A survey of fly and nematode small RNAs by deep sequencing

By

J. Graham Ruby

B.A., Biology (2001)
Northwestern University

SUBMITTED TO THE DEPARTMENT OF BIOLOGY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

AT THE

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2008

Signature of Author.....
Department of Biology
May 20, 2008

Certified by.....
David P. Bartel
Professor of Biology
Thesis Supervisor

Accepted by.....
Steve Bell
Professor of Biology
Chairman, Graduate Committee

A survey of fly and nematode small RNAs by deep sequencing

By

James Graham Ruby

Submitted to the Department of Biology on May 20, 2008
In Partial Fulfilment of the Requirements for the Degree of Doctor of Philosophy

ABSTRACT

Small RNAs of ~22 nt length play a variety of roles in the biology of animals by repressing the translation or stimulating the degradation of complementary messenger RNAs. Depending on the structure of their precursors, they can be categorized as either microRNAs (miRNAs) or small interfering RNAs (siRNAs). In animals, miRNAs derive from characteristic hairpins in primary transcripts through two sequential RNase III-mediated cleavages; Drosha cleaves near the base of the stem to liberate a pre-miRNA hairpin, then Dicer cleaves near the loop to generate a miRNA:miRNA* duplex.

Large-scale sequencing of cDNAs derived from endogenously expressed small RNAs is used here to examine the small RNAs of the nematode *Caenorhabditis elegans* and the fly *Drosophila melanogaster*, revealing a number of previously unidentified miRNA genes from each organism. These data also revealed a novel miRNA biogenesis pathway, the mirtron pathway, in which debranched introns mimic the structural features of pre-miRNAs to enter the miRNA-processing pathway without Drosha-mediated cleavage. Mirtrons were identified in both *D. melanogaster* and *C. elegans*, some of which exhibit patterns of sequence conservation suggesting important regulatory functions.

Sequencing was performed across a timecourse of *D. melanogaster* development, permitting refinement of preexisting miRNA annotations and providing insights into miRNA biogenesis and expression. Conserved miRNAs were typically expressed more broadly and robustly than nonconserved miRNAs, and miRNAs with more restricted expression tended to have fewer predicted targets. Insights were also provided into miRNA gene evolution. Finally, two possible sources of endogenous siRNAs were revealed: antisense transcription and endogenous hpRNAs.

Besides miRNAs, sequencing from *C. elegans* revealed thousands of endogenous siRNAs generated by RNA-directed RNA polymerases acting preferentially on spermatogenesis- and transposon-associated transcripts. A third class of nematode small RNAs, called 21U-RNAs, was also discovered. 21U-RNAs are precisely 21 nucleotides long and begin with a uridine but are diverse in their remaining 20 nucleotides. 21U-RNAs originate from >5700 genomic loci dispersed in two broad regions of chromosome IV. These loci share an upstream motif that enables accurate prediction of additional 21U-RNAs. The motif is conserved in other nematodes, presumably because of its importance for producing these diverse, autonomously expressed, small RNAs.

Thesis Advisor: David P. Bartel
Title: Professor

Acknowledgements

I would like to thank the members of the Bartel lab who have worked with me directly on a wide variety of projects over the years of graduate school. Erik Schultes and Ed Curtis, for help exploring the RNA world. Ben Lewis for early help exploring the application of computers to biological data analysis. Calvin Jan for exploring the 21U-RNAs, endogenous siRNAs, and mirtrons with me. Rosaria Chiang for exploring the small RNAs of mice with me.

I would like to thank the whole Bartel lab for continuous discussion and motivation, especially Sheq, Anna, Noah, Uli, Guapo, Andrew, Alex, Huili, Ramya, Axtell, Garcia, Vincent, Soraya, Aliaa, and I-hung.

I would like to thank my collaborators and colleagues in the Mello lab, especially Pedro, and in the Blelloch lab, especially Josh, for giving me the opportunity to contribute to their stories and for creating such a wonderful collaborative environment.

I would like to thank my roommates during graduate school, whose extensive biological expertise often made home as useful a place for scientific discussions as the lab itself: Brent, Shannon, James, Dan, Bryan, and Chris.

I would like to thank the Max Funk Institut. Their musical skill was accompanied by extensive biological expertise that often made band practice as useful a place for scientific discussions as the lab itself: Sheq, Chris, John, Phil, Lee, Bob, and Bill.

I would like to thank the members of my committee, past and present, for offering their time and expertise as I have moved along the trek of graduate school: Chris Burge, Phil Sharp, Terry Orr-Weaver, Bob Sauer, Amy Keating, Uttam RajBhandary, Gary Ruvkun, and Tom Maniatis.

I would like to thank the individuals alongside whom I have had the pleasure of teaching during graduate school, especially Melissa Kemp. I would also like to thank the instructors who put so much time and effort into the courses that I was able to teach: Claudette Gardel, Chris Burge, Amy Keating, and Mike Yaffe.

I would like to thank the instructors, and especially the TAs, for the courses that I have taken during graduate school, not only for the effort that they put forward in helping me learn the course material but also for the example of great teaching that they set for me. I would especially like to thank David Pritchard and Tucker Sylvestro, my TAs for 6.001 and 6.170, whose instruction is largely responsible for my productivity in recent years.

I would like to thank my parents for continuous support, encouragement, and love.

I would like to thank Dave Bartel for patience during the slow times and support during the fast times, across projects both great and small, and for prioritizing my intellectual development as we charted the course for my research.

Most of all, I would like to thank Sally McFall, Rick Morimoto, and Sue Fox for investing so much time and energy into my development as a scientist early, when I had enthusiasm but did not show skill or even particular promise. The only thing that they stood to gain for their efforts was the satisfaction of seeing me prosper. I have prospered, and the course of my life continues to be largely a product of their help and guidance. Thank you.

Table of contents

Chapter tables of contents

Additional electronic materials are included on the accompanying CD-ROM. Electronic materials are organized into directories by chapter affiliations.

Chapter 1

Introduction

Non-coding RNA genes

The term "gene" was originally defined abstractly as a general identifier for the "unit-factors, elements, or allelomorphs in the gametes" that carry heritable information from parent to offspring (Johannsen 1911). The modern concept of a gene is more specific. It derives from Beadle and Tatum's "one gene, one enzyme" hypothesis (Beadle and Tatum 1941) and reflects the definition given by Benzer for a cistron: "a [genomic] map segment, corresponding to a function which is unitary as defined by the *cis-trans* test applied to the heterocaryon" (Benzer 1957). In the framework of the central dogma of molecular biology, an enzyme is a protein, and a cistron is the DNA segment encoding that protein. The role of RNA is to shuttle information from the DNA of the gene towards the ultimate goal of protein production.

An appreciation for the role of RNA molecules beyond the framework of the central dogma, as mature gene products themselves, came early in the history of molecular biology. The importance of the ribosomal RNAs (rRNAs) in protein synthesis was recognized early, and they were later confirmed as not only structural but catalytic components of the ribosome (Nissen et al. 2000; Noller et al. 1992; Schweet et al. 1958). The *in vivo* role of the 'soluble RNAs' (sRNA) was described by the adapter hypothesis, asserting that a sequence of DNA nucleotides is translated into a sequence of amino acids by RNA converters, and motivating the rechristening of sRNAs as transfer RNAs (tRNAs) (Berg and Ofengand 1958; Haogland et al. 1958). It was not until the non-coding roles of tRNAs and rRNAs had been roughly outlined that the role of the short-

lived messenger RNAs (mRNAs) as an information-carrying intermediate was established (Jacob and Monod 1961).

The tRNAs and rRNAs proved that not all genes would carry information from the DNA of the genome to the endpoint of protein sequence specified by the central dogma. In fact, the dual abilities of RNA as both a functional gene product and as a template for its own replication fueled speculation that RNA had served as both alpha and omega of a minimized central dogma prior to evolution's discovery of translation, in the context of the so-called RNA world (Gilbert 1986). But in the context of contemporary biology, many more types of non-protein-coding RNAs (ncRNAs) would be discovered in the wake of tRNA and rRNA. These RNA gene products fulfill a similar spectrum of roles in the cell as their proteinaceous counterparts, including enzymatic catalysis, structural roles in macromolecular complexes, molecular recognition, and gene regulation. They include the small nuclear RNAs involved in splicing; the small nucleolar RNAs that guide the covalent modification of rRNA; the catalytic self-splicing introns and RNase P; the structural signal recognition particle RNAs, vault RNAs, and Y RNAs; the ligand-binding riboswitches; and the chromosome-coating RNAs Xist, POF, and roX. Discoveries made mostly in the past decade have expanded the set of known ncRNAs to include a cornucopia of small RNAs associated with the phenomena of RNA interference, described below.

The discovery of RNA interference

The *trans* suppression of endogenous genes by introduced homologous transgenes was first observed in plants, and was termed 'cosuppression' (Napoli et al. 1990; van der

Krol et al. 1990). The observation that viral infection can eliminate the expression of viral mRNA from nuclear transgenes without reducing the rates of transcription of those transgenes identified cosuppression as post-transcriptional, and further identified a role for this type of silencing in innate immunity (Lindbo et al. 1993). Distinction was made between two 'cosuppression' pathways, one induced by single-stranded RNA (ssRNA) and the other by double-stranded RNA (dsRNA) (Que et al. 1997). These two modes of cosuppression were later observed to have some distinct genetic requirements, with the requirements for silencing of infectious viruses matching those of dsRNA-induced repression (Beclin et al. 2002; Dalmay et al. 2000).

Gene silencing in *trans* was also discovered in the fungus *Neurospora crassa*, and in that context referred to as 'quelling' (Romano and Macino 1992). A role for RNA in quelling was identified through molecular characterization of *qde-1*, a gene required for quelling that encodes an RNA-dependent RNA polymerase (RdRP) (Cogoni and Macino 1997; Cogoni and Macino 1999; Makeyev and Bamford 2002). In plants, a *qde-1* homolog is required for cosuppression by ssRNA but not by dsRNA (Dalmay et al. 2000). The requirement for either dsRNA or an enzyme capable of manufacturing dsRNA in both quelling and cosuppression implicated dsRNA as the key signaling molecule of silencing in these systems.

The observation of gene silencing by dsRNA in *Caenorhabditis elegans* extended this phenomenon to animals (Fire et al. 1998). In this context, it was referred to as RNA interference (RNAi), and was subsequently identified in other animals, including *Drosophila melanogaster* (Kennerdell and Carthew 1998) and mouse (Wianny and Zernicka-Goetz 2000). The role of small RNAs as mediators of RNAi was established

through the observation that the dsRNA that induces RNAi is cleaved into ~21-23 nt fragments both *in vitro* and *in vivo* (Bernstein et al. 2001; Parrish et al. 2000; Zamore et al. 2000), and that those fragments are sufficient to induce cleavage of target mRNAs (Elbashir et al. 2001; Hammond et al. 2000). Those fragments are called small interfering RNAs (siRNAs).

The discovery of microRNAs

Even before the role of siRNAs as mediators of RNAi was established, the endogenously expressed small temporal RNAs (stRNAs) had been discovered and their role in gene silencing established. The *lin-4* and *let-7* genes of *C. elegans* comprised this class; both genes encode small RNAs ~22 nt long that derive from a hairpin precursor (Lee et al. 1993; Reinhart et al. 2000; Wightman et al. 1993). Both of these genes were identified genetically. Mutant *lin-4* nematodes exhibit numerous post-embryonic cell lineage reiterations with diverse phenotypic consequences, including malformation of the vulva, altered body shape, and extra larval molts (Chalfie et al. 1981; Horvitz and Sulston 1980). In *let-7* mutants, the hypodermal blast cells fail to fuse with seam cells and form an adult cuticle structure at the L4-to-adult molt, resulting in a superfluous fifth larval stage (Reinhart et al. 2000).

The catalog of known stRNA genes expanded dramatically with direct molecular cloning and sequencing of cDNAs generated from their small RNA gene products in *C. elegans* and *Drosophila* (Lagos-Quintana et al. 2001; Lau et al. 2001; Lee and Ambros 2001). The genes of this family were observed to all express small RNAs that derived from larger hairpin precursors. Subsequent sequencing efforts have continued to expand

the miRNA gene number in nematodes and flies, and also identified large numbers of stRNA genes in vertebrates, planarians, and plants (Mourelatos et al. 2002; Palakodeti et al. 2006; Reinhart et al. 2002). In response to the vast expansion of this gene class, the stRNAs were rechristened as microRNAs (miRNAs).

The mechanism of RNA interference

The primary siRNAs needed for RNAi are generated through cleavage of dsRNA by the RNase III enzyme Dicer (Bernstein et al. 2001; Grishok et al. 2001; Knight and Bass 2001). The RNA products of these cleavage reactions are left with monophosphates at their 5′ ends and hydroxyl groups at their 3′ ends (Elbashir et al. 2001). Dicer cleaves long dsRNAs into a series of ~21 nt duplexes with 2 nt 3′ overhangs, and either strand of each duplex can be incorporated as a single strand into the RNA-induced silencing complex (RISC) (Elbashir et al. 2001; Martinez et al. 2002; Nykanen et al. 2001). The relative stabilities of the two ends of each siRNA duplex determine which strand of that duplex is preferentially loaded into RISC, with preference given to the strand whose 5′ end is less stable (Khvorova et al. 2003; Schwarz et al. 2003). The loaded strand directs RISC to cleave the phosphodiester bond of a complementary target RNA between the bases pairing with nucleotides 10 and 11 of the siRNA (Elbashir et al. 2001). Alternatively, RISC will translationally repress an mRNA target with non-complementary nucleotides at those positions (Doench et al. 2003) (Figure 1A).

The ability of RISC to cleave target RNAs, called Slicer activity, is provided by the Argonaute protein at its core (Liu et al. 2004). The Argonaute protein family, whose members are defined by possession of Piwi and PAZ domains, spans all of the major

clades of life on Earth (Cerutti et al. 2000). It can generally be divided into two subfamilies, Ago and Piwi, each named after a representative member (Carmell et al. 2002). The Ago subfamily is named after the AGO1 protein of *Arabidopsis* that plays a role in dsRNA-mediated post-transcriptional gene silencing (PTGS) (Morel et al. 2002). The Piwi subfamily is named after the PIWI protein of *Drosophila* that is required for renewal of germ-line stem cells (Cox et al. 1998). There are also a number of Argonaute proteins in *C. elegans* that do not fit well into either subfamily and comprise a third Argonaute clade (Yigit et al. 2006). The Piwi domain is better conserved than the PAZ domain across the Argonautes, and its tertiary structure resembles that of RNase H enzymes (Carmell et al. 2002; Song et al. 2004). In accord with the catalytic role of its structural homolog, the Piwi domain of human Ago2 catalyzes slicing (Rivas et al. 2005), and Piwi domains in Argonautes from both major subfamilies from organisms of taxa as diverse as mammals, insects, nematodes, fungi, plants, and eubacteria maintain Slicer activity (Aoki et al. 2007; Baumberger and Baulcombe 2005; Gunawardane et al. 2007; Liu et al. 2004; Maiti et al. 2007; Meister et al. 2004; Miyoshi et al. 2005; Nishida et al. 2007; Saito et al. 2006; Yuan et al. 2005). However, some Argonaute proteins have maintained the Piwi domain but lost the ability to slice (Liu et al. 2004; Meister et al. 2004) (Figure 1B). This loss of catalysis likely reflects the diverse roles beyond slicing that Argonaute proteins are thought to play in gene regulation (Peters and Meister 2007).

The integration of one strand from an siRNA duplex into RISC (this strand is called the guide strand) requires separating it from the other strand (called the passenger strand). In *Drosophila*, the guide and passenger strands are initially distinguished by R2D2, a dsRNA-binding protein that attaches to the end of the siRNA duplex with more

double-stranded character in solution; the strand whose 3′ end is bound by R2D2 becomes the guide strand (Tomari et al. 2004). R2D2 binds to Dicer-2, and it is dispensable for dsRNA cleavage by Dicer-2 but is required for Dicer-2 to remain associated with siRNAs during RISC assembly (Liu et al. 2003). Dissociation of the guide and passenger strands occurs after transfer of the siRNA duplex to Ago2 (Nykanen et al. 2001). The energetic challenge of this dissociation is eased by cleavage of the passenger strand by Ago2/Slicer (Matranga et al. 2005; Miyoshi et al. 2005; Rand et al. 2005). In human cells, the dsRNA-binding protein TRBP plays a similar role to that played by R2D2 in *Drosophila* (Chendrimada et al. 2005). In *C. elegans*, RDE-4 is the dsRNA-binding protein that interacts with both DCR-1 and the RNAi-critical Argonaute protein RDE-1 (Tabara et al. 1999; Tabara et al. 2002). RDE-4 also recruits a helicase protein, DRH-1, to the *C. elegans* RISC (Tabara et al. 2002).

Like PTGS in plants and quelling in fungi but unlike RNAi in mammals or *Drosophila*, RNAi in *C. elegans* involves the amplification of silencing activity through the action of an RNA-dependent RNA polymerase (RdRP) (Smardon et al. 2000). Small RNAs, termed secondary siRNAs, are generated that complement the target mRNA 5′ of the region of homology to the double-stranded trigger (Sijen et al. 2001) (Figure 1C). Even without introduction of a double-stranded RNA trigger, endogenous siRNAs of this type are generated in vivo that complement expressed mRNAs and transposon messages (Ambros et al. 2003; Sijen and Plasterk 2003). Secondary siRNAs differ from primary siRNAs in that they carry 5′ triphosphates rather than monophosphates and they interact with distinct Argonaute and RNAi-related proteins (see chapter 2) (Aoki et al. 2007; Pak

and Fire 2007; Sijen et al. 2007; Yigit et al. 2006). However, they can target complementary messages for slicing just like primary siRNAs (Aoki et al. 2007).

In plants, suppression of endogenous messages by siRNAs via post-transcriptional slicing is supplemented by transcriptional repression via RNA-directed DNA methylation (RdDM) (Mette et al. 2000). Cytosines are methylated by this mechanism only at positions with direct complementarity to siRNAs through a process that depends on the nuclear-localized Argonaute protein AGO4 (Xie et al. 2004; Zilberman et al. 2003). RdDM induced by siRNAs has been reported in mammals as well (Morris et al. 2004), but a specific lack of RNAi-induced RdDM in mammals has also been reported (Svoboda et al. 2004). Also, while RdDM of promoters is a mechanism of transcriptional silencing, siRNAs targeted to mammalian promoters have also been reported to activate transcription through a process called RNA activation (RNAa) (Janowski et al. 2007; Li et al. 2006). Such contradictory reports on the effects of promoter-directed siRNAs make the possibilities of both RdDM and RNAa in mammals uncertain.

The biogenesis of microRNAs

Animal miRNAs are encoded by nuclear genes that are transcribed, then processed in multiple steps, to generate mature RNA species ~22 nt long. The primary miRNA transcript, or pri-miRNA (Lee et al. 2002), is generally a product of RNA polymerase II (pol II) and as such is 5′ capped and 3′ polyadenylated (Bracht et al. 2004; Cai et al. 2004; Lee et al. 2004b). Unlike the vast majority of animal protein-coding genes, many miRNA genes are polycistronic. Single continuous pri-miRNA transcripts have been detected for several sets of genomically clustered miRNA genes (Lee et al.

2002), and the generally correlated expression of genomically clustered miRNAs supports the existence of many more polycistronic miRNA messages (Baskerville and Bartel 2005; Lagos-Quintana et al. 2001; Lau et al. 2001; Sempere et al. 2004). In addition, a large number of miRNA genes reside within the introns of protein-coding mRNAs (Rodriguez et al. 2004). These intronic miRNAs can be excised from their unspliced host pre-mRNAs without adverse effects on the stability, further processing, or eventual translation of the mRNA (Kim and Kim 2007). While the vast majority of pri-miRNA transcripts are generated by pol II, a small number of endogenously-expressed miRNA genes are transcribed by RNA polymerase III (pol III) (Borchert et al. 2006). MicroRNA genes that are endogenously transcribed by pol II continue to be processed into active mature effective miRNAs when artificially expressed as pol III transcripts (Chen et al. 2004), indicating that the genesis of the pri-miRNA transcript has little influence on its ability to enter the miRNA biogenesis pathway.

The critical feature of the pri-miRNA that enables processing is its ability to form an RNA hairpin, with the mature miRNA deriving from one arm of that hairpin. This feature was immediately recognized as significant in the first miRNA genes to be identified (Lee et al. 1993; Reinhart et al. 2000; Wightman et al. 1993), and the subsequent proliferation of miRNA annotations through molecular cloning and sequencing demonstrated its ubiquity (Lagos-Quintana et al. 2001; Lau et al. 2001; Lee and Ambros 2001). While the presence of a hairpin precursor is qualitatively consistent across all miRNA genes, a quantitative, predictive model of the hairpin characteristics that are salient to miRNA processing is lacking. Many efforts have been made to develop such a model and apply it to the identification of novel miRNA genes, but in each case

the models have either relied in large part on non-structural features such as conservation for their predictive power, been woefully unsuccessful at accurately identifying novel miRNAs, or both (Altuvia et al. 2005; Ambros et al. 2003; Bentwich et al. 2005; Berezikov et al. 2005; Grad et al. 2003; Lai et al. 2003; Lim et al. 2003; Miranda et al. 2006; Nam et al. 2005; Ohler et al. 2004; Pfeffer et al. 2005; Sewer et al. 2005; Xie et al. 2005). Better understanding of the miRNA precursor structural requirements has been achieved through the identification and biochemical evaluation of components of the processing machinery.

Pri-miRNA hairpins are cleaved in the nucleus to liberate the miRNA precursor hairpin (pre-miRNA) that eventually gives rise to the mature miRNA (Lee et al. 2002) (Figure 1D). Two cleavages, one in each arm of the pri-miRNA hairpin, are coordinately catalyzed by the Microprocessor complex, with the RNase III enzyme Drosha at its core (Denli et al. 2004; Gregory et al. 2004; Lee et al. 2003b). Microprocessor, which contains multiple copies of Drosha and of its RNA-binding cofactor, DGCR8/Pasha, recognizes the transition between ssRNA and dsRNA at the base of the pri-miRNA hairpin and cleaves ~11 bp into the hairpin (Denli et al. 2004; Gregory et al. 2004; Han et al. 2004; Han et al. 2006; Lee et al. 2003b; Zeng and Cullen 2005). In addition to miRNA maturation, a role for the RNase activity of Drosha in the maturation of 5.8S rRNA has been suggested (Wu et al. 2000). DGCR8/Pasha is required to target the Drosha endonuclease activity to pri-miRNA substrates (Denli et al. 2004; Gregory et al. 2004; Han et al. 2004; Landthaler et al. 2004). The resulting pre-miRNA hairpins have the 2 nt 3′ overhangs, 5′ monophosphates, and 3′ hydroxyls that are typical of RNase III products (Basyuk et al. 2003).

Pre-miRNAs are recognized in the nucleus and transported to the cytoplasm in a RanGTP-dependent manner by Exportin-5 (Bohnsack et al. 2004; Lund et al. 2004; Yi et al. 2003) (Figure 1D). Exportin-5 recognizes the RNA minihelix structural motif, consisting of an RNA duplex of at least 17 bp that is flanked on one side by few or no unpaired nucleotides, in the presence of RanGTP but not RanGDP (Gwizdek et al. 2001; Gwizdek et al. 2003). Pre-miRNA hairpins adhere to these minihelix requirements, and shortening of the pre-miRNA stem or introduction of a 5′ overhang instead of the typical 3′ overhang results in loss of pre-miRNA nuclear export (Zeng and Cullen 2004).

The miRNA and siRNA biogenesis pathways intersect at the cytoplasmic RNase III enzyme Dicer, which cleaves both long dsRNA into ~21nt siRNAs and pre-miRNA hairpins to generate mature miRNAs (Bernstein et al. 2001; Grishok et al. 2001; Hutvagner et al. 2001; Ketting et al. 2001; Knight and Bass 2001; Lee et al. 2002) (Figure 1D). Unlike mammals and nematodes, *Drosophila* possesses two paralogous Dicers that partition the processing of small RNA precursors between them, with Dicer-1 primarily responsible for pre-miRNA cleavage and Dicer-2 primarily responsible for dsRNA cleavage (Lee et al. 2004c; Liu et al. 2003). Like Drosha, which partners with the RNA-binding protein DGCR8/Pasha to cleave pri-miRNAs, Dicer-1 partners with the RNA-binding protein Loquacious (also called R3D1) that is required for the processing of many pre-miRNAs and enhances dsRNA-mediated RNAi (Forstemann et al. 2005; Jiang et al. 2005; Liu et al. 2007; Saito et al. 2005). R2D2, which complexes with Dicer-2 and is required for RISC-loading of siRNAs, has no effect on pre-miRNA processing (Forstemann et al. 2005; Liu et al. 2003; Saito et al. 2005).

The duplex generated by Dicer cleavage of a pre-miRNA is analogous to an siRNA duplex. In this case, the guide strand is called the miRNA, and the passenger strand is called the miRNA star (miRNA*) (Lau et al. 2001; Lim et al. 2003). The loading of a miRNA into its silencing complex (miRISC) from this duplex parallels the loading of RISC with a guide siRNA (Figure 1D). In humans, the siRNA RISC-loading protein TRBP also loads miRNAs into miRISC (Chendrimada et al. 2005). In Drosophila, the loading of Dicer-2-generated siRNAs into Ago2 RISC is paralleled by the loading of Dicer-1-generated miRNAs into Ago1 RISC (Forstemann et al. 2007). Because miRNA/miRNA* duplexes are generally far from perfect, miRNA* strands are not generally cleavable by Slicer as are siRNA guide strands (Tomari et al. 2007). But for the same reason, their dissociation is also far less of an energetic challenge than that posed by siRNA duplexes. The core human RISC-loading complex, composed of Ago2, Dicer, and TRBP, is sufficient to process and load miRNAs provided as pre-miRNA substrates in an ATP-independent manner (MacRae et al. 2008; Maniataki and Mourelatos 2005).

Plant miRNAs have many of the same features as animal miRNAs. Both derive from hairpin precursors, both are generated by cleavage of a primary transcript by RNase III enzymes, and both form a complex with Argonaute proteins to mediate mRNA repression (Park et al. 2002; Reinhart et al. 2002; Vaucheret et al. 2004). However, several features distinguish the two biogenesis pathways, notably the generation of both cleavages by a single enzyme and the localization of the entire biogenesis process to the nucleus (Papp et al. 2003; Xie et al. 2004). These differences, among others, have fueled speculation that the miRNA biogenesis pathways may have arisen independently in

plants and animals and represent convergent usages of conserved RNAi process components (Bartel 2004).

Gene regulation by microRNAs

The mechanisms of gene regulation by microRNAs are diverse and are not yet fully understood. The primary siRNAs that derive from dsRNAs in RNAi direct the Slicer-mediated cleavage of their mRNA targets. Plant miRNAs share this mechanism of target repression, directing the cleavage of their near-perfect-complement targets (Rhoades et al. 2002). In contrast, animal miRNAs generally exhibit far less complementarity to their endogenous targets. The first natural miRNA/target interaction to be identified, between the *lin-4* miRNA and *lin-14* mRNA 3′ untranslated region (UTR) includes many mismatches and causes repression of translation without mRNA degradation (Wightman et al. 1993). Just as siRNAs can mimic miRNAs by repressing the translation of messages with imperfect complementarity, animal miRNAs are capable of mimicking siRNAs by guiding the cleavage of near-perfect-complement RNAs when such targets present themselves (Doench et al. 2003; Hutvagner and Zamore 2002). However, such complementarity is rare. There are a few animal miRNAs that target the cleavage of transcripts that derive from the opposite genomic strand as the miRNA (Davis et al. 2005), and only one example of a miRNA with such complementarity to a distal gene (Yekta et al. 2004).

The ability of miRNAs to inhibit translation via imperfect base pairing with the 3′ UTRs of target mRNAs has been demonstrated both *in vivo* (Brennecke et al. 2003; Wightman et al. 1993) and *in vitro* (Wang et al. 2006). However, the mechanism of such

inhibition remains largely elusive. Messenger RNAs continue to associate with polyribosomes while being repressed by miRNAs, as do miRNAs themselves and the miRISC component Argonaute, indicating that inhibition occurs after the initiation of translation (Maroney et al. 2006; Nottrott et al. 2006; Olsen and Ambros 1999; Seggerson et al. 2002). In addition, miRNAs repress the translation of messages with both cap- and IRES-dependent initiation, and the repressive effects of miRNAs are additive with those of a drug that inhibits initiation, further supporting a post-initiation model of repression (Petersen et al. 2006). Intriguingly, miRNAs can inhibit the cap-dependent translational initiation of synthesized mRNAs that are transfected into cells or introduced into cytoplasmic extracts (Humphreys et al. 2005; Mathonnet et al. 2007; Pillai et al. 2005; Thermann and Hentze 2007). Such inhibition of initiation likely results from Ago2 successfully competing with initiation factor eIF4E for binding to the m^7G cap of mRNAs that are rapidly introduced into the cytoplasm (Kiriakidou et al. 2007). The relevance of these interactions to repression of nuclear-derived mRNAs is unclear, as is the specific mechanism of post-initiation translational repression by miRNAs.

While translational inhibition is their primary method for gene repression, miRNAs can also modestly destabilize target mRNAs. Investigations of miRNA-mediated translational repression commonly reveal decreases in the abundance of target mRNAs in response to miRNA, though always of insufficient magnitudes to account for the observed losses of protein (Bagga et al. 2005; Olsen and Ambros 1999; Petersen et al. 2006). In addition, the blocking of endogenously-expressed miRNAs with complementary oligonucleotides and the transfection of cells with non-endogenous miRNAs both have subtle but wide-spread effects on mRNA levels, both indicating a role

for partial complementarity between miRNAs and the 3′ UTRs of down-regulated mRNAs (Krutzfeldt et al. 2005; Lim et al. 2005). However, as with translational inhibition, the mechanism of miRNA-mediated mRNA destabilization is unclear at present. The poly-A tails of target mRNAs are particularly destabilized by miRNAs both *in vivo* and *in vitro*, and the shortening of poly-A tails may contribute to the overall destabilization of mRNAs and/or reductions of their translation (Humphreys et al. 2005; Wakiyama et al. 2007; Wu et al. 2006). The localization of miRNA-silenced mRNAs to cytoplasmic processing bodies (P-bodies) may also contribute to mRNA decay and/or repression of translation (Liu et al. 2005; Sen and Blau 2005). P-bodies accumulate non-translating mRNAs and are sites of mRNA decay through the canonical pathway of 5′ decapping and 5′→3′ exonucleolytic degradation (Sheth and Parker 2003; Teixeira et al. 2005). MicroRNA-mediated silencing by both mRNA decay and translational repression depends at least in part on the defining P-body component GW182 (Behm-Ansmant et al. 2006; Rehwinkel et al. 2005). However, it remains uncertain whether these effects on mRNAs are causes or consequences of P-body association (Chu and Rana 2006).

