheng et al., 2010), polyadenylated RNAs of *Saccharomyces cerevisiae* (Kertesz et al., 2010), and known and newly discovered ncRNAs of mouse (Underwood et al., 2010). These initial studies validated high-throughput, sequencing-based, structure-mapping approaches as effective and efficient methods of interrogating RNA secondary structure on a global scale (Westhof and Romby, 2010). However, a comprehensive, whole-genome analysis of RNA secondary structure is still lacking for any metazoan. Furthermore, a thorough analysis of mRNA secondary-structure correlation between animals has never been accomplished, even though such a study has the potential to uncover RNAs with structures or substructures that may be functional.

Here, we use our high-throughput, sequencing-based, structure-mapping approach to comprehensively identify the paired (dsRNA) and unpaired (single-stranded RNA [ssRNA]) components of the *Drosophila melanogaster* and *Caenorhabditis elegans* transcriptomes, which allows us to interrogate the structural landscape in both animals more globally. From this analysis, we reveal that ssRNAs and dsRNAs are significantly correlated with specific epigenetic modifications in animals. We also uncover a sizable population of, to our knowledge, novel, highly base-paired RNAs, many of which likely encode lncRNAs with intricate and dynamic expression patterns. Additionally, we identify conserved features of mRNA secondary structure that indicate that RNA folding demarcates regions of protein translation. Our analysis also reveals that mRNA secondary structure surrounding miRNA binding sites is strikingly distinct in *Drosophila* and *C. elegans*, suggesting that target mRNA recognition and/or regulation by miRNAs is significantly different in various animals. Furthermore, we use a comparative genomics approach to identify and characterize RNA secondary structures that are correlated or anticorrelated between two organisms that are separated by >1 billion years of evolution. Interestingly, we find that mRNAs encoding proteins involved in chromatin related processes are overrepresented only in the transcripts with correlated secondary structure between animals. These results suggest that the secondary structure of mRNAs with highly correlated folding patterns has some function within these molecules. In total, our findings emphasize the importance of RNA folding and provide global evidence of widespread mRNA secondary structure correlation and anticorrelation between animals.

RESULTS

Genome-wide Characterization of the dsRNA and ssRNA Components of *Drosophila* and *C. elegans* Transcriptomes

Using a high-throughput, sequencing-based, structure-mapping approach, we characterized the paired (dsRNA) and unpaired (ssRNA) components of rRNA-depleted total RNA from *Drosophila* DL1 culture cells and *C. elegans* mixed-stage animals. As expected, we found that the majority of dsRNA sequencing reads from both *Drosophila* (220,216,161 total reads) and *C. elegans* (213,499,238 total reads) corresponded to known highly structured RNA classes (e.g., rRNAs, snRNAs, snoRNAs, etc.) and smRNA-producing loci (e.g., miRNAs) (Figures 1A and 1B). We also identified a large proportion of dsRNA reads that mapped to transposable elements (TEs) in *Drosophila*, consistent with previous reports that demonstrated active siRNA-mediated silencing of expressed TEs in *Drosophila* culture cells (Czech et al., 2008; Ghildiyal et al., 2008; Kawamura et al., 2008). Conversely, only an extremely small fraction (~1.0%) of dsRNA reads corresponded to TEs in *C. elegans* (Figure 1B). This is likely a consequence of highly efficient transposable element silencing in these animals (Sijen and Plasterk, 2003). In contrast, our ssRNA sequencing reads from both animals (224,124,379 and 218,364,316 total ssRNA reads for *Drosophila* and *C. elegans*, respectively) were enriched in protein-coding mRNAs and pseudogenes (Figures 1C and 1D).

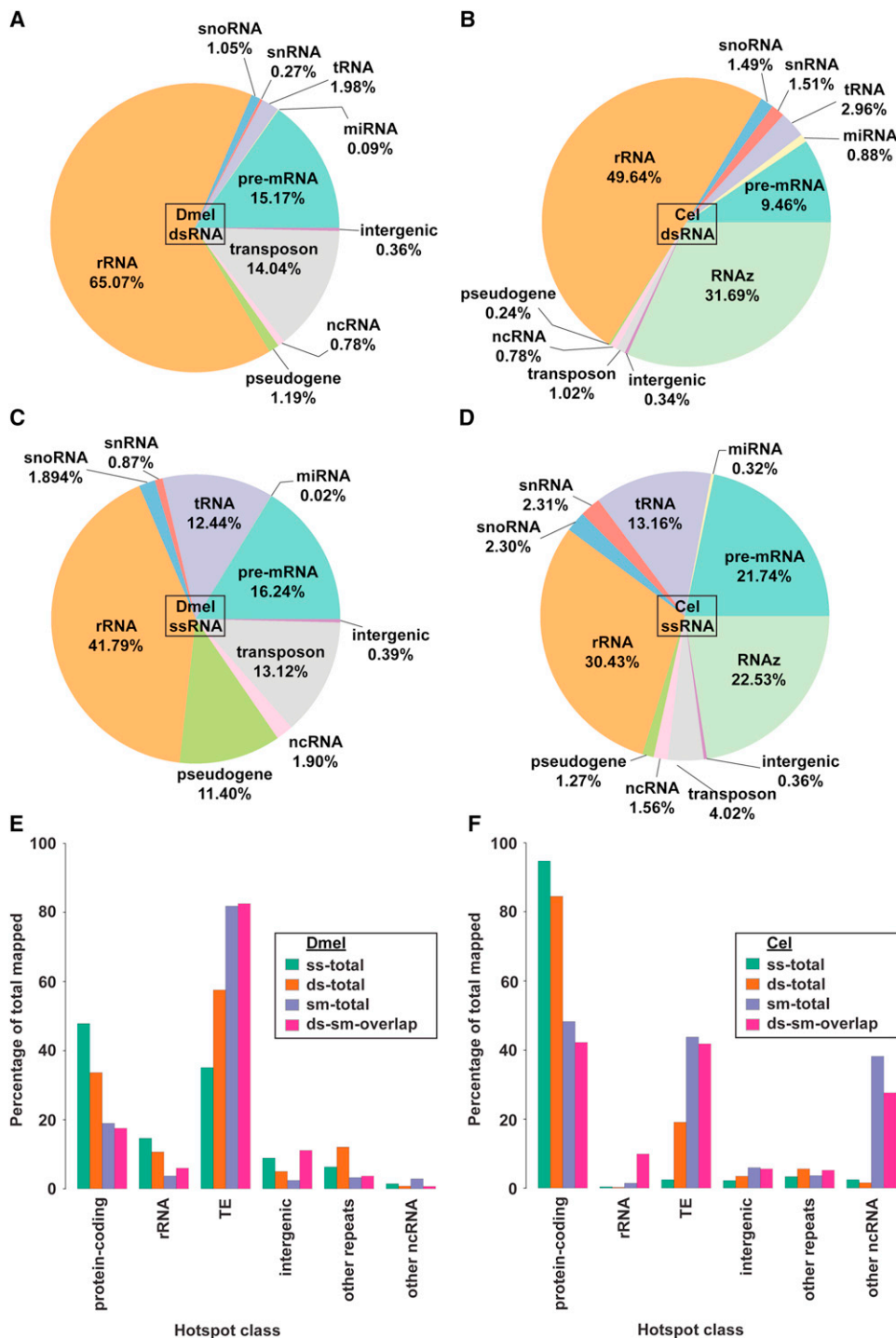


Figure 1. Characterization of the dsRNA and ssRNA Components of the *Drosophila* and *C. elegans* Transcriptomes

(A and B) Pie charts showing functional classification of dsRNA sequencing reads for *Drosophila* (labeled Dmel) and *C. elegans* (labeled Cel), respectively. (C and D) Pie charts showing functional classification of ssRNA sequencing reads for *Drosophila* (labeled Dmel) and *C. elegans* (labeled Cel), respectively. (E) Classification of dsRNA, ssRNA, and smRNA hot spots for *Drosophila*. Values are as a percentage of total hot spots for each type (e.g., dsRNA). Purple bars show the classification of overlapping ds- and smRNA hot spots, suggesting that these dsRNAs are the substrates for smRNA processing from these regions. (F) Classification of dsRNA, ssRNA, and smRNA hot spots for *C. elegans*. Values are as a percentage of total hot spots for each type (e.g., dsRNA). Purple bars show the classification of overlapping ds- and smRNA hot spots, suggesting that these dsRNAs are the precursors of smRNAs.

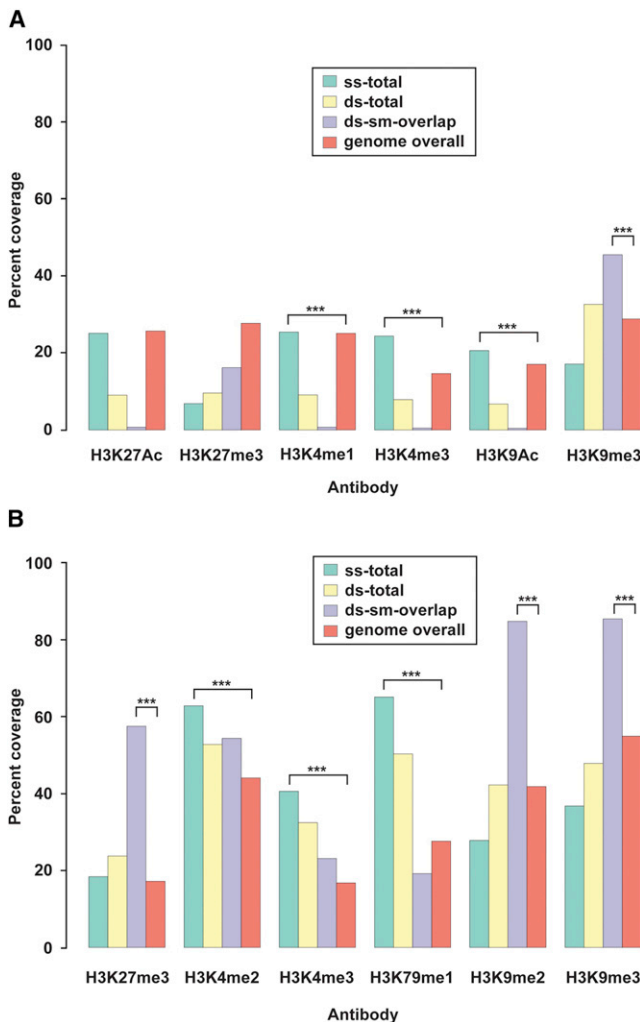


Figure 2. dsRNA, ssRNA, and smRNA Hot Spots Are Associated with Specific Epigenetic Modifications in Animals

(A) The percentage of ssRNA (green), dsRNA (yellow), and smRNA-producing dsRNA hot spots (purple), as well as bases in the entire genome (red), with specific histone modifications (as indicated in the figure) for *Drosophila*. Values are given as a percentage of all base positions for each hot spot class or the entire genome (control) that are associated with the given epigenetic mark. *** denotes p value $\rightarrow 0$.

(B) Same analysis as in (A), but for *C. elegans*.

Together, these results indicate that our approaches are interrogating the desired components of the transcriptome in the two animal systems.

Next, we used a geometric distribution-based approach to identify genomic regions that were significantly enriched in either paired RNAs (dsRNA hot spots) or unpaired RNAs (ssRNA hot spots) (Table S1 available online). This analysis identified 25,007 and 9,972 dsRNA hot spots and 19,464 and 7,068 ssRNA hot spots in *Drosophila* and *C. elegans*, respectively (Figure S1; Tables S2 and S3). As expected, the highly repetitive, transposon-rich pericentromeric regions of the *Drosophila* genome were found to be a rich source of dsRNA (Figure 1E; Figure S2). This is not surprising, because *cis* transcriptional silencing of

these regions is likely mediated by esiRNAs that require a dsRNA intermediate for their biogenesis (Fagegaltier et al., 2009). It is noteworthy that we did not observe a similarly strong bias of dsRNA hot spots in *C. elegans* transposable elements (Figure 1F). These findings suggest that the quantity of smRNA-producing dsRNAs required for transposon silencing is different between *C. elegans* and *Drosophila*, which could be because of a more efficient regulatory pathway or reduced abundance of repetitive elements in worms. Alternatively, differences in the animal materials used for these experiments (DL1 culture cells for *Drosophila* and mixed-stage worms for *C. elegans*) could also explain these findings. Further studies are necessary to test between these hypotheses.

This hot spot analysis also revealed that the majority of highly paired RNAs in *C. elegans*, and the second most abundant class in *Drosophila*, were protein-coding mRNAs. In both animals the majority of ssRNA hot spots corresponded to mRNAs, as expected (Figures 1E and 1F, green bars). In total, these results substantiate that dsRNA-seq and ssRNA-seq interrogate the desired portion of the transcriptome.

Highly Structured, Functional RNAs are Sources of smRNAs

The biogenesis of all known functional small silencing RNAs (e.g., miRNAs and esiRNAs) requires a dsRNA intermediate. Therefore, we determined the propensity of highly base-paired regions (dsRNA hot spots) to be processed into smRNAs by using corresponding smRNA-seq data (Figures 1E and 1F; Tables S2 and S3; see Extended Experimental Procedures for smRNA data analysis). We found that for both organisms, the highly base-paired regions within all interrogated RNA categories, including functional RNAs (e.g., rRNAs, snRNAs, snoRNAs, and tRNAs) and pre-mRNAs, were extremely likely to be processed into smRNAs (Figures 1E and 1F). Although these results were expected for transposable and repetitive elements, which are known to be smRNA biogenesis substrates in animals (Czech et al., 2008; Ghildiyal et al., 2008; Kawamura et al., 2008), it was surprising that functional RNAs are often processed into smRNAs, as intramolecular base-pairing interactions are intrinsic to their function. Overall, these results demonstrate that highly base-paired regions of animal rRNA, snRNA, snoRNA, and tRNA molecules are ideal smRNA biogenesis precursors, similar to what we previously observed for *Arabidopsis* (Zheng et al., 2010).

dsRNA and ssRNA Hot Spots Are Associated with Distinct Epigenetic Modifications

Recent studies have suggested that small dsRNAs may influence diverse patterns of epigenetic histone modification along both heterochromatic and euchromatic regions of animal genomes (Kouzarides, 2007; Moazed, 2009; modEncode Consortium et al., 2010). To this end, we examined the relationship between genome-wide histone modifications, determined previously with ChIP-seq (Kharchenko et al., 2010; modEncode Consortium et al., 2010) (Table S4), and *Drosophila* dsRNA and ssRNA hot spots (Figures 1E and 2A; Tables S2 and S3). We found that dsRNA hot spots in *Drosophila* were significantly enriched for the repressive, heterochromatic histone 3 lysine 9

(H3K9) trimethylation modification (p value $\rightarrow 0$, hypergeometric test), whereas ssRNA hot spots were enriched for the activating, euchromatic H3K4 trimethylation and H3K9 acetylation epigenetic marks (p value $\rightarrow 0$, hypergeometric test) (Figure 2A). Furthermore, smRNA-generating dsRNA hot spots were even more highly enriched for heterochromatic H3K9 trimethylation (p value $\rightarrow 0$, hypergeometric test).