While the mechanism of miRNA-mediated repression remains somewhat ambiguous at present, the requirements for such repression of an mRNA via base pairing with the 3′ UTR are better understood. The degree of complementarity between natural targets varies, but almost invariantly includes perfect complementarity at the 5′ end of the miRNA (Johnston and Hobert 2003; Lai 2002; Lee et al. 1993; Reinhart et al. 2000; Wightman et al. 1993). These ~7nt stretches of complementarity to the miRNA 'seed' sequence are the only target pairings that are consistently and significantly conserved, and are also the only pairings whose disruption results in a pronounced loss of repression

(Brennecke et al. 2005; Doench and Sharp 2004; Lewis et al. 2005; Lewis et al. 2003; Stark et al. 2003). Seed pairing in the 3′ UTR is also significantly enriched among those mRNAs whose stability is perturbed by introduction/sequestration of a miRNA to/from the cytoplasm (Krutzfeldt et al. 2005; Lim et al. 2005). Several aspects of the context of a seed match within a 3′ UTR contribute to the efficacy of the match as a target site, and like the seed match itself, these additional determinants are equally consequential when considering either the translational repression or the mRNA destabilization component of miRNA-mediated gene repression (Grimson et al. 2007).

Additional regulatory roles for small RNAs in biology

Because Argonaute proteins interact with small RNAs, their ubiquity, abundance, and diversity across all the kingdoms of life indicates many as-yet-uncharacterized roles for small RNAs in biology. Combined, the related phenomena of quelling, PTGS, and RNAi span only a small slice of the diversity of the biome, and miRNAs span an even smaller slice. Even within the organisms for which these processes have been characterized, the abundance of additional Argonaute-family genes indicates that there are additional roles played by small RNAs. For instance, in vertebrates and in *Drosophila*, Piwi interacts with a class of small RNAs called piwi-associated RNAs (piRNAs) that are longer than other small RNAs (typically >26 nt), are 2′ O-methylated at their 3′ end, and are highly diverse (Aravin et al. 2006; Brennecke et al. 2007; Girard et al. 2006; Grivna et al. 2006a; Horwich et al. 2007; Houwing et al. 2007; Lau et al. 2006; Saito et al. 2007). Clusters of piRNAs are derived haphazardly from one strand of genomic segments that typically extend more than 10 kb. In *Drosophila*, piRNAs repress

transposable elements, and they were originally classified based on their complementarity to genomic transposon repeat elements as repeat-associated siRNAs (rasiRNAs) (Aravin et al. 2003; Brennecke et al. 2007; Klenov et al. 2007). That role seems to be conserved for some but not all of the vertebrate piRNAs (Aravin et al. 2007), and their association with polysomes indicates a more general role for piRNAs in the regulation of translation (Grivna et al. 2006b).

Several eukaryotic lineages, including the frequently-used yeast model *Saccharomyces cerevisiae*, have lost core components of the RNAi machinery and with it the ability to silence genes post-transcriptionally in response to the introduction of homologous dsRNA (Cerutti and Casas-Mollano 2006). Nonetheless, small RNAs play an important role in the regulation of chromatin in these species. In fission yeast, maintenance of centromeric heterochromatin depends on Argonaute, Dicer, and RdRP proteins (Volpe et al. 2002). Small RNAs similar in size to siRNAs and miRNAs whose sequences match those of the centromeric repeats accumulate in *Schizosaccharomyces pombe* (Reinhart and Bartel 2002). They combine with the Argonaute protein Ago1 and other factors to form the RNA-induced initiation of transcriptional gene silencing complex (RITS complex), and they localize that complex to the centromeric repeats (Verdel et al. 2004). Introduction of homologous dsRNA is sufficient to induce chromatin-based silencing in *S. pombe* (Schramke and Allshire 2003). The similar requirement of RNAi components or small RNAs for some chromatin-based transcriptional silencing phenomenae in *Drosophila* (Pal-Bhadra et al. 1997; Pal-Bhadra et al. 2002) and *Arabidopsis* (Lippman et al. 2004) indicates broad conservation of this physiological role for small RNAs.

An extreme form of RNAi exists in ciliates in which the transcription of dsRNA in the micronucleus targets the destruction of homologous genomic DNA in the macronucleus (Chalker and Yao 2001; Yao et al. 2003). In *Tetrahymena*, DNA elimination is dependent on the Twi1p protein, a member of the Piwi subfamily of Argonautes, and the induction of Twi1p correlates with that of a set of abundant small RNAs (scnRNAs) (Mochizuki et al. 2002). Another form of RNAi, meiotic silencing of unpaired DNA (MSUD), exists alongside quelling in *Neurospora*. MSUD represses all homologous copies of any gene that is either present in one copy of the diploid zygotic genome but not the other, or whose loci in the two copies of the genome are not homologously placed (Shiu et al. 2001). This type of repression requires transcription of the gene to be silenced and depends on both an RdRP and a member of the Ago subfamily of Argonautes (Lee et al. 2003a; Lee et al. 2004a; Shiu et al. 2001).

Argonaute proteins are found in prokaryotes as well as eukaryotes, and span both archeabacteria (Song et al. 2004) and eubacteria (Yuan et al. 2005). Several of the prokaryotic Argonautes have been crystallized in order to better understand their eukayotic counterparts. However, their roles in their host organisms are not understood. The usefulness of those structures to analyses of eukaryotic Argonautes suggests that some of the mechanisms of small RNA action are conserved between prokaryotes and eukaryotes. Just as it has in eukaryotes, further exploration of the Argonaute proteins and their small RNA cofactors in prokaryotes will surely reveal more exciting biology.

Figure Legends

Figure 1. Aspects of small RNA biology. (A) Mechanisms of siRNA- and miRNA-mediated gene silencing include Slicing of the target mRNA between the bases pairing with nucleotides 10 and 11 of the siRNA (top) and translational repression, which requires only basepairing with the 7nt 'seed' at the 5′ end of the miRNA/siRNA (bottom). (B) Slicing activity is maintained across many distantly-related Argonaute proteins from diverse taxa but is also lost in some Argonautes. Examples are shown for which Slicer catalytic activity has been experimentally demonstrated to be present or absent. Tree showing the evolutionary relationships between Argonaute family members is based on previous analyses (Catalanotto et al. 2000; Yigit et al. 2006). (C) RNAi amplification in *C. elegans* results in the production of secondary siRNAs that complement the target mRNA 5′ but not 3′ of the regions of homology to the dsRNA trigger. (D) The canonical miRNA biogenesis pathway.

References

Altuvia, Y., P. Landgraf, G. Lithwick, N. Elefant, S. Pfeffer, A. Aravin, M.J. Brownstein, T. Tuschl, and H. Margalit. 2005. Clustering and conservation patterns of human microRNAs. *Nucleic Acids Res* **33:** 2697-2706.

Ambros, V., R.C. Lee, A. Lavanway, P.T. Williams, and D. Jewell. 2003. MicroRNAs and other tiny endogenous RNAs in *C. elegans*. *Curr Biol* **13:** 807-818.

Aoki, K., H. Moriguchi, T. Yoshioka, K. Okawa, and H. Tabara. 2007. In vitro analyses of the production and activity of secondary small interfering RNAs in C. elegans. *Embo J* **26:** 5007-5019.

Aravin, A., D. Gaidatzis, S. Pfeffer, M. Lagos-Quintana, P. Landgraf, N. Iovino, P. Morris, M.J. Brownstein, S. Kuramochi-Miyagawa, T. Nakano, M. Chien, J.J. Russo, J. Ju, R. Sheridan, C. Sander, M. Zavolan, and T. Tuschl. 2006. A novel class of small RNAs bind to MILI protein in mouse testes. *Nature* **442:** 203-207.

Aravin, A.A., M. Lagos-Quintana, A. Yalcin, M. Zavolan, D. Marks, B. Snyder, T. Gaasterland, J. Meyer, and T. Tuschl. 2003. The small RNA profile during *Drosophila melanogaster* development. *Dev Cell* **5:** 337-350.

Aravin, A.A., R. Sachidanandam, A. Girard, K. Fejes-Toth, and G.J. Hannon. 2007. Developmentally regulated piRNA clusters implicate MILI in transposon control. *Science* **316:** 744-747.

Bagga, S., J. Bracht, S. Hunter, K. Massirer, J. Holtz, R. Eachus, and A.E. Pasquinelli. 2005. Regulation by let-7 and lin-4 miRNAs results in target mRNA degradation. *Cell* **122:** 553-563.

Bartel, D.P. 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116:** 281-297.

Baskerville, S. and D.P. Bartel. 2005. Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA* **11:** 241-247.

Basyuk, E., F. Suavet, A. Doglio, R. Bordonne, and E. Bertrand. 2003. Human *let-7* stem-loop precursors harbor features of RNase III cleavage products. *Nucleic Acids Res* **31:** 6593-6597.

Baumberger, N. and D.C. Baulcombe. 2005. Arabidopsis ARGONAUTE1 is an RNA Slicer that selectively recruits microRNAs and short interfering RNAs. *Proc Natl Acad Sci U S A* **102:** 11928-11933.

Beadle, G.W. and E.L. Tatum. 1941. Genetic Control of Biochemical Reactions in Neurospora. *Proc Natl Acad Sci U S A* **27:** 499-506.

Beclin, C., S. Boutet, P. Waterhouse, and H. Vaucheret. 2002. A branched pathway for transgene-induced RNA silencing in plants. *Curr Biol* **12:** 684-688.

Behm-Ansmant, I., J. Rehwinkel, T. Doerks, A. Stark, P. Bork, and E. Izaurralde. 2006. mRNA degradation by miRNAs and GW182 requires both CCR4:NOT deadenylase and DCP1:DCP2 decapping complexes. *Genes Dev* **20:** 1885-1898.

Bentwich, I., A. Avniel, Y. Karov, R. Aharonov, S. Gilad, O. Barad, A. Barzilai, P. Einat, U. Einav, E. Meiri, E. Sharon, Y. Spector, and Z. Bentwich. 2005. Identification of hundreds of conserved and nonconserved human microRNAs. *Nat Genet* **37:** 766-770.

Benzer, S. 1957. The elemental units of heredity. In *A Symposium on the Chemical Basis of Heredity* (eds. W. McElroy and G. Bentley). The Johns Hopkins Press, Baltimore.

Berezikov, E., V. Guryev, J. van de Belt, E. Wienholds, R.H. Plasterk, and E. Cuppen. 2005. Phylogenetic shadowing and computational identification of human microRNA genes. *Cell* **120:** 21-24.

Berg, P. and E.J. Ofengand. 1958. An enzymatic mechanism for linking amino acids to RNA. *Proc Natl Acad Sci U S A* **44:** 78-86.

Bernstein, E., A.A. Caudy, S.M. Hammond, and G.J. Hannon. 2001. Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature* **409:** 363-366.

Bohnsack, M.T., K. Czaplinski, and D. Gorlich. 2004. Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs. *Rna* **10:** 185-191.

Borchert, G.M., W. Lanier, and B.L. Davidson. 2006. RNA polymerase III transcribes human microRNAs. *Nat Struct Mol Biol* **13:** 1097-1101.

Bracht, J., S. Hunter, R. Eachus, P. Weeks, and A.E. Pasquinelli. 2004. Trans-splicing and polyadenylation of *let-7* microRNA primary transcripts. *Rna* **10:** 1586-1594.

Brennecke, J., A.A. Aravin, A. Stark, M. Dus, M. Kellis, R. Sachidanandam, and G.J. Hannon. 2007. Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. *Cell* **128:** 1089-1103.

Brennecke, J., D.R. Hipfner, A. Stark, R.B. Russell, and S.M. Cohen. 2003. bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene hid in Drosophila. *Cell* **113:** 25-36.

Brennecke, J., A. Stark, R.B. Russell, and S.M. Cohen. 2005. Principles of microRNA-target recognition. *PLoS Biol* **3:** e85.

Cai, X., C.H. Hagedorn, and B.R. Cullen. 2004. Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *Rna* **10:** 1957-1966.

Carmell, M.A., Z. Xuan, M.Q. Zhang, and G.J. Hannon. 2002. The Argonaute family: tentacles that reach into RNAi, developmental control, stem cell maintenance, and tumorigenesis. *Genes Dev* **16:** 2733-2742.

Catalanotto, C., G. Azzalin, G. Macino, and C. Cogoni. 2000. Gene silencing in worms and fungi. *Nature* **404:** 245.

Cerutti, H. and J.A. Casas-Mollano. 2006. On the origin and functions of RNA-mediated silencing: from protists to man. *Curr Genet* **50:** 81-99.

Cerutti, L., N. Mian, and A. Bateman. 2000. Domains in gene silencing and cell differentiation proteins: the novel PAZ domain and redefinition of the Piwi domain. *Trends Biochem Sci* **25:** 481-482.

Chalfie, M., H.R. Horvitz, and J.E. Sulston. 1981. Mutations that lead to reiterations in the cell lineages of C. elegans. *Cell* **24:** 59-69.

Chalker, D.L. and M.C. Yao. 2001. Nongenic, bidirectional transcription precedes and may promote developmental DNA deletion in Tetrahymena thermophila. *Genes Dev* **15:** 1287-1298.

Chen, C.Z., L. Li, H.F. Lodish, and D.P. Bartel. 2004. MicroRNAs modulate hematopoietic lineage differentiation. *Science* **303:** 83-86.

Chendrimada, T.P., R.I. Gregory, E. Kumaraswamy, J. Norman, N. Cooch, K. Nishikura, and R. Shiekhattar. 2005. TRBP recruits the Dicer complex to Ago2 for microRNA processing and gene silencing. *Nature* **436:** 740-744.

Chu, C.Y. and T.M. Rana. 2006. Translation repression in human cells by microRNA-induced gene silencing requires RCK/p54. *PLoS Biol* **4:** e210.

Cogoni, C. and G. Macino. 1997. Isolation of quelling-defective (qde) mutants impaired in posttranscriptional transgene-induced gene silencing in Neurospora crassa. *Proc Natl Acad Sci U S A* **94:** 10233-10238.

Cogoni, C. and G. Macino. 1999. Gene silencing in Neurospora crassa requires a protein homologous to RNA-dependent RNA polymerase. *Nature* **399:** 166-169.

Cox, D.N., A. Chao, J. Baker, L. Chang, D. Qiao, and H. Lin. 1998. A novel class of evolutionarily conserved genes defined by piwi are essential for stem cell self-renewal. *Genes Dev* **12:** 3715-3727.

Dalmay, T., A. Hamilton, S. Rudd, S. Angell, and D.C. Baulcombe. 2000. An RNA-dependent RNA polymerase gene in Arabidopsis is required for

posttranscriptional gene silencing mediated by a transgene but not by a virus. *Cell* **101:** 543-553.

Davis, E., F. Caiment, X. Tordoir, J. Cavaille, A. Ferguson-Smith, N. Cockett, M. Georges, and C. Charlier. 2005. RNAi-mediated allelic trans-interaction at the imprinted Rtl1/Peg11 locus. *Curr Biol* **15:** 743-749.

Denli, A.M., B.B. Tops, R.H. Plasterk, R.F. Ketting, and G.J. Hannon. 2004. Processing of primary microRNAs by the Microprocessor complex. *Nature* **432:** 231-235.

Doench, J.G., C.P. Petersen, and P.A. Sharp. 2003. siRNAs can function as miRNAs. *Genes Dev* **17:** 438-442.

Doench, J.G. and P.A. Sharp. 2004. Specificity of microRNA target selection in translational repression. *Genes Dev* **18:** 504-511.

Elbashir, S.M., W. Lendeckel, and T. Tuschl. 2001. RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes Dev* **15:** 188-200.

Fire, A., S. Xu, M.K. Montgomery, S.A. Kostas, S.E. Driver, and C.C. Mello. 1998. Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* **391:** 806-811.

Forstemann, K., M.D. Horwich, L. Wee, Y. Tomari, and P.D. Zamore. 2007. Drosophila microRNAs are sorted into functionally distinct argonaute complexes after production by dicer-1. *Cell* **130:** 287-297.

Forstemann, K., Y. Tomari, T. Du, V.V. Vagin, A.M. Denli, D.P. Bratu, C. Klattenhoff, W.E. Theurkauf, and P.D. Zamore. 2005. Normal microRNA maturation and germ-line stem cell maintenance requires Loquacious, a double-stranded RNA-binding domain protein. *PLoS Biol* **3:** e236.

Gilbert, W. 1986. The RNA world. *Nature* **319:** 618.

Girard, A., R. Sachidanandam, G.J. Hannon, and M.A. Carmell. 2006. A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature* **442:** 199-202.

Grad, Y., J. Aach, G.D. Hayes, B.J. Reinhart, G.M. Church, G. Ruvkun, and J. Kim. 2003. Computational and experimental identification of *C. elegans* microRNAs. *Mol Cell* **11:** 1253-1263.

Gregory, R.I., K.P. Yan, G. Amuthan, T. Chendrimada, B. Doratotaj, N. Cooch, and R. Shiekhattar. 2004. The Microprocessor complex mediates the genesis of microRNAs. *Nature* **432:** 235-240.

Grimson, A., K.K. Farh, W.K. Johnston, P. Garrett-Engele, L.P. Lim, and D.P. Bartel. 2007. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell* **27:** 91-105.

Grishok, A., A.E. Pasquinelli, D. Conte, N. Li, S. Parrish, I. Ha, D.L. Baillie, A. Fire, G. Ruvkun, and C.C. Mello. 2001. Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control *C. elegans* developmental timing. *Cell* **106:** 23-34.

Grivna, S.T., E. Beyret, Z. Wang, and H. Lin. 2006a. A novel class of small RNAs in mouse spermatogenic cells. *Genes Dev* **20:** 1709-1714.

Grivna, S.T., B. Pyhtila, and H. Lin. 2006b. MIWI associates with translational machinery and PIWI-interacting RNAs (piRNAs) in regulating spermatogenesis. *Proc Natl Acad Sci U S A* **103:** 13415-13420.

Gunawardane, L.S., K. Saito, K.M. Nishida, K. Miyoshi, Y. Kawamura, T. Nagami, H. Siomi, and M.C. Siomi. 2007. A slicer-mediated mechanism for repeat-associated siRNA 5' end formation in Drosophila. *Science* **315:** 1587-1590.

Gwizdek, C., E. Bertrand, C. Dargemont, J.C. Lefebvre, J.M. Blanchard, R.H. Singer, and A. Doglio. 2001. Terminal minihelix, a novel RNA motif that directs polymerase III transcripts to the cell cytoplasm. Terminal minihelix and RNA export. *J Biol Chem* **276:** 25910-25918.

Gwizdek, C., B. Ossareh-Nazari, A.M. Brownawell, A. Doglio, E. Bertrand, I.G. Macara, and C. Dargemont. 2003. Exportin-5 mediates nuclear export of minihelix-containing RNAs. *J Biol Chem* **278:** 5505-5508.

Hammond, S.M., E. Bernstein, D. Beach, and G.J. Hannon. 2000. An RNA-directed nuclease mediates post-transcriptional gene silencing in Drosophila cells. *Nature* **404:** 293-296.

Han, J., Y. Lee, K.H. Yeom, Y.K. Kim, H. Jin, and V.N. Kim. 2004. The Drosha-DGCR8 complex in primary microRNA processing. *Genes Dev* **18:** 3016-3027.

Han, J., Y. Lee, K.H. Yeom, J.W. Nam, I. Heo, J.K. Rhee, S.Y. Sohn, Y. Cho, B.T. Zhang, and V.N. Kim. 2006. Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. *Cell* **125:** 887-901.

Haogland, M.B., M.L. Stephenson, J.F. Scott, L.I. Hecht, and P.C. Zamecnik. 1958. A soluble ribonucleic acid intermediate in protein synthesis. *Journal of Biological Chemistry* **231:** 241-257.

Horvitz, H.R. and J.E. Sulston. 1980. Isolation and genetic characterization of cell-lineage mutants of the nematode Caenorhabditis elegans. *Genetics* **96:** 435-454.

Horwich, M.D., C. Li, C. Matranga, V. Vagin, G. Farley, P. Wang, and P.D. Zamore. 2007. The Drosophila RNA methyltransferase, DmHen1, modifies germline piRNAs and single-stranded siRNAs in RISC. *Curr Biol* **17:** 1265-1272.

Houwing, S., L.M. Kamminga, E. Berezikov, D. Cronembold, A. Girard, H. van den Elst, D.V. Filippov, H. Blaser, E. Raz, C.B. Moens, R.H. Plasterk, G.J. Hannon, B.W. Draper, and R.F. Ketting. 2007. A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in Zebrafish. *Cell* **129:** 69-82.

Humphreys, D.T., B.J. Westman, D.I. Martin, and T. Preiss. 2005. MicroRNAs control translation initiation by inhibiting eukaryotic initiation factor 4E/cap and poly(A) tail function. *Proc Natl Acad Sci U S A* **102:** 16961-16966.

Hutvagner, G., J. McLachlan, A.E. Pasquinelli, E. Balint, T. Tuschl, and P.D. Zamore. 2001. A cellular function for the RNA-interference enzyme Dicer in the maturation of the *let-7* small temporal RNA. *Science* **293:** 834-838.

Hutvagner, G. and P.D. Zamore. 2002. A microRNA in a multiple-turnover RNAi enzyme complex. *Science* **297:** 2056-2060.

Jacob, F. and J. Monod. 1961. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* **3:** 318-356.

Janowski, B.A., S.T. Younger, D.B. Hardy, R. Ram, K.E. Huffman, and D.R. Corey. 2007. Activating gene expression in mammalian cells with promoter-targeted duplex RNAs. *Nat Chem Biol* **3:** 166-173.

Jiang, F., X. Ye, X. Liu, L. Fincher, D. McKearin, and Q. Liu. 2005. Dicer-1 and R3D1-L catalyze microRNA maturation in Drosophila. *Genes Dev* **19:** 1674-1679.

Johannsen, W. 1911. The genotype conception of heredity. *The American Naturalist* **45:** 129-159.

Johnston, R.J. and O. Hobert. 2003. A microRNA controlling left/right neuronal asymmetry in *Caenorhabditis elegans*. *Nature* **426:** 845-849.

Kennerdell, J.R. and R.W. Carthew. 1998. Use of dsRNA-mediated genetic interference to demonstrate that frizzled and frizzled 2 act in the wingless pathway. *Cell* **95:** 1017-1026.

Ketting, R.F., S.E. Fischer, E. Bernstein, T. Sijen, G.J. Hannon, and R.H. Plasterk. 2001. Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in *C. elegans*. *Genes Dev* **15:** 2654-2659.

Khvorova, A., A. Reynolds, and S.D. Jayasena. 2003. Functional siRNAs and miRNAs exhibit strand bias. *Cell* **115:** 209-216.

Kim, Y.K. and V.N. Kim. 2007. Processing of intronic microRNAs. *Embo J* **26:** 775-783.

Kiriakidou, M., G.S. Tan, S. Lamprinaki, M. De Planell-Saguer, P.T. Nelson, and Z. Mourelatos. 2007. An mRNA m7G cap binding-like motif within human Ago2 represses translation. *Cell* **129:** 1141-1151.

Klenov, M.S., S.A. Lavrov, A.D. Stolyarenko, S.S. Ryazansky, A.A. Aravin, T. Tuschl, and V.A. Gvozdev. 2007. Repeat-associated siRNAs cause chromatin silencing of retrotransposons in the Drosophila melanogaster germline. *Nucleic Acids Res* **35:** 5430-5438.

Knight, S.W. and B.L. Bass. 2001. A role for the RNase III enzyme DCR-1 in RNA interference and germ line development in Caenorhabditis elegans. *Science* **293:** 2269-2271.

Krutzfeldt, J., N. Rajewsky, R. Braich, K.G. Rajeev, T. Tuschl, M. Manoharan, and M. Stoffel. 2005. Silencing of microRNAs in vivo with 'antagomirs'. *Nature* **438:** 685-689.

Lagos-Quintana, M., R. Rauhut, W. Lendeckel, and T. Tuschl. 2001. Identification of novel genes coding for small expressed RNAs. *Science* **294:** 853-858.

Lai, E.C. 2002. Micro RNAs are complementary to 3' UTR sequence motifs that mediate negative post-transcriptional regulation. *Nat Genet* **30:** 363-364.

Lai, E.C., P. Tomancak, R.W. Williams, and G.M. Rubin. 2003. Computational identification of *Drosophila* microRNA genes. *Genome Biol* **4:** R42.

Landthaler, M., A. Yalcin, and T. Tuschl. 2004. The human DiGeorge syndrome critical region gene 8 and Its D. melanogaster homolog are required for miRNA biogenesis. *Curr Biol* **14:** 2162-2167.

Lau, N.C., L.P. Lim, E.G. Weinstein, and D.P. Bartel. 2001. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science* **294:** 858-862.

Lau, N.C., A.G. Seto, J. Kim, S. Kuramochi-Miyagawa, T. Nakano, D.P. Bartel, and R.E. Kingston. 2006. Characterization of the piRNA complex from rat testes. *Science* **313:** 363-367.

Lee, D.W., R.J. Pratt, M. McLaughlin, and R. Aramayo. 2003a. An argonaute-like protein is required for meiotic silencing. *Genetics* **164:** 821-828.

Lee, D.W., K.Y. Seong, R.J. Pratt, K. Baker, and R. Aramayo. 2004a. Properties of unpaired DNA required for efficient silencing in Neurospora crassa. *Genetics* **167:** 131-150.

Lee, R.C. and V. Ambros. 2001. An extensive class of small RNAs in *Caenorhabditis elegans*. *Science* **294:** 862-864.

Lee, R.C., R.L. Feinbaum, and V. Ambros. 1993. The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell* **75:** 843-854.

Lee, Y., C. Ahn, J. Han, H. Choi, J. Kim, J. Yim, J. Lee, P. Provost, O. Radmark, S. Kim, and V.N. Kim. 2003b. The nuclear RNase III Drosha initiates microRNA processing. *Nature* **425:** 415-419.

Lee, Y., K. Jeon, J.T. Lee, S. Kim, and V.N. Kim. 2002. MicroRNA maturation: stepwise processing and subcellular localization. *Embo J* **21:** 4663-4670.

Lee, Y., M. Kim, J. Han, K.H. Yeom, S. Lee, S.H. Baek, and V.N. Kim. 2004b. MicroRNA genes are transcribed by RNA polymerase II. *Embo J* **23:** 4051-4060.

Lee, Y.S., K. Nakahara, J.W. Pham, K. Kim, Z. He, E.J. Sontheimer, and R.W. Carthew. 2004c. Distinct roles for Drosophila Dicer-1 and Dicer-2 in the siRNA/miRNA silencing pathways. *Cell* **117:** 69-81.

Lewis, B.P., C.B. Burge, and D.P. Bartel. 2005. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120:** 15-20.

Lewis, B.P., I.H. Shih, M.W. Jones-Rhoades, D.P. Bartel, and C.B. Burge. 2003. Prediction of mammalian microRNA targets. *Cell* **115:** 787-798.

Li, L.C., S.T. Okino, H. Zhao, D. Pookot, R.F. Place, S. Urakami, H. Enokida, and R. Dahiya. 2006. Small dsRNAs induce transcriptional activation in human cells. *Proc Natl Acad Sci U S A* **103:** 17337-17342.

Lim, L.P., N.C. Lau, P. Garrett-Engele, A. Grimson, J.M. Schelter, J. Castle, D.P. Bartel, P.S. Linsley, and J.M. Johnson. 2005. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* **433:** 769-773.

Lim, L.P., N.C. Lau, E.G. Weinstein, A. Abdelhakim, S. Yekta, M.W. Rhoades, C.B. Burge, and D.P. Bartel. 2003. The microRNAs of *Caenorhabditis elegans*. *Genes Dev* **17:** 991-1008.

Lindbo, J.A., L. Silva-Rosales, W.M. Proebsting, and W.G. Dougherty. 1993. Induction of a Highly Specific Antiviral State in Transgenic Plants: Implications for Regulation of Gene Expression and Virus Resistance. *Plant Cell* **5:** 1749-1759.

Lippman, Z., A.V. Gendrel, M. Black, M.W. Vaughn, N. Dedhia, W.R. McCombie, K. Lavine, V. Mittal, B. May, K.D. Kasschau, J.C. Carrington, R.W. Doerge, V. Colot, and R. Martienssen. 2004. Role of transposable elements in heterochromatin and epigenetic control. *Nature* **430:** 471-476.

Liu, J., M.A. Carmell, F.V. Rivas, C.G. Marsden, J.M. Thomson, J.J. Song, S.M. Hammond, L. Joshua-Tor, and G.J. Hannon. 2004. Argonaute2 is the catalytic engine of mammalian RNAi. *Science* **305:** 1437-1441.

Liu, J., M.A. Valencia-Sanchez, G.J. Hannon, and R. Parker. 2005. MicroRNA-dependent localization of targeted mRNAs to mammalian P-bodies. *Nat Cell Biol* **7:** 719-723.

Liu, Q., T.A. Rand, S. Kalidas, F. Du, H.E. Kim, D.P. Smith, and X. Wang. 2003. R2D2, a bridge between the initiation and effector steps of the Drosophila RNAi pathway. *Science* **301:** 1921-1925.

Liu, X., J.K. Park, F. Jiang, Y. Liu, D. McKearin, and Q. Liu. 2007. Dicer-1, but not Loquacious, is critical for assembly of miRNA-induced silencing complexes. *Rna* **13:** 2324-2329.

Lund, E., S. Guttinger, A. Calado, J.E. Dahlberg, and U. Kutay. 2004. Nuclear export of microRNA precursors. *Science* **303:** 95-98.

MacRae, I.J., E. Ma, M. Zhou, C.V. Robinson, and J.A. Doudna. 2008. In vitro reconstitution of the human RISC-loading complex. *Proc Natl Acad Sci U S A* **105:** 512-517.

Maiti, M., H.C. Lee, and Y. Liu. 2007. QIP, a putative exonuclease, interacts with the Neurospora Argonaute protein and facilitates conversion of duplex siRNA into single strands. *Genes Dev* **21:** 590-600.

Makeyev, E.V. and D.H. Bamford. 2002. Cellular RNA-dependent RNA polymerase involved in posttranscriptional gene silencing has two distinct activity modes. *Mol Cell* **10:** 1417-1427.

Maniataki, E. and Z. Mourelatos. 2005. A human, ATP-independent, RISC assembly machine fueled by pre-miRNA. *Genes Dev* **19:** 2979-2990.

Maroney, P.A., Y. Yu, J. Fisher, and T.W. Nilsen. 2006. Evidence that microRNAs are associated with translating messenger RNAs in human cells. *Nat Struct Mol Biol* **13:** 1102-1107.

Martinez, J., A. Patkaniowska, H. Urlaub, R. Luhrmann, and T. Tuschl. 2002. Single-stranded antisense siRNAs guide target RNA cleavage in RNAi. *Cell* **110:** 563-574.

Mathonnet, G., M.R. Fabian, Y.V. Svitkin, A. Parsyan, L. Huck, T. Murata, S. Biffo, W.C. Merrick, E. Darzynkiewicz, R.S. Pillai, W. Filipowicz, T.F. Duchaine, and N. Sonenberg. 2007. MicroRNA inhibition of translation initiation in vitro by targeting the cap-binding complex eIF4F. *Science* **317:** 1764-1767.

Matranga, C., Y. Tomari, C. Shin, D.P. Bartel, and P.D. Zamore. 2005. Passenger-strand cleavage facilitates assembly of siRNA into Ago2-containing RNAi enzyme complexes. *Cell* **123:** 607-620.

Meister, G., M. Landthaler, A. Patkaniowska, Y. Dorsett, G. Teng, and T. Tuschl. 2004. Human Argonaute2 mediates RNA cleavage targeted by miRNAs and siRNAs. *Mol Cell* **15:** 185-197.

Mette, M.F., W. Aufsatz, J. van der Winden, M.A. Matzke, and A.J. Matzke. 2000. Transcriptional silencing and promoter methylation triggered by double-stranded RNA. *Embo J* **19:** 5194-5201.

Miranda, K.C., T. Huynh, Y. Tay, Y.S. Ang, W.L. Tam, A.M. Thomson, B. Lim, and I. Rigoutsos. 2006. A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. *Cell* **126:** 1203-1217.

Miyoshi, K., H. Tsukumo, T. Nagami, H. Siomi, and M.C. Siomi. 2005. Slicer function of Drosophila Argonautes and its involvement in RISC formation. *Genes Dev* **19:** 2837-2848.

Mochizuki, K., N.A. Fine, T. Fujisawa, and M.A. Gorovsky. 2002. Analysis of a piwi-related gene implicates small RNAs in genome rearrangement in tetrahymena. *Cell* **110:** 689-699.

Morel, J.B., C. Godon, P. Mourrain, C. Beclin, S. Boutet, F. Feuerbach, F. Proux, and H. Vaucheret. 2002. Fertile hypomorphic ARGONAUTE (ago1) mutants impaired in post-transcriptional gene silencing and virus resistance. *Plant Cell* **14:** 629-639.

Morris, K.V., S.W. Chan, S.E. Jacobsen, and D.J. Looney. 2004. Small interfering RNA-induced transcriptional gene silencing in human cells. *Science* **305:** 1289-1292.

Mourelatos, Z., J. Dostie, S. Paushkin, A. Sharma, B. Charroux, L. Abel, J. Rappsilber, M. Mann, and G. Dreyfuss. 2002. miRNPs: a novel class of ribonucleoproteins containing numerous microRNAs. *Genes Dev* **16:** 720-728.

Nam, J.W., K.R. Shin, J. Han, Y. Lee, V.N. Kim, and B.T. Zhang. 2005. Human microRNA prediction through a probabilistic co-learning model of sequence and structure. *Nucleic Acids Res* **33:** 3570-3581.

Napoli, C., C. Lemieux, and R. Jorgensen. 1990. Introduction of a Chimeric Chalcone Synthase Gene into Petunia Results in Reversible Co-Suppression of Homologous Genes in trans. *Plant Cell* **2:** 279-289.

Nishida, K.M., K. Saito, T. Mori, Y. Kawamura, T. Nagami-Okada, S. Inagaki, H. Siomi, and M.C. Siomi. 2007. Gene silencing mechanisms mediated by Aubergine piRNA complexes in Drosophila male gonad. *Rna* **13:** 1911-1922.

Nissen, P., J. Hansen, N. Ban, P.B. Moore, and T.A. Steitz. 2000. The structural basis of ribosome activity in peptide bond synthesis. *Science* **289:** 920-930.

Noller, H.F., V. Hoffarth, and L. Zimniak. 1992. Unusual resistance of peptidyl transferase to protein extraction procedures. *Science* **256:** 1416-1419.

Nottrott, S., M.J. Simard, and J.D. Richter. 2006. Human let-7a miRNA blocks protein production on actively translating polyribosomes. *Nat Struct Mol Biol* **13:** 1108-1114.

Nykanen, A., B. Haley, and P.D. Zamore. 2001. ATP requirements and small interfering RNA structure in the RNA interference pathway. *Cell* **107:** 309-321.

Ohler, U., S. Yekta, L.P. Lim, D.P. Bartel, and C.B. Burge. 2004. Patterns of flanking sequence conservation and a characteristic upstream motif for microRNA gene identification. *Rna* **10:** 1309-1322.

Olsen, P.H. and V. Ambros. 1999. The lin-4 regulatory RNA controls developmental timing in Caenorhabditis elegans by blocking LIN-14 protein synthesis after the initiation of translation. *Dev Biol* **216:** 671-680.

Pak, J. and A. Fire. 2007. Distinct populations of primary and secondary effectors during RNAi in C. elegans. *Science* **315:** 241-244.

Pal-Bhadra, M., U. Bhadra, and J.A. Birchler. 1997. Cosuppression in Drosophila: gene silencing of Alcohol dehydrogenase by white-Adh transgenes is Polycomb dependent. *Cell* **90:** 479-490.

Pal-Bhadra, M., U. Bhadra, and J.A. Birchler. 2002. RNAi related mechanisms affect both transcriptional and posttranscriptional transgene silencing in Drosophila. *Mol Cell* **9:** 315-327.

Palakodeti, D., M. Smielewska, and B.R. Graveley. 2006. MicroRNAs from the Planarian Schmidtea mediterranea: a model system for stem cell biology. *Rna* **12:** 1640-1649.

Papp, I., M.F. Mette, W. Aufsatz, L. Daxinger, S.E. Schauer, A. Ray, J. van der Winden, M. Matzke, and A.J. Matzke. 2003. Evidence for nuclear processing of plant micro RNA and short interfering RNA precursors. *Plant Physiol* **132:** 1382-1390.

Park, W., J. Li, R. Song, J. Messing, and X. Chen. 2002. CARPEL FACTORY, a Dicer homolog, and HEN1, a novel protein, act in microRNA metabolism in Arabidopsis thaliana. *Curr Biol* **12:** 1484-1495.

Parrish, S., J. Fleenor, S. Xu, C. Mello, and A. Fire. 2000. Functional anatomy of a dsRNA trigger: differential requirement for the two trigger strands in RNA interference. *Mol Cell* **6:** 1077-1087.

Peters, L. and G. Meister. 2007. Argonaute proteins: mediators of RNA silencing. *Mol Cell* **26:** 611-623.

Petersen, C.P., M.E. Bordeleau, J. Pelletier, and P.A. Sharp. 2006. Short RNAs repress translation after initiation in mammalian cells. *Mol Cell* **21:** 533-542.

Pfeffer, S., A. Sewer, M. Lagos-Quintana, R. Sheridan, C. Sander, F.A. Grasser, L.F. van Dyk, C.K. Ho, S. Shuman, M. Chien, J.J. Russo, J. Ju, G. Randall, B.D. Lindenbach, C.M. Rice, V. Simon, D.D. Ho, M. Zavolan, and T. Tuschl. 2005. Identification of microRNAs of the herpesvirus family. *Nat Methods* **2:** 269-276.

Pillai, R.S., S.N. Bhattacharyya, C.G. Artus, T. Zoller, N. Cougot, E. Basyuk, E. Bertrand, and W. Filipowicz. 2005. Inhibition of translational initiation by Let-7 MicroRNA in human cells. *Science* **309:** 1573-1576.

Que, Q., H.Y. Wang, J.J. English, and R.A. Jorgensen. 1997. The Frequency and Degree of Cosuppression by Sense Chalcone Synthase Transgenes Are Dependent on Transgene Promoter Strength and Are Reduced by Premature Nonsense Codons in the Transgene Coding Sequence. *Plant Cell* **9:** 1357-1368.

Rand, T.A., S. Petersen, F. Du, and X. Wang. 2005. Argonaute2 cleaves the anti-guide strand of siRNA during RISC activation. *Cell* **123:** 621-629.

Rehwinkel, J., I. Behm-Ansmant, D. Gatfield, and E. Izaurralde. 2005. A crucial role for GW182 and the DCP1:DCP2 decapping complex in miRNA-mediated gene silencing. *Rna* **11:** 1640-1647.

Reinhart, B.J. and D.P. Bartel. 2002. Small RNAs correspond to centromere heterochromatic repeats. *Science* **297:** 1831.

Reinhart, B.J., F.J. Slack, M. Basson, A.E. Pasquinelli, J.C. Bettinger, A.E. Rougvie, H.R. Horvitz, and G. Ruvkun. 2000. The 21-nucleotide *let-7* RNA regulates developmental timing in *Caenorhabditis elegans*. *Nature* **403:** 901-906.

Reinhart, B.J., E.G. Weinstein, M.W. Rhoades, B. Bartel, and D.P. Bartel. 2002. MicroRNAs in plants. *Genes Dev* **16:** 1616-1626.

Rhoades, M.W., B.J. Reinhart, L.P. Lim, C.B. Burge, B. Bartel, and D.P. Bartel. 2002. Prediction of plant microRNA targets. *Cell* **110:** 513-520.

Rivas, F.V., N.H. Tolia, J.J. Song, J.P. Aragon, J. Liu, G.J. Hannon, and L. Joshua-Tor. 2005. Purified Argonaute2 and an siRNA form recombinant human RISC. *Nat Struct Mol Biol* **12:** 340-349.

Rodriguez, A., S. Griffiths-Jones, J.L. Ashurst, and A. Bradley. 2004. Identification of mammalian microRNA host genes and transcription units. *Genome Res* **14:** 1902-1910.

Romano, N. and G. Macino. 1992. Quelling: transient inactivation of gene expression in Neurospora crassa by transformation with homologous sequences. *Mol Microbiol* **6:** 3343-3353.

Saito, K., A. Ishizuka, H. Siomi, and M.C. Siomi. 2005. Processing of pre-microRNAs by the Dicer-1-Loquacious complex in Drosophila cells. *PLoS Biol* **3:** e235.

Saito, K., K.M. Nishida, T. Mori, Y. Kawamura, K. Miyoshi, T. Nagami, H. Siomi, and M.C. Siomi. 2006. Specific association of Piwi with rasiRNAs derived from retrotransposon and heterochromatic regions in the Drosophila genome. *Genes Dev* **20:** 2214-2222.

Saito, K., Y. Sakaguchi, T. Suzuki, H. Siomi, and M.C. Siomi. 2007. Pimet, the Drosophila homolog of HEN1, mediates 2'-O-methylation of Piwi- interacting RNAs at their 3' ends. *Genes Dev* **21:** 1603-1608.

Schramke, V. and R. Allshire. 2003. Hairpin RNAs and retrotransposon LTRs effect RNAi and chromatin-based gene silencing. *Science* **301:** 1069-1074.

Schwarz, D.S., G. Hutvagner, T. Du, Z. Xu, N. Aronin, and P.D. Zamore. 2003. Asymmetry in the assembly of the RNAi enzyme complex. *Cell* **115:** 199-208.

Schweet, R., H. Lamfrom, and E. Allen. 1958. The Synthesis of Hemoglobin in a Cell-Free System. *Proc Natl Acad Sci U S A* **44:** 1029-1035.

Seggerson, K., L. Tang, and E.G. Moss. 2002. Two genetic circuits repress the Caenorhabditis elegans heterochronic gene lin-28 after translation initiation. *Dev Biol* **243:** 215-225.

Sempere, L.F., S. Freemantle, I. Pitha-Rowe, E. Moss, E. Dmitrovsky, and V. Ambros. 2004. Expression profiling of mammalian microRNAs uncovers a subset of brain-expressed microRNAs with possible roles in murine and human neuronal differentiation. *Genome Biol* **5:** R13.

Sen, G.L. and H.M. Blau. 2005. Argonaute 2/RISC resides in sites of mammalian mRNA decay known as cytoplasmic bodies. *Nat Cell Biol* **7:** 633-636.

Sewer, A., N. Paul, P. Landgraf, A. Aravin, S. Pfeffer, M.J. Brownstein, T. Tuschl, E. van Nimwegen, and M. Zavolan. 2005. Identification of clustered microRNAs using an ab initio prediction method. *BMC Bioinformatics* **6:** 267.

Sheth, U. and R. Parker. 2003. Decapping and decay of messenger RNA occur in cytoplasmic processing bodies. *Science* **300:** 805-808.

Shiu, P.K., N.B. Raju, D. Zickler, and R.L. Metzenberg. 2001. Meiotic silencing by unpaired DNA. *Cell* **107:** 905-916.

Sijen, T., J. Fleenor, F. Simmer, K.L. Thijssen, S. Parrish, L. Timmons, R.H. Plasterk, and A. Fire. 2001. On the role of RNA amplification in dsRNA-triggered gene silencing. *Cell* **107:** 465-476.

Sijen, T. and R.H. Plasterk. 2003. Transposon silencing in the *Caenorhabditis elegans* germ line by natural RNAi. *Nature* **426:** 310-314.

Sijen, T., F.A. Steiner, K.L. Thijssen, and R.H. Plasterk. 2007. Secondary siRNAs result from unprimed RNA synthesis and form a distinct class. *Science* **315:** 244-247.

Smardon, A., J.M. Spoerke, S.C. Stacey, M.E. Klein, N. Mackin, and E.M. Maine. 2000. EGO-1 is related to RNA-directed RNA polymerase and functions in germ-line development and RNA interference in *C. elegans*. *Curr Biol* **10:** 169-178.

Song, J.J., S.K. Smith, G.J. Hannon, and L. Joshua-Tor. 2004. Crystal structure of Argonaute and its implications for RISC slicer activity. *Science* **305:** 1434-1437.

Stark, A., J. Brennecke, R.B. Russell, and S.M. Cohen. 2003. Identification of *Drosophila* microRNA targets. *PLoS Biol* **1:** E60.

Svoboda, P., P. Stein, W. Filipowicz, and R.M. Schultz. 2004. Lack of homologous sequence-specific DNA methylation in response to stable dsRNA expression in mouse oocytes. *Nucleic Acids Res* **32:** 3601-3606.

Tabara, H., R.J. Hill, C.C. Mello, J.R. Priess, and Y. Kohara. 1999. pos-1 encodes a cytoplasmic zinc-finger protein essential for germline specification in C. elegans. *Development* **126:** 1-11.

Tabara, H., E. Yigit, H. Siomi, and C.C. Mello. 2002. The dsRNA binding protein RDE-4 interacts with RDE-1, DCR-1, and a DExH-box helicase to direct RNAi in C. elegans. *Cell* **109:** 861-871.

Teixeira, D., U. Sheth, M.A. Valencia-Sanchez, M. Brengues, and R. Parker. 2005. Processing bodies require RNA for assembly and contain nontranslating mRNAs. *Rna* **11:** 371-382.

Thermann, R. and M.W. Hentze. 2007. Drosophila miR2 induces pseudo-polysomes and inhibits translation initiation. *Nature* **447:** 875-878.

Tomari, Y., T. Du, B. Haley, D.S. Schwarz, R. Bennett, H.A. Cook, B.S. Koppetsch, W.E. Theurkauf, and P.D. Zamore. 2004. RISC assembly defects in the Drosophila RNAi mutant armitage. *Cell* **116:** 831-841.

Tomari, Y., T. Du, and P.D. Zamore. 2007. Sorting of Drosophila small silencing RNAs. *Cell* **130:** 299-308.

van der Krol, A.R., L.A. Mur, M. Beld, J.N. Mol, and A.R. Stuitje. 1990. Flavonoid genes in petunia: addition of a limited number of gene copies may lead to a suppression of gene expression. *Plant Cell* **2:** 291-299.

Vaucheret, H., F. Vazquez, P. Crete, and D.P. Bartel. 2004. The action of ARGONAUTE1 in the miRNA pathway and its regulation by the miRNA pathway are crucial for plant development. *Genes Dev* **18:** 1187-1197.

Verdel, A., S. Jia, S. Gerber, T. Sugiyama, S. Gygi, S.I. Grewal, and D. Moazed. 2004. RNAi-mediated targeting of heterochromatin by the RITS complex. *Science* **303:** 672-676.

Volpe, T.A., C. Kidner, I.M. Hall, G. Teng, S.I. Grewal, and R.A. Martienssen. 2002. Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. *Science* **297:** 1833-1837.

Wakiyama, M., K. Takimoto, O. Ohara, and S. Yokoyama. 2007. Let-7 microRNA-mediated mRNA deadenylation and translational repression in a mammalian cell-free system. *Genes Dev* **21:** 1857-1862.

Wang, B., T.M. Love, M.E. Call, J.G. Doench, and C.D. Novina. 2006. Recapitulation of short RNA-directed translational gene silencing in vitro. *Mol Cell* **22:** 553-560.

Wianny, F. and M. Zernicka-Goetz. 2000. Specific interference with gene function by double-stranded RNA in early mouse development. *Nat Cell Biol* **2:** 70-75.

Wightman, B., I. Ha, and G. Ruvkun. 1993. Posttranscriptional regulation of the heterochronic gene *lin-14* by *lin-4* mediates temporal pattern formation in *C. elegans*. *Cell* **75:** 855-862.

Wu, H., H. Xu, L.J. Miraglia, and S.T. Crooke. 2000. Human RNase III is a 160-kDa protein involved in preribosomal RNA processing. *J Biol Chem* **275:** 36957-36965.

Wu, L., J. Fan, and J.G. Belasco. 2006. MicroRNAs direct rapid deadenylation of mRNA. *Proc Natl Acad Sci U S A* **103:** 4034-4039.

Xie, X., J. Lu, E.J. Kulbokas, T.R. Golub, V. Mootha, K. Lindblad-Toh, E.S. Lander, and M. Kellis. 2005. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* **434:** 338-345.

Xie, Z., L.K. Johansen, A.M. Gustafson, K.D. Kasschau, A.D. Lellis, D. Zilberman, S.E. Jacobsen, and J.C. Carrington. 2004. Genetic and functional diversification of small RNA pathways in plants. *PLoS Biol* **2:** E104.

Yao, M.C., P. Fuller, and X. Xi. 2003. Programmed DNA deletion as an RNA-guided system of genome defense. *Science* **300:** 1581-1584.

Yekta, S., I.H. Shih, and D.P. Bartel. 2004. MicroRNA-directed cleavage of HOXB8 mRNA. *Science* **304:** 594-596.

Yi, R., Y. Qin, I.G. Macara, and B.R. Cullen. 2003. Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. *Genes Dev* **17:** 3011-3016.

Yigit, E., P.J. Batista, Y. Bei, K.M. Pang, C.C. Chen, N.H. Tolia, L. Joshua-Tor, S. Mitani, M.J. Simard, and C.C. Mello. 2006. Analysis of the C. elegans Argonaute family reveals that distinct Argonautes act sequentially during RNAi. *Cell* **127:** 747-757.

Yuan, Y.R., Y. Pei, J.B. Ma, V. Kuryavyi, M. Zhadina, G. Meister, H.Y. Chen, Z. Dauter, T. Tuschl, and D.J. Patel. 2005. Crystal structure of A. aeolicus argonaute, a site-specific DNA-guided endoribonuclease, provides insights into RISC-mediated mRNA cleavage. *Mol Cell* **19:** 405-419.

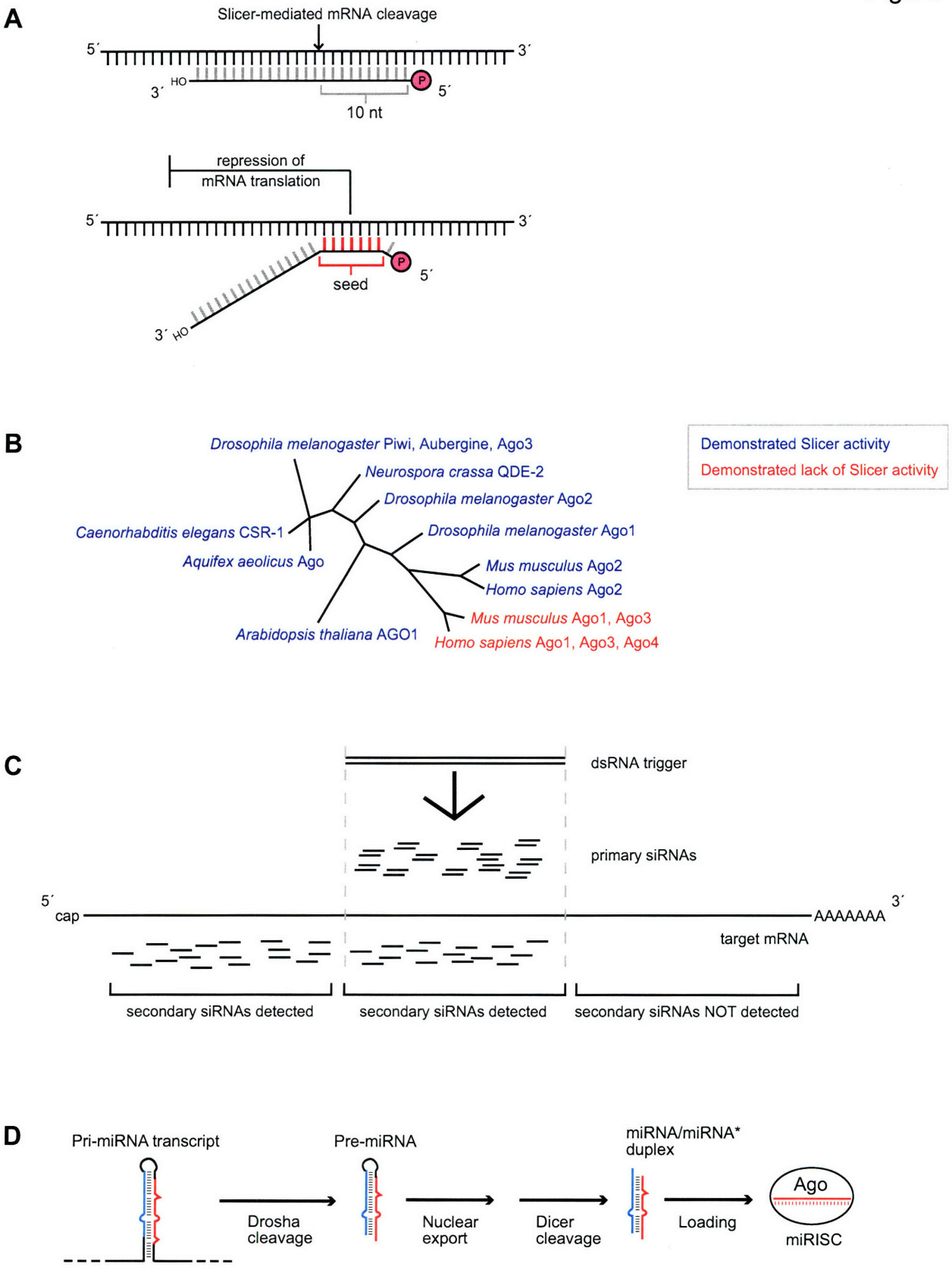
Zamore, P.D., T. Tuschl, P.A. Sharp, and D.P. Bartel. 2000. RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. *Cell* **101:** 25-33.

Zeng, Y. and B.R. Cullen. 2004. Structural requirements for pre-microRNA binding and nuclear export by Exportin 5. *Nucleic Acids Res* **32:** 4776-4785.

Zeng, Y. and B.R. Cullen. 2005. Efficient processing of primary microRNA hairpins by Drosha requires flanking nonstructured RNA sequences. *J Biol Chem* **280:** 27595-27603.

Zilberman, D., X. Cao, and S.E. Jacobsen. 2003. ARGONAUTE4 control of locus-specific siRNA accumulation and DNA and histone methylation. *Science* **299:** 716-719.

Figure 1


A
Slicer-mediated mRNA cleavage
5′
3′
3′
HO
P
5′
10 nt
repression of
mRNA translation
5′
3′
P
5′
seed
3′
HO
B
Demonstrated Slicer activity
Demonstrated lack of Slicer activity
Drosophila melanogaster Piwi, Aubergine, Ago3
Neurospora crassa QDE-2
Drosophila melanogaster Ago2
Caenorhabditis elegans CSR-1
Drosophila melanogaster Ago1
Aquifex aeolicus Ago
Mus musculus Ago2
Homo sapiens Ago2
Mus musculus Ago1, Ago3
Arabidopsis thaliana AGO1
Homo sapiens Ago1, Ago3, Ago4
C
dsRNA trigger
primary siRNAs
5′
cap
AAAAAAA
3′
target mRNA
secondary siRNAs detected
secondary siRNAs detected
secondary siRNAs NOT detected
D
Pri-miRNA transcript
Pre-miRNA
miRNA/miRNA*
duplex
Drosha
cleavage
Nuclear
export
Dicer
cleavage
Loading
Ago
miRISC

Chapter 2

Large-Scale Sequencing Reveals 21U-RNAs and Additional MicroRNAs and Endogenous siRNAs in *Caenorhabditis elegans*

J. Graham Ruby[1,2], Calvin Jan[1,2], Christopher Player[1], Michael J. Axtell[1], William Lee[3], Chad Nusbaum[3], Hui Ge[1], David P. Bartel[1,2]

[1]Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA, 02142
[2]Howard Hughes Medical Institute and Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, 02139
[3]Broad Institute of MIT and Harvard, Cambridge, MA, 02142

J.G.R. performed the computational analysis excluding siRNA gene mountain analysis, which was performed by C.P. and H.G. M.J.A., W.L., and C.N. sequenced the libraries. C.J. performed molecular analyses. J.G.R. and D.P.B. wrote the manuscript.

Electronic supplemental files are provided on the accompanying CD-ROM, under the directory "Chapter 2". All files are ASCI text; .html files are best opened with a web browser, and .fa are fasta-formatted text that is best viewed with a text editor.

Published as:
JG Ruby, C Jan, C Player, MJ Axtell, W Lee, C Nusbaum, H Ge, and DP Bartel. (2006) Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. *Cell.* 127:1193-1207.

Abstract

We sequenced ~400,000 small RNAs from *Caenorhabditis elegans*. Another 18 microRNA (miRNA) genes were identified, thereby extending to 112 our tally of confidently identified miRNA genes in *C. elegans*. Also observed were thousands of endogenous siRNAs generated by RNA-directed RNA polymerases acting preferentially on transcripts associated with spermatogenesis and transposons. In addition, a third class of nematode small RNAs, called 21U-RNAs, was discovered. 21U-RNAs are precisely 21 nucleotides long, begin with a uridine 5′-monophosphate but are diverse in their remaining 20 nucleotides, and appear modified at their 3′-terminal ribose. 21U-RNAs originate from more than 5700 genomic loci dispersed in two broad regions of chromosome IV—primarily between protein-coding genes or within their introns. These loci share a large upstream motif that enables accurate prediction of additional 21U-RNAs. The motif is conserved in other nematodes, presumably because of its importance for producing these diverse, autonomously expressed, small RNAs (dasRNAs).

Introduction

RNAs ~22 nt in length play gene-regulatory roles in numerous eukaryotic lineages, including plants, animals, and fungi (Bartel 2004; Nakayashiki 2005). The first endogenous ~22-nt RNAs discovered in eukaryotes were the *lin-4* and *let-7* RNAs, both of which were found by mapping mutant *C. elegans* loci (Lee et al. 1993; Reinhart et al. 2000). The mature *lin-4* and *let-7* RNAs are each processed from a hairpin formed within their respective primary transcripts. Through molecular cloning and sequencing, many small RNAs with the potential to arise from foldback structures characteristic of the *lin-4*

and *let-7* hairpins were identified, including more than 50 from *C. elegans*, thereby establishing a class of endogenous RNAs called miRNAs (Lagos-Quintana et al. 2001; Lau et al. 2001; Lee and Ambros 2001). Additional miRNAs have been identified in *C. elegans* by cloning, genetics, or computational prediction supported by experimentation (Ambros et al. 2003; Grad et al. 2003; Johnston and Hobert 2003; Lim et al. 2003; Ohler et al. 2004).

In addition to the miRNA, a less abundant species, known as the miRNA star (miRNA*), derives from the miRNA hairpin precursor (Lau et al. 2001; Lim et al. 2003). The miRNA and miRNA* species pair to each other with ~2-nt 3′ overhangs. In animals, this miRNA:miRNA* duplex is generated by the sequential action of Drosha and Dicer RNAse-III endonucleases (Grishok et al. 2001; Hutvagner et al. 2001; Lee et al. 2003). Drosha cleaves at sites near the base of the stem, thereby liberating a 60- to 70-nt fragment comprising the majority of the hairpin, which Dicer then cleaves at sites near the loop (Han et al. 2006; Lee et al. 2003). The miRNA strand of the resulting miRNA:miRNA* duplex is then loaded into a silencing complex, which contains at its core a member of the Argonaute family of proteins (Hutvagner and Zamore 2002; Mourelatos et al. 2002).

Once incorporated into the silencing complex, the miRNA serves as a guide to direct the post-transcriptional repression of protein-coding messages. Most important for target recognition is pairing to the miRNA seed, defined as the 6-nt segment comprising nucleotides two through seven, counting from the 5′ terminus of the miRNA (Brennecke et al. 2005; Doench and Sharp 2004; Lewis et al. 2005; Lewis et al. 2003). When comparing related miRNAs, the seed is also the most conserved portion of the RNA, and

C. elegans miRNAs can be grouped into families based largely on their shared seed sequences (Ambros et al. 2003; Lim et al. 2003).

Other types of endogenous small RNAs have been found within libraries made from *C. elegans*. Those that are antisense to *C. elegans* mRNAs have been classified as small interfering RNAs (siRNAs), with the idea that they might be processed from long double-stranded RNA (dsRNA) and might direct the silencing of complementary mRNAs (Ambros et al. 2003; Lim et al. 2003). Other cloned and sequenced ~22-nt RNAs do not appear to correspond to protein-coding regions, do not have the potential to arise from hairpins characteristic of miRNA precursors, and yet are expressed at sufficiently high levels to be detected on RNA blots. These have been annotated as tiny non-coding RNAs (tncRNAs; (Ambros et al. 2003). In flies and mammals, other distinct classes of small RNAs have been reported, including repeat-associated siRNAs (rasiRNAs; (Aravin et al. 2003; Vagin et al. 2006) and Piwi-interacting RNAs (piRNAs; (Aravin et al. 2006; Girard et al. 2006; Lau et al. 2006).

Recent advances in high-throughput sequencing technology have allowed for a more complete assessment of the global small RNA population in plants (Lu et al. 2005). Here, we applied high-throughput pyrosequencing methods (Margulies et al. 2005) to the discovery of small RNAs expressed in mixed-staged *C. elegans*. Our results reshape the list of known miRNAs by reporting newly identified miRNA genes, defining the processing of most previously annotated miRNAs, refining the termini of some, and raising new questions as to the authenticity of others. In addition, we describe thousands of endogenous siRNAs that appear to be RNA-templated products of activities acting preferentially on messages associated with spermatogenesis and transposons. We also

describe the 21U-RNAs, which originate from an estimated 12,000-16,000 genomic loci dispersed between and within protein-coding genes in two broad regions of chromosome IV. These loci each have a conserved upstream motif, which we propose specifies the production of 21U-RNAs from thousands of non-coding transcripts.

Results

Our library of small RNAs isolated from mixed-staged *C. elegans* was previously constructed so as to represent only those RNAs with 5′ monophosphate and 3′ hydroxyl groups, the termini expected of miRNAs and siRNAs (Lau et al. 2001). Standard sequencing of this and similar libraries previously yielded sequences of 4078 small RNA clones that match the *C. elegans* genome (Lau et al. 2001; Lim et al. 2003). High-throughput pyrosequencing (Margulies et al. 2005) of the library yielded 394,926 sequence reads that perfectly matched the worm genome. Of those, 80% matched annotated miRNA hairpins. Another 6.4% matched other annotated non-coding RNA genes, such as rRNA and tRNA, and were present at similar frequencies for each length from 18 to 28 nt, which was the pattern expected for degradation fragments of these non-coding RNAs. Another 9.3% corresponded to 21U-RNAs, and at least 0.7% corresponded to endogenous siRNAs that were antisense to annotated exons. The remaining sequences included what appeared to be endogenous siRNAs that were antisense to annotated introns, mRNA/intron degradation fragments, and a small contingent of uncharacterized sequences.