A similar analysis in *C. elegans* using previously published ChIP-chip data (Celniker et al., 2009; Gerstein et al., 2010; Liu et al., 2011) showed that both dsRNA and ssRNA hot spots were significantly enriched (p value $\rightarrow 0$, hypergeometric test) for activating, euchromatic H3K4 dimethylation and trimethylation and H3K79 monomethylation (Table S4), which is likely a consequence of most worm hot spots being encompassed by protein-coding mRNAs. Additionally, our analysis revealed an overrepresentation of the repressive H3K9 dimethylation and trimethylation and H3K27 trimethylation in dsRNA hot spots (p value $\rightarrow 0$, hypergeometric test), especially those that generate smRNAs (Figure 2B). The observation that H3K27 trimethylation is significantly correlated with smRNA-generating dsRNA hot spots is similar to what was previously observed in *Tetrahymena* (Liu et al., 2007). However, this correlation was not observed for *Drosophila* (Figure 2), suggesting that smRNAs could direct this epigenetic modification to specific genomic locations, or vice versa, in *C. elegans* but not in flies. Overall, these results reveal that ssRNAs and dsRNAs are significantly correlated with specific epigenetic modifications in animals and indicate that the well-studied smRNA-mediated chromatin modification pathways in plants and yeast (Lister et al., 2008; Volpe et al., 2002) are likely conserved across higher eukaryotes.

Identification of Highly Base-Paired RNAs in Two Metazoans

Our dsRNA hot spot analysis also revealed 1,203 and 223 transcription units in *Drosophila* and *C. elegans*, respectively (Figures 1E and 1F; Tables S2 and S3), that do not overlap with any known or previously identified transcription units (Gerstein et al., 2010; Graveley et al., 2011), unannotated transposable/repetitive elements, or simple repeats. These *Drosophila* and *C. elegans* RNAs likely have biological relevance because when compared to the genomes of 14 other insects and five other worm species, respectively, it was observed that they were under purifying selection (p value $< 1.5 \times 10^{-12}$ and p value $< 1.1 \times 10^{-4}$, respectively; see Figure S3). Additionally, we found that 861 (in *Drosophila*) and four (in *C. elegans*) of these transcripts overlapped with regions of the genome that produced a significant amount of smRNAs (Figures 3B, 3C, 4B, and 4C; Tables S2 and S3), suggesting that they could function as esiRNA biogenesis precursors.

To validate our sequencing data and further interrogate the newly identified transcription units, we characterized several of these RNAs by RT-PCR in a panel of *Drosophila* tissues and developmental stages (Figure 3; Figure S3). We confirmed that all four of the RNAs tested were expressed in the culture cell line used for the initial analysis of paired and unpaired RNAs. Interestingly, three of the four (75%) transcripts exhibited tissue- and developmental-stage-specific expression patterns (Figure 3D; Figure S4), and one of the RNAs (h1529 on chromo-

some X) demonstrated differently sized RNAs in specific *Drosophila* tissues and developmental stages (Figures 3C and 3D). We also used RT-PCR to confirm the expression of three newly identified (Figures 4A, 4C, 4D, and 4F) and three previously identified (Figures 4B and 4E; Figure S5; Gerstein et al., 2010) highly base-paired RNAs in mixed-stage *C. elegans*. Furthermore, we determined the spatiotemporal expression patterns of a number of these transcripts by using single-molecule fluorescence in situ hybridization (FISH) (Figure 4G; Figure S6; Raj et al., 2008). We found that one transcript displayed expression in a large subset of cells at the 41-cell stage, with two cells harboring particularly bright spots consistent with an accumulation of multiple RNAs at the site of transcription itself (Raj et al., 2006; Vargas et al., 2005). At later stages, expression of this transcript is restricted to a few cells (Figure 4G). We observed similar dynamic patterns of expression in other dsRNAs (Figure S6), suggesting that the abundance of many of these RNAs is subject to cell-specific regulation.

Using dsRNA and ssRNA-seq Data to Develop Experimentally-Derived Models of mRNA Secondary Structure on a Genome-wide Scale

The secondary structure of all eukaryotic mRNA molecules is dictated by specific base-pairing interactions that are encoded within their nucleotide sequence (Cooper et al., 2009; Cruz and Westhof, 2009; Sharp, 2009). Prior to the development of high-throughput, sequencing-based, structure-mapping approaches (Kertesz et al., 2010; Underwood et al., 2010; Zheng et al., 2010), most RNA secondary structure models had been predicted through sequence comparisons (e.g., Infernal: <http://infernal.janelia.org/>), energy dynamics of base pairing (e.g., the RNAfold program of the Vienna package: <http://www.tbi.univie.ac.at/~ivo/RNA/>), or enzymatic and chemical experiments (Cruz and Westhof, 2009; Westhof and Romby, 2010). Here, we used the combination of dsRNA-seq (paired regions) and ssRNA-seq (unpaired regions) data to produce experimentally derived structural models of all *Drosophila* and *C. elegans* mRNAs detected in this study (see Experimental Procedures). As we observed previously for *Arabidopsis* (Zheng et al., 2010), experimentally determined mRNA secondary structures exhibit striking base-pairing differences in comparison to computationally predicted structures. Many regions that were expected by RNAfold to form large, single-stranded loops and open regions were more highly paired in our models, and vice versa (see Figure 5A, http://gregorylab.bio.upenn.edu/annoj_ce/ and http://gregorylab.bio.upenn.edu/annoj_dm/).

To validate our structural models, we characterized highly base-paired regions of several *Drosophila* mRNAs (as determined by our methodology) (Figure 5A) by RT-PCR after digestion with a single-stranded or double-stranded RNase. We expected that the selected mRNA regions would be sufficiently intact for RT-PCR amplification after treatment with the single-stranded RNase but not the double-stranded RNase. As predicted, the regions of mRNA molecules determined to be highly base paired were amplified after treatment with the ssRNase (Figure 5B). Conversely, we could not amplify these same regions after treatment with the dsRNase, which implies that they were degraded by this enzyme.

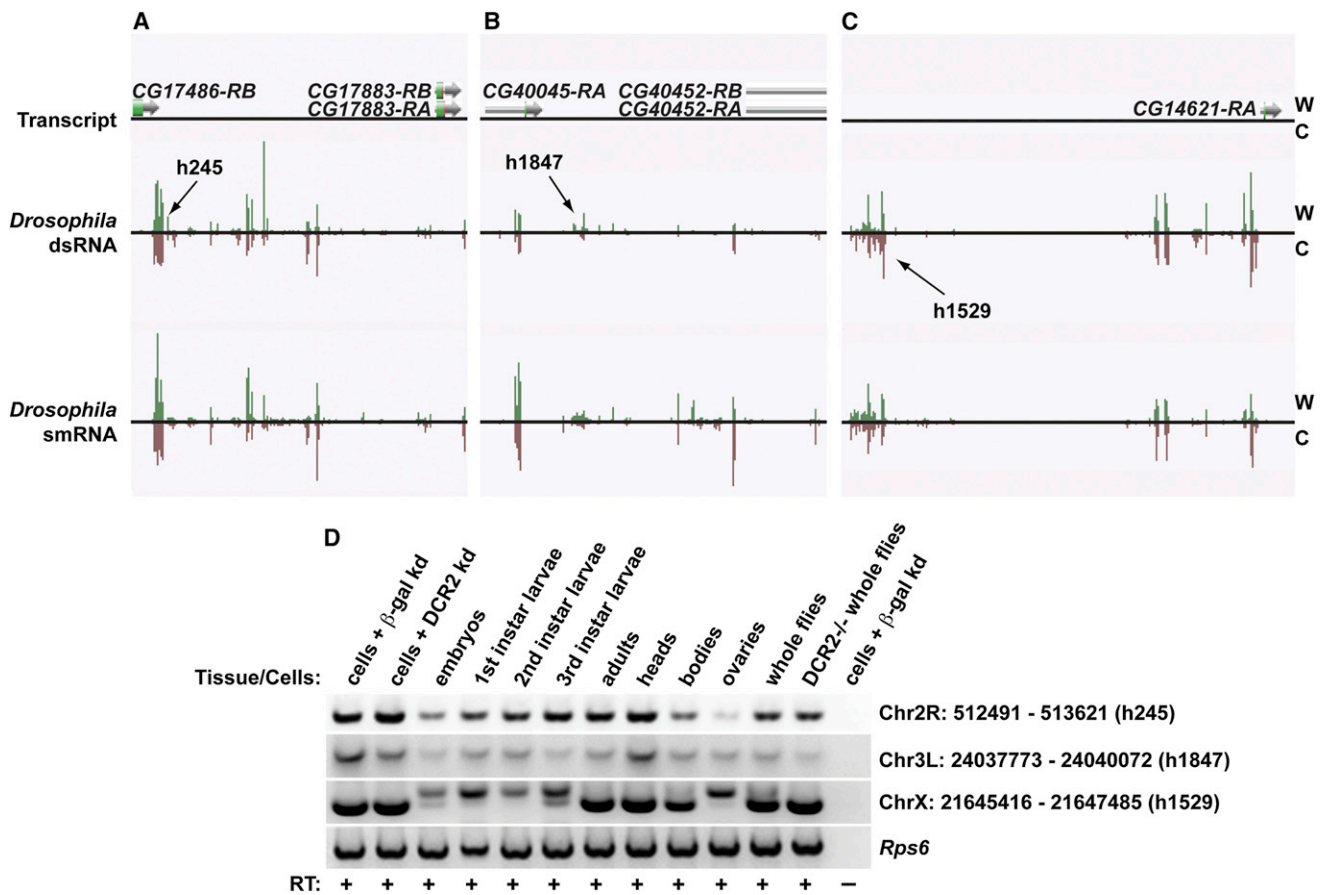


Figure 3. Highly Base-Paired RNAs in *Drosophila*

(A–C) Three examples of intergenic, highly base-paired transcripts (screenshots from our *Drosophila* RNA-seq browser, http://gregorylab.bio.upenn.edu/anno_j_dm/). W (green bars) and C (red bars) indicate signal from Watson and Crick strands, respectively. (A) An intergenic dsRNA hot spot found between CG17486 and CG17883. (B) A base-paired RNA between CG40045 and CG40452. (C) An intergenic dsRNA hot spot between CG32820 and CG14621. (D) Random-primed RT-PCR analysis of base-paired RNAs in multiple tissues and developmental stages of *Drosophila*. *Rps6* serves as a loading control.

It is worth noting that our methodology reveals the pairing status of RNA molecules in the absence of cellular proteins. This may result in models of secondary structure that do not perfectly reflect the fully folded RNA molecule in cells. However, the results of this study as well as a number of previous reports (Kertesz et al., 2010; Underwood et al., 2010; Zheng et al., 2010) have demonstrated that the models of mRNA secondary structure produced through high-throughput, structure-mapping approaches are highly informative. Overall, our results suggest that the sequencing data sets, models of RNA secondary structure, and analyses that have resulted from this study will contribute positively to future work aimed at illuminating the numerous functions of RNA secondary structure in eukaryotes.

Identification of Structural Features in Animal mRNAs that Potentially Affect Translation and miRNA-Mediated Regulation

To identify specific patterns within the secondary structure of animal mRNAs, we examined the average structure score, which is the normalized log-ratio of dsRNA- to ssRNA-seq reads,

across the coding region (CDS) and both 5' and 3' UTRs of detected *Drosophila* and *C. elegans* protein-coding transcripts (Figure 6A). For both animals, we identified significant decreases in the structure score near the start and stop codons of the CDS (p value = 1.5e-12 and 4.8e-6 for *Drosophila*, p value = 3.5e-3 and 3.9e-2 for *C. elegans*, respectively), revealing a considerably reduced tendency for base-pairing and increased accessibility of the RNA at the regions where protein translation begins and ends (Figure 6A). This was also observed for yeast mRNAs (Kertesz et al., 2010; Kozak, 2005), indicating that it may be a general feature of eukaryotic protein-coding transcripts. Somewhat surprisingly, we also found that one or both UTRs were, on average, much more highly structured than the coding region in animal mRNAs (p value < 2.2e-16 for both 5' and 3' UTR in *Drosophila*; p value = 0.54 and p value = 1.97e-06 for 5' and 3' UTR, respectively, in *C. elegans*). The inverse was observed for yeast transcripts, where the CDS is more structured than UTRs (Kertesz et al., 2010), suggesting that animal UTRs are enriched for RNA secondary structures, which might act as regulatory sites or interacting regions for RNA-binding proteins. Taken

together, these results reveal that there are conserved structural patterns within protein-coding mRNAs of animals, and suggest that these features may affect protein translation.

We hypothesized that the significant secondary structure we identified in the 3' UTRs of *C. elegans* and *Drosophila* protein-coding transcripts may be a consequence of animal mRNAs attempting to mask miRNA binding sites located in these regions, especially given that this mechanism is active in multicellular animals but not budding yeast. To test this possibility, we examined the site-specific average structure scores at positions within TargetScan-predicted miRNA target sites (Ruby et al., 2006, 2007) and the 50 bp of sequence up- and downstream of these regions (Figure 6B). This analysis revealed significantly decreased base-pairing across the entire length of miRNA binding sites in *C. elegans* target mRNAs. Further analysis confined to miRNA target sites experimentally determined to be bound by ALG-1 (the ARGONAUTE (AGO) protein at the core of *C. elegans* miRISC) (Zisoulis et al., 2010), uncovered similarly decreased base-pairing within the seed region of miRNA target sites. Conversely, we observed a significant increase in secondary structure specifically within the seed-pairing region of *Drosophila* target transcript miRNA binding sites (Figure 6C). These findings suggest that miRISC complexes encounter extremely different secondary structures during target mRNA interaction in these two animals. It is worth noting that we did not find significant differences in overall 3' UTR structure for transcripts with and without miRNA target sites, indicating that the structural constraints imposed by miRNA-mediated silencing are local to the target site.

The differences in secondary structure at miRNA binding sites between *Drosophila* and *C. elegans* led us to hypothesize that RNA folding across these regulatory regions may affect miRISC interaction. To test this idea, we determined if increasing secondary structure had a negative affect on *C. elegans* ALG-1 binding using previously published CLIP-seq data (Zisoulis et al., 2010). This analysis revealed a significant ($p = 0.03$) inverse correlation ($r = -0.23$) between ALG-1 binding affinity and miRNA target site structure (Figure 6D). In total, these results indicate that increased mRNA secondary structure at miRNA target sites can adversely affect miRISC interaction with regulatory targets.