Previously annotated miRNAs

Our previous sequencing of small RNA libraries from *C. elegans* discovered, refined, or confirmed the identities of 80 miRNAs (Lau et al. 2001; Lim et al. 2003). All 80 were observed in the new set of high-throughput reads, at relative frequencies similar to those observed previously (Table S1). As exemplified by *lin-4* (Figure 1A-B), these 80 miRNA genes were typically represented by one dominantly sequenced species, the miRNA, as well as a sequence from the opposing arm of the hairpin, the miRNA* (Table S1, Supplemental text). In addition, sequences were sometimes observed that matched the portion of the transcript in between the miRNA and miRNA* (Figure 1A-B; Table S1).

On average, the miRNA* species was present at about 1.0% the frequency of the miRNA. When paired to the miRNA it generally exhibited the 3′ overhangs typical of miRNA hairpin processing (Lee et al. 2003; Lim et al. 2003). Identifying the dominant miRNA* species for many of the miRNAs, together with information on end heterogeneity, provided useful data for considering the specificity and precision of Drosha and Dicer processing. For example, the observed miRNA 5′ ends were far more homogenous (99.5% identical) than the miRNA* 5′ ends (91% identical), which were more homogenous than the miRNA 3′ ends (85% identical) and miRNA* 3′ ends (77% identical). About half of the 5′ heterogeneity was from reads that were longer than the dominant species, implicating imprecise Drosha/Dicer processing as the major cause of heterogeneity at this end. Greater 3′ heterogeneity was attributed to three factors: 1) less precise Drosha/Dicer processing, as indicated by templated nucleotides extending beyond the dominant species, 2) preferential degradation at the 3′ end, and 3) addition of untemplated nucleotides to the 3′ ends of miRNA and miRNA* species. The more

precise cut at the miRNA 5′ end, compared to the miRNA* 5′ end, presumably reflected selective pressure for accurately defining the miRNA seed. Cleavage by either Drosha or Dicer appeared equally consistent when that cut would set the seed. The observation that when Dicer set the seed it was more precise than Drosha disfavored models in which Dicer simply measures from the Drosha cuts and suggested that additional determinants are employed when needed to more accurately define Dicer cleavage.

Examining the dominant mature miRNA sequences revealed that 1.33% 3′ ends were extended by a single untemplated nucleotide, with U being the preferred untemplated nucleotide (54%, Table S1). A second untemplated nucleotide appeared with greater efficiency (4% of those already extended by one untemplated nucleotide) and with greater preference for U (73%). Similar efficency and U preference was observed for a third nucleotide. The untemplated uridylation of miRNAs was reminiscent of that reported for unmethylated small RNAs in *Arabidopsis* (Li et al. 2005).

As expected, the high-throughput reads also included some annotated *C. elegans* miRNAs that were not among the 80 previously sequenced from our libraries. Thirteen such previously annotated miRNA hairpins gave rise to high-throughput reads (Table S1). All 13 were originally identified computationally and then experimentally supported by northern blotting and/or a PCR-based assay (Lim et al. 2003; Ohler et al. 2004). For five of these, the 5′ terminus did not match the one previously annotated, an observation with ramifications for the experimental validation of computational candidates (Supplemental text). No reads matched 19 of the *C. elegans* miRNA hairpins annotated in miRBase (Supplemental text). Of these 19, one was the *lys-6* miRNA, which had been

identified genetically and appears to be expressed in only a few cells (Johnston and Hobert 2003).

Newly Identified miRNAs

In a search for additional miRNAs, we evaluated reads that fell within potential miRNA-like hairpins, considering the following criteria: 1) the pairing characteristics of the hairpin; 2) the expression of the candidate, as measured by the abundance of sequence reads sharing the same 5′ terminus; 3) evolutionary conservation, as evaluated by the apparent conservation of the hairpin in *C. briggsae* and grouping of the miRNA candidate into a family based on its seed sequence; 4) the absence of annotation suggesting non-miRNA biogenesis; and 5) the presence of reads corresponding to the predicted miRNA* species. The observation of both a candidate miRNA and a candidate miRNA* in a set of reads provides particularly compelling evidence for Dicer-like processing from an RNA hairpin. As illustrated for miR-786 (Figure 1C-E), seven newly identified genes satisfied all of our criteria (Table1). Eleven others satisfied a subset of the criteria deemed sufficient for confident annotation as miRNAs. Three additional candidates that were sequenced more than once were, from our perspective, borderline cases and therefore not annotated here as miRNAs (Supplemental text).

Sequencing frequencies of the all newly and previously sequenced miRNAs are illustrated (Figure 1F). Seven newly identified genes were near another miRNA gene and on the same genomic strand (Table 1), an arrangement implying processing from a common polycystronic transcript (Lagos-Quintana et al. 2001; Lau et al. 2001). Seven newly identified genes added to previously known *C. elegans* miRNA families, in that

they shared the same seed (Table 2). For example, miR-793, miR-794, and miR-795 all added to the *let-7*/48/84/241 family. Four other newly identified genes shared seeds with miRNAs annotated in distant species, thereby extending the scope of families previously identified in insects or vertebrates to the nematode lineage (Table 2).

21U-RNAs

After accounting for the miRNAs and other types of annotated non-coding RNAs, the remaining reads were dominated by 21mers with 5′ uridines. We refer to the bulk of these as '21U-RNAs'. The vast majority of RNAs with these properties mapped to two broad but distinct regions of chromosome IV, one spanning chromosomal coordinates 4.5M to 7.0M, the other spanning 13.5M to 17.2M (Figure 2A). A few mapped to a third region, which spanned coordinates 9M to 9.7M of chromosome IV. The ~34,300 21U-RNA reads that derived from these three regions contained 5,454 unique sequences (Figure 2B), for which 5,302 loci were unambiguously mapped because their sequences were unique in the assembly. Many of these loci were represented by single reads in our set, suggesting the existence of more members of this small RNA class than were directly observed. Nonetheless, most of the 21U-RNA loci (67%) were represented by two or more identical reads, indicating that the 34,300 reads captured a non-trivial portion of the 21U-RNA diversity.

Four 21U-RNAs were sequenced more than 200 times, including 21UR-1 (pUGGUACGUACGUUAACCGUGC), which was represented by 521 reads and detectable on RNA blots. This 21U-RNA was sensitive to alkaline hydrolysis and phosphatase treatment, and was a suitable substrate for RNA ligase—the expected

properties of an RNA with a 5′ monophosphate (Figure 3C and S1). 21UR-1 was also resistant to periodate treatment (Figure 3C), indicating that its 3′ nucleotide was missing the *cis* diol and suggesting modification at either the 2′ or 3′ oxygen of this nucloetide, as reported for small RNAs in plants and rasiRNAs in flies (Li et al. 2005; Vagin et al. 2006).

The 21U-RNAs mapped to both strands of the DNA, but overlapped with each other or with other sequenced small RNAs on the opposing DNA strand less frequently than would be expected by chance given a random distribution, thereby providing no evidence for a dsRNA precursor. WormBase-annotated genes were somewhat less abundant within the 21U-RNA-rich portions of chromosome IV (mean ± s.d. of 93 ± 28 genes per 500 kb) compared to the genome as a whole (116 ± 26 genes per 500 kb). The vast majority of the 21U-RNAs mapped either between genes or within introns, with no preference for the sense or antisense orientiation among intronic matches. Only 2.5% of the 21U-RNA loci overlapped annotated exons, a substantial depletion versus the total fraction of the regions overlapping exons (~21%), and the read abundance of sense versus antisense exonic matches was nearly even (~750 and ~810, respectively). Overall, the genomic data suggested that the 21U-RNA loci are maintained independently of other genetic elements, with informational constraints that can conflict with those of other genes.

The ~34,300 21U-RNA reads in our set of high-throughput reads came from a mixed-staged library, raising the question of which stage(s) in development the 21U-RNAs might accumulate. Our previous effort (Lim et al. 2003) included reads from this mixed-stage library as well as reads from a larval stage L1 library, a dauer (dormant L3)

library and a mixed-staged library made from *him-8* mutant worms (which are enriched in males). Revisiting the 4078 reads from that earlier study revealed that 125 represented 21U-RNAs: 79 from mixed stage, 8 from dauer, 10 from L1, and 28 from *him-8*. Normalizing to the read counts of miRNAs with constant expression throughout larval development, the *him-8* library was ~2-fold enriched in 21U-RNAs compared to the wild-type mixed-stage library, whereas the L1 and dauer libraries were ~2- and ~3-fold depleted, respectively. The presence of 21U-RNAs in both L1 worms and dauer L3 worms implies their presence throughout much of worm development.

Two Sequence Motifs Associated with 21U-RNA Loci

Other than the U at their 5′ termini, the 21U-RNAs shared little sequence identity. Indeed, the composition of the four nucleotides was more equivalent for the 21U-RNAs than for their broader genomic contexts, which were A-T rich. However, the 21U-RNA genomic loci did share two upstream sequence motifs, one much larger than the other (Figure 3). The large motif was 34 bp and centered on an 8-nt core consensus sequence, CTGTTTCA. The small motif had a core sequence of YRNT, in which the T corresponded to the 5′ U of the 21U-RNA. The two subdomains of the motif were separated by a spacer typically 19–21 bp (Figure 3B).

A position-specific scoring matrix based on the combined properties of the two motifs was used to predict 21U-RNAs on *C. elegans* chromosome IV. With a score cut-off that correctly predicted 77% of the sequenced 21U-RNAs, 10,807 loci were identified on both strands of chromosome IV. The density of genomic matches to the motifs corresponded well to that of known 21U-RNA loci, demonstrating the specificity of our

motif-scanning procedure (Figure 2B and C). As illustrated for a 100-kb region of chromosome 4, this correspondence held at high-resolution views (Figure 2D). As a test of sensitivity, we cross-checked the 10,807 predictions with an independent set of 245,420 *C. elegans* small RNA reads provided by Andrew Fire (personal communication) and found that nearly half (46%) of the 21U-RNAs uniquely identified in this independent dataset had been predicted (see Methods). We suggest that the correspondence of 21U-RNAs predicted through motif scanning with those detected by sequencing reflected the function of the motifs in specifying 21U-RNA production in the animal.

Discovery of the upstream motif allowed assessment of the other properties ascribed to 21U-RNAs (Figure S2). Nearly all of the motif-associated 21mer reads (99.8%) began with a U, and 98.5% derived from the defined 21U-rich regions. Over 99% of the motif-associated reads were 21 nt or less, with those that were shorter (5.4%) likely corresponding to 3′ degradation products.

To explore the potential conservation of 21U-RNAs, we scanned all the *C. briggsae* genomic contigs (Stein et al. 2003) for motif matches. Each *C. briggsae* contig with a high concentration of motifs (≥ 75 per 100 kb) was syntenous with one of the three 21U-rich regions of *C. elegans* chromosome IV (Figure 2A and B). We conclude that any roles that the motifs might play in the biogenesis of 21U-RNAs have been conserved in the ~100 million years since the divergence of these two nematode species (Coghlan and Wolfe 2002). The 21U-RNAs themselves, in contrast, showed little evidence for conservation. Of the >10,000 21U-RNA sequences predicted on chromosome IV of *C.*

elegans and the >11,000 sequences similarly predicted in *C. briggsae*, not a single sequence was shared between the two species.

Endogenous siRNAs

Of the remaining sequences with perfect matches to the *C. elegans* genome, some were antisense to known protein-coding transcripts. In fact, a larger number matched the antisense strand of spliced mRNAs (2934 reads, 2378 unique sequences; Figure 4A) than matched the sense strand (2150 reads, 1800 unique sequences; Figure 4B). As done previously (Ambros et al. 2003; Lau et al. 2001; Lim et al. 2003), we classified the RNAs matching the antisense strand as candidate endogenous siRNAs, which for simplicity we refer to herein as siRNAs. RNAs that matched the sense strand also might include endogenous siRNAs, but as they likely include other hydrolysis products, we refer to them as sense RNAs.

For different *C. elegans* libraries, the proportion of miRNAs to siRNAs varies greatly; our libraries contain 100-times more miRNAs than siRNAs, whereas the Ambros library contains roughly equal numbers of the two (Ambros et al. 2003; Lim et al. 2003). The large difference suggests that most *C. elegans* siRNAs lack the 5′ monophosphate required by our cloning protocol (Ambros et al. 2003). Perhaps they are short RNA-dependent RNA polymerase (RdRP) products that have retained their 5′ triphosphate. Consistent with this idea, we detected a population of endogenous ~22mers that were suitable substrates for an in vitro 5′-capping reaction requiring a 5′ di- or triphosphate (Figure 4C). These sequences would be underrepresented in our library, although not totally absent if some molecules lost their γ and β phosphates or were transcribed with an

initiating nucleoside monophosphate rather than nucleoside triphosphate, as has been observed for other RNA polymerases (Martin and Coleman 1989; Ranjith-Kumar et al. 2002).

Recognizing the siRNAs of our library were likely depleted in the major subclass of endogenous siRNAs, we proceeded with their analysis. Their length distribution had prominent peaks at 21, 22, and 26 nt (Figure 4A and B). Comparison to the length distribution of reads matching tRNA and rRNA indicated that the 26mer siRNA population was distinct, rather than the shoulder of a larger, more broadly distributed population. A preference for a 5′ G, observed previously for siRNAs (Ambros et al., 2003), was persistent across all lengths of endogenous siRNAs but strongest among 26mers. A 26mer siRNA sequenced 9 times had a 5′ monophosphate (siR26-1, pGCAAGAUGGAAAAGUUUGAGAUUCCG; Figure S1). As observed for the 21U-RNA, this siRNA was resistant to periodate treatment, again suggesting modification at either the 2′ or 3′ oxygen of the 3′ nucleotide (Figure 3C). With so many classes of plant and animal small RNAs now shown to be resistant to periodate oxidation, metazoan miRNAs appear increasingly unusual in not being modified at their 3′ residue.

Despite being spread out over a large number of genes, dense clusters of siRNAs were observed at some genomic loci (Figure 4D, Table S3). Examination of surrounding sequence revealed that siRNAs did not exclusively match annotated exons. For example, some also matched annotated introns. Nonetheless, more than 40 of the unique sequences represented by our reads did not match the genomic DNA but instead spanned splice junctions (exemplified in Figure 4E), implying that these RNAs were produced by an RdRP acting on a spliced transcript. Because these junction-spanning siRNAs had the

length distribution and preference for a 5′ G characteristic of the siRNAs in general, it is reasonable propose that the remainder of the siRNAs were also RdRP products and that at least some of the RdRP activity was nuclear and thus could act on both spliced and unspliced templates.

Correlations with siRNAs supported the idea that the biogenesis or function of some sense RNAs was linked to that of the siRNAs. The overlap of siRNA-complemented genes was greater with genes matching sense RNAs (24%) than with genes picked using SAGE data to control for expression (16%; p-value <0.01, chi-square test). Among the sense-antisense pairs with at least 1-nt overlap at their genomic loci, 30% maximally overlapped (exemplified by all four sense reads in Figure 4D), which was 5-fold higher than expected by chance. For 47% of the sense-antisense pairs involving 26mers, the most common configuration placed the 5′ nucleotide of the sense read across from nucleotide 23 of a 26mer siRNA (exemplified by three sense reads in Figure 4D), which was 20-fold higher than chance expectation.

To gain insight into the biological consequences of siRNAs, we examined the functional categorization of genes they complemented. In addition to the enrichment for matching transposon genes, observed previously (Lee et al. 2006), the siRNAs had a high propensity to match sperm-enriched genes (Supplemental text). This propensity was particularly striking for the 26mer siRNAs, 55% of which matched sperm-enriched genes.

Discussion

112 Confidently Identified *C. elegans* miRNAs

The set of miRNA genes represented in our high-throughput reads included 93 previously annotated genes, plus 18 newly discovered genes (Table S1). The notable exception was the *lsy-6* miRNA, a genetically identified miRNA thought to be transcribed in only one to nine cells (Johnston and Hobert 2003). The absence of *lsy-6* in a set that included 37,225 reads of miR-52 illustrated the extreme diversity in metazoan miRNA expression. This difference can be attributed solely to the specific expression of *lys-6* in cells that are few in number and small in volume; we estimated that *lsy-6* RNA should have been ~100,000 times less abundant than a miRNA expressed in most cells of the worm (Supplemental text). Clearly, more reads must be sequenced before all the miRNAs expressed during the course of nematode development will be catalogued.

Although the unsaturated status of our sequencing project prohibited any definitive judgments about miRNA annotations that were not represented by our reads, our observations were informative for evaluating the confidence in those annotations and the data originally used to justify them. These considerations increased the number of annotated genes whose authenticity is in doubt (Supplemental text). Nonetheless, the 18 newly identified miRNA genes enabled the number of confidently identified *C. elegans* miRNAs to be revised upwards to 112, which included the 111 represented in our high-throughput reads, plus *lsy-6*. Currently annotated loci with reasonable prospects of eventually joining the list include *mir-273*, for which reverse-genetic functional data has been reported (Chang et al. 2004). Our three borderline candidates also might eventually be added (Supplemental text). These include one that was represented by only five reads and lacked conservation or miRNA* evidence, and two that might be considered "young" miRNAs, potential Drosha/Dicer substrates that might have recently emerged from short

inverted duplications and have not had sufficient time to acquire the mismatches usually observed in miRNA hairpins (Table S1). Our results also prompted re-evaluation of miRNA gene-number estimates in worms (Supplemental text).

The 112 confidently identified *C. elegans* miRNA genes arose from 83 genomic clusters, ranging from one to seven genes per cluster (Table S2). When grouped according to their seeds, they fell into 63 families, 58 (92%) of which have apparent orthologs in *C. briggsae* and 31 (49%) of which have counterparts in much more distantly related lineages, such as flies, fish and mammals (Tables 2, S2, and S5). The 31 families with counterparts in flies or vertebrates encompassed most (64 of 112) of the *C. elegans* genes. The newly identified and revised miRNA sequences provided the opportunity to improve and expand the current set of predicted miRNA targets in *C. elegans* (Chan et al. 2005; Lall et al. 2006). Accordingly, the TargetScanS algorithm was used to predict conserved regulatory targets, which can be viewed at TargetScan.org.

Endogenous siRNA Biogenesis and Targeting

Our library-construction protocol appears to exclude the vast majority of the *C. elegans* siRNA molecules, which we suspect have 5′ triphosphates. Nonetheless, high-throughput sequencing generated more candidate siRNAs than observed previously, enabling insights into endogenous siRNA taxonomy, biogenesis, and function.

Many of the previously annotated tncRNAs fell into clusters of reads that resembled the siRNA clusters, and many of these tncRNA-containing clusters overlapped annotated mRNA exons (Table S4; compare to Table S3). Furthermore, the known factors required for tncRNA biogenesis and endogenous siRNA biogenesis are similar

(Lee et al. 2006). Considering these similarities and reasoning that any minor differences reported between the biogenesis requirements of particular tncRNAs and siRNAs are likely to be no greater than those between different siRNAs, we propose that the tncRNAs do not represent a class of *C. elegans* RNAs separate from the endogenous siRNAs. Nonetheless, the endogenous siRNAs of *C. elegans* are not a monolithic class and appear to be combination of classes whose taxonomy includes an abundant shorter class underrepresented in our library, presumably because of 5′ triphosphates, and a newly identified ~26-nt class with 5′ monophosphates and modified 3′ termini.

Many of the small RNAs classified as *C. elegans* endogenous siRNAs have strong links with RNAi-mediated gene silencing. For example, they are enriched in matches to transposons, and their accumulation decreases in mutant worms that are defective in RNAi (Lee et al. 2006). Thus, their classification as siRNAs is appropriate. However, they differ from canonical siRNAs in that they lacked some of the classical features of Dicer products: most appear to lack a 5′ monophosphate; their length distribution (Figure 4A) largely differed from the 23-nt RNAs previously described for *C. elegans* exogenous siRNAs (Ketting et al. 2001), and their overlapping ends were uncharacteristic of Dicer processing (Figure 4D, Table S3), which should yield non-overlapping ends when the RNAs are in phase with each other. We conclude that endogenous siRNAs biogenesis in nematodes involves little, if any, sequential Dicer processing of long dsRNA, which is perhaps unexpected given the facility by which *C. elegans* utilizes long dsRNA for exogenous RNAi (Fire et al. 1998), the Dicer-dependence of some tncRNAs and siRNAs (Lee et al. 2006), and the models of transitive RNAi in worms, in which siRNAs serve as primers for the production of additional siRNAs (Sijen et al. 2001; Tijsterman et al.

2002). Instead, we propose that most endogenous *C. elegans* siRNAs are generated by unprimed RdRP activities insufficiently processive to generate long dsRNAs suitable for successive cleavage events, and are thus reminiscent of short antisense RNAs generated by *Neurospora* QDE-1 (Makeyev and Bamford 2002). Because longer dsRNA is mobile in worms (Feinberg and Hunter 2003), shorter polymerization might ensure that the endogenous silencing is cell autonomous. If only a single siRNA was made from each RdRP product, then the 5′ terminus of each siRNA could be determined by the nucleotide used to initiate synthesis of the antisense strand, which we suspect is predominantly a GTP.

Recognizing that there could be multiple endogenous RNAi pathways in worms, we draw a speculative model focusing on the 26mer siRNAs and the propensity of their 23rd residues to pair with sense RNA 5′ termini (Figure 5). A 26mer siRNA is synthesized without priming by an RdRP, initiating with a G across from a C in the template transcript (step 1). The siRNA guides an endonuclease to cleave the template between residues that pair to nucleotides 23 and 24 of the siRNA (step 2). The cleaved template triggers a second round of unprimed siRNA synthesis, which starts across from the C residue closest to the cleavage site (step 3). Steps 2 and 3 repeat, generating the phased pattern of siRNAs that overlap in cases where C residues lie close to the cleavage site. Degradation of the ~26-nt sense fragments proceeds in the 3′ to 5′ direction, but is slowed by pairing to the siRNA, thereby leading to accumulation of sense reads that fully pair to the siRNAs (step 4). Once liberated from the sense fragment, the siRNA might pair to a second transcript (step 5) and target its cleavage, thereby initiating another series of siRNA-synthesis and target-cleavage events. Although Dicer is not necessarily at the

heart of this model, siRNA accumulation would still be Dicer-dependent if Dicer was required for either the initial mRNA cleavage or subsequent cleavages that trigger unprimed synthesis. A requirement of PIR-1 to remove the siRNA γ- and β-phosphates might explain both the importance of this presumed RNA phosphatase for siRNA production (Duchaine et al. 2006) and the monophosphate at the 5′ terminus of 26mer siRNAs.

Endogenous siRNAs have previously been implicated in transposon silencing (Lee et al. 2006; Sijen and Plasterk 2003). We found that endogenous siRNAs, particularly 26mers, also had a propensity to match spermatogenesis-associated messages. Worms deficient in EGO-1, a nuclear RdRP, have delayed spermatogenesis-to-oogenesis transition (Smardon et al. 2000), tempting speculation that EGO-1 produces the endogenous siRNAs that silence sperm-enriched genes, thereby hastening the transition to oogenesis.

21U-RNAs: Diverse, Autonomously Expressed, Small RNAs

21U-RNAs are 21-nt RNAs that begin with a U and derive from thousands of loci in several broad regions of chromosome IV. The conservation in *C. briggsae* of the upstream motifs, presumably involved in 21U-RNA biogenesis, suggests that production of 21U-RNAs has an important biological function even if the RNA product itself might not. Such function might include opening of chromatin structure or changes to nucleosome phasing induced upon transcription of the 21U-RNA loci.

The more uniform nucleotide composition of 21U-RNA sequences versus their surrounding sequence, considered together with the diversity and lack of sequence

conservation within the set of 21U-RNAs, suggested that evolutionary pressure is maximizing their sequence complexities rather than maintaining their sequence identities. If 21U-RNAs act by base pairing with a complementary nucleic acid strand, then this increased complexity would enable a higher degree of pairing specificity for the 21U-RNA sequences (important for both targeting and preventing off-targeting) than would be possible using the less uniform nucleotide composition of neighboring sequence. Their 21-nt length and 5′ phosphate are both features of small RNAs that associate with Argonaute protein family members to target gene repression (Tomari and Zamore 2005), suggesting that the 21U-RNAs might do the same, and perhaps target the chromatin from which they derive. The regions defined by the 21U-RNA loci were vast, and contained many protein-coding genes, with a wide variety of functions and expression patterns. Which of those functions that the 21U-RNAs might be influencing, if indeed they act locally, is unclear.

Equally mysterious as 21U-RNA function are aspects of their biogenesis. The large and small motifs might together serve as a promoter, driving expression of each 21U-RNA, with the AT-rich region at the 3′ end of the larger motif acting as a TATAA box. Or perhaps the motifs serve as a signal for targeting the cleavage of a larger transcript. The larger motif could serve as a promoter for a transcript that is processed at the site of the smaller motif. If the 21U-RNA primary transcript were to begin at the 5′ end of the mature 21U-RNA, the transcribing polymerase would either have to prefer incorporation of UMP to that of UTP at the 5′ end, or the 21U-RNA would have to be post-transcriptionally processed to remove the γ- and β-phosphates of the 5′-terminal UTP.

In our favored scenarios for 21U-RNA production, each locus represents an independent transcription unit, that is, each could be classified as an individual non-coding RNA gene. From this perspective, the discovery of the 21U-RNA loci dramatically increased the number of known nematode genes. A minimum of 5772 loci produced the observed reads (when also considering the 21U-RNA loci unique to reads provided by A Fire), and we estimate there to be 12,000-16,000 total loci (Supplemental text). Nonetheless, the common upstream motif and broad clustering of 21U-RNA loci in the genome both suggest that these genes do not function alone, but instead act concurrently to produce some aggregate effect. This scenario presents some fascinating evolutionary questions: How do selective pressures act to maintain the motifs present at each of the thousands of individual 21U-RNA loci and, when they fail to do so, how do new loci emerge within the same broad regions of chromosome IV to replace those that are lost?

Another intriguing biogenesis question entails how the 3′ ends of the 21U-RNAs are defined. The absence of a discernable motif at or near the 3′ end suggests that it is defined in reference to the position of the 5′ end. This hypothesis requires a biochemical mechanism for precisely counting 21 ribonucleotides of any sequence. The known activity with closest precision in counting this number of ribonucloetides is Dicer-catalyzed cleavage. However, *C. elegans* Dicer is thought to produce 23mer RNAs (Ketting et al. 2001), and Dicer products have a size diversity exceeding that of 21U-RNAs, even when processing dsRNA without mismatches (Zamore et al. 2000). Furthermore, we saw no evidence of 21-nt RNAs arising from the opposing RNA strand—no analog to the siRNA passenger strand. Even without conventional Dicer

processing, counting 21 nt to determine the 3′ terminus in reference to the 5′ terminus is easiest to imagine if it occurs in the context of a double helix, presumably while the transcript is still paired to its DNA (or RNA) template.

21U-RNAs clearly represent a unique class of small RNAs. They are far more diverse than miRNAs, and unlike siRNAs and piRNAs, which are expressed in tight clusters, the 21U-RNAs appear to be autonomously expressed. We suggest that other types of diverse, autonomously expressed, small RNAs (dasRNAs) might be found in other species. The deep sequencing of small RNAs in species beyond *C. elegans* will provide important information for addressing this possibility.

Experimental Procedures

Library preparation. Five runs of high-throughput pyrophosphate sequencing (Margulies et al. 2005) were performed, the first at Broad Institute and the next four at 454 Life Sciences (Branford, CT, USA). Primary RT-PCR DNA generated previously (Lau et al. 2001) was prepared for sequencing using three different methods. For runs 1 and 2, it was amplified as in (Lau et al. 2001) but substituting pATCGTAGGCACCTGAGA for the 5′ PCR primer and stopping the PCR during the linear phase of amplification. The amplified DNA was purified by phenol/chloroform extraction then native PAGE. Sequencing runs 1 and 2 began with the standard blunt-end ligation step and yielded 283,557 and 298,625 reads, respectively. For run 3, the PCR reaction was smaller (1 x 100 μl) and used primers GCCTCCCTCGCGCCATCAGTATCGTAGGCACCTGAGA and GCCTTGCCAGCCCGCTCAGTATTGATGGTGCCTACAG, which added sequences

enabling the blunt-end ligation step of the protocol to be bypassed. This reaction was purified by phenol/chloroform extraction and denaturing (urea) PAGE and yielded 235,632 reads. For runs 4 and 5, PCR DNA was amplified as in run 3 but the second primer was replaced with A_{30}/iSp18/GCCTTGCCAGCCCGCTCAGTATTGATGGTGCCTACAG (IDT, Inc., Coralville, IA). The 18-atom spacer prevented Taq polymerase from using the poly-A portion of the primer as a template (Williams and Bartel 1995). PCR product (40 μl) was denatured (85° C, 10 minutes, formamide loading dye), and the differently sized strands were purified on a 90% formamide, 8% acrylamide gel, yielding single-stranded DNA suitable for the emulsion PCR reaction of the sequencing procedure. Sequencing of the longer strand yielded 196,083 reads (run 4), and the shorter yielded 110,299 reads (run 5). Although runs 4 and 5 yielded fewer reads than the other runs, the diversity of reads matching the genome was comparable.

Read processing. The 1,124,196 individual sequence reads were processed in four steps: 1) 9-nt segments of each linker that immediately flanked the small RNA-derived sequence were found in 850,870 reads (181,668 unique small RNA sequences); the remaining reads were discarded. 2) Each unique sequence was compared to annotated *C. elegans* miRNA hairpins (miRBase 7.0)(Griffiths-Jones 2004), and those ≥10 nt and with perfect matches over their entire length were set aside (1002 sequences, 317,694 reads; Table S1). 3) Sequences with perfect matches to the *E. coli* genome (Hayashi et al. 2001) as found by BLAST (Altschul et al. 1990) were discarded (20,845 sequences, 176,719 reads). 4) Sequences were compared to the WormBase WS120 assembly of the *C.*

elegans genome using BLAST, and those with perfect hits (no gaps or mismatches across their entire length) were retained (23,109 sequences, 77,232 reads). Up to 50 perfect hits to the *C. elegans* genome were recorded per query sequence. In downstream analyses, sequence and read counts were normalized to the number of genomic loci (Supplemental Text). Sequences spanning splice junctions were identified from those without matches in the *E. coli* or *C. elegans* genomes using BLAST to search annotated *C. elegans* cDNAs (Kent and Zahler 2000).

21U-RNA upstream motifs. 21U-RNA loci were defined as those whose sequences perfectly matched 21-nt reads beginning with a 5′ T and fell into regions of chromosome IV whose matching normalized reads were dominated by these two properties. Motifs were defined using alignments of genomic sequence surrounding the 21U-RNA loci, with each locus equally weighted. The motif scoring matrix was constructed using $\log_2$-odds ratios of nucleotide frequencies at positions in the alignments (foreground) to genomic nucleotide frequencies (background). Predicted 21U-RNA loci were those scoring ≥15.5 (Supplemental Text).