Identification and Characterization of Significantly Correlated and Anticorrelated mRNA Secondary Structures

Although it is well accepted that mRNA maturation and regulation require that these molecules be folded into precise secondary structures, a comprehensive study to assess RNA folding functionality has never been done. To identify potentially functional mRNA secondary structures, we compared levels of sequence and structure correlation in a set of 2,223 orthologous transcripts between *Drosophila* and *C. elegans* (Table S5) (Lyne et al., 2007). Briefly, we calculated the correlation between structure profiles for each transcript pair at positions of homology to determine structure correlation (see Experimental Procedures). As expected, sequence and structure similarity were reasonably correlated ($r = 0.35$) within the entire set of 2,223 orthologous pairs (Figure 7A). Using a binomial model of RNA folding cor-

relation, we identified 736 orthologous transcript pairs that exhibited significant correlation (546 total) or anticorrelation (190 total) of their secondary structure profiles (hereafter referred to as “positively correlated” and “negatively correlated” orthologous sets, respectively) (Figure 7A, red and green, respectively). Interestingly, positively correlated orthologous pairs (see Figures 7B and 7C for an example) were enriched for mRNAs that encode proteins with functions in chromatin-related biology (chromatin organization), but there was no specific enrichment within the negatively correlated transcripts (Figure 7D, top two boxes). Furthermore, we also found that transcript pairs demonstrating high sequence but not structural similarity were not enriched for chromatin-related processes (Figure 7D, bottom two boxes). These results indicate that the observed functional enrichment within the 546 orthologous transcript pairs with significantly correlated secondary structure is not merely due to high sequence similarity. In total, these findings suggest that the structure of the 546 mRNAs with positively correlated folding patterns between these two metazoans has some function within these RNA molecules.

DISCUSSION

Here, we report a simultaneous genome-wide study of RNA secondary structure in two metazoans, *Drosophila* and *C. elegans*. Our analysis revealed a large population of dsRNAs and ssRNAs, many of which likely have distinct regulatory roles in animals (Figures 1–4). For instance, we found significant correlations between dsRNA hot spots, especially those that likely generate smRNAs, and heterochromatic histone modifications in both animals (Figure 2), indicating a potential role for these RNAs in the deposition and/or maintenance of silencing epigenetic marks. In fact, our findings provide evidence for a mechanism of smRNA-mediated transcriptional gene silencing widely conserved across eukaryotes. Additionally, we identified a strong correlation between ssRNA hot spots and activating, euchromatic histone modifications (Figure 2), suggesting that actively transcribed mRNAs prefer unpaired RNA structures. This is not overly surprising given that a decrease in RNA secondary structure in protein-coding transcripts has the potential to allow for more efficient translation (Kozak, 2005). We also characterized a set of highly base-paired transcripts, many of which are evolutionarily conserved and display intricate and dynamic expression patterns during animal development (Figures 3 and 4; Figures S3–S6). Future experiments will be aimed at addressing the functionality of these RNAs.

Given the ability to characterize RNA secondary structure globally in two animals, we examined the average folding patterns of protein-coding transcripts for both animals. This analysis uncovered conserved mRNA structural features (Figures 5 and 6). For example, we revealed a significant decrease in mRNA secondary structure at both the start and stop codons of the CDS in the two animals. These findings indicate that specific folding patterns demarcate the protein-coding region of mRNAs, suggesting that secondary structure has a regulatory effect on protein translation (Figure 6A). Specifically, decreased pairing at the translation start site could increase ribosome binding efficiency, and thereby modulate translation,

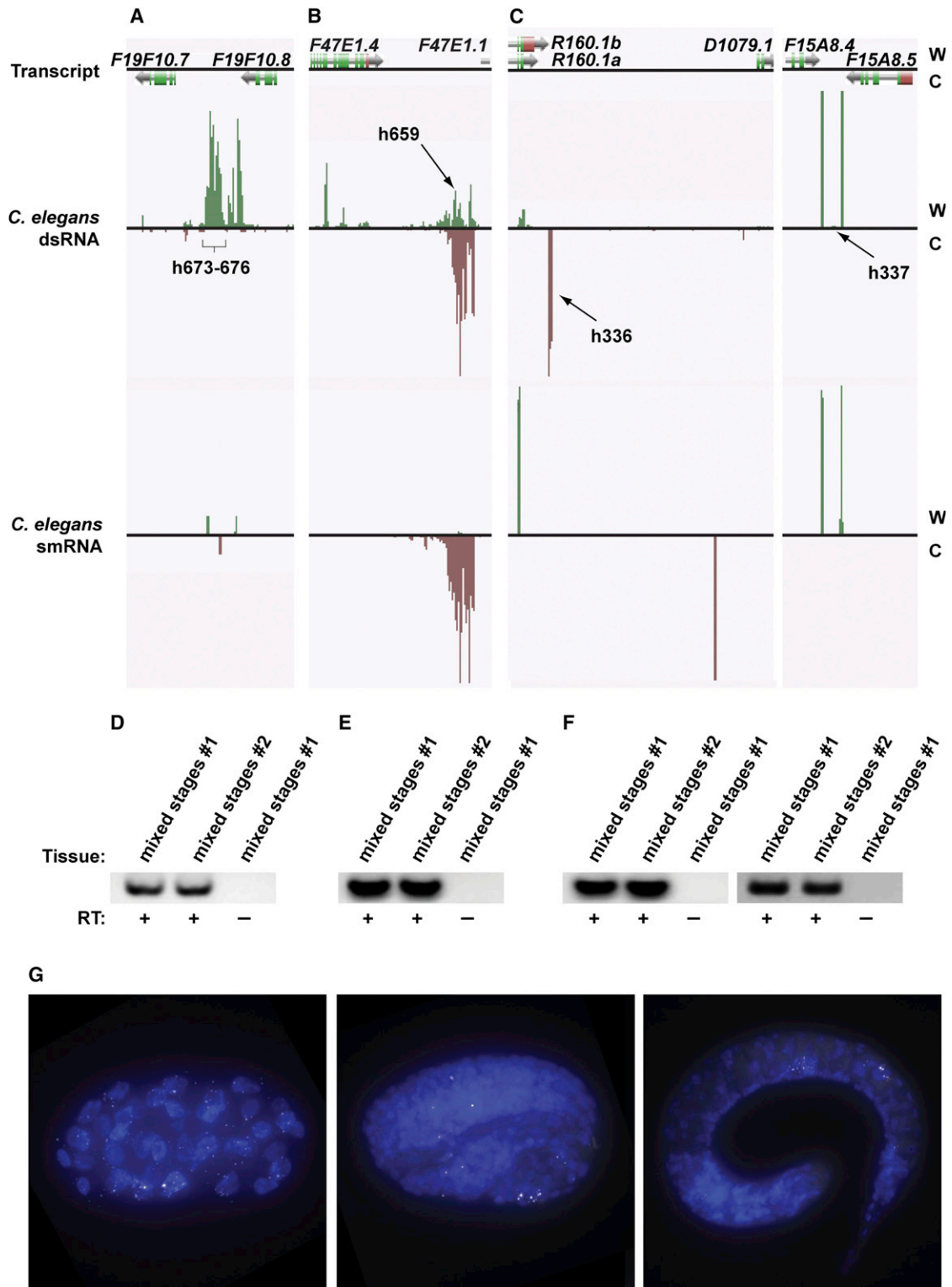


Figure 4. Highly Base-Paired RNAs in *C. elegans*

(A–C) Three examples of intergenic, highly base-paired transcripts (screenshots from our *C. elegans* RNA-seq browser, http://gregorylab.bio.upenn.edu/annoj_ce/). W (green bars) and C (red bars) indicate signal from Watson and Crick strands, respectively. (A) Four intergenic dsRNA hot spots found between *F19F10.7* and *F19F10.8*. (B) A highly base-paired RNA between *F47E1.4* and *F47E1.1*. This transcript was recently identified via high-throughput RNA profiling (Gerstein et al., 2010). (C) Two intergenic dsRNA hot spots between *R160.1b* and *D1079.1* or *F15A8.4* and *F15A8.5*, respectively.

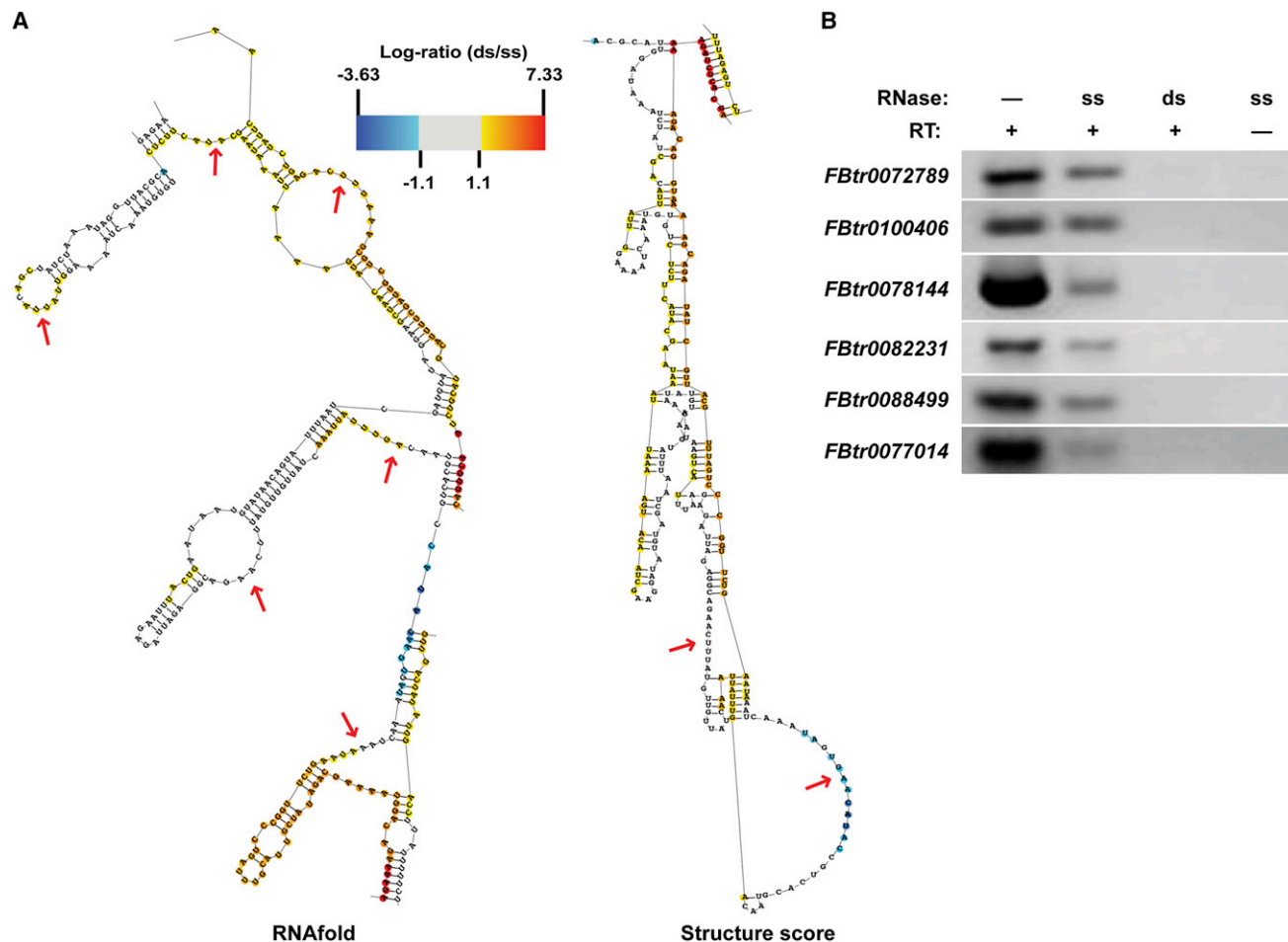


Figure 5. A Genome-wide Approach to Experimentally Interrogate RNA Secondary Structure in Eukaryotes

(A) Model of secondary structure for the *Drosophila* *FBtr0100406* transcript determined by default RNAfold (left, labeled RNAfold) or our high-throughput sequencing-based, structure-mapping approach (right, labeled Structure score). The region of this RNA interrogated in (B) is shown in this figure. The heatmap indicates the normalized log-ratio of dsRNA-seq to ssRNA-seq reads (see [Experimental Procedures](#)) at each base position. Red arrows indicate regions of the RNA model where ≥ 7 nt are unpaired.

(B) Random-primed RT-PCR analysis of dsRNA hot spots from *FBtr0072789*, *FBtr0100406*, *FBtr0078144*, *FBtr0082231*, *FBtr0088499*, and *FBtr0077014* after treatment of total RNA samples with either a single-stranded or double-stranded RNase. Samples that were not treated with reverse transcriptase or either RNase serve as controls for this experiment.

as has been suggested in yeast ([Kertesz et al., 2010](#)). Similarly, the secondary structure surrounding the translation termination site could serve as a signal for the ribosome to disengage from the transcript. Additionally, we found that on average one or both of the 5' and 3' UTRs are more highly structured than the coding region of animal transcripts ([Figure 6A](#)), suggesting that these noncoding portions of animal mRNAs contain highly folded structures that could serve as regulatory signals or interaction sites for RNA-binding proteins. To address the secondary structure of one class of potential regulatory regions in the 3'

UTRs of animal transcripts, we analyzed miRNA target sites. This analysis revealed significant and distinct patterns of base-pairing across 3' UTR-localized miRNA binding sites in *Drosophila* and *C. elegans* ([Figures 6B](#) and [6C](#)), indicating that miRISC complexes encounter extremely different secondary structures during interactions with target mRNAs in these two animals. Specifically, we found increased secondary structure within the seed region of *Drosophila* miRNA binding sites, suggesting that more energy is needed for miRISC:target RNA binding, as numerous base-pairing interactions will need to be

(D–F) Random-primed RT-PCR analysis of base-paired RNAs from mixed stage *C. elegans* that are pictured in A–C. (D), (E), and (F) correspond to (A), (B), and (C), respectively. For (F) the two RT-PCR analyses correspond to h336 and h337 in (C), respectively.

(G) FISH images of dsRNA hot spots chrIV_1804 – 1806 taken at single molecule resolution (RNA in white, nuclei stained with DAPI in blue). Images are maximum merges of a series of optical sections at a variety of developmental stages (41-cell stage, left panel; pretzel stage, middle panel; L1, right panel). Scale bars are 5 μ m long.