An independent set of 245,420 *C. elegans* small RNA pyrosequencing reads was provided by Andrew Fire (personal communication). Processing as described above yielded 1475 21U-RNA sequences representing 7985 reads. 344 sequences were not present in our dataset. Of those, 157 (46%) matched predicted 21U-RNA loci of chromosome IV, which was a smaller portion than for sequences unique to any of our five datasets (64%, 65%, 66%, 69%, and 72%), indicating that some information represented in our motif model originated from peculiarities of our training set.

Nonetheless, of the 4.7 million 21mers beginning with a T from within those three regions, motif scanning predicted that only 0.1% were loci of unsequenced 21U-RNAs. Thus, correctly predicting almost half of the unique sequences from an independent set of reads (versus 0.1% if those sequences were picked randomly) indicated that most of the information in our model reflected the biological requirements of the motif.

siRNA methods. Exon coordinates were from WormBase gene annotations (release WS120, 3/1/2004). Counts matching the sense and antisense strands of exons, excluding loci classified as 21U-RNAs, were normalized to the number of genomic loci. Splicing variants were collapsed, leaving 1720 siRNA-complemented genes and 1346 sense RNA-matched genes. To account for expression, SAGE data from the Genome BC *C. elegans* Gene Expression Consortium (http://elegans.bcgsc.bc.ca) was used to select control cohorts (Supplemental text).

Molecular analyses. For alkaline hydrolysis, mixed-stage *C elegans* total RNA (40 μg) was incubated in 0.1 M KOH (90° C, 10 minutes) then neutralized with TrisHCl. Periodate oxidation and β elimination were as described (Kemper 1976). For enzymatic analyses, 800 μg of total RNA were gel purified, and one fortieth was used to cap with the remainder divided equally for five treatments. Phosphatase (50U CIP, NEB) and re-phosphorylation (20U T4 polynucleotide kinase, NEB) were performed according to manufacturer. RNA ligations were as in the second ligation step of the library construction (Lau et al. 2001). Capping was with vaccinia guanylyl transferase (Ambion) and α-^{32}P GTP per manufacturer's instructions. The 26mer marker was an in vitro

transcribed version of siR26-1. Northerns were as described (Lau et al. 2001), except 21U-1 and siR26-1 were hybridized to LNA probes (Exiqon) as described (Vagin et al. 2006).

Figure Legends

Figure 1. Distribution of reads across the *lin-4* and *mir-786* hairpins

(A) The sequence of the *lin-4* hairpin is depicted above its bracket-notation secondary structure as determined by RNAfold (Hofacker et al. 1994) and above the prior annotation of the mature *lin-4* miRNA (Lee et al. 1993), as refined by Lau et al. (2001). Below, each of the small RNA sequences that matched the *lin-4* hairpin is listed, with the number of reads representing each sequence shown. The dominant miRNA sequence is red; the dominant miRNA* species is blue; and the loop-containing sequence is green. Reads from the other previously annotated miRNA hairpins are provided (Table S1).

(B) The *lin-4* predicted hairpin, with the dominant species highlighted as in (A). Lines indicate inferred sites of Drosha and Dicer cleavage.

(C) The sequence of the *mir-786* hairpin depicted as in (A). Reads from the other newly identified miRNA hairpins are provided (Table S1).

(D) An alignment of the *mir-786* hairpin sequence with that of its inferred ortholog in *C. briggsae*. The dominant miRNA and miRNA* species are highlighted as in (A), and *C. briggsae* residues differing from those of *C. elegans* are in grey.

(E) The *C. elegans* and *C. briggsae mir-786* hairpins, depicted as in (B) with residues colored as in (D).

(F) Cumulative plot of *C. elegans* miRNAs with the indicated pyrosequencing frequency; blue, 53 miRNAs sequenced in Lau et al. (2001); cyan, 27 miRNAs first sequenced in Lim et al., (2003); orange, 31 miRNAs first sequenced in the current study (including 13 from previously annotated miRNA hairpins).

Figure 2. Observed and predicted 21U-RNAs from thousands of loci across two broad regions of *C. elegans* chromosome IV

(A) Observed small RNA reads from chromosome IV. All normalized reads were counted in 100-kb bins (orange). The subset of normalized reads that were precisely 21 nt long and began with U were also counted (green). Grey shading is explained in (B).

(B) Observed and predicted 21U RNA loci on chromosome IV. Loci that matched one or more 21U-RNA read were counted in 100-kb bins (blue). The same was done for 21U-RNA loci predicted by scanning for the associated motifs (pink). Sections of the chromosome shaded in grey are syntenic to *C. briggsae* contigs with a high density (≥ 75 per 100 kb) of the 21U-RNA-associated motifs.

(C) Observed and predicted 21U-RNA loci on other chromosomes. Coloring as in (B). The asterisk above chromosome I indicates the position of the ribosomal repeats, which are collapsed in the genome assembly; ribosomal RNA fragments mapped to this region, some of which were 21 nt with a 5′ U.

(D) Representative 100-kb fragment of a region that gives rise to 21U-RNAs. Shown are the 146 loci corresponding to observed 21U RNA reads (blue) and the 257 predicted loci (pink) from coordinates 14.4–14.5M (WormBase, build WS120). Shown also are WormBase-annotated genes.

Figure 3. The 21U-RNA sequence motifs and small RNA chemical reactivity

(A) The large and small motifs found upstream of 21U-RNA loci, depicted as a sequence motif (Crooks et al. 2004). The T at position 1 corresponds to the 5′ U of the 21U RNA.

(B) The distribution of distances between the large and small motifs.

(C) Chemical reactivity of small RNAs. Total RNA (40 μg) was treated as indicated and analyzed by RNA blot, probing first for 21U-1, then stripping and reprobing for siR26-1, then miR-52.

Figure 4. Many reads antisense to known or predicted mRNAs

(A) The length and initial nucleotide distribution of the antisense reads.

(B) The length and initial nucleotide distribution of the sense reads.

(C) A population of ~22mer RNAs with terminal 5′ di- or triphosphates. Those RNAs with 5′ di- or triphosphates were selectively radiolabeled in a capping reaction that used α-^{32}P GTP and compared to the indicated 5′ phosphorylated (5′ P) or capped size standards by 15% PAGE.

(D) Portions of two WormBase-annotated protein-coding genes aligned with small RNA reads that matched the sense (blue) and antisense (orange) strands. One hundred siRNA clusters, each comprising from 4 to 61 antisense reads, are shown in Table S3.

(E) Examples of siRNAs that did not match the genome but did match the splice junctions (vertical lines) of mature mRNAs.

Figure 5. Speculative model for endogenous RNAi in worms, illustrated using the F55C9.3 transcript (blue) and sequenced siRNAs (orange) from Figure 4D. Small arrowheads indicate the transcript cleavage sites. See discussion for explanation.

Acknowledgments

We thank W. Johnston for assistance in preparing DNA for high-throughput sequencing, W. Brockman and P. Alvarez for base calling of sequencing run #1, and C. Perbost and others at 454 Life Sciences for sequencing runs #2–5. We also thank S. Bagby, A. Grishok, A. Grimson, A. Mallory and H. Vaucheret for useful comments on the manuscript. The SAGE data were produced at the Michael Smith Genome Sciences Centre with funding from Genome Canada. Supported by the Prix Louis D from the Institut de France and a grant from the NIH (D.P.B). D.P.B is an HHMI Investigator.

Accessions

All sequences with linker matches were deposited in the Gene Expression Omnibus (GSE5990). The 21U-RNA sequences were deposited in GenBank (EF044580-EF050033). MicroRNA sequences were submitted to miRBase (Griffiths-Jones 2004). 21U-RNA, siRNA, and sense RNA sequences are also provided in FASTA format as Supplemental Materials (21U-RNA.fa, siRNA.fa, and senseRNA.fa).

References

Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. 1990. Basic local alignment search tool. *J Mol Biol* **215:** 403-410.

Ambros, V., R.C. Lee, A. Lavanway, P.T. Williams, and D. Jewell. 2003. MicroRNAs and other tiny endogenous RNAs in *C. elegans*. *Curr Biol* **13:** 807-818.

Aravin, A., D. Gaidatzis, S. Pfeffer, M. Lagos-Quintana, P. Landgraf, N. Iovino, P. Morris, M.J. Brownstein, S. Kuramochi-Miyagawa, T. Nakano, M. Chien, J.J. Russo, J. Ju, R. Sheridan, C. Sander, M. Zavolan, and T. Tuschl. 2006. A novel class of small RNAs bind to MILI protein in mouse testes. *Nature* **442:** 203-207.

Aravin, A.A., M. Lagos-Quintana, A. Yalcin, M. Zavolan, D. Marks, B. Snyder, T. Gaasterland, J. Meyer, and T. Tuschl. 2003. The small RNA profile during *Drosophila melanogaster* development. *Dev Cell* **5:** 337-350.

Bartel, D.P. 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116:** 281-297.

Brennecke, J., A. Stark, R.B. Russell, and S.M. Cohen. 2005. Principles of microRNA-target recognition. *PLoS Biol* **3:** e85.

Chan, C.S., O. Elemento, and S. Tavazoie. 2005. Revealing posttranscriptional regulatory elements through network-level conservation. *PLoS Comput Biol* **1:** e69.

Chang, S., R.J. Johnston, Jr., C. Frokjaer-Jensen, S. Lockery, and O. Hobert. 2004. MicroRNAs act sequentially and asymmetrically to control chemosensory laterality in the nematode. *Nature* **430:** 785-789.

Coghlan, A. and K.H. Wolfe. 2002. Fourfold faster rate of genome rearrangement in nematodes than in *Drosophila*. *Genome Res* **12:** 857-867.

Crooks, G.E., G. Hon, J.M. Chandonia, and S.E. Brenner. 2004. WebLogo: a sequence logo generator. *Genome Res* **14:** 1188-1190.

Doench, J.G. and P.A. Sharp. 2004. Specificity of microRNA target selection in translational repression. *Genes Dev* **18:** 504-511.

Duchaine, T.F., J.A. Wohlschlegel, S. Kennedy, Y. Bei, D. Conte, Jr., K. Pang, D.R. Brownell, S. Harding, S. Mitani, G. Ruvkun, J.R. Yates, 3rd, and C.C. Mello. 2006. Functional proteomics reveals the biochemical niche of *C. elegans* DCR-1 in multiple small-RNA-mediated pathways. *Cell* **124:** 343-354.

Feinberg, E.H. and C.P. Hunter. 2003. Transport of dsRNA into cells by the transmembrane protein SID-1. *Science* **301:** 1545-1547.

Fire, A., S. Xu, M.K. Montgomery, S.A. Kostas, S.E. Driver, and C.C. Mello. 1998. Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* **391:** 806-811.

Girard, A., R. Sachidanandam, G.J. Hannon, and M.A. Carmell. 2006. A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature* **442:** 199-202.

Grad, Y., J. Aach, G.D. Hayes, B.J. Reinhart, G.M. Church, G. Ruvkun, and J. Kim. 2003. Computational and experimental identification of *C. elegans* microRNAs. *Mol Cell* **11:** 1253-1263.

Griffiths-Jones, S. 2004. The microRNA Registry. *Nucleic Acids Res* **32:** D109-111.

Grishok, A., A.E. Pasquinelli, D. Conte, N. Li, S. Parrish, I. Ha, D.L. Baillie, A. Fire, G. Ruvkun, and C.C. Mello. 2001. Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control *C. elegans* developmental timing. *Cell* **106:** 23-34.

Han, J., Y. Lee, K.H. Yeom, J.W. Nam, I. Heo, J.K. Rhee, S.Y. Sohn, Y. Cho, B.T. Zhang, and V.N. Kim. 2006. Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. *Cell* **125:** 887-901.

Hayashi, T., K. Makino, M. Ohnishi, K. Kurokawa, K. Ishii, K. Yokoyama, C.G. Han, E. Ohtsubo, K. Nakayama, T. Murata, M. Tanaka, T. Tobe, T. Iida, H. Takami, T. Honda, C. Sasakawa, N. Ogasawara, T. Yasunaga, S. Kuhara, T. Shiba, M. Hattori, and H. Shinagawa. 2001. Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Res* **8:** 11-22.

Hofacker, I.L., W. Fontana, P.F. Stadler, L.S. Bonhoeffer, M. Tacker, and P. Schuster. 1994. Fast folding and comparison of RNA secondary structures. *Monatshefte Fur Chemie* **125:** 167-188.

Hutvagner, G., J. McLachlan, A.E. Pasquinelli, E. Balint, T. Tuschl, and P.D. Zamore. 2001. A cellular function for the RNA-interference enzyme Dicer in the maturation of the *let-7* small temporal RNA. *Science* **293:** 834-838.

Hutvagner, G. and P.D. Zamore. 2002. A microRNA in a multiple-turnover RNAi enzyme complex. *Science* **297:** 2056-2060.

Johnston, R.J. and O. Hobert. 2003. A microRNA controlling left/right neuronal asymmetry in *Caenorhabditis elegans*. *Nature* **426:** 845-849.

Kemper, B. 1976. Inactivation of parathyroid hormone mRNA by treatment with periodate and aniline. *Nature* **262:** 321-323.

Kent, W.J. and A.M. Zahler. 2000. The intronerator: exploring introns and alternative splicing in *Caenorhabditis elegans*. *Nucleic Acids Res* **28:** 91-93.

Ketting, R.F., S.E. Fischer, E. Bernstein, T. Sijen, G.J. Hannon, and R.H. Plasterk. 2001. Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in *C. elegans*. *Genes Dev* **15:** 2654-2659.

Lagos-Quintana, M., R. Rauhut, W. Lendeckel, and T. Tuschl. 2001. Identification of novel genes coding for small expressed RNAs. *Science* **294:** 853-858.

Lall, S., D. Grun, A. Krek, K. Chen, Y.L. Wang, C.N. Dewey, P. Sood, T. Colombo, N. Bray, P. Macmenamin, H.L. Kao, K.C. Gunsalus, L. Pachter, F. Piano, and N. Rajewsky. 2006. A genome-wide map of conserved microRNA targets in *C. elegans*. *Curr Biol* **16:** 460-471.

Lau, N.C., L.P. Lim, E.G. Weinstein, and D.P. Bartel. 2001. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science* **294:** 858-862.

Lau, N.C., A.G. Seto, J. Kim, S. Kuramochi-Miyagawa, T. Nakano, D.P. Bartel, and R.E. Kingston. 2006. Characterization of the piRNA complex from rat testes. *Science* **313:** 363-367.

Lee, R.C. and V. Ambros. 2001. An extensive class of small RNAs in *Caenorhabditis elegans*. *Science* **294:** 862-864.

Lee, R.C., R.L. Feinbaum, and V. Ambros. 1993. The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell* **75:** 843-854.

Lee, R.C., C.M. Hammell, and V. Ambros. 2006. Interacting endogenous and exogenous RNAi pathways in *Caenorhabditis elegans*. *Rna* **12:** 589-597.

Lee, Y., C. Ahn, J. Han, H. Choi, J. Kim, J. Yim, J. Lee, P. Provost, O. Radmark, S. Kim, and V.N. Kim. 2003. The nuclear RNase III Drosha initiates microRNA processing. *Nature* **425:** 415-419.

Lewis, B.P., C.B. Burge, and D.P. Bartel. 2005. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120:** 15-20.

Lewis, B.P., I.H. Shih, M.W. Jones-Rhoades, D.P. Bartel, and C.B. Burge. 2003. Prediction of mammalian microRNA targets. *Cell* **115:** 787-798.

Li, J., Z. Yang, B. Yu, J. Liu, and X. Chen. 2005. Methylation protects miRNAs and siRNAs from a 3'-end uridylation activity in *Arabidopsis*. *Curr Biol* **15:** 1501-1507.

Lim, L.P., N.C. Lau, E.G. Weinstein, A. Abdelhakim, S. Yekta, M.W. Rhoades, C.B. Burge, and D.P. Bartel. 2003. The microRNAs of *Caenorhabditis elegans*. *Genes Dev* **17:** 991-1008.

Lu, C., S.S. Tej, S. Luo, C.D. Haudenschild, B.C. Meyers, and P.J. Green. 2005. Elucidation of the small RNA component of the transcriptome. *Science* **309:** 1567-1569.

Makeyev, E.V. and D.H. Bamford. 2002. Cellular RNA-dependent RNA polymerase involved in posttranscriptional gene silencing has two distinct activity modes. *Mol Cell* **10:** 1417-1427.

Margulies, M., M. Egholm, W.E. Altman, S. Attiya, J.S. Bader, L.A. Bemben, J. Berka, M.S. Braverman, Y.J. Chen, Z. Chen, S.B. Dewell, L. Du, J.M. Fierro, X.V. Gomes, B.C. Godwin, W. He, S. Helgesen, C.H. Ho, G.P. Irzyk, S.C. Jando, M.L. Alenquer, T.P. Jarvie, K.B. Jirage, J.B. Kim, J.R. Knight, J.R. Lanza, J.H. Leamon, S.M. Lefkowitz, M. Lei, J. Li, K.L. Lohman, H. Lu, V.B. Makhijani, K.E. McDade, M.P. McKenna, E.W. Myers, E. Nickerson, J.R. Nobile, R. Plant, B.P. Puc, M.T. Ronan, G.T. Roth, G.J. Sarkis, J.F. Simons, J.W. Simpson, M. Srinivasan, K.R. Tartaro, A. Tomasz, K.A. Vogt, G.A. Volkmer, S.H. Wang, Y. Wang, M.P. Weiner, P. Yu, R.F. Begley, and J.M. Rothberg. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437:** 376-380.

Martin, C.T. and J.E. Coleman. 1989. T7 RNA polymerase does not interact with the 5'-phosphate of the initiating nucleotide. *Biochemistry* **28:** 2760-2762.

Mourelatos, Z., J. Dostie, S. Paushkin, A. Sharma, B. Charroux, L. Abel, J. Rappsilber, M. Mann, and G. Dreyfuss. 2002. miRNPs: a novel class of ribonucleoproteins containing numerous microRNAs. *Genes Dev* **16:** 720-728.

Nakayashiki, H. 2005. RNA silencing in fungi: mechanisms and applications. *FEBS Lett* **579:** 5950-5957.

Ohler, U., S. Yekta, L.P. Lim, D.P. Bartel, and C.B. Burge. 2004. Patterns of flanking sequence conservation and a characteristic upstream motif for microRNA gene identification. *Rna* **10:** 1309-1322.

Ranjith-Kumar, C.T., L. Gutshall, M.J. Kim, R.T. Sarisky, and C.C. Kao. 2002. Requirements for de novo initiation of RNA synthesis by recombinant flaviviral RNA-dependent RNA polymerases. *J Virol* **76:** 12526-12536.

Reinhart, B.J., F.J. Slack, M. Basson, A.E. Pasquinelli, J.C. Bettinger, A.E. Rougvie, H.R. Horvitz, and G. Ruvkun. 2000. The 21-nucleotide *let-7* RNA regulates developmental timing in *Caenorhabditis elegans*. *Nature* **403:** 901-906.

Sijen, T., J. Fleenor, F. Simmer, K.L. Thijssen, S. Parrish, L. Timmons, R.H. Plasterk, and A. Fire. 2001. On the role of RNA amplification in dsRNA-triggered gene silencing. *Cell* **107:** 465-476.

Sijen, T. and R.H. Plasterk. 2003. Transposon silencing in the *Caenorhabditis elegans* germ line by natural RNAi. *Nature* **426:** 310-314.

Smardon, A., J.M. Spoerke, S.C. Stacey, M.E. Klein, N. Mackin, and E.M. Maine. 2000. EGO-1 is related to RNA-directed RNA polymerase and functions in germ-line development and RNA interference in *C. elegans*. *Curr Biol* **10:** 169-178.

Stein, L.D., Z. Bao, D. Blasiar, T. Blumenthal, M.R. Brent, N. Chen, A. Chinwalla, L. Clarke, C. Clee, A. Coghlan, A. Coulson, P. D'Eustachio, D.H. Fitch, L.A. Fulton, R.E. Fulton, S. Griffiths-Jones, T.W. Harris, L.W. Hillier, R. Kamath, P.E. Kuwabara, E.R. Mardis, M.A. Marra, T.L. Miner, P. Minx, J.C. Mullikin, R.W. Plumb, J. Rogers, J.E. Schein, M. Sohrmann, J. Spieth, J.E. Stajich, C. Wei, D. Willey, R.K. Wilson, R. Durbin, and R.H. Waterston. 2003. The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics. *PLoS Biol* **1:** E45.

Tijsterman, M., R.F. Ketting, K.L. Okihara, T. Sijen, and R.H. Plasterk. 2002. RNA helicase MUT-14-dependent gene silencing triggered in *C. elegans* by short antisense RNAs. *Science* **295:** 694-697.

Tomari, Y. and P.D. Zamore. 2005. Perspective: machines for RNAi. *Genes Dev* **19:** 517-529.

Vagin, V.V., A. Sigova, C. Li, H. Seitz, V. Gvozdev, and P.D. Zamore. 2006. A distinct small RNA pathway silences selfish genetic elements in the germline. *Science* **313:** 320-324.

Williams, K.P. and D.P. Bartel. 1995. PCR product with strands of unequal length. *Nucleic Acids Res* **23:** 4220-4221.

Zamore, P.D., T. Tuschl, P.A. Sharp, and D.P. Bartel. 2000. RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. *Cell* **101:** 25-33.

Table 1. Eighteen newly identified miRNAs in *C. elegans*. Reads for miR-789-1 and miR-789-2 cannot be distinguished.

miRNA	Sequence	miRNA reads	miRNA* reads	*C. briggsae* ortholog	Fly or vertabrate family members	Genomic cluster parter
miR-784	UGGCACAAUCUGCGUACGUAGA	11	1	Yes		
miR-785	UAAGUGAAUUGUUUUGUGUAGA	14	2	Yes	Yes	miR-359
miR-786	UAAUGCCCUGAAUGAUGUUCAAU	80	3	Yes	Yes	miR-240
miR-787	UAAGCUCGUUUUAGUAUCUUUCG	32		Yes	Yes	
miR-788	UCCGCUUCUAACUUCCAUUUGCAG	667	10	Yes		
miR-789-1	UCCCUGCCUGGGUCACCAAUUGU	63	1	Yes		
miR-789-2	UCCCUGCCUGGGUCACCAAUUGU	63		Yes		
miR-790	CUUGGCACUCGCGAACACCGCG	16	5	Yes	Yes	miR-228
miR-791	UUUGGCACUCCGCAGAUAAGGCA	1	1	Yes	Yes	miR-230
miR-792	UUGAAAUCUCUUCAACUUUCAGA	4		Yes	Yes	
miR-793	UGAGGUAUCUUAGUUAGACAGA	73			Yes	
miR-794	UGAGGUAAUCAUCGUUGUCACU	5			Yes	miR-795
miR-795	UGAGGUAGAUUGAUCAGCGAGCUU	4			Yes	miR-794
miR-796	UGGAAUGUAGUUGAGGUUAGUAA	9			Yes	
miR-797	UAUCACAGCAAUCACAAUGAGAAGA	12			Yes	miR-247
miR-798	UAAGCCUUACAUAUUGACUGA	33				
miR-799	UGAACCCUGAUAAAGCUAGUGG	36				
miR-800	CAAACUCGGAAAUUGUCUGCCG	12	3			

Table 2 *C. elegans* miRNA families, with the corresponding known miRNAs in other animals. Families sorted alphabetically by seed are listed in Table S2, and newly reported *C. briggsae* orthologs are listed in Table S5.

Seed	*C. elegans*	*C. briggsae*	*D. melanogaster*	*D. rerio*	Mammal
CCCUGA	lin-4/237	lin-4	miR-125	miR-125a/b/c	miR-125a/b,mmu-miR-351
UUUGUA	lsy-6	lsy-6			
GAGGUA	let-7/48/84/241/793/ 794/795	let-7/48/84/241	let-7	let-7a/b/c/d/e/f/g/h/i	let-7a/b/c/d/e/f/g/i/ 98/202
GGAAUG	miR-1/796	miR-1	miR-1	miR-1/206	miR-1/206
AUCACA	miR-2/43/250/797	miR-43	miR-2a/b/c/6/11/ 13a/b/308		
GGCAGU	miR-34	miR-34	miR-34	miR-34	miR-34a/c/449
CACCGG	miR-35/36/37/38/39/ 40/41/42	miR-35/36/38/39/40/41			
GACUAG	miR-44/45/61/247	miR-44/45/61	miR-279/286		
GUCAUG	miR-46/47	miR-46/47	miR-281		
AGCACC	miR-49/83	miR-49/83	miR-285	miR-29a/b	miR-29a/b/c
GAUAUG	miR-50/62/90	miR-50/62/90		miR-190	miR-190
ACCCGU	miR-51/52/53/54/55/56	miR-51/52/55	miR-100	miR-99/100	miR-99b/100,hsa-miR-99a
ACCCUG	miR-57	miR-57		miR-10a/b/c/d	miR-10a,hsa-miR-10b
GAGAUC	miR-58/80/81/82	miR-58/80/81/82	bantam		
CGAAUC	miR-59	miR-59			
AUUAUG	miR-60	miR-60			
AUGACA	miR-63/64/65/66/229	miR-64			
CACAAC	miR-67	miR-67	miR-307	miR-220	
AAUACG	miR-70	miR-70			
GAAAGA	miR-71	miR-71			
GGCAAG	miR-72/73/74	miR-73/74	miR-31a/b		mmu-miR-31
UAAAGC	miR-75/79	miR-75/79	miR-4		
UCGUUG	miR-76	miR-76			
UCAUCA	miR-77	miR-77			
GGAGGC	miR-78				
ACAAAG	miR-85	miR-85			
AAGUGA	miR-86/785	miR-86/785			
UGAGCA	miR-87/233	miR-87/233/356	miR-87		
AAGGCA	miR-124	miR-124	miR-124	miR-124	hsa-miR-506,mmu-miR-124a
AUGGCA	miR-228	miR-228		miR-183	miR-183
UAUUAG	miR-230	miR-230			
AAGCUC	miR-231/787	miR-231/787			
AAAUGC	miR-232/357	miR-232/357	miR-277		
UAUUGC	miR-234	miR-234		miR-137	mmu-miR-137
AUUGCA	miR-235	miR-235	miR-92a/b/310/311/ 312/313	miR-25/92a/b/363	miR-25/32/92,hsa-miR-367
AAUACU	miR-236	miR-236	miR-8	miR-200b/c/429	miR-200b/c/429
UUGUAC	miR-238/239a/b	miR-239a	miR-305		
ACUGGC	miR-240	miR-240		miR-193a/b	miR-193
UGCGUA	miR-242	miR-242			
GGUACG	miR-243				
CUUUGG	miR-244	miR-244	miR-9a/b/c	miR-9	miR-9
UUGGUC	miR-245	miR-245		miR-133a/b/c	
UACAUG	miR-246	miR-246			
UACACG	miR-248.1				
ACACGU	miR-248.2	miR-248			
CACAGG	miR-249	miR-249			
UAAGUA	miR-251/252	miR-251			
UAGUAG	miR-253	miR-253			
GCAAAU	miR-254	miR-254			
AACUGA	miR-255	miR-255			
AAUCUC	miR-259	miR-259	miR-304	miR-216a/b	miR-216
UUGUUU	miR-355	miR-355			
UUGGUA	miR-358	miR-358			
CACUGG	miR-359	miR-359	miR-3/309/318		
AUCAUC	miR-392	miR-392			
GGCACA	miR-784	miR-784			
AAUGCC	miR-786	miR-786		miR-365	miR-365
CCGCUU	miR-788	miR-788			
CCCUGC	miR-789-1/-2	miR-789a/b			
UUGGCA	miR-790/791	miR-791	miR-263b	miR-96/182	miR-96/182
UGAAAU	miR-792	miR-792		miR-203a/b	
AAGCCU	miR-798				
GAACCC	miR-799				
AAACUC	miR-800				

Figure 1

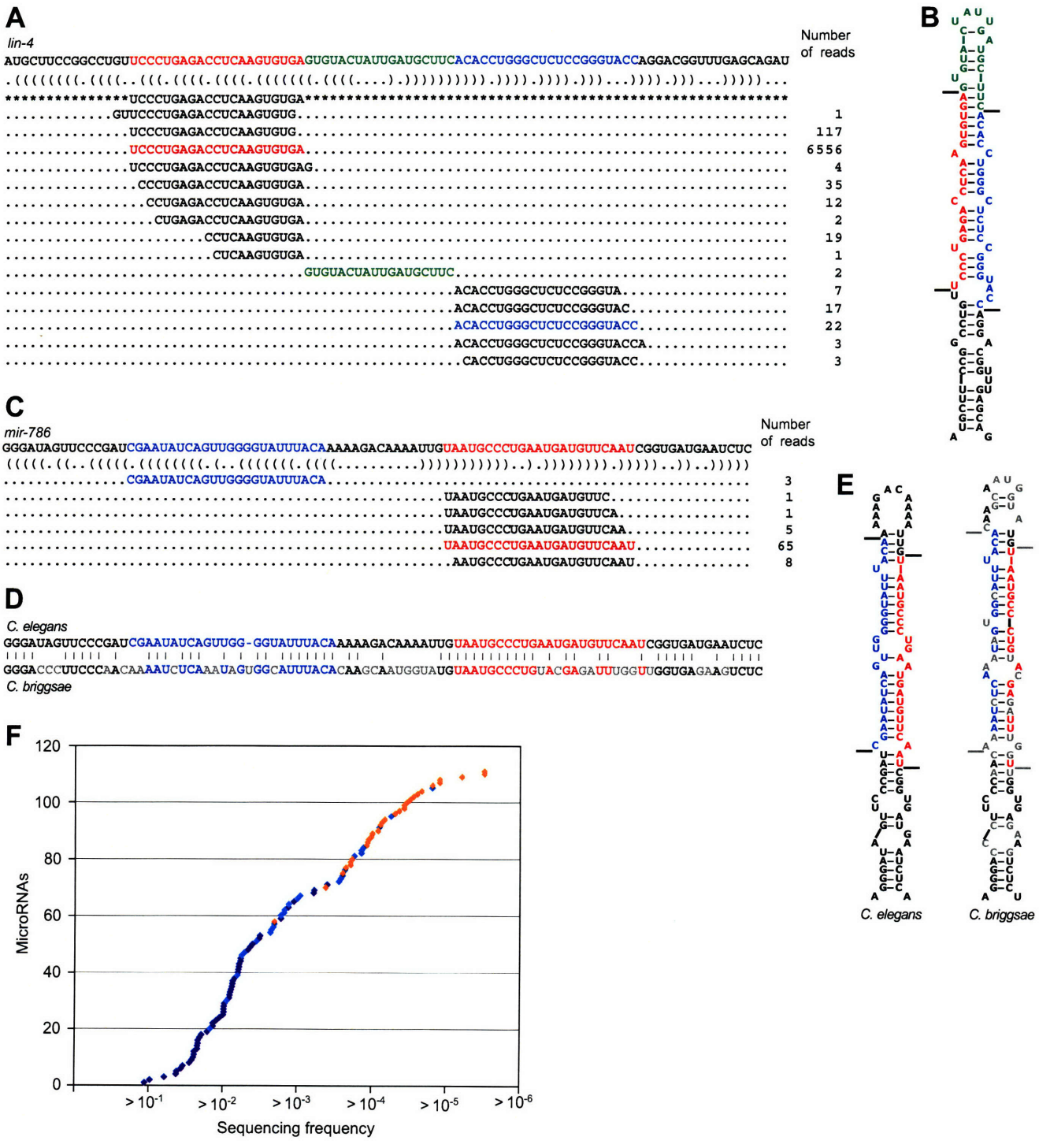

A
lin-4
Number of reads
AUGCUUCCGGCCUGUUCCCUGAGACCUCAAGUGUGAGUGUACUAUUGAUGCUUCACACCUGGGCUCUCCGGGUACCAGGACGGUUUGAGCAGAU
1
117
6556
4
35
12
2
19
1
2
7
17
22
3
3
B
C
mir-786
Number of reads
GGGAUAGUUCCCGAUCGAAUAUCAGUUGGGGUAUUUACAAAAAGACAAAAUUGUAAUGCCCUGAAUGAUGUUCAAUCGGUGAUGAAUCUC
3
1
1
5
65
8
D
C. elegans
GGGAUAGUUCCCGAUCGAAUAUCAGUUGG-GGUAUUUACAAAAAGACAAAAUUGUAAUGCCCUGAAUGAUGUUCAAUCGGUGAUGAAUCUC
GGGACCCUUCCCAACAAAAUCUCAAAUAGUGGCAUUUACACAAGCAAUGGUAUGUAAUGCCCUGUACGAGAUUUGGUUGGUGAGAAGUCUC
C. briggsae
E
C. elegans
C. briggsae
F
MicroRNAs
120
100
80
60
40
20
0
> 10-1
> 10-2
> 10-3
> 10-4
> 10-5
> 10-6
Sequencing frequency