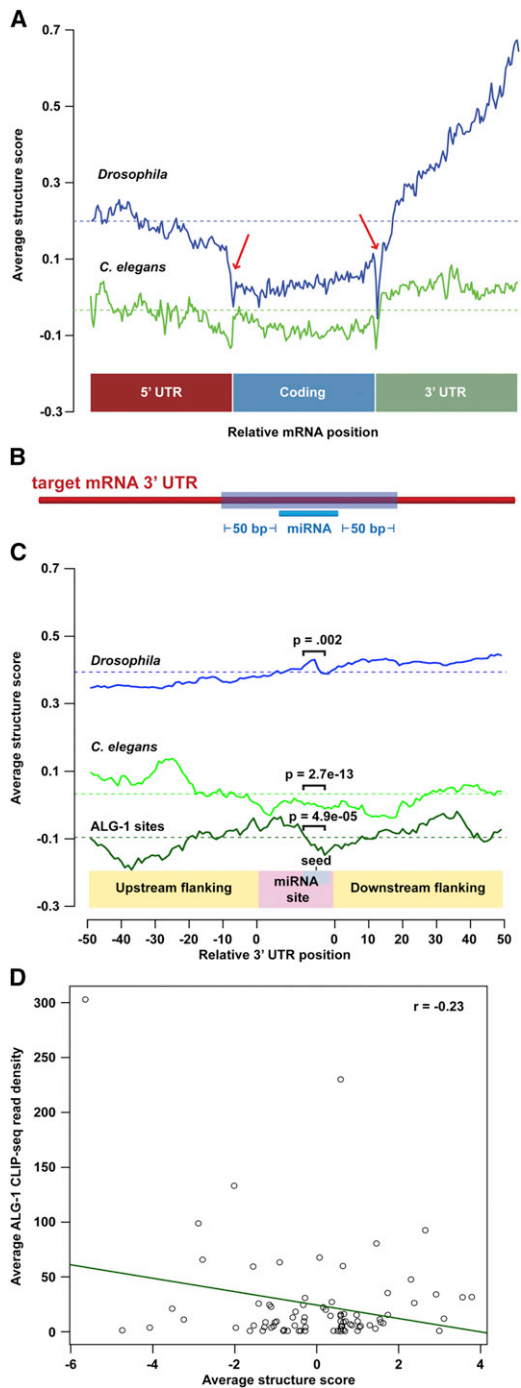


Figure 6. A Global View of mRNA Secondary Structure

(A) The average “structure score” plotted over the 5’ UTR, CDS, and 3’ UTR of all protein coding transcripts for *Drosophila* (blue line) and *C. elegans* (green line). The overall average for the entire transcript is shown as a dotted line. Red arrows highlight significant dips in secondary structure that occurred at the junctions between the UTRs and the coding region.

(B) Model depicting our analysis of RNA secondary structure at miRNA binding sites in target mRNAs.

(C) The average structure score across miRNA target sites and for 50 bp up- and downstream flanking regions of each site in *Drosophila* (blue line), *C. elegans* (green line), and experimentally identified *C. elegans* ALG-1 binding

interrupted for a functional interaction (Figure 6C). Conversely, *C. elegans* miRNA binding sites are on average much more accessible to interaction with miRISC (Figure 6C), suggesting that less energy is needed for regulatory complex:RNA binding. Taken together, these findings suggest that target mRNA recognition and/or regulation by miRISC is significantly different in *C. elegans* and *Drosophila*, and it would not be surprising if these processes are variable between other animals.

In support of a potential negative interaction between RNA folding at miRNA binding sites and miRISC interaction, we identified a significant negative correlation between increasing RNA secondary structure and *C. elegans* ALG-1 binding affinity at miRNA interaction sites (Figure 6D). Specifically, we found that ALG-1 binding affinity was increased for miRNA interaction sites with decreased levels of secondary structure, whereas the opposite was true for those regions that were more highly base-paired. (Figure 6D). In total, these results indicate that mRNA secondary structure at miRNA interaction sites has an adverse effect on miRISC interaction with regulatory targets.

Finally, we used a comparative genomics approach to identify RNA structures that are correlated or anticorrelated between the two animals (Figures 7A–7D), providing global evidence of mRNA secondary structure correlation between animals. Interestingly, the identification of RNA secondary structures that are specifically correlated (and likely functional) in protein-coding transcripts now makes it possible to test the functions of such sequences in mRNA maturation, stability, regulation, and/or protein interaction. In total, our results suggest that RNA secondary structure has effects on a myriad of cellular processes, including epigenetic chromatin modification, protein translation, miRNA-mediated posttranscriptional control, and regulation of mRNA stability, processing, and/or regulation. Additionally, we have established a useful framework for future comparative studies of RNA secondary structure.

EXPERIMENTAL PROCEDURES

Further details on the animal materials, experimental procedures, high-throughput sequencing, and processing, mapping, and analysis of Illumina GA and HiSeq sequence reads are provided in the [Extended Experimental Procedures](#).

dsRNA-seq, ssRNA-seq, and smRNA-seq Library Preparation

Briefly, total RNA was subjected to two rounds of rRNA depletion (Ribominus, Invitrogen, Carlsbad, CA) and then treated with a single-strand specific ribonuclease (RNase One, Promega, Madison, WI) for dsRNA-seq or with a double-strand specific ribonuclease (RNase V1, ABI, Foster City, CA) for ssRNA-seq, all per manufacturer’s instructions. The RNA samples are then used as the substrate for sequencing library construction using the Small RNA Sample Prep v1.5 kit (Illumina, San Diego, CA) as per manufacturer’s instructions. For *D. melanogaster*, we obtained 220,216,161 raw dsRNA reads (118,174,585 nonredundant [NR] sequences with 1.86 clones on average) and 224,124,379 raw ssRNA reads (144,624,885 NR, 1.55 copies each). For *C. elegans*, we produced 213,499,238 raw dsRNA reads (115,065,848 NR,

sites (Zisoulis et al., 2010) (dark green line). The overall structure score average for the entire ~122 bp region is shown as a dotted line.

(D) The average structure score (x axis) is plotted against average ALG-1 CLIP-seq tag density for experimentally identified miRNA binding sites that interact with this component of miRISC (Zisoulis et al., 2010) (y axis).

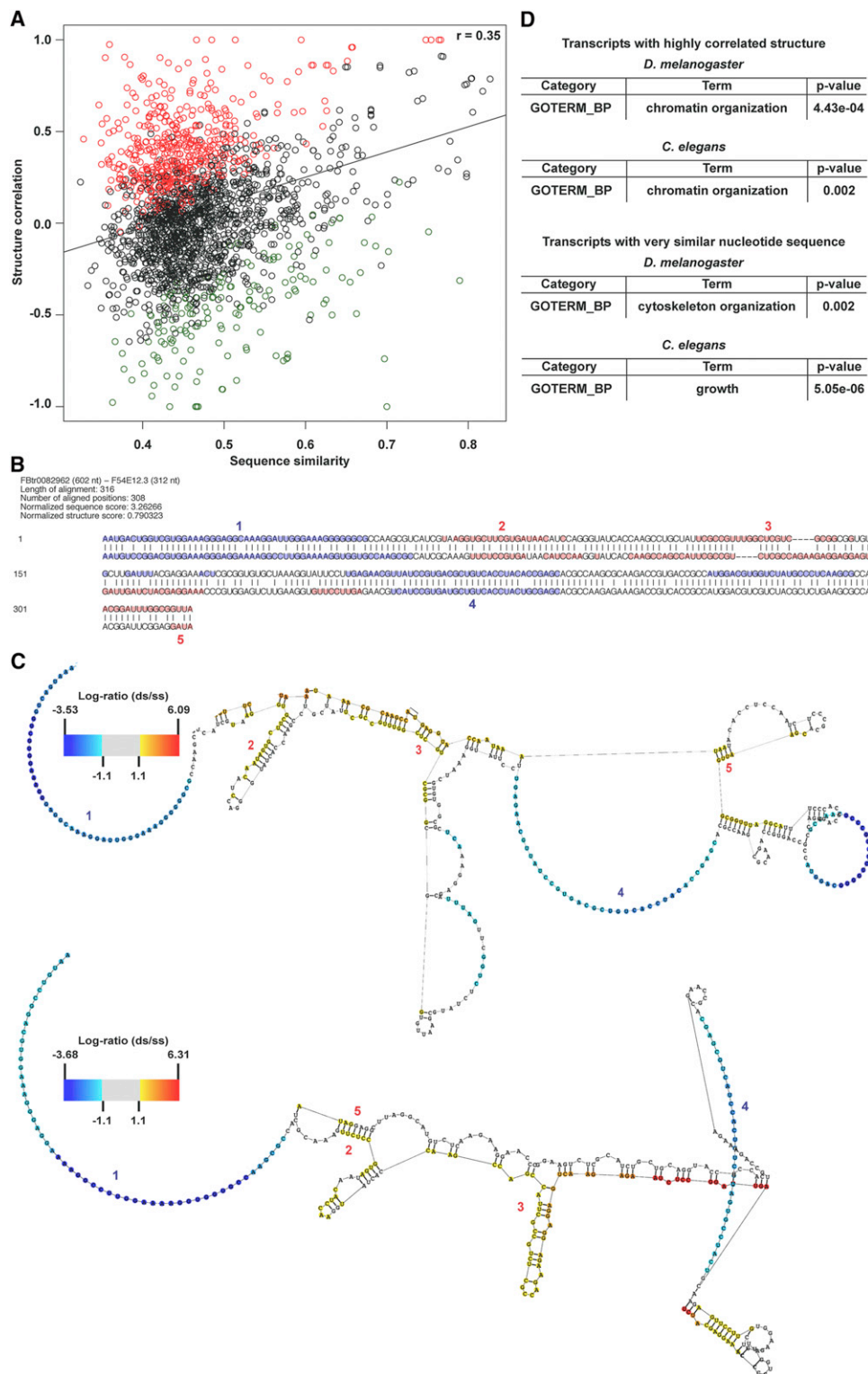


Figure 7. Highly Correlated and Anticorrelated mRNA Secondary Structures between *Drosophila* and *C. elegans*

(A) Plot of sequence similarity (x axis) versus structure correlation (y axis) for a set of 2,223 orthologous transcript pairs between *Drosophila* and *C. elegans*. Structure correlation scores range from -1 to 1 , with higher values indicating tendency of paired and unpaired nucleotides in one organism to also be the same in the other, whereas lower values indicate an opposite trend. Red and green markers indicate significantly correlated and anticorrelated structures, respectively.

1.86 copies each) and 218,364,316 raw ssRNA reads (116,488,694 NR, 1.87 copies each). smRNA-seq libraries were produced using the small RNA Sample Prep v1.5 kit (Illumina, San Diego, CA), as per manufacturer's instructions. We obtained a total of 7,597,106 and 10,594,321 smRNA reads from the *Drosophila* and *C. elegans* libraries, respectively. For more detailed methodology see [Supplemental Experimental Procedures](#).

Identification of Hot Spots in the *Drosophila* and *C. elegans* Genomes

To identify dsRNA and ssRNA hot spots in the two animal genomes, we utilized a geometric distribution-based approach. smRNA hot spots were identified using a Poisson distribution-based method as previously described (Zheng et al., 2010). For more detailed methodology see [Supplemental Experimental Procedures](#).

Functional Analysis of *Drosophila* and *C. elegans* Hot Spots

Existing gene annotations were downloaded from FlyBase (r5.22) and WormBase (WS205). Propensity of smRNA processing was assayed by overlapping dsRNA hot spots with our identified smRNA hot spots. Histone modification data (ChIP-seq for *D. melanogaster* and ChIP-chip for *C. elegans*) were downloaded from modENCODE (<http://www.modencode.org/>) (Celniker et al., 2009; modEncode Consortium et al., 2010). All of the comparative genomics data, including the multiple alignment files for "insects-15-way" and "worm-6-way," as well as the pre-calculated conservation scores from the "phastCons" program (Siepel et al., 2005) were downloaded from the UCSC Genome Browser (release dm3 and ce6 for *Drosophila* and *C. elegans*, respectively). For more detailed methodology, see [Supplemental Experimental Procedures](#).

Identification of Base-Paired Transcripts

dsRNA hot spots were first classified as intergenic if they did not overlap with any known annotations from FlyBase (r5.22) and Wormbase (WS205), for *Drosophila* and *C. elegans*, respectively. These transcripts were then filtered for transcription units identified in a series of recent RNA-seq profiling experiments (Gerstein et al., 2010; Graveley et al., 2011). For a more detailed methodology, see [Supplemental Experimental Procedures](#).

RNA Secondary Structure Prediction and Analysis of mRNA Secondary Structure Patterns

The dsRNA- and ssRNA-seq read coverages were separately normalized to the total number of reads for each transcript. Then, the log-ratio of dsRNA- to ssRNA-seq normalized read coverage was calculated at each base position to derive a "structure score," a normalized log-ratio of dsRNA- to ssRNA-seq reads. Positions with a structure score greater than 1.1 were constrained as paired ("|" in the structural constraint input), positions with a structure score less than -1.1 were constrained as unpaired ("x" in the structural constraint input), and all other positions were left unconstrained ("." in the structural constraint input). These thresholds were motivated by the distribution of structure scores under the null model where both dsRNA- and ssRNA-seq reads are uniformly distributed. In this case, 5% of all base positions are called as either paired or unpaired, even though the data are uninformative for secondary structure, and our threshold of 1.1 can therefore be loosely interpreted as a 5% FDR control. For all analyses involving an average structure score, positions with a score of 0 were ignored. Significance tests were

performed using R under an assumed Student's T distribution. TargetScan predictions were downloaded from TargetScanFly (http://www.targetscan.org/fly_12/) and TargetScanWorm (http://www.targetscan.org/worm_12/) using Predicted Conserved Targets.

Correlated and Anticorrelated Secondary Structure in Pre-mRNAs

For 2,223 orthologous pairs, the EMBOSS package's water program was used to generate sequence alignments and determine sequence similarity scores for each transcript pair. To compute the structure correlation, we used a score based on the number of positions constrained as paired or unpaired by our structure-mapping approach. Significance was assessed using a binomial model to call paired or unpaired positions. For more detailed methodology, see [Supplemental Experimental Procedures](#).

ACCESSION NUMBERS

All dsRNA-seq, ssRNA-seq, and smRNA-seq data (*Drosophila* and *C. elegans*) from our analyses were deposited into the GEO database under the accession number GSE29571.

SUPPLEMENTAL INFORMATION

Supplemental Information includes six figures, six tables, Extended Experimental Procedures, and Extended Results and can be found with this article online at [doi:10.1016/j.celrep.2011.10.002](https://doi.org/10.1016/j.celrep.2011.10.002).

LICENSING INFORMATION

This is an open-access article distributed under the terms of the Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 Unported License (CC-BY-NC-ND; <http://creativecommons.org/licenses/by-nc-nd/3.0/legalcode>).

ACKNOWLEDGMENTS

The authors thank Drs. Nancy Bonini and Matthew Willmann for critical reading of the manuscript; Hetty Rodriguez for technical assistance; and Rebecca T. Cook for assistance with preparing digital artwork. This work was supported by grant IRG-78-002-30 from the American Cancer Society, the Penn Genome Frontiers Institute and a grant from the Pennsylvania Department of Health (B.D.G.), and a Burroughs-Wellcome Fund Career Award at the Scientific Interface (A.R.). The Department of Health specifically disclaims responsibility for any analyses, interpretations, or conclusions.

Received: June 27, 2011

Revised: September 26, 2011

Accepted: October 21, 2011

Published online: January 26, 2012

REFERENCES

Almeida, R., and Allshire, R.C. (2005). RNA silencing and genome regulation. *Trends Cell Biol.* 15, 251–258.

(B) Sequence alignment for the orthologous pair *FBtr0082962* (*Drosophila*, top) – *F54E12.3* (*C. elegans*, bottom) that has highly correlated secondary structure. Red and blue circles indicate positions determined to be paired or unpaired by our high-throughput, structure-mapping approach, respectively. This screenshot was obtained from our alignment browser at http://gregorylab.bio.upenn.edu/anno/structures/index_aln.php.

(C) mRNA secondary structure models for the same orthologous transcript pair (*FBtr0082962* (*Drosophila*, top) – *F54E12.3* (*C. elegans*, bottom)) determined using our methodology. The heatscale for each model indicates the normalized log-ratio of dsRNA-seq to ssRNA-seq reads (see [Experimental Procedures](#)) at each base position.

(D) (Top two boxes) The most significantly enriched biological process (and corresponding p value) for all orthologous transcript pairs with highly correlated secondary structure in *Drosophila* (top) and *C. elegans* (bottom). It is worth noting that no enrichment was identified for orthologous transcript pairs with significantly anti-correlated folding patterns. (Bottom two boxes) The most significantly enriched biological process (and corresponding p value) for orthologous transcript pairs that exhibit only highly similar nucleotide sequences in *Drosophila* (top) and *C. elegans* (bottom).

- Aravin, A.A., and Hannon, G.J. (2008). Small RNA silencing pathways in germ and stem cells. *Cold Spring Harb. Symp. Quant. Biol.* 73, 283–290.
- Bartel, D.P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 116, 281–297.
- Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigó, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., Thurman, R.E., et al; ENCODE Project Consortium; NISC Comparative Sequencing Program; Baylor College of Medicine Human Genome Sequencing Center; Washington University Genome Sequencing Center; Broad Institute; Children's Hospital Oakland Research Institute. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447, 799–816.
- Brierley, I., Pennell, S., and Gilbert, R.J. (2007). Viral RNA pseudoknots: versatile motifs in gene expression and replication. *Nat. Rev. Microbiol.* 5, 598–610.
- Buratti, E., Muro, A.F., Giombi, M., Gherbassi, D., Iaconcig, A., and Baralle, F.E. (2004). RNA folding affects the recruitment of SR proteins by mouse and human polypurinic enhancer elements in the fibronectin EDA exon. *Mol. Cell. Biol.* 24, 1387–1400.
- Carthew, R.W., and Sontheimer, E.J. (2009). Origins and mechanisms of miRNAs and siRNAs. *Cell* 136, 642–655.
- Celniker, S.E., Dillon, L.A., Gerstein, M.B., Gunsalus, K.C., Henikoff, S., Karpen, G.H., Kellis, M., Lai, E.C., Lieb, J.D., MacAlpine, D.M., et al; modENCODE Consortium. (2009). Unlocking the secrets of the genome. *Nature* 459, 927–930.
- Chekulaeva, M., and Filipowicz, W. (2009). Mechanisms of miRNA-mediated post-transcriptional regulation in animal cells. *Curr. Opin. Cell Biol.* 21, 452–460.
- Cooper, T.A., Wan, L., and Dreyfuss, G. (2009). RNA and disease. *Cell* 136, 777–793.
- Cruz, J.A., and Westhof, E. (2009). The dynamic landscapes of RNA architecture. *Cell* 136, 604–609.
- Czech, B., Malone, C.D., Zhou, R., Stark, A., Schlingeheyde, C., Dus, M., Perrimon, N., Kellis, M., Wohlschlegel, J.A., Sachidanandam, R., et al. (2008). An endogenous small interfering RNA pathway in *Drosophila*. *Nature* 453, 798–802.
- Fagegaltier, D., Bougé, A.L., Berry, B., Poisot, E., Sismeiro, O., Coppée, J.Y., Théodore, L., Voinnet, O., and Antoniewski, C. (2009). The endogenous siRNA pathway is involved in heterochromatin formation in *Drosophila*. *Proc. Natl. Acad. Sci. USA* 106, 21258–21263.
- Gerstein, M.B., Lu, Z.J., Van Nostrand, E.L., Cheng, C., Arshinoff, B.I., Liu, T., Yip, K.Y., Robilotto, R., Rechtsteiner, A., Ikegami, K., et al; modENCODE Consortium. (2010). Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. *Science* 330, 1775–1787.
- Ghildiyal, M., Seitz, H., Horwich, M.D., Li, C., Du, T., Lee, S., Xu, J., Kittler, E.L., Zapp, M.L., Weng, Z., and Zamore, P.D. (2008). Endogenous siRNAs derived from transposons and mRNAs in *Drosophila* somatic cells. *Science* 320, 1077–1081.
- Girard, A., and Hannon, G.J. (2008). Conserved themes in small-RNA-mediated transposon control. *Trends Cell Biol.* 18, 136–148.
- Graveley, B.R., Brooks, A.N., Carlson, J.W., Duff, M.O., Landolin, J.M., Yang, L., Artieri, C.G., van Baren, M.J., Boley, N., Booth, B.W., et al. (2011). The developmental transcriptome of *Drosophila melanogaster*. *Nature* 471, 473–479.
- Kawamura, Y., Saito, K., Kin, T., Ono, Y., Asai, K., Sunohara, T., Okada, T.N., Siomi, M.C., and Siomi, H. (2008). *Drosophila* endogenous small RNAs bind to Argonaute 2 in somatic cells. *Nature* 453, 793–797.
- Kertesz, M., Wan, Y., Mazor, E., Rinn, J.L., Nutter, R.C., Chang, H.Y., and Segal, E. (2010). Genome-wide measurement of RNA secondary structure in yeast. *Nature* 467, 103–107.
- Kharchenko, P.V., Alekseyenko, A.A., Schwartz, Y.B., Minoda, A., Riddle, N.C., Ernst, J., Sabo, P.J., Larschan, E., Gorchakov, A.A., Gu, T., et al. (2010). Comprehensive analysis of the chromatin landscape in *Drosophila melanogaster*. *Nature* 471, 480–485.
- Kim, V.N., Han, J., and Siomi, M.C. (2009). Biogenesis of small RNAs in animals. *Nat. Rev. Mol. Cell Biol.* 10, 126–139.
- Klasens, B.I., Das, A.T., and Berkhout, B. (1998). Inhibition of polyadenylation by stable RNA secondary structure. *Nucleic Acids Res.* 26, 1870–1876.
- Kouzarides, T. (2007). Chromatin modifications and their function. *Cell* 128, 693–705.
- Kozak, M. (2005). Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene* 361, 13–37.
- Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H., and Ecker, J.R. (2008). Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* 133, 523–536.
- Liu, T., Rechtsteiner, A., Egelhofer, T.A., Vielle, A., Latorre, I., Cheung, M.S., Ercan, S., Ikegami, K., Jensen, M., Kolasinska-Zwierz, P., et al. (2011). Broad chromosomal domains of histone modification patterns in *C. elegans*. *Genome Res.* 21, 227–236.
- Liu, Y., Taverna, S.D., Muratore, T.L., Shabanowitz, J., Hunt, D.F., and Allis, C.D. (2007). RNAi-dependent H3K27 methylation is required for heterochromatin formation and DNA elimination in *Tetrahymena*. *Genes Dev.* 21, 1530–1545.
- Lyne, R., Smith, R., Rutherford, K., Wakeling, M., Varley, A., Guillier, F., Janssens, H., Ji, W., McClaren, P., North, P., et al. (2007). FlyMine: an integrated database for *Drosophila* and *Anopheles* genomics. *Genome Biol.* 8, R129.
- Moazed, D. (2009). Small RNAs in transcriptional gene silencing and genome defence. *Nature* 457, 413–420.
- Roy, S., Ernst, J., Kharchenko, P.V., Kheradpour, P., Negre, N., Eaton, M.L., Landolin, J.M., Bristow, C.A., Ma, L., et al; modENCODE Consortium. (2010). Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science* 330, 1787–1797.
- Montange, R.K., and Batey, R.T. (2008). Riboswitches: emerging themes in RNA structure and function. *Annu. Rev. Biophys.* 37, 117–133.
- Okamura, K., Chung, W.J., Ruby, J.G., Guo, H., Bartel, D.P., and Lai, E.C. (2008). The *Drosophila* hairpin RNA pathway generates endogenous short interfering RNAs. *Nature* 453, 803–806.
- Raj, A., Peskin, C.S., Tranchina, D., Vargas, D.Y., and Tyagi, S. (2006). Stochastic mRNA synthesis in mammalian cells. *PLoS Biol.* 4, e309.
- Raj, A., van den Bogaard, P., Rifkin, S.A., van Oudenaarden, A., and Tyagi, S. (2008). Imaging individual mRNA molecules using multiple singly labeled probes. *Nat. Methods* 5, 877–879.
- Raker, V.A., Mironov, A.A., Gelfand, M.S., and Pervouchine, D.D. (2009). Modulation of alternative splicing by long-range RNA structures in *Drosophila*. *Nucleic Acids Res.* 37, 4533–4544.
- Ruby, J.G., Jan, C., Player, C., Axtell, M.J., Lee, W., Nusbaum, C., Ge, H., and Bartel, D.P. (2006). Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. *Cell* 127, 1193–1207.
- Ruby, J.G., Stark, A., Johnston, W.K., Kellis, M., Bartel, D.P., and Lai, E.C. (2007). Evolution, biogenesis, expression, and target predictions of a substantially expanded set of *Drosophila* microRNAs. *Genome Res.* 17, 1850–1864.
- Sharp, P.A. (2009). The centrality of RNA. *Cell* 136, 577–580.
- Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15, 1034–1050.
- Sijen, T., and Plasterk, R.H. (2003). Transposon silencing in the *Caenorhabditis elegans* germ line by natural RNAi. *Nature* 426, 310–314.
- Underwood, J.G., Uzilov, A.V., Katzman, S., Onodera, C.S., Mainzer, J.E., Mathews, D.H., Lowe, T.M., Salama, S.R., and Haussler, D. (2010). FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. *Nat. Methods* 7, 995–1001.
- Vargas, D.Y., Raj, A., Marras, S.A., Kramer, F.R., and Tyagi, S. (2005). Mechanism of mRNA transport in the nucleus. *Proc. Natl. Acad. Sci. USA* 102, 17008–17013.

- Volpe, T.A., Kidner, C., Hall, I.M., Teng, G., Grewal, S.I., and Martienssen, R.A. (2002). Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. *Science* 297, 1833–1837.
- Warf, M.B., and Berglund, J.A. (2010). Role of RNA structure in regulating pre-mRNA splicing. *Trends Biochem. Sci.* 35, 169–178.
- Watanabe, T., Totoki, Y., Toyoda, A., Kaneda, M., Kuramochi-Miyagawa, S., Obata, Y., Chiba, H., Kohara, Y., Kono, T., Nakano, T., et al. (2008). Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature* 453, 539–543.
- Westhof, E., and Romby, P. (2010). The RNA structureome: high-throughput probing. *Nat. Methods* 7, 965–967.
- Zarudnaya, M.I., Kolomiets, I.M., Potyahaylo, A.L., and Hovorun, D.M. (2003). Downstream elements of mammalian pre-mRNA polyadenylation signals: primary, secondary and higher-order structures. *Nucleic Acids Res.* 31, 1375–1386.
- Zheng, Q., Ryvkin, P., Li, F., Dragomir, I., Valladares, O., Yang, J., Cao, K., Wang, L.S., and Gregory, B.D. (2010). Genome-wide double-stranded RNA sequencing reveals the functional significance of base-paired RNAs in *Arabidopsis*. *PLoS Genet.* 6, e1001141.
- Zisoulis, D.G., Lovci, M.T., Wilbert, M.L., Hutt, K.R., Liang, T.Y., Pasquinelli, A.E., and Yeo, G.W. (2010). Comprehensive discovery of endogenous Argonaute binding sites in *Caenorhabditis elegans*. *Nat. Struct. Mol. Biol.* 17, 173–179.

EXTENDED RESULTS

The smRNA Component of the *Drosophila* and *C. elegans* Transcriptomes

We utilized smRNA-seq to characterize the populations of smRNA molecules from *Drosophila melanogaster* culture cells and *Caenorhabditis elegans* mixed stage N2 worms. Using this approach, we obtained a total of 7,597,106 and 10,594,321 smRNA reads from the *Drosophila* and *C. elegans* libraries, respectively. We found that a majority of smRNA sequencing reads map to miRNAs, consistent with previous findings that miRNAs are abundant and functionally important in these animals (Ambros, 2004; Bushati and Cohen, 2007; Carthew and Sontheimer, 2009). We also found that the TE-mapping reads are very abundant in the *Drosophila* smRNA-seq library, in contrast to the *C. elegans* smRNA-seq library (Figures 1E and 1F), again suggesting that transposons in *Drosophila* S2 cells are very active for producing smRNAs.

To obtain highly confident subsets of smRNA components, we also systematically identified smRNA hot spots of both *Drosophila* and *C. elegans* genomes using a modified version of a Poisson distribution-based statistical approach (Heisel et al., 2008). As a result, we revealed 13,983 and 2,099 smRNA hot spots within the *Drosophila* and *C. elegans* genomes, respectively. Interestingly, we found that the majority of smRNA hot spots in *Drosophila* are located in TEs, whereas in *C. elegans* many smRNA hot spots also come from protein-coding transcripts and all classes of noncoding RNAs (Figures 1E and 1F). These findings are in complete correspondence with our dsRNA hot spot analysis (Figures 1E and 1F), and provide further support for the hypothesis that TE-derived dsRNA precursors and their corresponding smRNA products are critical effectors of RNA silencing pathways directed at silencing transposons and repetitive elements in *Drosophila*.

EXTENDED EXPERIMENTAL PROCEDURES

Animal Materials

Drosophila DL1 culture cells and *C. elegans* mixed stage N2 worms were used for all experiments in this study.

dsRNA-seq Library Preparation

40 μ g of total RNA (13.33 μ g from each of three biological replicates) was subjected to two rounds (1X Ribominus) of rRNA depletion as per manufacturer's instructions (Ribominus, Invitrogen (Carlsbad, CA)). Next, these rRNA-depleted RNA samples were treated with a single-strand specific ribonuclease as per manufacturer's instructions (RNase One, Promega (Madison, WI)) in structure buffer (10mM Tris pH7, 100mM KCl, 10mM MgCl₂). dsRNA was then purified using a phenol:chloroform extraction. The purified dsRNA sample was subjected to a fragmentation reaction (Fragmentation Reagents, Applied Biosystems (Foster City, CA)) as per manufacturer's instructions. To resolve the dsRNAs after single-stranded RNase treatment and fragmentation, they were treated with T4 polynucleotide kinase (T4 PNK, New England Biolabs (Cambridge, MA)) as previously described (Wang and Shuman, 2002). The fragmented RNA sample was then used as the substrate for sequencing library construction using the Small RNA Sample Prep v1.5 kit (Illumina, San Diego, CA) as per manufacturer's instructions.

ssRNA-seq Library Preparation

40 μ g of total RNA (13.33 μ g from each of three biological replicates) was subjected to two rounds (1X Ribominus) of rRNA depletion as per manufacturer's instructions (Ribominus, Invitrogen (Carlsbad, CA)). Next, these rRNA-depleted RNA samples were treated with a double-strand specific ribonuclease as per manufacturer's instructions (RNase V1, Applied Biosystems, Foster City, CA) in structure buffer (10 mM Tris [pH 7], 100 mM KCl, 10 mM MgCl₂). ssRNA was then purified using a phenol:chloroform extraction. The purified ssRNA sample was subjected to a fragmentation reaction (Fragmentation Reagents, Applied Biosystems (Foster City, CA)) as per manufacturer's instructions. To resolve the ssRNAs after double-stranded RNase treatment and fragmentation, they were treated with T4 polynucleotide kinase (T4 PNK, New England Biolabs (Cambridge, MA)) as previously described (Wang and Shuman, 2002). The fragmented RNA sample was then used as the substrate for sequencing library construction using the Small RNA Sample Prep v1.5 kit (Illumina, San Diego, CA) as per manufacturer's instructions.