Figure 2

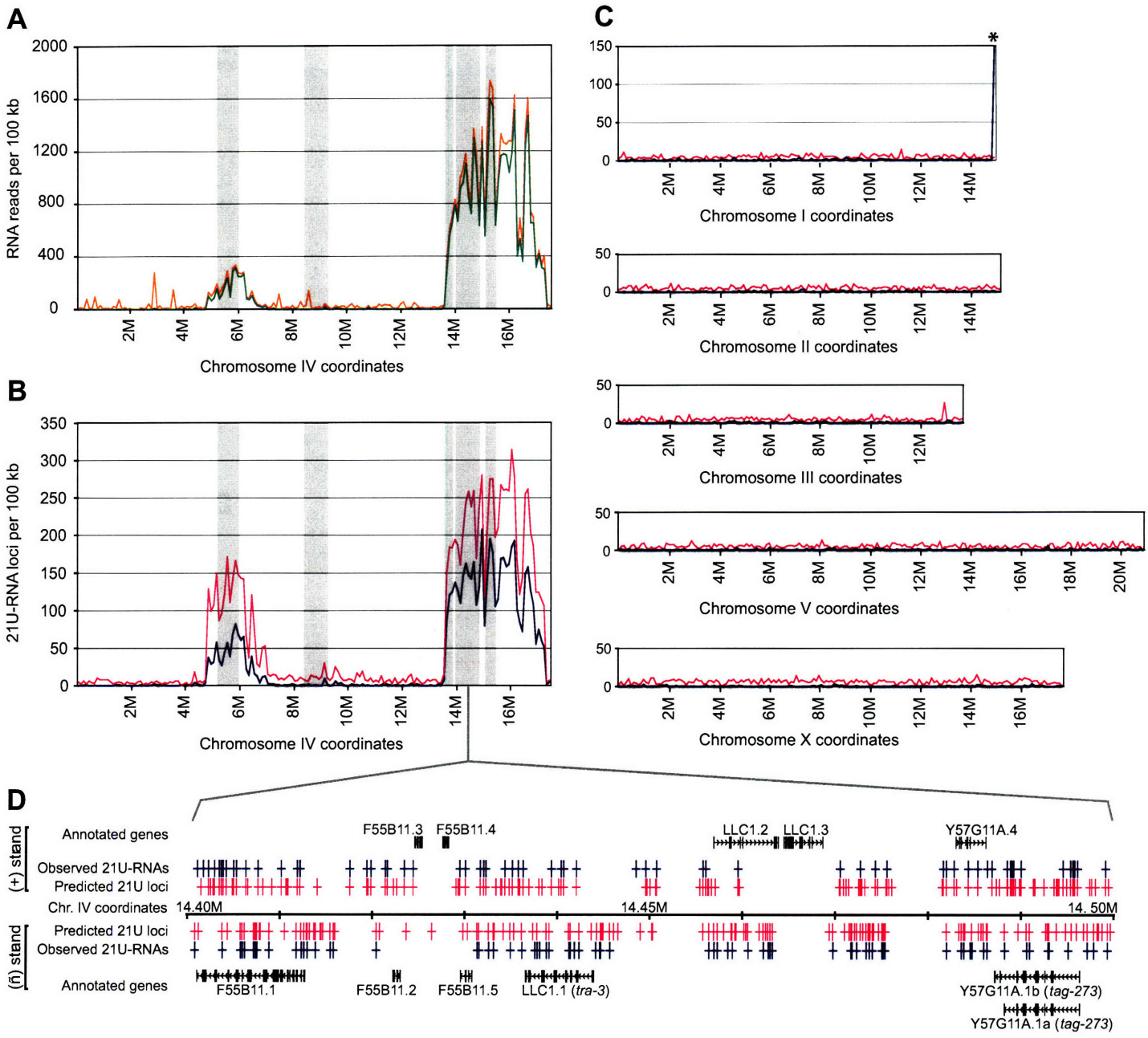
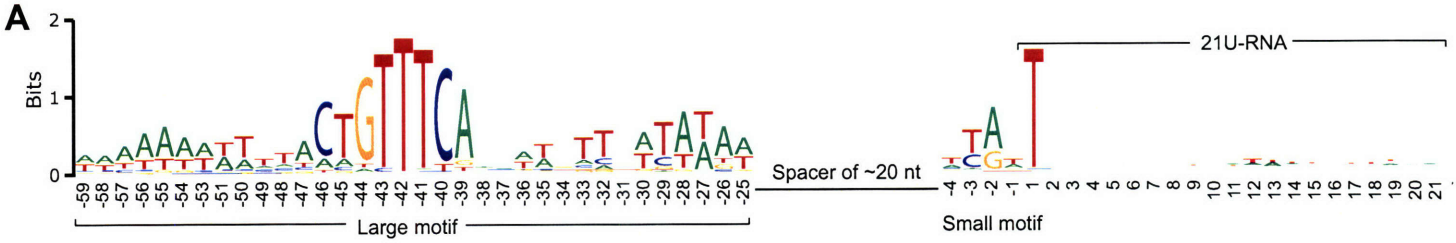
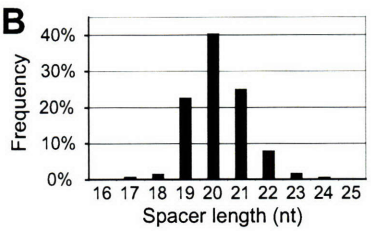

A
RNA reads per 100 kb
2000
1600
1200
800
400
0
2M
4M
6M
8M
10M
12M
14M
16M
Chromosome IV coordinates
B
21U-RNA loci per 100 kb
350
300
250
200
150
100
50
0
Chromosome IV coordinates
C
150
100
50
*
Chromosome I coordinates
Chromosome II coordinates
Chromosome III coordinates
Chromosome V coordinates
18M
20M
Chromosome X coordinates
D
(+) stand
(-) stand
Annotated genes
Observed 21U-RNAs
Predicted 21U loci
Chr. IV coordinates
14.40M
14.45M
14. 50M
F55B11.3
F55B11.4
LLC1.2
LLC1.3
Y57G11A.4
F55B11.1
F55B11.2
F55B11.5
LLC1.1 (tra-3)
Y57G11A.1b (tag-273)
Y57G11A.1a (tag-273)

Figure 3


A
Bits
21U-RNA
Spacer of ~20 nt
Large motif
Small motif


B
Frequency
40%
30%
20%
10%
0%
16 17 18 19 20 21 22 23 24 25
Spacer length (nt)

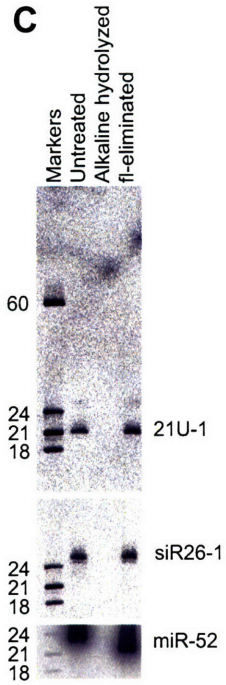
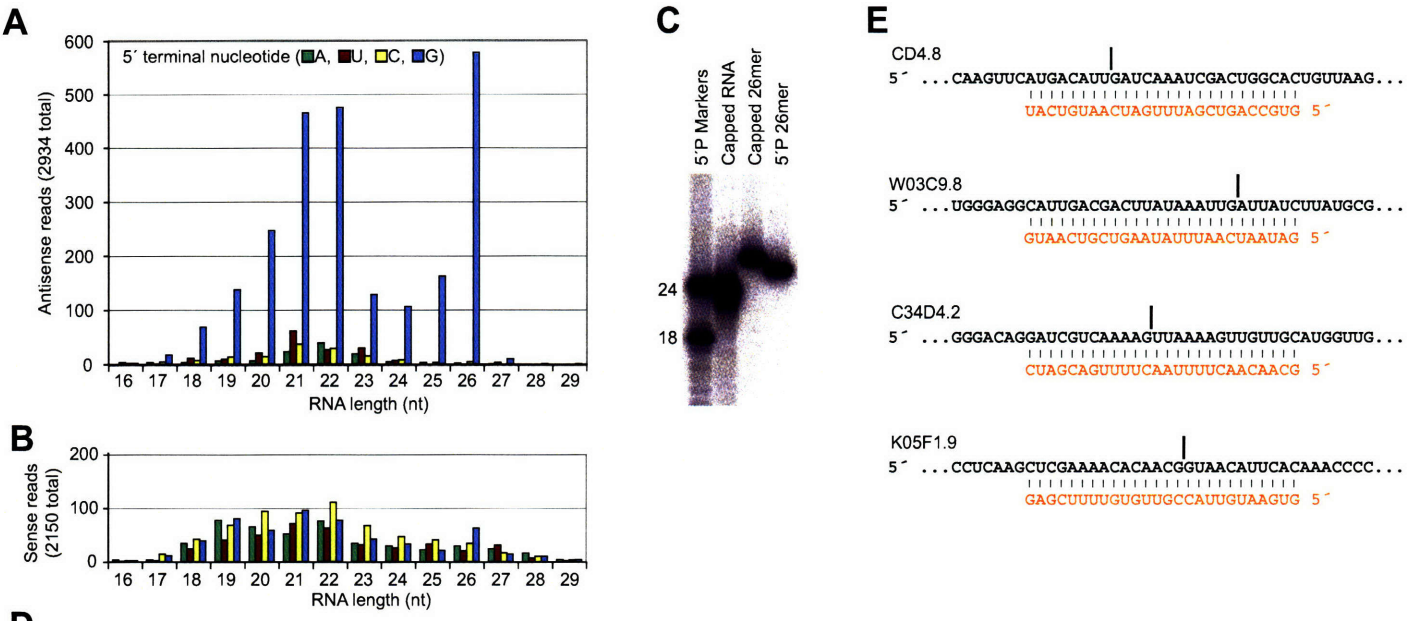

C
Markers
Untreated
Alkaline hydrolyzed
fl-eliminated
60
24
21
18
21U-1
siR26-1
miR-52

Figure 4



D

F55C9.3

```
5′ CAGTAGAAGAACATTTCCGCAGACGCTACAACGTGGAGCTAAGACACCCGAAACTAAGACTATTCCATTTCACCAACGAGACGCAGGCAAAATCACC 3′
5′ ..........................UACAACGUGGAGCUAAGACACCC.......................................................... 3′
5′ ..........................UACAACGUGGAGCUAAGAC............................................................. 3′
3′ GUCAUCUUCUUGUAAAGGCGUCUGCG.......................................................................... 5′
3′ ........................GCGAUGUUGCACCUCGAUUCUGUGGG.......................................................... 5′
3′ ..................................................GCUUUGAUUCUGAUAAGGU.................................... 5′
3′ ..................................................GCUUUGAUUCUGAUAAGGUAAAGUGG............................. 5′
3′ ..........................................................................UGGUUGCUCUGCGUCCGUUUUAGUGG 5′
3′ ..........................................................................GGUUGCUCUGCGUCCGUUUUAGUGG 5′
3′ ..........................................................................GUUGCUCUGCGUCCGUUUUAGUGG 5′
```

T16A1.9

```
5′ AAGACATCGAAGAAGTACTGGAAGCCATACTCAAATCGGGAATGAAGCTCAAAATGGAAAAATCCAATTTTGGACACGCAAAGATCAAGTTCCTAGGATACATTGTCTTGGAAGAAGGCATACAGACAGATCCGAAAAC 3′
5′ ...ACAUCGAAGAAGUACUGG.................................................................................................... 3′
5′ ........................CAUACUCAAAUCGGGAAU.............................................................................. 3′
3′ UUCUGUAGCUUCUUCAUGACCUUCGG.......................................................................................... 5′
3′ .UCUGUAGCUUCUUCAUGACC................................................................................................... 5′
3′ ........................GGUAUGAGUUUAGCCCUUACUUCGAG.................................................................... 5′
3′ ..................................................UUACCUUUUUAGGUUAAAACCUGUGCG.................................... 5′
3′ ..................................................UACCUUUUUAGGUUAAAACCUGUGCG..................................... 5′
3′ ..................................................ACCUUUUUAGGUUAAAACCUGUGCG..................................... 5′
3′ ..................................................CCUUUUUAGGUUAAAACCUGUGCG..................................... 5′
3′ ..........................................................................UCUAGUUCAAGGAUCCUAUGUAACAG.................. 5′
3′ ..........................................................................AACCUUCUUCCGUAUGUCUGUCUAGG...... 5′
3′ ..........................................................................CUUCCGUAUGUCUGUCUAGGCUUUUG 5′
```

Figure 5


5´ ...CAGUAGAAGAACAUUUCCGCAGACGCUACAACGUGGAGCUAAGACACCCGAAACUAAGACUAUUCCAUUUCACCAACGAGACGCAGGCAAAAUCACC... 3´
Synthesize siRNA (RDRP)
UGGUUGCUCUGCGUCCGUUUUAGUGG
Cleave target transcript (endonuclease)
Synthesize siRNA (RDRP)
GCUUUGAUUCUGAUAAGGUAAAGUGG
Cleave target
Synthesize siRNA
GCGAUGUUGCACCUCGAUUCUGUGGG
Cleave target
Synthesize siRNA
Degrade sense fragment
(3´ exonuclease)
GUCAUCUUCUUGUAAAGGCGUCUGCG
Cleave target ...
Target second trancript
Cleave target ...

Supplemental text

Previously sequenced miRNAs

Our previous sequencing of small RNA libraries from *C. elegans* discovered, refined, or confirmed the identities of 80 miRNAs (Lau et al. 2001; Lim et al. 2003). These include the *lin-4* and *let-7* RNAs (whose termini were refined in (Lau et al. 2001)) as well as 25 miRNAs that were concurrently reported by others (Ambros et al. 2003; Grad et al. 2003; Lee and Ambros 2001). Of the 80 miRNAs previously sequenced from our libraries, all were observed in the new set of high-throughput reads.

The reads matching these 80 hairpins had a consistent set of characteristics. There was usually (70 of 80 hairpins) one dominantly abundant sequence, the miRNA, with a smaller number of overlapping sequences representing 5′- and 3′-end heterogeneity. For 9 of the 10 exceptions, heterogeneity among the dominant sequences was only found at the 3′ end; the remaining case, miR-248, displayed heterogeneity among the dominant sequences at the 5′ end. In addition, there was typically (70 of 80 hairpins) a set of reads from the opposing arm of the hairpin, which corresponded to the miRNA* species (Lau et al. 2001) and exhibited the 3′ overhangs typical of miRNA hairpin processing (Lee et al. 2003; Lim et al. 2003). On average, the miRNA* species was present at about 1.0% the frequency of the miRNA, although this was somewhat variable. At one extreme was miR-239b, which had a similar number of reads from each arm (37 and 25 reads for the 5′ and 3′ arms, respectively). At the opposite extreme, none of the 7121 miR-81 reads were from the star arm. For 21 of 80 hairpins, sequences were observed that matched the portion of the transcript in between the miRNA and miRNA*. Reads for these byproducts of Dicer processing of the pre-miRNA, whose lengths in some cases

approximated that of a miRNA, were typically much less abundant than the miRNA* reads. In the case of *mir-44* and *mir-45*, two genes with identical mature miRNA sequences but different hairpin sequences, reads matching the unique miRNA* and loop-containing sequences provided the sole evidence that both genes are expressed.

The high sequencing coverage (with a median of 2059 reads per miRNA, Supplemental Table 1) enabled refinement of the annotated miRNA species for 13 of these 80 miRNAs. The refinement was nearly always at the 3′ terminus of the mature miRNA, as expected based on the heterogeneity of metazoan miRNA termini (Lagos-Quintana et al. 2001; Lau et al. 2001). The exceptions were miR-42 and miR-248, which had been initially annotated based on the sequencing of only one and two clones, respectively (Lau et al. 2001; Lim et al. 2003). The revised miR-42 annotation extended the miRNA 5′ terminus by one nucleotide (Supplemental Table S1). Based on the seed model for miRNA target recognition, this shift of a single nucleotide at the 5′ terminus would dramatically influence the identities of predicted targets. In the case of miR-248, a dominant 5′ terminus was not observed; 19 of the 42 reads matched the previously annotated sequence (re-annotated as miR-248.2), whereas the other 23 reads were extended by one nucleotide at their 5′ terminus, suggesting that the predicted targets of miR-248 should be expanded to include also matches to the extended species (annotated as miR-248.1).

Other previously annotated miRNAs found in our high-throughput reads

RNAs that have not been cloned and sequenced from our libraries have also been annotated as miRNAs. Thirteen such previously annotated miRNA hairpins gave rise to

high-throughput reads (range, 2 to 159 reads/hairpin; Supplemental Table 1). All 13 were initially identified computationally and experimentally supported by northern blotting and/or a PCR assay, in which a miRNA-specific primer is used to preferentially amplify complementary members of the library, with subsequent cloning and sequencing of the amplicon (Lim et al. 2003; Ohler et al. 2004). For five of these (miR-239b, miR-250, miR-252, miR-253, and miR-358), the 5′ terminus of the high-throughput reads did not match the one mapped in the previous PCR-cloning-sequencing assay (Supplemental Table 1). Although we consider these five miRNA genes to now be validated based on high-throughput reads, the question remains as to what was amplified and cloned previously. We note that Chan *et al.* (2005) hypothesize that miR-253 is misannotated and that they correctly predict the dominant RNA species from this gene.

When considering reads that mapped to the miRNA hairpins, the mature miRNA was generally defined based on the position of the most abundant 5′ end from our high-throughput reads. However, an exception was in the case of miR-259, which had a similar number of reads from each arm (29 and 38 reads for the 5′ and 3′ arms, respectively); conservation criteria suggested that the functional miRNA was from the 5′ arm.

The previously annotated miRNAs missing in our high-throughput reads

As described above, 93 of the 114 *C. elegans* miRNA hairpins annotated in miRBase (version 7.0 (Griffiths-Jones 2004)) were represented in our set of high-throughput reads. Below we discuss the absence of reads representing the remaining 21 loci.

lsy-6

The absence of *lys-6* reads was anticipated based on our sequencing depth and the very small volume of cells thought to express *lsy-6* miRNA. Estimating their aggregate volume to be between 30 and 600 femtoliters (one to nine cells, each with a cell body 3–4 microns in diameter and a doubling of the volume to account for cytoplasm in their axons and dentrites; O. Hobert, personal communication) and the volume of the worm at ~10 nanoliters (~1.0-1.5 mm length, ~50 micron diameter), we suggest that *lsy-6* RNA should have been ~100,000 times less abundant than a miRNA expressed in most cells of the worm. By this rough estimate we would not expect to observe a *lsy-6* read until we reach a depth of coverage that yields 100,000 reads for a broadly expressed miRNA. Because the most frequently sequenced miRNA, miR-52, had 37,225 reads, 3-fold deeper coverage would be needed before a *lsy-6* read might be expected. Furthermore, not all RNAs in the same cell are expressed at the same concentration; the abundance disparity would increase further if the *lsy-6* RNA concentration in those few cells that do express it was lower than that of other miRNAs in the same cell.

mir-353, *mir-354*, *mir-356*, and *mir-360*

Although the unsaturated status of our sequencing prohibited any definitive judgments about the veracity of miRNA annotations that were not represented by our reads, our observations were informative for evaluating the confidence in those annotations and the data that was originally used to justify them. For example, a lesson was drawn from those miRNAs whose annotated 5′ end differed from the dominant 5′ end in the reads, which we take to be the true 5′ end. Most of those original 5′ ends had been identified

using miRNA-specific PCR amplification followed by cloning and sequencing (Lim et al. 2003; Ohler et al. 2004). That the original 5′ end sometimes (5 of 12 cases) differed from the authentic miRNA (or miRNA*) called into question what was actually amplified, cloned and sequenced, and showed that this method for experimental validation of computational predictions can sometimes yield false positives. The observation of experimental false positives decreased our confidence in the authenticity of miR-353, miR-354, miR-356, and miR-360, the four candidates that had been experimentally supported by this assay but which were not represented in our set of high-throughput reads. Nonetheless, they could be miRNAs, especially miR-356, which shares a seed with miR-87 and miR-233.

mir-264 through *mir-273*

Other annotated miRNAs from MirBase 7.0 not supported by the high-throughput reads included all 10 RNAs (miR-264 through miR-273) uniquely reported in (Grad et al. 2003). Each of these 10 had been predicted computationally and supported experimentally by PCR amplification using a candidate-specific primer (Grad et al. 2003). The types of false-positive experimental results described above, generated by PCR amplification and sequencing, would be expected with more prevalence when using PCR amplification alone. Furthermore, the observation that our set of high-throughput reads included many of the previously unsequenced miRNAs uniquely proposed by other computational efforts, yet this same set included none of the ten candidates uniquely proposed by (Grad et al. 2003), together with the discussion presented in Ohler et al (2004), supports the idea that most of these ten are not miRNAs.

mir-256, *mir-257*, *mir-258*, *mir-260*, *mir-261*, and *mir-262*

The remaining annotated miRNAs not supported by the high-throughput reads are the six RNAs (miR-256 through miR-258 and miR-260 through miR-262) uniquely reported in (Ambros et al. 2003). One is annotated based on similarity to miR-1 and northern blotting (miR-256), and five are annotated based on cloning and sequencing using a protocol that did not depend on a 5′ monophosphate, with further support from northern blotting. Two (miR-256 and miR-162) were described as having unusual hairpins and possibly not miRNAs when they were first proposed (Ambros et al. 2003). For two (miR-260 and miR-262), we found sequences matching the annotated hairpins, but those sets of sequences did not resemble the sets of sequences typically obtained from miRNA hairpins and were instead reminiscent of siRNA clusters (Supplemental Table 1, Figure 5). Thus, the data from high-throughput sequencing, together with the discussion presented in Ohler et al (2004), supported the idea that these five RNAs cloned in Ambros et al (2003) are not miRNAs and instead represent endogenous siRNAs.

None of the 1 candidate miRNAs recently predicted by Chan et al (2005) were among our newly identified miRNAs.

C. briggsae orthologs of the newly identified miRNAs

Putative orthologs were found in the genome sequence of *C. briggsae* for ten of the 18 newly identified miRNAs. For some of those, the level of conservation between the hairpins and their *C. briggsae* counterparts was low compared to that typically exhibited

by the previously annotated miRNAs, and was often scant beyond the miRNA seed region. Nonetheless, according to current understanding of miRNA targeting (Brennecke et al. 2005; Farh et al. 2005; Lewis et al. 2005), the functions of these counterpart miRNAs should be maintained nearly to the same degree as those of the more extensively conserved miRNAs. It seems that, for some miRNAs, the limited scope of selective pressure across miRNA hairpins will make the identification of orthologous hairpins very noisy in species as divergent as *C. elegans* and *C. briggsae*. Thus the absence of a reported ortholog for a miRNA gene does not imply that the miRNA is species-specific; the degree of divergence between the orthologs may simply be too great for us to have detected the corresponding sequence in *C. briggsae*.

Family Designations

When assigning miRNAs to families, we did not require that all members of the family derive from the same arm of the hairpin. However, most were from the same arm, as would be expected if most members of the same family shared common ancestry. Seven of the 22 *C. elegans* families with multiple members included miRNAs derived from inconsistent sides of their precursor hairpins. For example, mature miR-72 is from the 5′ arm, whereas mature miR-73 and miR-74 are from the 3′ arms. In all but one family from Table 2, examples from the same arm were found in each species with members, which implies that even in cases of very limited sequence identity, family members in *C. elegans* have common ancestry with those in flies and vertebrates. The exception was *C. elegans*/*C. briggsae* miR-67 (3′ arm), *Drosophila melanogaster* miR-307 (3′ arm), and *Danio rerio* miR-220 (5′ arm).

Three borderline candidates

Three additional candidates that were sequenced more than once were, from our perspective, borderline cases and therefore were not annotated here as miRNAs. One candidate with potential to derive from a hairpin with pairing characteristic of metazoan miRNAs (hairpin-1, Supplemental Table 1) was represented by only five reads, all from the same arm of the hairpin, and lacked detectable conservation. In two additional cases (hairpin-2 and hairpin-3, with 5 and 84 reads, respectively), candidates were not annotated as miRNAs because the hairpins exhibited more extensive pairing than has been observed for metazoan miRNA hairpins (Supplemental Table 1). The reads from hairpin-3 clearly derived from both arms of the hairpin and shared a consistent length (22-23 nucleotides), both features reminiscent of miRNAs. However, perhaps because of the uniform pairing within the hairpin, they derived from a motley set of registers up and down the length of the hairpin—a pattern uncharacteristic of previously described miRNAs.

The number of *C. elegans* miRNAs

The lower overall conservation in *C. briggsae* for the newly identified miRNAs was the expected result if previous studies found most of the *C. elegans* miRNAs with extensive conservation in nematodes. Even if most of the conserved miRNAs have been found, it will always be possible to speculate that many non-conserved miRNAs, each expressed only in a few cells or only in special conditions or circumstances, remain to be discovered. Our increased difficulty of finding orthologs for the newly-reported miRNAs

(Table 1), together with the minimal scope of conservation in those orthologs that we did identify, provided the first indication that *C. elegans* miRNAs expressed at very low levels also tend to be less conserved. A similar phenomenon is observed in vertebrates, and makes it impossible to estimate meaningful upper limits on gene number by extrapolating from previous computational studies (Bartel 2004). However, the trajectory of new miRNA discovery versus depth of sequencing seems to indicate that there are not many more miRNAs to be found in *C. elegans* (Figure 1G). Our initial 330 reads captured 55 miRNAs (Lau et al. 2001); increasing sequencing coverage by one order of magnitude, to 4078, captured another 35 genes (Lim et al. 2003); yet increasing coverage by nearly two additional orders of magnitude, to 394,926, captured only another 31 genes (18 newly identified genes, plus 13 that were previously annotated but not cloned). It will be interesting to see if this downward trend continues or reverses with even greater sequencing coverage.

Supplemental Text Describing 21U-RNAs

We estimate the total number of *C. elegans* 21U-RNA loci to be between 12,000 and 16,000. The 10,800 predicted loci of chromosome IV captured 77% of the sequenced 21U-RNAs. Based on the low frequency of predictions on other chromosomes (Fig. 2C), the number of false-positives in the set of 10,800 would be less than 1000 loci. Using 1000 as the number of false positives and 0.77 as the specificity yields lower estimate of ~12,000 total loci. Over 6000 predicted 21U-RNAs were not validated by our reads. These predictions captured only 46% of the 21U-RNAs that were unique to the dataset from A. Fire. This suggests that the predictions not yet validated might represent only

half of the remaining 21U-RNAs. Subtracting the 1000 false positives, then doubling the number of predictions not yet validated yields 10,000 loci not yet validated. Adding these to the 5600 observed 21U-RNA loci suggests an upper estimate of 16,000.

Supplemental Text Describing Genes Corresponding to Endogenous siRNAs

In a compendium of microarray experiments, *C. elegans* genes with positively correlated changes in expression across many environmental conditions and mutant backgrounds are grouped into collections called mountains, many of which are enriched with genes from certain functional categories (Kim et al. 2001). We compared the distribution of siRNA-complemented genes among the mountains to that of all genes included in this compendium (Table S6). All the mountains of transposon-enriched genes and all but one of the mountains of germline-enriched genes were overrepresented in the set of siRNA-complemented genes. An enrichment for endogenous siRNAs matching transposases had been observed previously and is consistent with the proposal that the RNAi machinery directly silences transposable elements (Lee et al. 2006; Sijen and Plasterk 2003).

We further explored the overrepresentation of germline-enriched genes using published expression annotations from microarray analysis of mutant worms (*glp-4, fem-3(gf),* and *fem-1(lf)*), in which germline cells do not proliferate, only give rise to sperm, or only give rise to oocytes, respectively (Reinke et al. 2000). Of the siRNA-complemented genes in this microarray analysis, 10.4% were annotated as germline-intrinsic, 16.5% were annotated as sperm enriched, and 3.0% were annotated as oocyte enriched, an enhancement over the 4.3%, 5.5%, and 2.2% representation of all genes in each of those three categories, respectively. When considering only the genes

complementing 26mer siRNAs, the fraction that was sperm enriched was even more striking—increasing from 16.5% to 55%. The observed germline enrichment among those genes complemented by siRNAs might have reflected preferred targeting of germline genes by endogenous siRNAs. However, the observed dependence of siRNA production on mRNA templates implied that the abundance of siRNAs that complemented any particular mRNA should have scaled linearly with both the mRNA length and abundance. To evaluate the contribution of length/abundance biases among germline-expressed genes to the results presented above, we ranked mRNAs identified by serial analysis of gene expression (SAGE) of mixed stage *C. elegans* (http://elegans.bcgsc.bc.ca) according to the products of their tag counts and lengths, and used the top-ranked genes as a control set to repeat our comparison. Of the control genes included in the microarray analysis, 9.5%, 0.6% and 3.4% were annotated as germline-intrinsic, sperm enriched, and oocyte enriched, respectively. The comparison of the siRNA-complemented set with the control set showed that among germline enriched genes, sperm-enriched genes were more than 20-fold over-represented in the siRNA set (p-value <0.01, chi-square test), consistent with a potential regulatory role of siRNAs for sperm-enriched genes. Analogous results were obtained for transposon genes. In contrast, the enrichment of germline intrinsic genes and oocyte-enriched genes might have been due to abundance and length biases.