High-Throughput Sequencing and Sequence Read Processing and Mapping

dsRNA-, ssRNA-, and smRNA-seq libraries were sequenced using the Illumina Genome Analyzer IIx (GAIIx) and HiSeq2000 as per manufacturer's instructions (Illumina Inc., San Diego, CA). Sequence information was extracted from the image files with the Illumina (San Diego, CA.) base calling software package. Prior to alignment, the sequencing reads were reduced to a list of nonredundant (NR) sequences to minimize the computational requirement in all following procedures. Then, NR-sequences for which a 3' adaptor sequence was observed were truncated up to the junction with the adaptor sequence. The dsRNA-seq, ssRNA-seq, and smRNA-seq reads were then aligned to the corresponding *Drosophila* (UCSC dm3 assembly) or *C. elegans* (UCSC ce6 assembly) genome.

Balanced Preprocessing Pipeline for Mapping dsRNA-seq/ssRNA-seq/smRNA-seq Reads

Our sequencing libraries contained a significant portion of reads in which no 3'-adaptor sequences can be found. To maintain sequence reads that both had discernible 3'-adaptor sequences (short reads), as well as reads without 3'-adaptors, a balanced

pipeline was developed by dividing reads into “trimmed” and “untrimmed” categories according to whether they have detectable 3'-adaptor sequences or not, respectively. To begin, all reads were reduced to nonredundant (NR) sequences to minimize the computational requirement for subsequent analysis steps. Then, in order to detect 3'-adaptor sequences, all NR-sequences were aligned to the Illumina 3'-adaptor version 1.5 sequence using the “cross-match” program from the Phrap/Cross_match/Swat package (<http://www.phrap.org/phredphrapconsed.html>). The alignment parameters for cross-match were carefully tuned to maintain all alignments with less than 6% mismatches. The cross-match alignment results were parsed using in-house Perl scripts. All NR-sequences that aligned to ≥ 6 bp of the 3'-adaptor sequence at their 3' end were defined as short reads (with “detectable” 3'-adaptor sequence) and were subsequently trimmed at the adaptor-sequence boundary. The remaining NR-sequences (without detectable 3'-adaptors) remained “untrimmed.”

All trimmed and untrimmed NR-sequences were then aligned to their respective genomes (dm3 for *D. melanogaster* and ce6 for *C. elegans*) using cross-match, again with the parameters set to maintain all alignments at $\leq 6\%$ mismatches. Alignment results for trimmed or untrimmed inputs were then parsed independently using in-house Perl scripts. More specifically, alignments for trimmed sequences were required to extend to the ends of the query sequences, whereas alignments for untrimmed sequences were only required to extend to the imaginary positions of undetectable 3'-adaptors (<6 bp from the 3' end of the sequence). The true lengths of the untrimmed sequences were also determined in this step by computing the most-frequent aligned length of all possible alignments to the respective genome. As an additional alignment step, unmapped untrimmed sequences were forcibly trimmed to 60nt and then realigned to the respective genome; this last step was motivated by the empirical observation that a large population of cloned inserts was present at approximately 60nt. Finally, all trimmed and untrimmed NR-sequences as well as their genomic loci information were combined to form the final dataset using in-house Perl scripts. The dsRNA-, ssRNA-, and smRNA-seq libraries were all independently processed using this balanced pipeline.

Evaluation of Sequencing Coverage by dsRNA-seq and ssRNA-seq

To evaluate the genome-wide coverage of structured and unstructured regions by the dsRNA-seq and ssRNA-seq methodologies, respectively, we randomly sampled 19 subsets containing 95% to 5% of total mapped dsRNA-seq or ssRNA-seq reads (merged between GAllx and HiSeq2000 runs) in 5% increments. The genomic locations of these subsets of mapped reads were also filtered from the complete set of reference loci. Finally, the base coverage (unit: bp) for all random subsets, as well as the reference total dataset, was calculated for each class of RNA molecule (e.g., rRNA, tRNA, etc.) and for the overall genome. The relative base coverage was defined as the fraction of bases covered in each subset compared to the total covered bases of the reference whole dataset (Figures S1A, S1C, S1E, and S1G). The relative base coverage of all dsRNA and ssRNA hot spots was calculated by the same method, with the exception that only reads located in hot spots were used in the analysis (Figures S1B, S1D, S1F, and S1H).

Estimating the False Discovery Rate of dsRNA-seq and ssRNA-seq

The actual false discovery rates (FDRs) of dsRNA-seq and ssRNA-seq are determined by the enzyme efficiency of the RNases used (here we used RNaseONE and RNase V1). However, we could get an estimate of the upper bound of the FDR values by evaluating the proportion of dsRNA reads that have complementary reads in the same library (and ssRNA reads that do not have complementary reads in the same library). To begin, we aligned all the genome-mapped reads back to themselves using the NCBI-BLASTN program; the word-size parameter was set to 6 to improve the searching sensitivity of short Illumina reads. By seeking reads with complementary reads (align-length $\geq 50\%$ of the read, identity $\geq 85\%$, mismatches $\leq 10\%$, and gaps $\leq 5\%$), we then defined the true dsRNA sequences as those with one or more complementary reads in the same library. Conversely, true ssRNA sequences were defined as those with no complementary reads in the same library. It is noteworthy that the FDR values calculated by this method are very likely to be an overestimate, since the random fragmentation step in the dsRNA-seq assay can often result in the inability to obtain a complementary read with $\geq 50\%$ overlap even if both reads of the complementary pair are present in the sequencing library. Furthermore, for the lowly expressed dsRNA molecules we could easily sequence only one member of the complementary pair by random chance, while missing the other. Therefore, our estimated FDR for dsRNA-seq likely serves as a conservative upper bound for the technology. Reads that did not meet the respective criteria for dsRNA-seq and ssRNA-seq were discarded for all subsequent steps.

Classification and Characterization of Sequencing Reads

To classify dsRNA-seq, ssRNA-seq, and smRNA-seq reads, GFF-formatted annotation files for all *D. melanogaster* and *C. elegans* genetic elements (protein-coding mRNAs, all noncoding RNAs (rRNAs, tRNAs, miRNAs, pseudogenes, transposable elements, etc.)) were downloaded from FlyBase (r5.22) and WormBase (WS205), respectively. Annotations were then reformatted using in-house Perl scripts and loaded into a local MySQL database. NR-sequences were then classified and annotated according to their genomic locations. It is of note that some of our sequencing reads overlap multiple genetic elements and were correspondingly counted toward all pertinent genetic elements for classification purposes.

Identification of dsRNA Hot Spots in *D. melanogaster* and *C. elegans*

To identify dsRNA hot spots in the *D. melanogaster* and *C. elegans* genomes, dsRNA-seq reads were used to identify contiguous dsRNAs (dsRNA contigs), as well as the remaining ssRNA-regions between these base-paired regions. Then, the lengths for dsRNA and ssRNA regions on each chromosome are assumed to both follow a geometric distribution,

$$P(X_i = k) = (1 - P_i)^{k-1} p_i$$

$$P(X_i \leq k) = 1 - (1 - p_i)^k,$$

where X_i is the dsRNA/ssRNA region length and P_i is the probability for a DNA base to be in dsRNA/ssRNA status for chromosome k . This assumption holds as long as P_i is constant for every chromosome and the dsRNA/ssRNA status of a DNA base is only dependent on the status of its proceeding base. Using this model, the lengths of dsRNA/ssRNA regions X_i are fitted, the parameters P_i are estimated, and the confidence intervals of dsRNA/ssRNA region lengths are determined for each chromosome. Finally, dsRNA hot spots are identified by selecting “merged” dsRNA contigs with lengths that are longer than statistically expected for that chromosome (see Table S1 for confidence intervals for all *D. melanogaster* and *C. elegans* chromosomes used in this analysis). It is of note that “merged” dsRNA regions are combined dsRNA contigs that are separated by ssRNA regions that are shorter than statistically expected as determined by the confidence intervals for ssRNA on that chromosome.

Identification of ssRNA Hot Spots

Identification of ssRNA hot spots was performed as for dsRNA hot spots, but with the roles of dsRNA and ssRNA reversed. In other words, ssRNA hot spots have a length that is longer than statistically expected, and are separated by dsRNA regions that are shorter than statistically expected (see Table S1).

Identification of smRNA Hot Spots

smRNA hot spots were identified differently than the dsRNA and ssRNA hot spots in order to take into account expression abundance information of all sequenced smRNA molecules. To begin, consecutive smRNAs were identified on each chromosome and then pre-grouped into smRNA clusters (smRNA contigs). Next, a derived “per-smRNA site” abundance (PSS-abundance) was calculated for all smRNA clusters as $N_r/L_c \times \bar{X}_s$, where N_r and L_c are the total number of cloned reads and length for this smRNA cluster, respectively, and \bar{X}_s stands for the average length of all smRNA reads. Then the derived PSS-abundance on each chromosome is assumed to follow a Poisson distribution:

$$P(X_i = k) = \frac{\lambda_i^k}{k!} \cdot e^{-\lambda_i}$$

$$P(X_i \leq k) = e^{-\lambda_i} \sum_{j=0}^k \frac{\lambda_i^j}{j!} = \frac{\Gamma([k+1], \lambda_i)}{[k]!}$$

where X_i is the derived PSS-abundance and λ_i is the expected number of smRNA reads per smRNA-site on chromosome k . Thus, the derived PSS-abundance data are fitted to this Poisson distribution model, the parameters λ_i are estimated, and the confidence intervals for PSS-abundance of all smRNA clusters are estimated for each chromosome. Finally, smRNA hot spots were identified as smRNA clusters with significantly high PSS-abundance.

Merging of GAllx and HiSeq2000 Data

For all subsequent analyses, mapped NR-sequences, genomic loci, and dsRNA and ssRNA hot spots were merged between the GAllx and HiSeq2000 runs. Hot spot identification was performed prior to merging due to the different read lengths obtained.

Classification and Functional Characterization of Hot Spots

dsRNA, ssRNA, and smRNA hot spots were classified and annotated in the same way as sequencing reads (see above). Propensity for dsRNA hot spots to generate smRNAs was examined by finding smRNA hot spots either contained within or partly overlapping with dsRNA hot spots.

Various histone modification ChIP-seq (for *D. melanogaster*) and ChIP-chip (for *C. elegans*) data were downloaded from modENCODE (<http://www.modencode.org>). Specific datasets are listed in Table S4. For ChIP-seq data, genomic intervals from modENCODE were directly uploaded to a local MySQL database. For ChIP-chip data, ChIPOTie v1.11 (Buck et al., 2005) was used to identify genomic intervals of enriched histone modifications. Genomic intervals of significantly enriched histone modifications were then overlapped with the locations of dsRNA and ssRNA hot spots.

Comparative genomics analysis of dsRNA and ssRNA hot spots was performed as previously described (Zheng et al., 2010), except using “insects-15-way” and “worm-6-way” multiple alignments. The insects-15-way genomes used were *D. melanogaster* (dm3), *D. simulans* (droSim1, Apr. 2005), *D. sechellia* (droSec1, Oct. 2005), *D. yakuba* (droYak2, Nov. 2005), *D. erecta* (droEre2, Feb. 2006), *D. ananassae* (droAna3, Feb. 2006), *D. pseudoobscura* (dp4, Feb. 2006), *D. persimilis* (droPer1, Oct. 2005), *D. willistoni* (droWil1, Feb. 2006), *D. virilis* (droVir3, Feb. 2006), *D. mojavensis* (droMoj3, Feb. 2006), *D. grimshawi* (droGri2, Feb. 2006), *A. gambiae* (anoGam1, Feb. 2003), *A. mellifera* (apiMel3, May 2005), and *T. castaneum* (triCas2, Sep. 2005). The worm-6-way genomes used were *C. elegans* (ce6, May 2008), *C. remanei* (ceaRem3, May 2007), *C. briggsae* (cb3, Jan 2007), *C. brenneri* (caePb1, Feb 2008), *C. japonica* (caeJap1, Mar 2008), and *P. pacificus* (priPac1, May 2007).

Identification and Characterization of Base-Paired Transcripts

Newly identified base-paired transcripts were identified as dsRNA hot spots with no known annotation from either FlyBase (r5.22) or WormBase (WS205), nor any overlap with exons identified in a series of recent transcriptome profiling experiments (Gerstein et al., 2010; Graveley et al., 2011). A full breakdown of newly identified transcripts is given in Tables S2 and S3. RT-PCR analyses of *D. melanogaster* dsRNA hot spots were performed as previously described (Zheng et al., 2010).

FISH

In preparation for FISH on *C. elegans*, we harvested embryos and larvae from synchronized and unsynchronized cultures of N2 worms. We fixed, permeabilized, and performed single molecule FISH on *C. elegans* embryos and larvae as previously described (Raj et al., 2010; Raj et al., 2008). We determined the concentration of probe empirically, ending up with roughly the same concentration per fluorescently labeled oligonucleotide as used previously (Raj et al., 2010; Raj et al., 2008).

Analysis of Secondary Structure in miRNA-Mediated Gene Regulation

ALG-1 binding sites were downloaded from the UCSC Genome Browser (Zisoulis et al., 2010), and miRNA target sites were then predicted within these binding sites using TargetScan (v5.0) (http://www.targetscan.org/worm_12/). Average structure scores were calculated as previously described in the Experimental Procedures section.

Folding and Analysis of mRNA Secondary Structure Correlation

Orthologous transcript pairs were downloaded from FlyMine v26.0 (Comparative Genomics module) (<http://www.flymine.org/>) and aligned using the EMBOSS package's water program. A filtered set of 2,223 orthologs was obtained by requiring at least 10% of all aligned base positions to be called as paired or unpaired in both structures (i.e., log-ratio of dsRNA/ssRNA-seq reads to be greater than 1.1 or less than -1.1). To compute the structure correlation, we first generated vectors of structure scores (log-ratio of dsRNA/ssRNA-seq reads) for each transcript at nongap positions in the respective alignment. Next, we turned these structure scores into a paired/unpaired profile using the above thresholds of 1.1 and -1.1 . Given these profiles for a pair of orthologous transcripts, we can then count the number of positions at which both transcripts are paired or unpaired (num_same), as well as the number of positions at which one of the orthologs is paired and the other is unpaired or vice versa (num_opposite). To test for significantly correlated or anticorrelated pairs, we used a binomial model Bin(n, p) as the null distribution. The parameter p was taken as equal to the sequence similarity for the given ortholog pair in order to account for sequence effects (i.e., tendency for more similar sequences to show higher structure correlation). Significance values were then calculated using the R package's binom.test() function. Finally, as a plotable score for structure correlation (see Figure 7A, y axis), we used the following: $S = (2 * \text{num_same} / (\text{num_same} + \text{num_opposite})) - 1$, $-1 \leq S \leq 1$.

dsRNA and ssRNA RT-PCR Analysis

RNaseONE ssRNase digestion (dsRNA selection) was performed on three 20 μg total RNA samples from *Drosophila* DL1 culture cells as per manufacturer's instructions. Following digestion, these three samples were pooled together and purified using a phenol:chloroform extraction. To obtain ssRNA, a dsRNase digestion (RNase V1, (Ambion, Foster City, CA)) was carried on three 20 μg total RNA samples from *Drosophila* DL1 culture cells as per manufacturer's instructions. Following digestion, these three samples were pooled together and purified using a phenol:chloroform extraction. Random-primed RT PCR analyses were performed on these digested samples using primers listed in Table S6. This experiment was repeated three times and a representative example can be seen in Figure 5C.

AnnoJ and RNA Structure Browser

The AnnoJ Genome Browser is a REST-based genome annotation visualization program built using Web 2.0 technology. Licensing information and documentation are available at <http://www.annoj.org>. We have developed a structure browser enhancement for AnnoJ that enables visualization of the mRNA secondary structure models produced as described above. To do this, each predicted structural model was rendered as a SVG plot using Vienna (<http://www.tbi.univie.ac.at/~ivo/RNA/>) RNAplot. Reads and other features of interest such as validated regions for mRNAs were then added to the SVG file. Users can visualize the models of secondary structure for an annotated transcript by selecting the corresponding genomic interval on AnnoJ (RNA structures track) or by entering its accession number. Ortholog alignments can also be visualized on the Ortholog alignments track.

SUPPLEMENTAL REFERENCES

- Ambros, V. (2004). The functions of animal microRNAs. *Nature* 431, 350–355.
- Buck, M.J., Nobel, A.B., and Lieb, J.D. (2005). ChIPOTile: a user-friendly tool for the analysis of ChIP-chip data. *Genome Biol.* 6, R97.
- Bushati, N., and Cohen, S.M. (2007). microRNA functions. *Annu. Rev. Cell Dev. Biol.* 23, 175–205.
- Carthew, R.W., and Sontheimer, E.J. (2009). Origins and Mechanisms of miRNAs and siRNAs. *Cell* 136, 642–655.
- Gerstein, M.B., Lu, Z.J., Van Nostrand, E.L., Cheng, C., Arshinoff, B.I., Liu, T., Yip, K.Y., Robilotto, R., Rechtsteiner, A., Ikegami, K., et al; modENCODE Consortium. (2010). Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. *Science* 330, 1775–1787.

- Graveley, B.R., Brooks, A.N., Carlson, J.W., Duff, M.O., Landolin, J.M., Yang, L., Artieri, C.G., van Baren, M.J., Boley, N., Booth, B.W., et al. (2011). The developmental transcriptome of *Drosophila melanogaster*. *Nature* 471, 473–479.
- Heisel, S.E., Zhang, Y., Allen, E., Guo, L., Reynolds, T.L., Yang, X., Kovalic, D., and Roberts, J.K. (2008). Characterization of unique small RNA populations from rice grain. *PLoS ONE* 3, e2871.
- Raj, A., Rifkin, S.A., Andersen, E., and van Oudenaarden, A. (2010). Variability in gene expression underlies incomplete penetrance. *Nature* 463, 913–918.
- Raj, A., van den Bogaard, P., Rifkin, S.A., van Oudenaarden, A., and Tyagi, S. (2008). Imaging individual mRNA molecules using multiple singly labeled probes. *Nat. Methods* 5, 877–879.
- Wang, L.K., and Shuman, S. (2002). Mutational analysis defines the 5'-kinase and 3'-phosphatase active sites of T4 polynucleotide kinase. *Nucleic Acids Res.* 30, 1073–1080.
- Zheng, Q., Ryzkin, P., Li, F., Dragomir, I., Valladares, O., Yang, J., Cao, K., Wang, L.S., and Gregory, B.D. (2010). Genome-wide double-stranded RNA sequencing reveals the functional significance of base-paired RNAs in *Arabidopsis*. *PLoS Genet.* 6, 6.
- Zisoulis, D.G., Lovci, M.T., Wilbert, M.L., Hutt, K.R., Liang, T.Y., Pasquinelli, A.E., and Yeo, G.W. (2010). Comprehensive discovery of endogenous Argonaute binding sites in *Caenorhabditis elegans*. *Nat. Struct. Mol. Biol.* 17, 173–179.

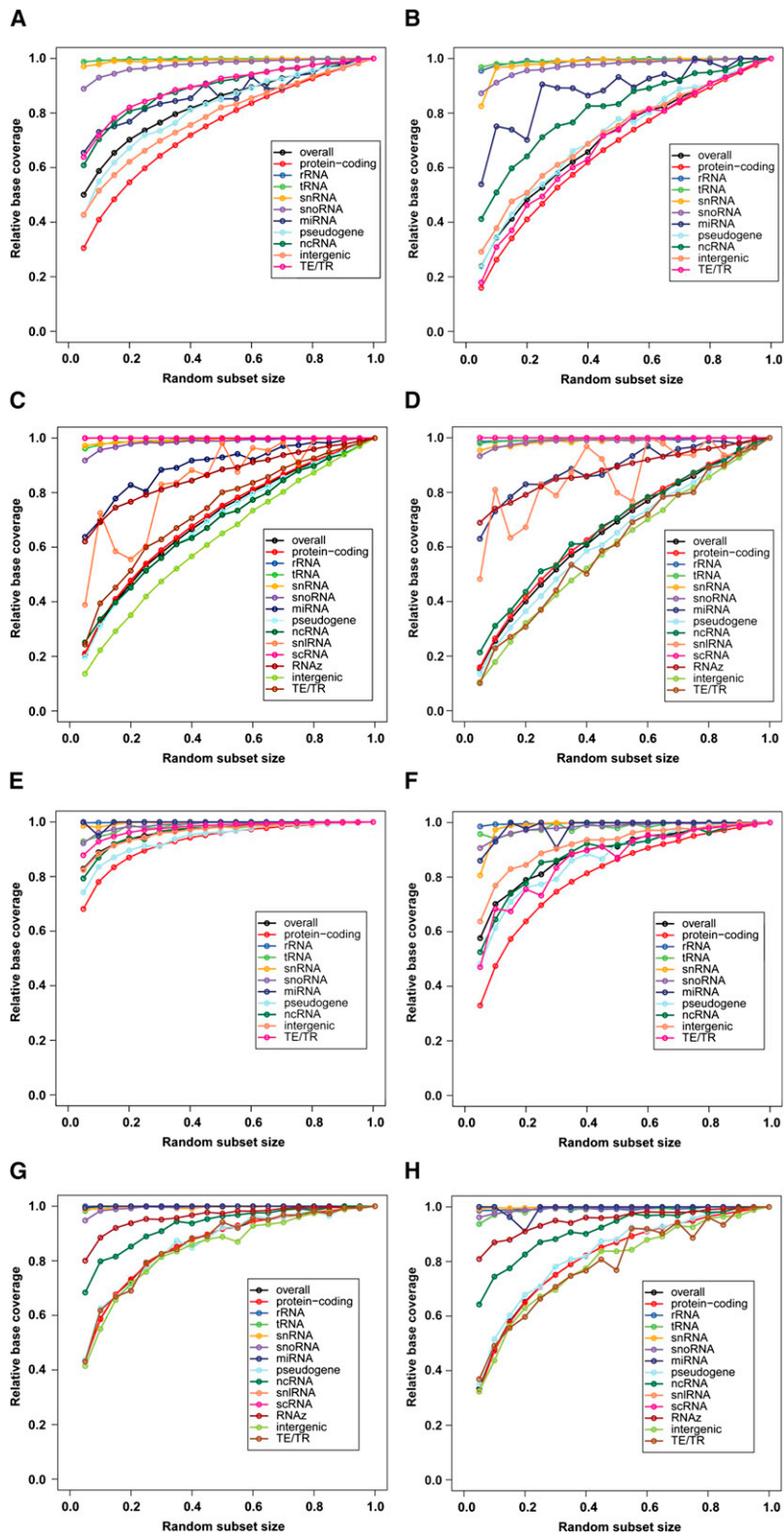


Figure S1. Saturation Analysis of dsRNA-seq and ssRNA-seq; Related to Figures 1 and 2

(A) The relative dsRNA-seq coverage overall (black line) and for 10 classes of RNA molecules (colored lines as specified in legend) as the library subset size changes from 5% to 95% (in 5% increments) for *Drosophila* dsRNA-seq data.

-
- (B) The relative highly base-paired RNA (dsRNA hot spot) coverage overall (black line) and for 10 classes of RNA molecules (colored lines as specified in legend) as the library subset size changes from 5% to 95% (in 5% increments) for *Drosophila*.
- (C) The relative dsRNA-seq coverage overall (black line) and for 13 classes of RNA molecules (colored lines as specified in legend) as the library subset size changes from 5% to 95% (in 5% increments) for *C. elegans* dsRNA-seq data.
- (D) The relative highly base-paired RNA (dsRNA hot spot) coverage overall (black line) and for 13 classes of RNA molecules (colored lines as specified in legend) as the library subset size changes from 5% to 95% (in 5% increments) for *C. elegans*.
- (E) The relative ssRNA-seq coverage overall (black line) and for 10 classes of RNA molecules (colored lines as specified in legend) as the library subset size changes from 5% to 95% (in 5% increments) for *Drosophila* ssRNA-seq data.
- (F) The relative highly unpaired RNA (ssRNA hot spot) coverage overall (black line) and for 10 classes of RNA molecules (colored lines as specified in legend) as the library subset size changes from 5% to 95% (in 5% increments) for *Drosophila*.
- (G) The relative ssRNA-seq coverage overall (black line) and for 13 classes of RNA molecules (colored lines as specified in legend) as the library subset size changes from 5% to 95% (in 5% increments) for *C. elegans* ssRNA-seq data.
- (H) The relative highly unpaired RNA (ssRNA hot spot) coverage overall (black line) and for 13 classes of RNA molecules (colored lines as specified in legend) as the library subset size changes from 5% to 95% (in 5% increments) for *C. elegans*.

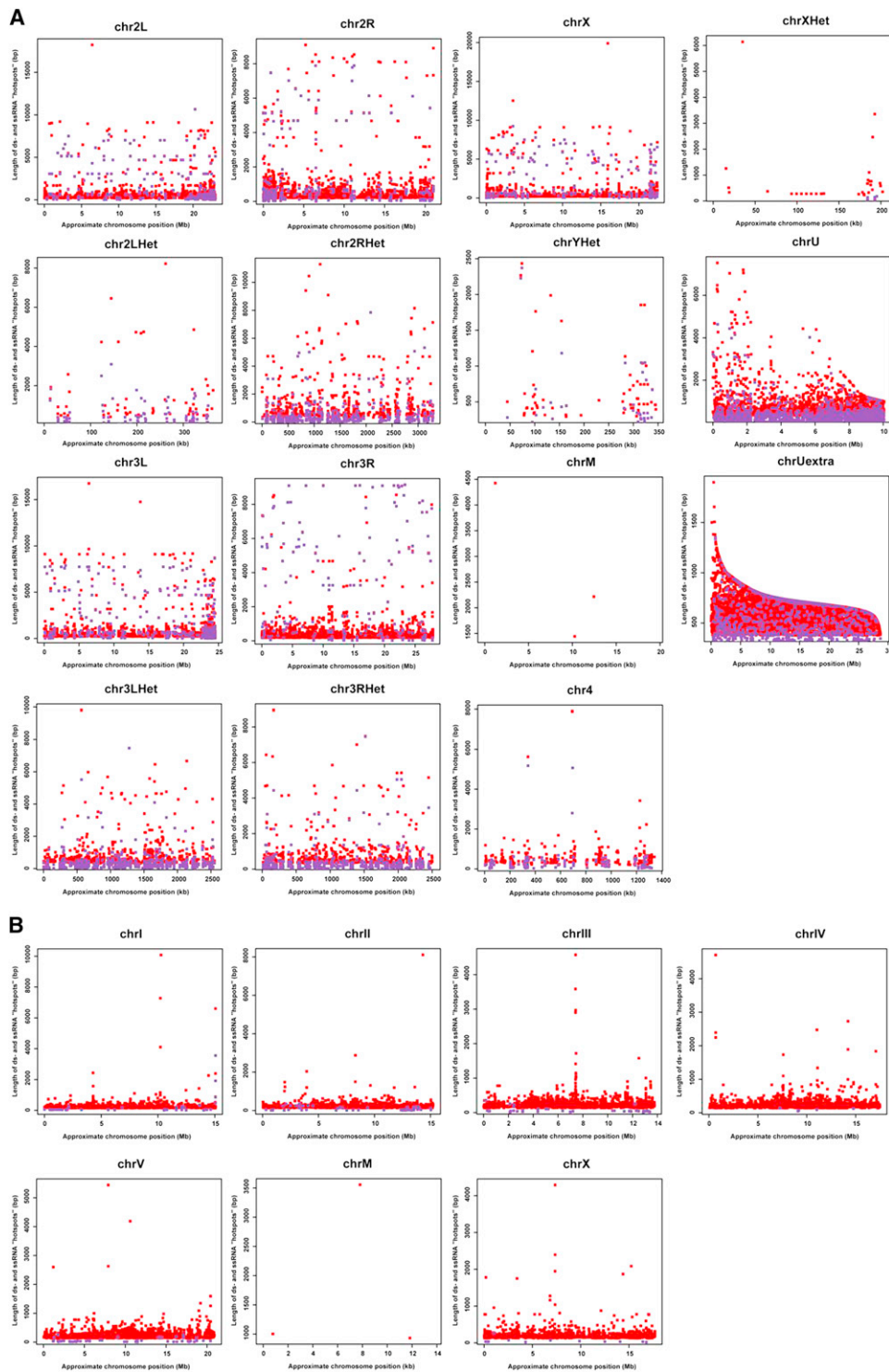


Figure S2. Chromosomal Distribution of dsRNA and ssRNA Hot Spots; Related to Figures 1, 2, 3, and 4

(A) The distribution of dsRNA (red) and ssRNA (purple) hot spots along the length of all *Drosophila* chromosomes. Red and purple dots denote specific hot spots. (B) The distribution of dsRNA (red) and ssRNA (purple) hot spots along the length of all *C. elegans* chromosomes. Red and purple dots denote specific hot spots.