Experimental Procedures

Genome hit normalization. We considered unique sequences to have a unique character sequence when compared to all other sequences of equivalent length. The numbers of

unique sequences and reads were both normalized to the number of perfect BLAST-derived matches to the *C. elegans* genome. Each hit to a region of interest was counted independently; thus, if one unique sequence representing 28 sequence reads matched three loci in the *C. elegans* genome, two of which were sense matches to exons, we would report that this contributed (# sense hits to exons)/(# total hits to the genome) = .67 counts to the number of unique sequences matching the sense strand of exons, and (# reads) * (# sense hits to exons) / (# total hits to the genome) = 18.67 counts to the number of reads matching the sense strand of exons.

Defining genomic regions rich in 21U-RNAs. Using data from read set 1, a Markov model was constructed with 3 states: "S" (giving rise to small 21U-RNA species of interest), "N" (giving rise to no small RNAs), and "F" (false-positive, giving rise to small RNAs with a variety of sizes and 5´ nucleotides). Emissions of four types were observed for non-overlapping 100nt blocks: 1) no reads mapping to that block; 2) one or more 21U-RNA read, no other reads; 3) no 21U-RNA reads, one or more other read; 4) one or more 21U-RNA read and one or more other read. Emission probabilities were generated by training on regions selected manually to resemble each state ("S": chrIV, 15.5-16.5Mb; "N": chrIV, 10.5-11.5Mb; "F": chrI, 15,000,000-15,080,200). Transition probabilities were set manually to reflect the observed sizes of clumps of 21U-RNA reads between coordinates 16.25Mb-17.25Mb on chrIV ("S": 70kb; "N": 40kb; "F" set equal to "S"). Initial state probablilites were also set manually ("S": 0.3; "N": 0.6, "F": 0.1). That model was parsed over non-overlapping 100nt blocks of the entire genome using the Viterbi algorithm, and the resulting parse defined regions rich in 21U-RNAs in

downstream analysis. Briefly, the parse yielded 20 "S" regions spanning chrIV:4,834,600-6,994,100; one "S" region spanning chrIV:9,131,000-9,132,300; and 22 "S" regions spanning chrIV:13,599,100-17,262,800. No portions of any other chromosome were parsed to state "S".

Defining and detecting the 21U-RNA upstream motifs. The sequence motifs found upstream of the 21U-RNA loci were defined by examining the sequences upstream of each 21U-RNA genomic locus. 21U-RNA loci were defined as those mapping to regions of chromosome IV that had been parsed to the "S" state as described above. The small motif was derived by aligning the surrounding genomic sequence based on the 5′ nucleotide of the 21U-RNA. The distribution of distances between the small motifs (whose position is fixed relative to the 21U-RNA sequence) and the large motifs was determined by plotting the frequency of perfect matches to the core sub-motif 'GTTTC' across the sequence upstream of the 21U-RNAs. The large motif was derived from an alignment of the upstream genomic sequences constructed based on searches for the expanded sub-motif 'CTGTTTCA'. Matches were sought with 0, 1, 2, 3, or 4 mismatches to the expanded sub-motif, in that order. For each number of allowed mismatches, the expanded sub-motif was sought in each position allowed by the aforementioned distance distribution, in descending order of the distance frequency. Alignments for the large motif were centered based on the first matches to the expanded sub-motif.

The scoring matrix for detecting the 21U-RNA upstream motifs was constructed using $\log_2$-odds ratios. For each position in the large motif, the foreground frequency of

each nucleotide was calculated based on the counts of that nucleotide in the large motif alignment described above. Foreground nucleotide frequencies for the small motif were calculated similarly using the small motif alignment described above. The background nucleotide frequencies were estimated to be 34% for A and T, 16% C and G, based on the properties of the sequences surrounding the 21U-RNAs and their associated motifs. For each nucleotide N at each position, pseudocounts were added, and scores derived, according to the following formula:

$$\text{Score} = \log_2\left(\frac{(f_N * R) + (b_N * P)}{b_N(R + P)}\right)$$

where f_N was the foreground (observed) frequency of N at the given position, b_N was the background frequency of N, R was the total number of real (observed) counts, and P was the total number of pseudocounts. We set P equal to the square root of R.

Scores for the distance between the two motifs were based on the counts of the minimal sub-motif at various positions relative to the 5′ end of the 21U-RNA, as described above. Foreground frequencies were calculated as number of counts for distances ranging from 16 to 25 nucleotides (inclusive) divided by the total number of counts in that range. The background frequency model described even probabilities across the full range. Pseudocounts were added to each foreground count as described for nucleotides, and a scoring matrix was calculated in the same manner as for nucleotide identities.

21U-RNAs were predicted for a given stretch of genomic sequence by applying the scoring matrix for the large motif to each position on each strand of the genomic sequence. For each position, the maximum sum of the distance and small motif scores

over all allowed distances from the large motif was determined, and was added to the score for the large motif. If the sum of the scores was ≥ 15.5, the 21U-RNA was predicted using the location of the small motif to define the 5′ end and assuming a length of 21nt.

Length and 5′ identity of 21U-RNAs. To assess the length distributions of 21U-RNAs, all genomic loci with matching reads were scored using the motif matrix described above, and normalized counts were binned according to those scores. For this assessment, the position of the small motif was fixed according to the 5′ end of the read, and its score added to the maximal combined large motif and distance matrix scores. The fractions of <21 nt, 21 nt, and >21 nt reads with motifs were derived from the total counts with scores >0. 5′nucleotide identity was similarly assessed, except that the two positions of the small motif scoring matrix corresponding to the 1st and 2nd nucleotides of the read did not contribute to the motif score.

siRNA methods. Matches to the sense and antisense strands of exons were found by comparing the coordinates of BLAST hits representing perfect matches to the *C. elegans* genome to the coordinates of exons as annotated in the Sanger gene set accompanying assembly ce2 downloaded from UCSC (Karolchik et al. 2003). This gene annotation set is derived from the WormBase gene annotations (www.wormbase.org; release WS120, 3/1/2004). Counts matching the sense and antisense strands of exons were determined using the genome hit normalization scheme described above. Genes with at least a fraction of a normalized match to any of the small RNA sequences that had not been

classified as 21U-RNAs comprised the lists of siRNA-complemented genes and sense RNA-matched genes, depending on the orientation of the BLAST hit. Splicing variants were collapsed, leaving 1720 siRNA-complemented genes and 1346 sense RNA-matched genes.

For each mountain presented in (Kim et al. 2001), the fraction of siRNA-complemented genes included in the mountain (out of 1503 siRNA-complemented genes included in that analysis) was compared to the fraction of all genes from the topomap (following the collapse of splice variants) included in the mountain. Significant enrichment was identified using Chi square tests with a p-value threshold of 0.01. Similarly, the fraction of siRNA-complemented genes annotated as germline intrinsic, sperm enriched, or oocyte enriched in reference (Reinke et al. 2000) was compared to the fraction of all genes included in that microarray analysis so annotated, and significant enrichment was identified using Chi square tests as described above.

The SAGE control set was constructed using mixed-stage SAGE data with WS140 gene names obtained from the Genome BC *C. elegans* Gene Expression Consortium (http://elegans.bcgsc.bc.ca) with the sequence quality filter set to 0.99. From that dataset, ambiguous and unmatched tags were removed and splicing variants were collapsed. Genes were ranked according to the product of tag counts and gene lengths, and the top 1720 comprised the control set. The fraction of siRNA-complemented genes annotated as germline intrinsic, sperm enriched, or oocyte enriched in reference (Reinke et al. 2000) was compared to the fraction of SAGE control genes so annotated, and significant enrichment was identified using Chi square tests as described above.

Acknowledgment: The SAGE data were produced at the Michael Smith Genome Sciences Centre with funding from Genome Canada.

References

Ambros, V., R.C. Lee, A. Lavanway, P.T. Williams, and D. Jewell. 2003. MicroRNAs and other tiny endogenous RNAs in *C. elegans*. *Curr Biol* **13:** 807-818.

Bartel, D.P. 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116:** 281-297.

Brennecke, J., A. Stark, R.B. Russell, and S.M. Cohen. 2005. Principles of microRNA-target recognition. *PLoS Biol* **3:** e85.

Farh, K.K., A. Grimson, C. Jan, B.P. Lewis, W.K. Johnston, L.P. Lim, C.B. Burge, and D.P. Bartel. 2005. The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. *Science* **310:** 1817-1821.

Grad, Y., J. Aach, G.D. Hayes, B.J. Reinhart, G.M. Church, G. Ruvkun, and J. Kim. 2003. Computational and experimental identification of *C. elegans* microRNAs. *Mol Cell* **11:** 1253-1263.

Griffiths-Jones, S. 2004. The microRNA Registry. *Nucleic Acids Res* **32:** D109-111.

Karolchik, D., R. Baertsch, M. Diekhans, T.S. Furey, A. Hinrichs, Y.T. Lu, K.M. Roskin, M. Schwartz, C.W. Sugnet, D.J. Thomas, R.J. Weber, D. Haussler, and W.J. Kent. 2003. The UCSC Genome Browser Database. *Nucleic Acids Res* **31:** 51-54.

Kim, S.K., J. Lund, M. Kiraly, K. Duke, M. Jiang, J.M. Stuart, A. Eizinger, B.N. Wylie, and G.S. Davidson. 2001. A gene expression map for *Caenorhabditis elegans*. *Science* **293:** 2087-2092.

Lagos-Quintana, M., R. Rauhut, W. Lendeckel, and T. Tuschl. 2001. Identification of novel genes coding for small expressed RNAs. *Science* **294:** 853-858.

Lau, N.C., L.P. Lim, E.G. Weinstein, and D.P. Bartel. 2001. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science* **294:** 858-862.

Lee, R.C. and V. Ambros. 2001. An extensive class of small RNAs in *Caenorhabditis elegans*. *Science* **294:** 862-864.

Lee, R.C., C.M. Hammell, and V. Ambros. 2006. Interacting endogenous and exogenous RNAi pathways in *Caenorhabditis elegans*. *Rna* **12:** 589-597.

Lee, Y., C. Ahn, J. Han, H. Choi, J. Kim, J. Yim, J. Lee, P. Provost, O. Radmark, S. Kim, and V.N. Kim. 2003. The nuclear RNase III Drosha initiates microRNA processing. *Nature* **425:** 415-419.

Lewis, B.P., C.B. Burge, and D.P. Bartel. 2005. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120:** 15-20.

Lim, L.P., N.C. Lau, E.G. Weinstein, A. Abdelhakim, S. Yekta, M.W. Rhoades, C.B. Burge, and D.P. Bartel. 2003. The microRNAs of *Caenorhabditis elegans*. *Genes Dev* **17:** 991-1008.

Ohler, U., S. Yekta, L.P. Lim, D.P. Bartel, and C.B. Burge. 2004. Patterns of flanking sequence conservation and a characteristic upstream motif for microRNA gene identification. *Rna* **10:** 1309-1322.

Reinke, V., H.E. Smith, J. Nance, J. Wang, C. Van Doren, R. Begley, S.J. Jones, E.B. Davis, S. Scherer, S. Ward, and S.K. Kim. 2000. A global profile of germline gene expression in *C. elegans*. *Mol Cell* **6:** 605-616.

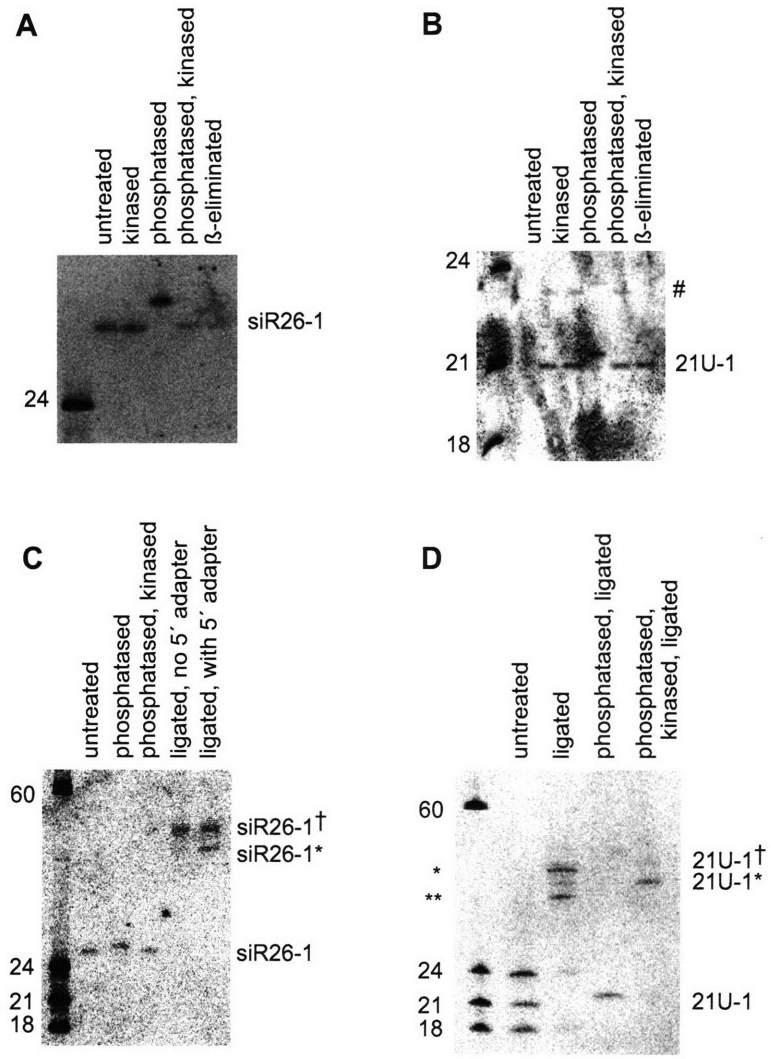
Sijen, T. and R.H. Plasterk. 2003. Transposon silencing in the *Caenorhabditis elegans* germ line by natural RNAi. *Nature* **426:** 310-314.

Tables S1-S5 are found on the accompanying CD-ROM.

Table S6. Gene mountains (Kim et al., 2001) enriched in siRNA-complemented genes

Mountain	Functional desciption	Percent of siRNA-complemented genes	Percent of all genes
Mount 4	Sperm-enriched genes; protein kinases; protein phospatases; major sperm proteins	21%	6.5%
Mount 5		12.4%	5.1%
Mount 7	Germ line-enriched; oocyte; meiosis; mitosis	9.1%	4.4%
Mount 11	Germ line-enriched; oocyte; meiosis; mitosis; histone H1; retinoblastoma complex	7.7%	3.2%
Mount 18	Germ line; oocyte; biosynthesis; protein synthesis	1.7%	1.0%
Mount 20	Germ line-enriched; biosynthesis; protein synthesis	2.5%	0.87%
Mount 23	Protein expression; energy generation	2.1%	0.76%
Mount 25	Mariner transposase	1.7%	0.55%
Mount 32	Nucleosomal histones	0.47%	0.13%
Mount 33	Tc1 transposon	1.0%	0.14%
Mount 37	Tc3 transposon	0.67%	0.06%

Supplemental Figure 1



Supplemental Figure S1. Enzymatic probing of two small endogenous RNAs indicates they both have 5´ monophosphates.

(A) Evidence for a phosphate group on a 26mer siRNA. Total RNA from mixed-stage C. elegans was subjected to the indicated treatments, separated on a 60 cm 17% PAGE gel, then blotted and probed for siR26-1. The shift with phosphatase treatment and return to normal mobility upon 5´ phosphorylation indicated the presence of a phosphate group. These data ruled out phosphorylation at more than one site, and suggested a 5´ monophosphate, although a 5´ di- or triphosphate, or a monophophate at another position were difficult to exclude based on mobility alone. As also shown in Figure 3C, resistance to periodate oxidation/ ß-elimination indicated modification on either the 2´ or 3´ hydroxyl of the 3´ terminal ribose.

(B) Evidence for a phosphate group on a 21U-RNA. Total RNA from mixed-stage C. elegans was subjected to the indicated treatments, separated on a 60 cm 17% PAGE gel, then blotted and probed for 21U-1. The shift with phosphatase treatment and return to normal mobility upon 5´ phosphorylation indicated the presence of a phosphate group. As in panel A, these data ruled out phosphorylation at more than one site, and suggested a 5´ monophosphate, although a 5´ di- or triphosphate, or a monophophate at another position were difficult to exclude based on mobility alone. As also shown in Figure 3C, resistance to periodate oxidation/ß-elimination indicated modification on either the 2´ or 3´ hydroxyl of the 3´ terminal ribose. Cross hybridization to a 23-nt species was detected (#). This cross-hybridizing RNA contained a 2´, 3´ diol, as indicated by susceptibility to periodate oxidation and ß-elimination.

(C) Evidence for a 5´ monophosphate on a 26mer siRNA. Small RNAs from mixed-stage C. elegans were gel purified and then subjected to the indicated treatments. Small RNAs were phosphatased, phosphatased and rephosporylated, or tested as substrates in ligation reactions identical to the second ligation step of the library construction (Lau et al., 2001), either in the presence or absence of the 17-nt 5´ adapter. The samples were resolved on a 15 cm 15% PAGE gel then blotted and probed for siR26-1. As in panel A, phosphatase treatment retarded the mobility of siR26-1, indicating the presence of a phosphate group. Rephosphorylation restored mobility to that of untreated siR26-1. In the absence of the 17-nt 5´ adapter, siR26-1 could potentially ligate on either its 5´ or 3´ end to endogenous ~22-nt RNAs (siR26-1†), such microRNAs, which are known to be abundant and have ligation-compatible ends. The 17mer, nonphosphorylated its 5´ terminus, was capable of ligating only at its 3´ terminus. Thus the ligation product appearing only in the presence of the adapter (siR26-1*) indicated that siR26-1 contained a terminal 5´ monophosphate.

(D) Evidence for a 5´ monophosphate on a 21U RNA. Small RNAs from mixed-stage C. elegans were gel purified using 24mer and 18mer 5´ 32P-labeled markers. Trace amounts of radiolabeled markers were carried forward in the purified sample (observed in the untreated lane). Untreated RNA, phosphatase-treated RNA, and phosphatased and rephosphorylated RNA were tested as substrates in ligation reactions identical to the second ligation step of the library construction (Lau et al., 2001), and samples were resolved on a 15 cm 15% PAGE gel then blotted and probed for 21U-1. The untreated 21U-1 was fully competent for ligation to the 17-nt 5´ adapter (21U-1*). Because the 17mer was capable of ligating only at its 3´ terminus, the ligation to 21U-1 confirmed that the phosphate group implicated in the analysis of panel B was a 5´ monophosphate. As expected, the trace size markers were also suitable ligation substrates (* and **), and their signal disappeared after removing the 5´ 32P by phosophatase treatment. A faint upper band (21U-1†) likely corresponded to ligation to miRNAs.

Figure S2

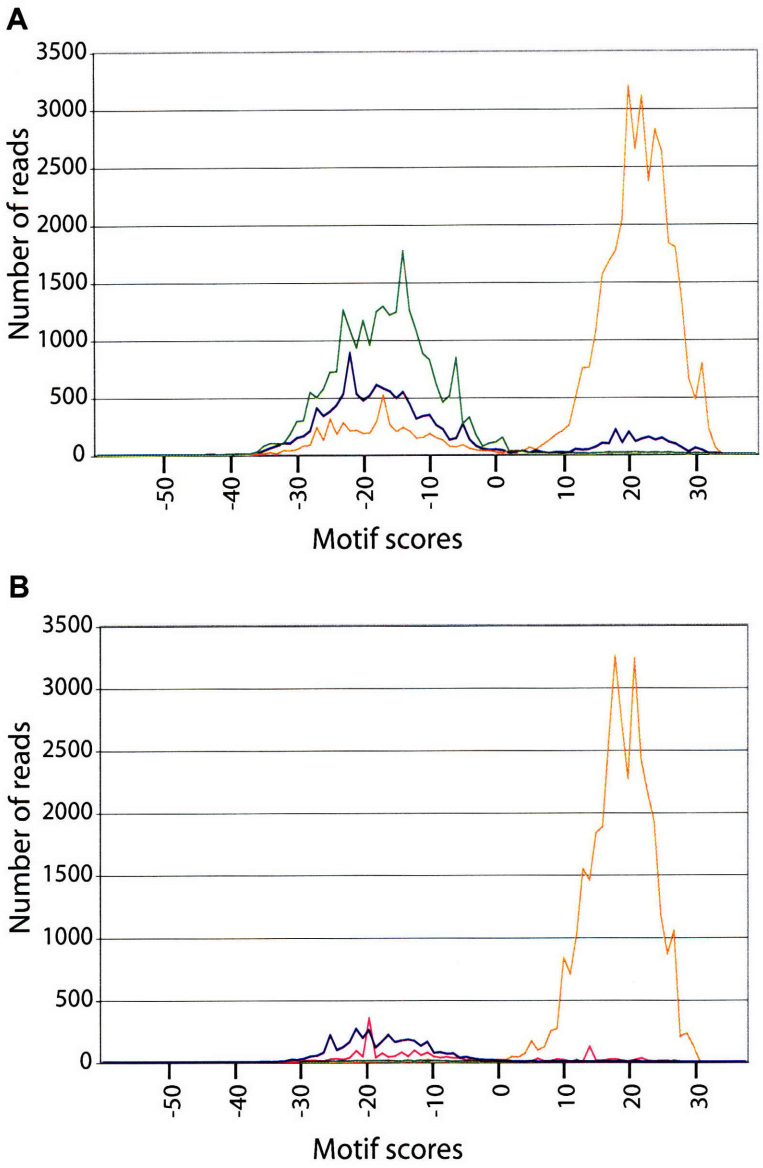


Figure S2. The properties of the 21U-RNAs correlate very strongly with each other.
(A) Analysis of read length. For binned 21U-RNA upstream motif scores (using integer bins), read frequencies are plotted for three length categories: <21 nt (blue), 21 nt (orange), or >21 nt (green). Higher scores are more likely to have been derived from the foreground (motif) model than the background model. For the positive-scoring loci, the large enrichment of <21-nt reads over >21-nt reads suggests that the <21-nt reads might be the result of degradation. Degradation from the 5′ end of a 21U-RNA would negatively affect scores because it would often change the 5′ nucleotide and would increase the distance between the RNA and the upstream motifs. The observation that positive score distributions of <21-nt reads and 21-nt reads tightly correlated implied that length heterogeneity is primarily at the 3′ end of 21U-RNAs.
(B) Analysis of 5′ nucleotide and chromosomal location of 21-nt reads. For binned 21U-RNA upstream motif scores, read frequencies are plotted for four categories: reads that begin with U and fall within the 21U-rich regions of chromosome IV (orange), reads that do not begin with U but do fall within the 21U-rich regions of chromosome IV (green), reads that begin with U but fall outside the 21U-rich regions of chromosome IV (pink), reads that do not begin with U and fall outside the 21U-rich regions of chromosome IV (blue). For this panel, the 5′ nucleotides of each read were excluded from the motif scoring. 21U-rich regions were defined as in the text. Very few (0.2%) 21-nt reads lacking a 5′ U were associated with motifs.

Chapter 3

Intronic microRNA precursors that bypass Drosha processing

J. Graham Ruby*, Calvin H. Jan*, David P. Bartel

Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA, 02142
Howard Hughes Medical Institute and Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, 02139

*These authors contributed equally to this work.

J.G.R. performed all the computational analysis. C.H.J. performed the experimental analysis. All authors contributed to the design of the study and preparation of the manuscript.

Published as:

JG Ruby*, CH Jan*, and DP Bartel. (2007) Intronic microRNA precursors that bypass Drosha processing. *Nature*. 448:83-86.

* these authors contributed equally to this work.

MicroRNAs (miRNAs) are ~23-nt endogenous RNAs that often repress the expression of complementary messenger RNAs (Bartel 2004). In animals, miRNAs derive from characteristic hairpins in primary transcripts through two sequential RNase III-mediated cleavages; Drosha cleaves near the base of the stem to liberate a ~60-nt pre-miRNA hairpin, then Dicer cleaves near the loop to generate a miRNA:miRNA* duplex (Lee et al. 2003; Tomari and Zamore 2005). From that duplex, the mature miRNA is incorporated into the silencing complex. Here, we identified an alternative pathway for miRNA biogenesis in which certain debranched introns mimic the structural features of pre-miRNAs to enter the miRNA-processing pathway without Drosha-mediated cleavage. We call these pre-miRNAs/introns "mirtrons" and have identified 14 mirtrons in *Drosophila melanogaster* and another four in *Caenorhabditis elegans* (including the reclassification of *mir-62*). Some of these have been selectively maintained during evolution with patterns of sequence conservation suggesting important regulatory functions in the animal. The abundance of introns comparable in size to pre-miRNAs appears to have created a context favourable for the emergence of mirtrons in flies and nematodes, suggesting that other lineages with many similarly sized introns probably also have mirtrons and that the mirtron pathway could have provided an early avenue for the emergence of miRNAs before the advent of Drosha.

While examining sequencing data of small RNAs from *D. melanogaster* (Ruby 2007), we observed clusters of small RNAs originating from the outer edges of an annotated 56-nt intron (Fig. 1a). These sets of reads (each read representing an

independently-sequenced cDNA) had properties similar to those observed previously for miRNA/miRNA* duplexes (Ruby et al. 2006), in that each set had a more consistent 5′ than 3′ terminus, and the two sets were complementary to each other, with the dominantly abundant species of each set forming 2-nt 3′ overhangs when paired to each other. Moreover, the sequence and predicted secondary structure of the intron were conserved in a pattern resembling that of pre-miRNAs (Lim et al. 2003) (Fig. 1b, c). We annotated this locus as *mir-1003*.

Despite these clearly miRNA-like properties, semblance to canonical miRNA primary transcripts (pri-miRNAs) stopped abruptly at the borders of the intron. Pairing at the base of the hairpin did not extend beyond the miRNA/miRNA* duplex, i.e. beyond the splice sites. In place of extended pairing, which is needed for pri-miRNA cleavage by Drosha (Han et al. 2006), the intron had conserved canonical splice sites (Fig. 1a), leading to the model that this miRNA did not arise from a canonical miRNA biogenesis pathway but instead arose from an alternative pathway in which splicing, rather than Drosha, defined the pre-miRNA (Fig. 1d). Consistent with this model, spliced lariats linearized by the lariat debranching enzyme bear 5′ monophosphates(Ruskin and Green 1985) and 3′ hydroxyls (Padgett et al. 1984), the same moieties found in pre-miRNAs (Hutvagner et al. 2001).

Thirteen additional pre-miRNAs/introns, termed mirtrons, were found in a search of other loci with similar properties (*mir-1004~1016*, Table S1). The most abundant RNA species from each of the 14 mirtrons, annotated as the mature miRNA, derived from the 3′ arm of its hairpin. Such bias was consistent with the known 5′ nucleotide biases of miRNAs, which frequently begin with a U and rarely with a G (Lau et al. 2001). The

near-ubiquitous intronic 5′ G, together with other requirements at intron 5′ ends (Lim and Burge 2001), would place unfavourable constraints on miRNAs deriving from the 5′ arm of a mirtron, whereas the species from the 3′ arm would have more freedom. As expected, the species from the 3′ arms, like canonical miRNAs, usually had a 5′ U (12/14 mirtrons).

To test whether the small RNAs from mirtrons were functional miRNAs or inactive degradation intermediates, we assessed the gene-silencing capacities of miR-1003 and miR-1006 in *Drosophila* S2 cells. In animals, extensive complementarity leads to cleavage of the target mRNA, but posttranscriptional repression is more commonly mediated by less extensive complementary, primarily involving pairing to a 5′ region of the miRNA known as the miRNA seed (Bartel 2004). miR-1003 and miR-1006 repressed reporter genes with perfectly complementary sites, with the repression levels approaching that observed for the *let-7* miRNA and an analogous reporter (Fig. 1e). In addition, both mirtronic miRNAs repressed reporter genes containing *Drosophila* UTR fragments with seed-based matches typical of metazoan miRNA targets. Conservation of the miR-1003 and miR-1006 seeds (Fig. 1d, Table S1) suggested an *in vivo* role for such mirtron-mediated repression; target predictions for conserved mirtronic miRNAs are provided at targetscan.org.

Having established that mirtrons can direct miRNA-like gene repression, we tested the dependence of mirtron processing on splicing and debranching. A mutant *mir-1003* with a substitution that impaired splicing (3′ mut) failed to generate detectible pre- or mature miR-1003 (Fig. 2a, b) and displayed significantly less silencing activity (Fig. 1e). Mutations disrupting the 5′ splice site (5′ mut) also impaired splicing and miR-1003

accumulation (Fig 2a, b). Coexpressing a mutant U1 snRNA (U1-3G) that had compensatory changes designed to restore splice site recognition(Lo et al. 1994) restored splicing of *mir-1003* 5′ mut (Fig. 2b). Rescuing splicing also restored the levels of pre- and mature miR-1003 (Fig. 2b). These results demonstrated that splicing was required for mirtron maturation and function, which contrasts with the splicing-independent biogenesis of canonical miRNAs found within introns (Kim and Kim 2007).

We next used RNAi knockdown experiments to examine the trans-factor requirements for miR-1003 and miR-1006 biogenesis in *Drosophila* cells. As predicted by our model, in which mirtrons enter the miRNA biogenesis pathway after splicing and debranching, targeting the lariat debranching enzyme reduced the amount of pre- and mature mirtronic miRNAs without impeding canonical miRNA maturation (Fig. 2c, d). For each mirtron, a probe to the 5′ end of the intron (probe 1) detected both the pre-miRNA hairpin and the accumulating lariat, whereas a probe to the 3′ end of the intron (probe 2) detected the pre-miRNA but failed to detect the lariat, presumably due to overlap with the branch-point (Fig. S1a). Altered relative mobility on gels with different polyacrylamide densities confirmed detection of the mirtron lariat (Fig. S1b). The debranching knockdown results, together with those of the splice-site mutations and rescue, demonstrated that the intron lariat was an intermediate on the pathway of mirtronic miRNA biogenesis.

Knockdown of other miRNA biogenesis factors further supported our model. As expected if debranched mirtrons enter the later steps of the miRNA pathway rather than the siRNA pathway (Tomari and Zamore 2005), knockdown of Dicer-1 or its partner, Loquacious, increased the ratio of pre- to mature mirtronic miRNA, whereas knockdown

of Dicer-2 or its partner, R2D2, did not (Fig. 2c, d). Knockdown of Drosha decreased pre- and mature *let-7* RNA accumulation with little effect on mature miR-1003 or miR-1006 accumulation and a modest effect on mirtronic pre-miRNAs (Fig. 2c, d). The more modest effect on mirtronic pre- and mature miRNAs supported the idea that mirtronic pre-miRNAs are not Drosha cleavage products. The decrease of mirtronic pre-miRNA would be explained if Drosha bound mirtronic pre-miRNAs, stabilized them from degradation, and perhaps facilitated their loading into the nuclear export machinery. The decrease could also reflect increased Dicer-1 accessibility in the Drosha knockdown due to reduced substrate competition from endogenous pre-miRNAs. In this case, simultaneous knockdown of Dicer-1 and Drosha would lead to a more substantial accumulation of pre-miRNAs derived from mirtrons than from canonical miRNAs, as was observed for pre-miR-1003 and pre-miR-1006 compared to *let-7* pre-miRNA (Fig. 2c, d).