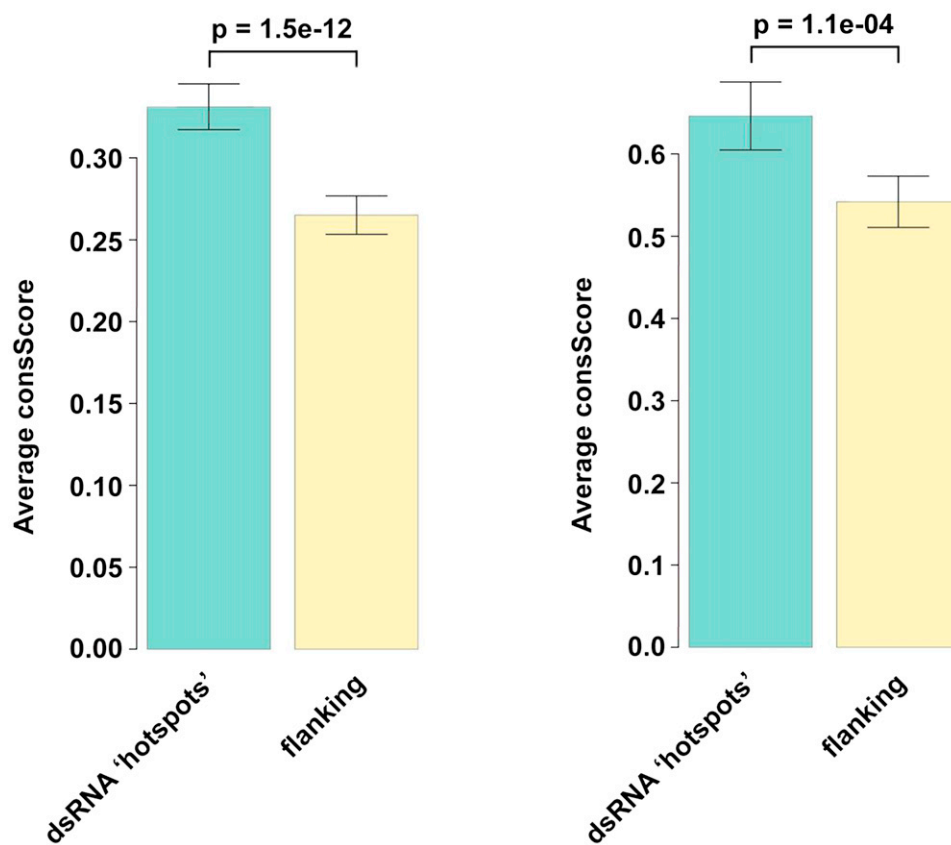


Figure S3. Identification of Highly Conserved RNAs in *Drosophila* and *C. elegans*; Related to Figures 3 and 4

(A and B) The average conservation scores (consScore) calculated using a comparative genomics analysis (see [Extended Experimental Procedures](#)) of dsRNA hot spots (green bars) or their flanking regions (yellow bars) in intergenic regions of (A) *Drosophila* or (B) *C. elegans*.

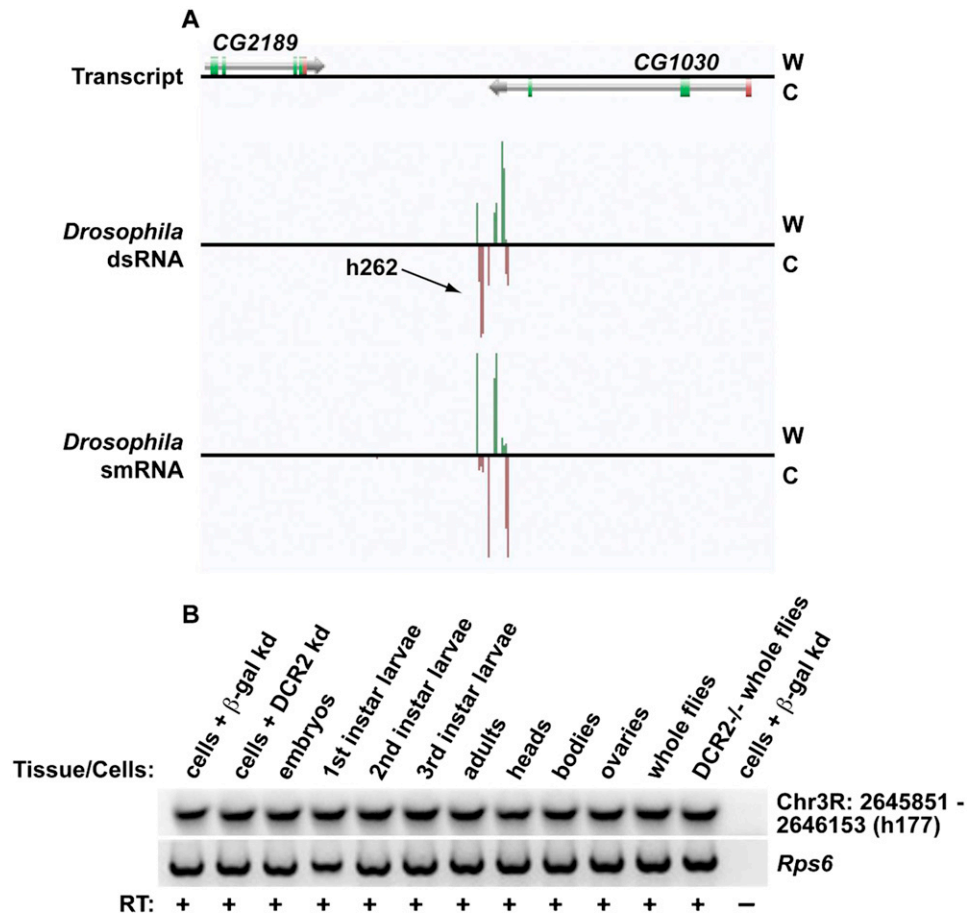


Figure S4. Identification of Highly Structured RNAs in *Drosophila*; Related to Figure 3

(A and B) Another example of an intergenic, highly base-paired transcript (screenshots from our *Drosophila* RNA-seq browser, http://tesla.pcbi.upenn.edu/anno_j_dm/). W (green bars) and C (red bars) indicate signal from Watson and Crick strands, respectively. (A) An intergenic dsRNA hot spot found between CG2189 and CG1030. (B) Random-primed RT-PCR analysis of base-paired RNAs in multiple tissues and developmental stages of *Drosophila*. *Rps6* serves as a loading control.

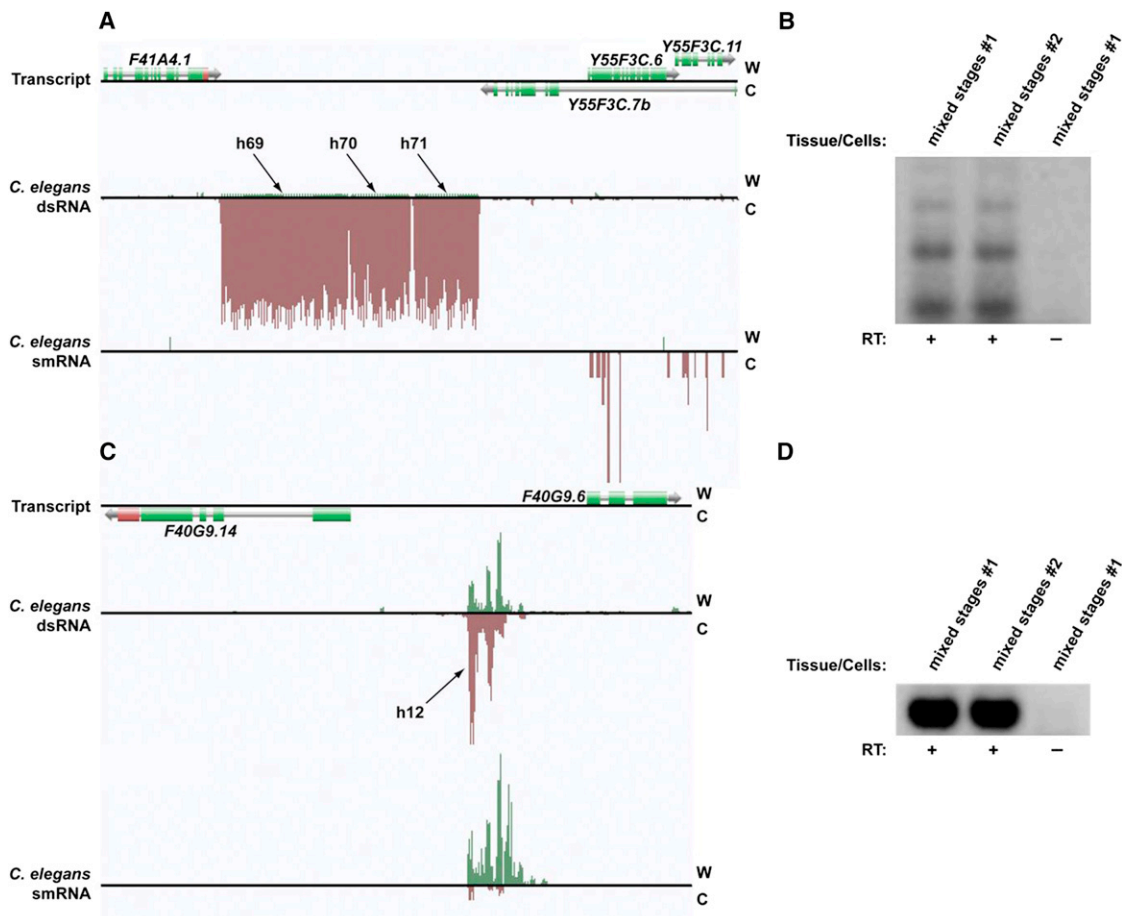


Figure S5. Identification of Highly Structured RNAs in *C. elegans*; Related to Figure 4

(A and C) Two examples of intergenic, highly base-paired transcripts (screenshots from our *C. elegans* RNA-seq browser, http://tesla.pcbi.upenn.edu/annoj_ce/). W (green bars) and C (red bars) indicate signal from Watson and Crick strands, respectively. (A) Three intergenic dsRNA hot spots found between *F41A4.1* and *Y55F3C.7b* (h69 – h71). (C) A highly base-paired transcript found between *F40G9.14* and *F40G9.6*. (B) Random-primed RT-PCR analysis of a base-paired RNA (h69) from mixed stage *C. elegans* that is pictured in (A). (D) Random-primed RT-PCR analysis of a new transcript (h12) from mixed stage *C. elegans* that is pictured in (C). These transcripts were also recently identified via high-throughput RNA profiling (Gerstein et al., 2010).

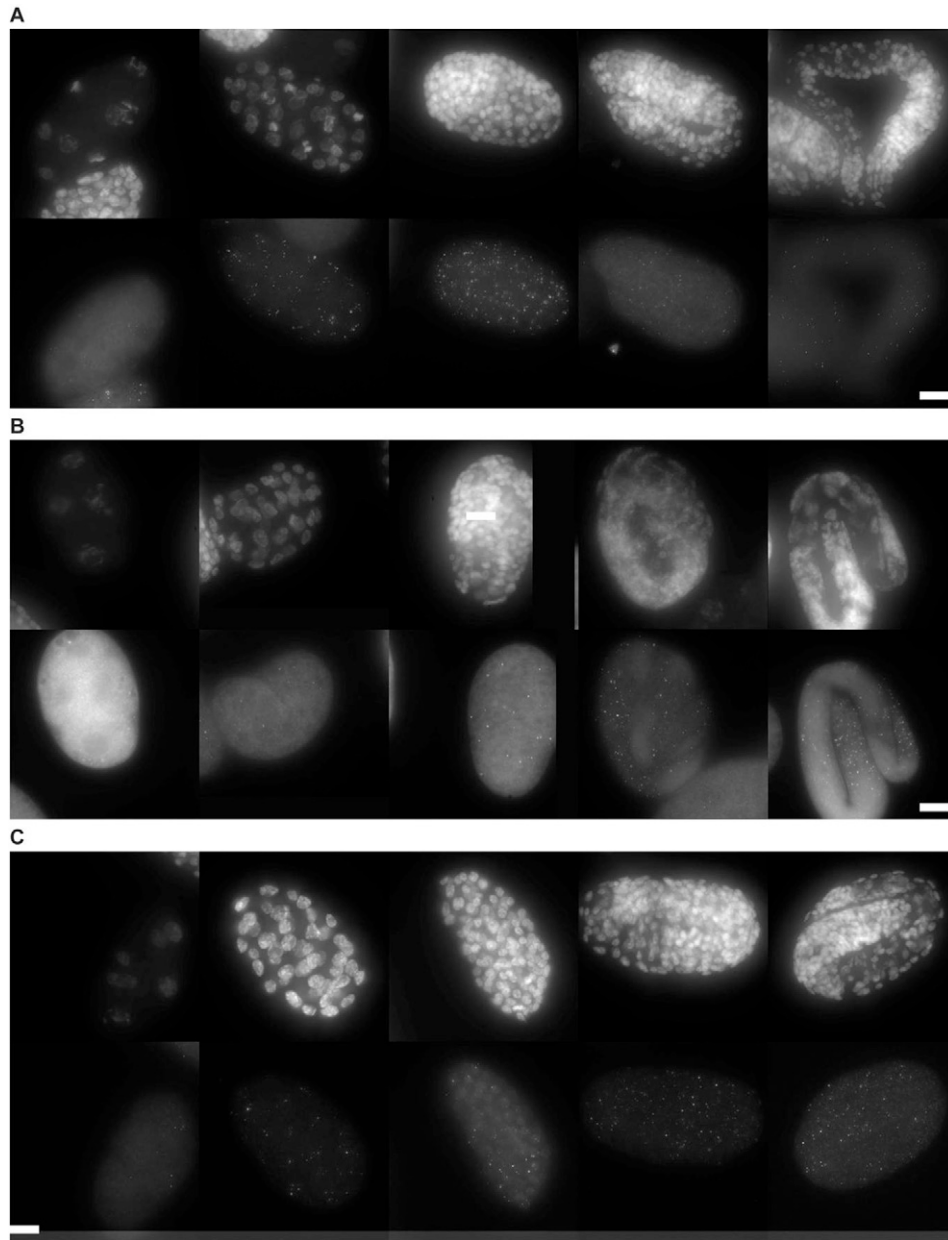


Figure S6. Characterization of Highly Structured RNAs in *C. elegans*; Related to Figure 4

(A–C) FISH images of 3 highly base-paired RNAs of *C. elegans* (chrV_h1921 in A, chrV_h2006 in B, and chrI_h719 in C) taken at single molecule resolution at a variety of developmental stages. The top panels show the nuclei (stained with DAPI), whereas the bottom panels show maximum merges of a series of optical sections of the RNA labeled with probes coupled to the TMR fluorophore. Notice that the images contain spots of variable intensity. The dimmer spots most likely represent single dsRNA molecules (based on a comparison of spot intensity to previous acquired data (Raj et al., 2008), whereas the brighter spots mostly likely arise from the accumulation of multiple dsRNAs. We believe these agglomerations are most likely located at the site of transcription, given that we see at most 1 or two per cell and that they are located within the nucleus. All scale bars are 5 μm long.