The distribution of intron lengths, which varies widely in different organims (Lim and Burge 2001; Yandell et al. 2006), would influence the probability of new mirtrons arising during evolution. The introns of *Drosophila* share a similar length distribution with the annotated pre-miRNAs, producing a context particularly well suited to the emergence to mirtrons (Fig. 3a, c). *C. elegans* also has a substantial number of pre-miRNA-sized introns. Indeed, examination of prior miRNA annotations revealed that *mir-62*, which produces a highly conserved nematode miRNA that was among the very first to be cloned in animals (Lau et al. 2001; Lee and Ambros 2001), had mirtron-like properties (Fig. 3b). Like the mirtrons of *D. melanogaster*, the base pairing capacity of the sequence surrounding pre-miR-62 ended at the border of the host intron, and the most

abundant miRNA 3′ terminus corresponded to the 3′ splice site (with the single read whose 3′ terminus extended into the 3′ exon attributable to untemplated nucleotide addition to the miRNA 3′ end (Ruby et al. 2006)). A directed search of *C. elegans* small RNA sequences (Ruby et al. 2006) revealed three more mirtrons, annotated here as *mir-1018~1020* (Table S2).

Even if only a very small portion of debranched introns can form secondary structures resembling those of pre-miRNAs, the abundance of pre-miRNA-sized introns in flies and nematodes would allow a large absolute number of candidate mirtrons to emerge over evolutionary timescales. Whether they persist as functional mirtrons depends on the selective advantage conferred to the host organism as a consequence of their gene-repression activities. This model for mirtron emergence predicts that, at any historical point, some introns will be processed as mirtrons that provide no advantage to the organism but have yet to be eliminated by natural selection or neutral drift. Accordingly, some but not all processed *D. melanogaster* mirtrons were significantly more conserved in *D. pseudoobscura* than were most small introns, and the same trend was observed for *C. elegans* mirtrons in *C. briggsae*, although their numbers were small (Fig. 3d). The three most conserved *D. melanogaster* mirtrons (*mir-1003/1006/1010*) gave rise to more reads than 27%, 16%, and 4% of the non-mirtronic miRNAs conserved to *D. pseudoobscura*, respectively (Ruby et al. 2007), while the most conserved *C. elegans* mirtron (*mir-62*) gave rise to more reads than 52% of the non-mirtronic miRNAs conserved to *C. briggsae* (Ruby et al. 2006).

Compared to flies and nematodes, mammals have few pre-miRNA-sized introns (Lim and Burge 2001; Yandell et al. 2006) (Fig. 3a), perhaps explaining why we found

no mirtrons among the annotated mammalian miRNAs (Griffiths-Jones 2004). Nonetheless, high-throughput sequencing of mammalian small RNAs might yet reveal mirtrons. In plants, miRNA processing could similarly bypass one of the RNase III cleavages, although plant mirtrons have not yet been identified (Bartel 2004; Griffiths-Jones 2004). Moreover, lineages with long introns might have other types of intronic miRNAs that bypass Drosha-mediated cleavage. This possibility was raised by *mir-1017*, whose putative pre-miRNA 5′ end, but not 3′end, matched the 5′ splice site of its host intron (Table S1). In contrast to true mirtrons, miRNAs of this type would depend on a nuclease to cleave their extensive 3′overhangs, as observed for the U14 snRNA derived from an intron of *hsc70* (Leverette et al. 1992). This mechanism, together with that of mirtron processing, would enable miRNAs to emerge in any organism with both splicing and posttranscriptional RNA silencing, even those lacking the specialized RNase III enzyme Drosha or its plant counterpart, DICER-LIKE1 (Bartel 2004). In this scenario, miRNAs might have emerged in ancient eukaryotes prior to the advent of modern miRNA biogenesis pathways.

Methods

Computational methods. *D. melanogaster* small RNAs were from 2,075,098 high-throughput pyrosequencing reads (Ruby 2007) and are available at the GEO. *C. elegans* small RNA sequences were from reference (Ruby et al. 2006). Introns were defined according to FlyBase v4.2 *D. melanogaster* gene annotations (Grumbling and Strelets 2006). *C. elegans* introns were defined using annotations and genomic sequence from WormBase (release WS120) (Stein et al. 2001). *Mus musculus* introns were defined

using NCBI RefSeq annotations (Pruitt et al. 2005) applied to the March 2005 release of the mouse genome available through UCSC (mm6) (Karolchik et al. 2003). RNA secondary structures were predicted using RNAfold (Hofacker et al. 1994). *D. melanogaster* intron conservation was assessed based on an 9-species multiZ alignment (Blanchette et al. 2004) of *D. melanogaster, D. simulans, D. yakuba, D. ananassae, D. pseudoobscura, D. virilis, D. mojavensis, Anopheles gambiae*, and *Apis mellifera* genomes, generated at UCSC (Karolchik et al. 2003). Percent nucleotide identity between *D. melanogaster* and *D. pseudoobscura* introns was calculated as the number of identity matches between the two orthologous introns in the multiZ alignment divided by the length of the longer intron. Introns not aligned between those two species were not tallied. *C. elegans* intron conservation was similarly determined using multiZ alignment of the *C. elegans* and *C. briggsae* (WormBase cb25.agp8) (Stein et al. 2001) genomes generated at UCSC (Karolchik et al. 2003). Pre-miRNA lengths were calculated using miRBase v9.1 hairpin annotations (Griffiths-Jones 2004). Secondary structures were generated using RNAfold (Hofacker et al. 1994), and the miRNA* position was inferred based from the annotated miRNA, assuming 2-nt 3′ overhangs. Pre-miRNA lengths were the sum of the miRNA length, the miRNA* length, and the length of intervening sequence.

Plasmids. Minigenes containing *mir-1003* and *mir-1006* and flanking exons were PCR amplified from genomic DNA. Minigenes for *mir-1006* and *mir-1003* were cloned into pMT-puro with the indicated sites to make expression plasmids pCJ19 and pCJ20, respectively. *let-7* was amplified from genomic DNA with primers 474 bp upstream and

310 bp downstream of the *let-7* hairpin and cloned into pMT-puro to make pCJ24. Similar minigenes replaced EGFP in p2032 (Brennecke et al. 2005) to give pCJ31 (*mir-1006*), pCJ30 (*mir-1003*), or pCJ32 (*let-7*). U1a snRNA and U1a-3G snRNA expression constructs were constructed essentially as described (Lo et al. 1994). Sequences of inserts in pCJ19 (pMT-puro_*mir-1006*), pCJ20 (pMT-puro_*mir-1003*), pCJ24 (pMT-puro_*let-7*), pCJ30 (p2032_*mir-1003*), pCJ31 (p2032_*mir-1006*), and pCJ32 (p2032_*let-7*) are provided (tableS3.FASTA). Quikchange site-directed mutagenesis (Stratagene, CA) was used to make 3′ splice site mutations with the indicated primers: *mir-1003* 3′ mut (CCTCTCACATTTACATATTCACGACGCCGTGAGCTGC and GCAGCTCACGGCGTCGTGAATATGTAAATGTGAGAGG), and *mir-1006* 3′ mut (GGTACAATTTAAATTCGATTTCTTATTCATGCGTGCAATACCAGTTGATC and GATCAACTGGTATTGCACGCATGAATAAGAAATCGAATTTAAATTGTACC). Similarly, *mir-1003* 5′ mut was made with the following mutagenic primers: (GCTGCGCAGAACGTGGGCATCTGGATGTGGTTGGC and GCCAACCACATCCAGATGCCCACGTTCTGCGCAGC; CCTCTCACATTTACATGTTCACAGGCGCCGTGAG and CTCACGGCGCCTGTGAACATGTAAATGTGAGAGG).

Luciferase-reporter inserts were made by annealing oligonucleotides with their reverse complements, leaving overhangs for the indicated restriction sites (lowercase): let-7-ps (gagctcACTATACAACCTACTACCTCAactagt), let-7-psm (gagctcACTATACAACCTACAAGCACAactagt), miR-1003-ps (gagctcCTGTGAATATGTAAATGTGAGAactagt), miR-1003-psm (gagctcCTGTGAATATGTAAAAGAGTGAactagt), miR-1006-ps

(gagctcCTATGAATAAGAAATCGAATTTAactagt), and miR-1006-psm (gagctcCTATGAATAAGAAATCCATTATAactagt). Annealed oligos were ligated into SacI/SpeI cleaved pIS2 (ref. (Lim et al. 2005)) . These plasmids were linearized with HindIII, polished with Klenow enzyme to create blunt ends, and digested with NotI to excise the *Renilla* luciferase gene with the modified UTR from the remainder of pIS2. The gel-purified *Renilla* gene fragment was then ligated into pMT-puro between EcoRV and NotI sites for copper-induced expression in S2 cells.

Cell culture and RNAi. S2-SFM cells were adapted from S2 cells to grow in Drosophila Serum Free Media (SFM) by passaging into increasing amounts of SFM (0%, 25%, 50%, 75%, 90%, 100%), then grown in SFM supplemented with 2 mM L-glutamine at 25°C in a humidified incubator. 5 μg of pCJ19 or pCJ20 were transfected into a 60 mm plate containing $2.5x10^6$ S2 cells with FuGENE HD. Cells were grown for 3 days, split 1:10, and selected for 3 weeks in 10 μg/ml puromycin prior to experimentation, then maintained in 5 μg/ml puromycin.

Templates for dsRNA were amplified by PCR and extended to have convergent T7 promoters. 400 μl PCR reactions were phenol/chloroform extracted, ethanol precipitated, and used as template for 400 μl T7 transcriptions. Transcription reactions were treated with 20U of DNase I for 15 minutes. The transcription products were then extracted in phenol:chloroform (5:1 pH 5.3) and ethanol precipitated. RNA was resuspended, desalted over Sephadex G-300, then heated to 75°C for 10 minutes and slow cooled to room temperature. Yield and quality were assessed by agarose gel and UV absorbance. The sense sequence of each dsRNA is listed in the supplemental FASTA file (Table S3).

S2 cells were soaked in 10μg/ml dsRNA in SFM. 500,000 cells were plated per well of a 24 well plate and soaked for 2 days, split 1:4, soaked another 2 days, expanded into 6 well plates, then soaked for three days. MicroRNA expression was induced by addition of 500 μM CuSO4 to the growth media, and RNA harvested 12 hours later with TRI reagent.

Northern blots were performed as described(Ruby et al. 2006), using the following oligonucleotides (purchased from IDT) as probes for the indicated RNA species ('+' precedes LNA bases): ACTATACAACCTACTACCTCA (*let-7*), C+TGT+GAA+TAT+GTA+AAT+GTG+AGA (*mir-1003* probe 1), CCAACCACATCCAGATACCCACC (*mir-1003* probe 2), C+TAT+GAA+TAA+GAA+ATC+GAA+TTT+A (*mir-1006* probe 1), TTTACGCATTTCAATTTCAAACTCAC (*mir-1006* probe 2), TTGCGTGTCATCCTTGCGCAGG (U6).

RT-PCR. 500 ng mirtron plasmids were cotransfected with 500 ng either U1 or GFP carrier plasmid using 3 μl FuGENE HD per well of a 12 well plate. 24 hours post-transfection, mirtron expression was induced for 36 hours in the presence of 500 μM CuSO4. Total RNA was extracted with TRI-reagent, and 4 μg were treated with DNase using the DNA-free kit (Ambion, TX). 500 ng DNA-free RNA were reverse-transcribed with oligo-dT(16) and Superscript III (Invitrogen, CA) per manufacturers instructions. 1 μl cDNA was used as a template for PCR using exonic primers (ATAAAGCCGATAAGCGTGCG and CGTCCTTGTGCGTCTCCTCC) flanking *mir-*

1003. After 24 cycles of PCR, 10 µl of the reaction was resolved on an ethidium-stained 1.5% agarose gel and visualized by UV illumination.

Quantitative RT-PCR was performed on an ABI 7000 Real-Time PCR system with ABI Power SYBR Green reagents. First-strand synthesis was performed as above. The following primer pairs were used to amplify the specified mRNA:
Actin 5c (CCCATCTACGAGGGTTATGC, TTGATGTCACGGACGATTTC); Drosha (TCACCATCCACGAGCTAGAC, ACGAAACGCGGAAAGAAGTG); Dicer-1 (GCCATTGAAGCATGACATTG, AAATCCCTCCTTGCCGATAG); Loquacious (CGATTACCGAGTGGATACGG, CAAAGGAATCGGTGGAAAAG); Dicer-2 (GGCCACGAAACTTAAAGAGC, TGTGGAAAGGACACCATGAC); R2D2 (GACGGAGGGTACGTCTGTAAA, AGCAGTTGGATTTTACGCAAG); CG7942 (TTATCCCTGCCAGCACCTAC, CCTCTACATGAGGCGTTTCC).
Ct and baseline were detected by ABI 7000 SDS software. Actin5C was used to calculate the ΔCt, and ΔΔCt was calculated by subtracting the ΔCt from that of the GFP dsRNA treated samples; the relative abundance was calculated as 1 / (2^(ΔΔCt). Geometric mean ± standard deviation are shown for three replicate wells.

Luciferase assays. S2-SFM cells were plated 300,000 cells/ml in 96 well plates. After 24 hours, cells were cotransfected with 96 ng microRNA-expressing plasmid, 4 ng perfect-site reporter and 2 ng firefly reporter per well using FuGENE HD (3 µl lipid per µg DNA). Expression of *Renilla* luciferase was induced 24 hours post-transfection with 500 µM CuSO4. Luciferase assays were performed 24 hours post-induction with the Dual-Glo Luciferase system (Promega, WI) on a Tecan Safire2 plate reader. The ratio of

Renilla:firefly luciferase activity was measured for each well. To calculate fold repression, the ratio of *Renilla*:firefly for reporters with mutant sites was divided by the ratio of *Renilla*:firefly for reporters with wild-type sites. These values were also obtained in the presence of a plasmid expressing a non-cognate miRNA, and fold repression for the cognate miRNA was normalized to that of the non-cognate.

Figure legends

Figure 1. Introns that form pre-miRNAs. **a**, *D. melanogaster mir-1003* with corresponding reads from high-throughput sequencing (Ruby 2007). The miRNA (red), miRNA* (blue) and splice sites (green lines) are indicated, with predicted secondary structure shown in bracket notation (Hofacker et al. 1994). **b**, Conservation of *mir-1003* across seven *Drosophila* species (Blanchette et al. 2004; Karolchik et al. 2003), coloured as in (**a**), and also indicating consensus splice sites (Lim and Burge 2001) (green) and nucleotides differing from *D. melanogaster* (grey). **c**, Predicted secondary structures of representative debranched pre-miR-1003 orthologs, coloured as in (**b**). **d**, Model for convergence of the canonical and mirtronic miRNA biogenesis pathways (see text). **e**, MicroRNA regulation of luciferase reporters in S2 cells. Plotted is the ratio of repression for wild-type versus mutated sites, normalized to that with the indicated non-cognate miRNA. Bar colour represents the cotransfected miRNA expression plasmid; coloured lines below indicate the cognate miRNA for the specified reporter. Error bars represent the third largest and smallest values from 12 replicates (four independent experiments, each with three transfections; * $P < 0.01$, ** $P < 0.0001$, Wilcoxon rank-sum test).

Figure 2. Mirtrons are spliced as introns and diced as pre-miRNAs. **a,** Schematic of splice-site mutations. **b,** Base pairing between the indicated U1a and *mir-1003* RNAs (left), and RT-PCR and Northern-blot analyses of mir-1003 variants from (a). **c,** Northern blots analyzing *let-7* and *miR-1003* maturation in cells treated with double-stranded RNAs (dsRNAs) corresponding to indicated genes. Shown are results from one membrane, sequentially stripped and probed for *let-7* RNA, pre-miR-1003/lariat (probe 1), pre-miR-1003/miR-1003 (probe 2), and U6. Previously validated dsRNAs were used(Dorner et al. 2006; Forstemann et al. 2005), except for debranching enzyme (DBR), for which two unique dsRNAs were used. Knockdowns were confirmed by monitoring mRNA level and protein function (Fig. S2). Quantification of band intensities is provided (Table S3). * marks the lariat. **d,** Analysis of miR-1006 processing, as in (**c**).

Figure 3. Emergence and conservation of mirtrons in species with appropriately-sized introns. **a**, Distributions of intron (orange) and pre-miRNA (green) lengths from the indicated species. Introns and pre-miRNAs were binned by length. **b**, Intron and associated reads of *C. elegans* miR-62 (Ruby et al. 2006), coloured as in Figure 1a. Reads with untemplated nucleotides added at their 3′ terminus are shown below. **c,** Distributions of pre-miRNA (green) and mirtron (grey) lengths from *D. melanogaster* and *C. elegans*. **d,** Conservation of all 40-90 nt introns (orange) versus mirtrons (grey) from *D. melanogaster* (% identity shared with *D. pseudoobscura*) and *C. elegans* (% identity shared with *C. briggsae*).

Acknowledgements

We are grateful to P. Sharp, T. Baker, and members of the Bartel laboratory for discussions. We thank W. Johnston for assistance with molecular cloning, P. Zamore and R. Green for dsRNA plasmids, S. Cohen for GFP and firefly luciferase *Drosophila* expression plasmids, and D. Sabitini for pMT-puro. Supported by a grant from the NIH. C.H.J. is a NSF graduate research fellow. D.P.B. is an investigator of the Howard Hughes Medical Institute. Small RNA sequences were deposited in the Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo/ accessions GPL5061 and GSE7448).

References

Bartel, D.P. 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116:** 281-297.

Blanchette, M., W.J. Kent, C. Riemer, L. Elnitski, A.F. Smit, K.M. Roskin, R. Baertsch, K. Rosenbloom, H. Clawson, E.D. Green, D. Haussler, and W. Miller. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* **14:** 708-715.

Brennecke, J., A. Stark, R.B. Russell, and S.M. Cohen. 2005. Principles of microRNA-target recognition. *PLoS Biol* **3:** e85.

Dorner, S., L. Lum, M. Kim, R. Paro, P.A. Beachy, and R. Green. 2006. A genomewide screen for components of the RNAi pathway in Drosophila cultured cells. *Proc Natl Acad Sci U S A* **103:** 11880-11885.

Forstemann, K., Y. Tomari, T. Du, V.V. Vagin, A.M. Denli, D.P. Bratu, C. Klattenhoff, W.E. Theurkauf, and P.D. Zamore. 2005. Normal microRNA maturation and germ-line stem cell maintenance requires Loquacious, a double-stranded RNA-binding domain protein. *PLoS Biol* **3:** e236.

Griffiths-Jones, S. 2004. The microRNA Registry. *Nucleic Acids Res* **32:** D109-111.

Grumbling, G. and V. Strelets. 2006. FlyBase: anatomical data, images and queries. *Nucleic Acids Res* **34:** D484-488.

Han, J., Y. Lee, K.H. Yeom, J.W. Nam, I. Heo, J.K. Rhee, S.Y. Sohn, Y. Cho, B.T. Zhang, and V.N. Kim. 2006. Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. *Cell* **125:** 887-901.

Hofacker, I.L., W. Fontana, P.F. Stadler, L.S. Bonhoeffer, M. Tacker, and P. Schuster. 1994. Fast folding and comparison of RNA secondary structures. *Monatshefte Fur Chemie* **125:** 167-188.

Hutvagner, G., J. McLachlan, A.E. Pasquinelli, E. Balint, T. Tuschl, and P.D. Zamore. 2001. A cellular function for the RNA-interference enzyme Dicer in the maturation of the *let-7* small temporal RNA. *Science* **293:** 834-838.

Karolchik, D., R. Baertsch, M. Diekhans, T.S. Furey, A. Hinrichs, Y.T. Lu, K.M. Roskin, M. Schwartz, C.W. Sugnet, D.J. Thomas, R.J. Weber, D. Haussler, and W.J. Kent. 2003. The UCSC Genome Browser Database. *Nucleic Acids Res* **31:** 51-54.

Kim, Y.K. and V.N. Kim. 2007. Processing of intronic microRNAs. *Embo J* **26:** 775-783.

Lau, N.C., L.P. Lim, E.G. Weinstein, and D.P. Bartel. 2001. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science* **294:** 858-862.

Lee, R.C. and V. Ambros. 2001. An extensive class of small RNAs in *Caenorhabditis elegans*. *Science* **294:** 862-864.

Lee, Y., C. Ahn, J. Han, H. Choi, J. Kim, J. Yim, J. Lee, P. Provost, O. Radmark, S. Kim, and V.N. Kim. 2003. The nuclear RNase III Drosha initiates microRNA processing. *Nature* **425:** 415-419.

Leverette, R.D., M.T. Andrews, and E.S. Maxwell. 1992. Mouse U14 snRNA is a processed intron of the cognate hsc70 heat shock pre-messenger RNA. *Cell* **71:** 1215-1221.

Lim, L.P. and C.B. Burge. 2001. A computational analysis of sequence features involved in recognition of short introns. *Proc Natl Acad Sci U S A* **98:** 11193-11198.

Lim, L.P., N.C. Lau, P. Garrett-Engele, A. Grimson, J.M. Schelter, J. Castle, D.P. Bartel, P.S. Linsley, and J.M. Johnson. 2005. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* **433:** 769-773.

Lim, L.P., N.C. Lau, E.G. Weinstein, A. Abdelhakim, S. Yekta, M.W. Rhoades, C.B. Burge, and D.P. Bartel. 2003. The microRNAs of *Caenorhabditis elegans*. *Genes Dev* **17:** 991-1008.

Lo, P.C., D. Roy, and S.M. Mount. 1994. Suppressor U1 snRNAs in Drosophila. *Genetics* **138:** 365-378.

Padgett, R.A., M.M. Konarska, P.J. Grabowski, S.F. Hardy, and P.A. Sharp. 1984. Lariat RNA's as intermediates and products in the splicing of messenger RNA precursors. *Science* **225:** 898-903.

Pruitt, K.D., T. Tatusova, and D.R. Maglott. 2005. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **33:** D501-504.

Ruby, J.G., C. Jan, C. Player, M.J. Axtell, W. Lee, C. Nusbaum, H. Ge, and D.P. Bartel. 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., Kellis, M., Bartel, D.P., Lai, E.C.,. 2007. Biogenesis, expression, and target predictions for an expanded set of microRNA genes in *Drosophila*. *Genome Research* **In press**.

Ruskin, B. and M.R. Green. 1985. An RNA processing activity that debranches RNA lariats. *Science* **229:** 135-140.

Stein, L., P. Sternberg, R. Durbin, J. Thierry-Mieg, and J. Spieth. 2001. WormBase: network access to the genome and biology of Caenorhabditis elegans. *Nucleic Acids Res* **29:** 82-86.

Tomari, Y. and P.D. Zamore. 2005. Perspective: machines for RNAi. *Genes Dev* **19:** 517-529.

Yandell, M., C.J. Mungall, C. Smith, S. Prochnik, J. Kaminker, G. Hartzell, S. Lewis, and G.M. Rubin. 2006. Large-scale trends in the evolution of gene structures within 11 animal genomes. *PLoS Comput Biol* **2:** e15.

Figure 1

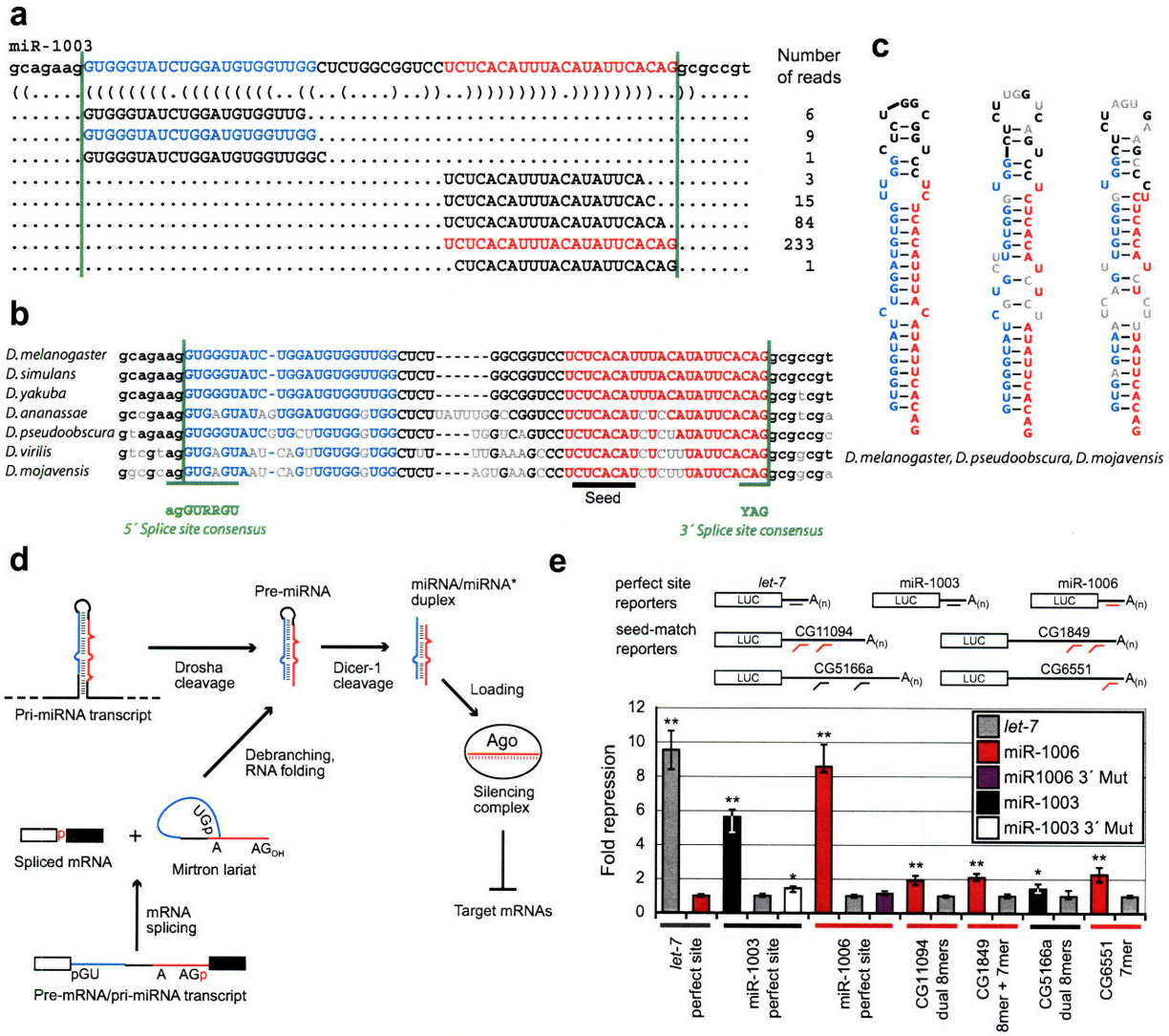

a
miR-1003
Number of reads
b
D. melanogaster
D. simulans
D. yakuba
D. ananassae
D. pseudoobscura
D. virilis
D. mojavensis
Seed
agGURRGU
5´ Splice site consensus
YAG
3´ Splice site consensus
c
D. melanogaster, D. pseudoobscura, D. mojavensis
d
Pre-miRNA
miRNA/miRNA* duplex
Drosha cleavage
Dicer-1 cleavage
Loading
Pri-miRNA transcript
Debranching, RNA folding
Ago
Silencing complex
Spliced mRNA
Mirtron lariat
mRNA splicing
Target mRNAs
Pre-mRNA/pri-miRNA transcript
e
perfect site reporters
let-7
miR-1003
miR-1006
seed-match reporters
CG11094
CG1849
CG5166a
CG6551
Fold repression
let-7
miR-1006
miR1006 3´ Mut
miR-1003
miR-1003 3´ Mut
let-7 perfect site
miR-1003 perfect site
miR-1006 perfect site
CG11094 dual 8mers
CG1849 8mer + 7mer
CG5166a dual 8mers
CG6551 7mer

Figure 2

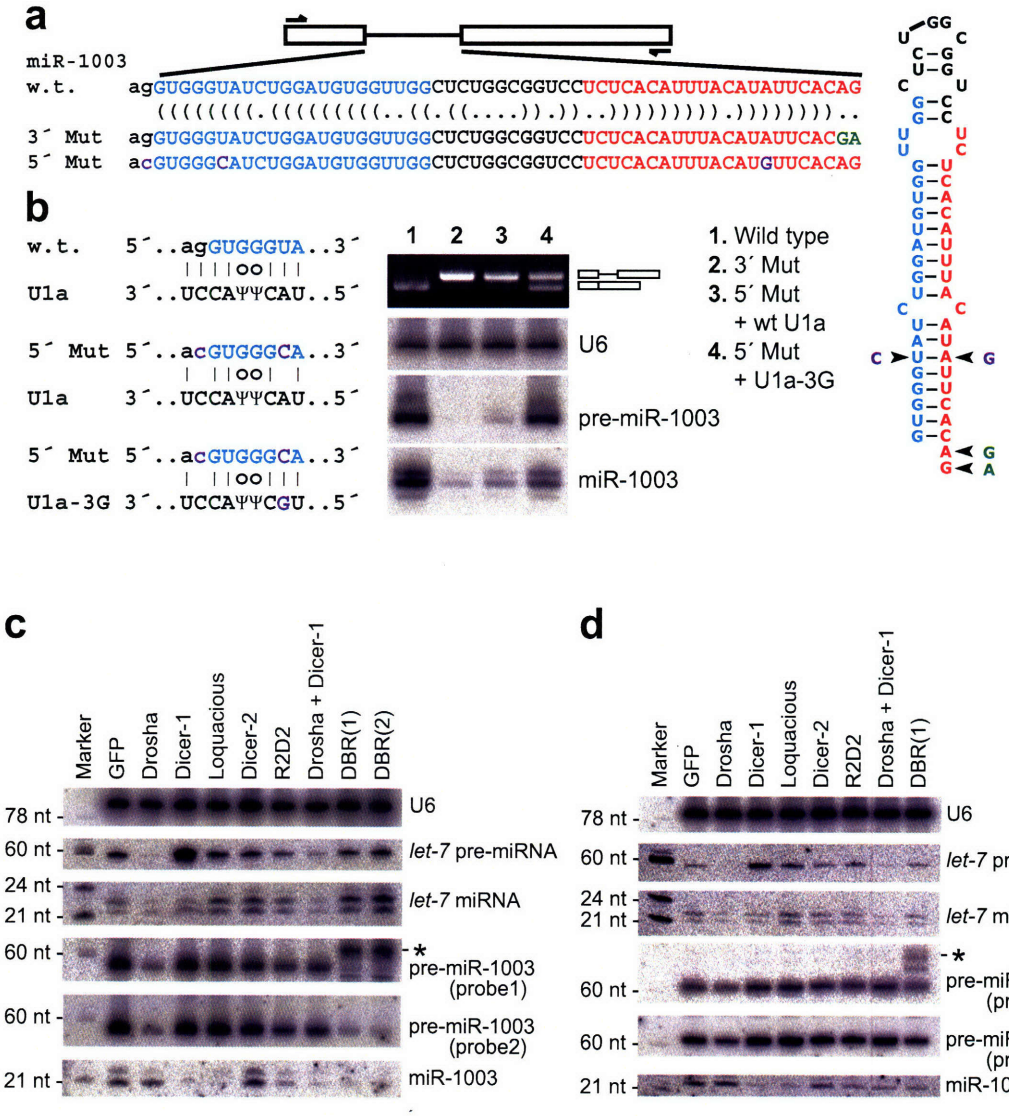
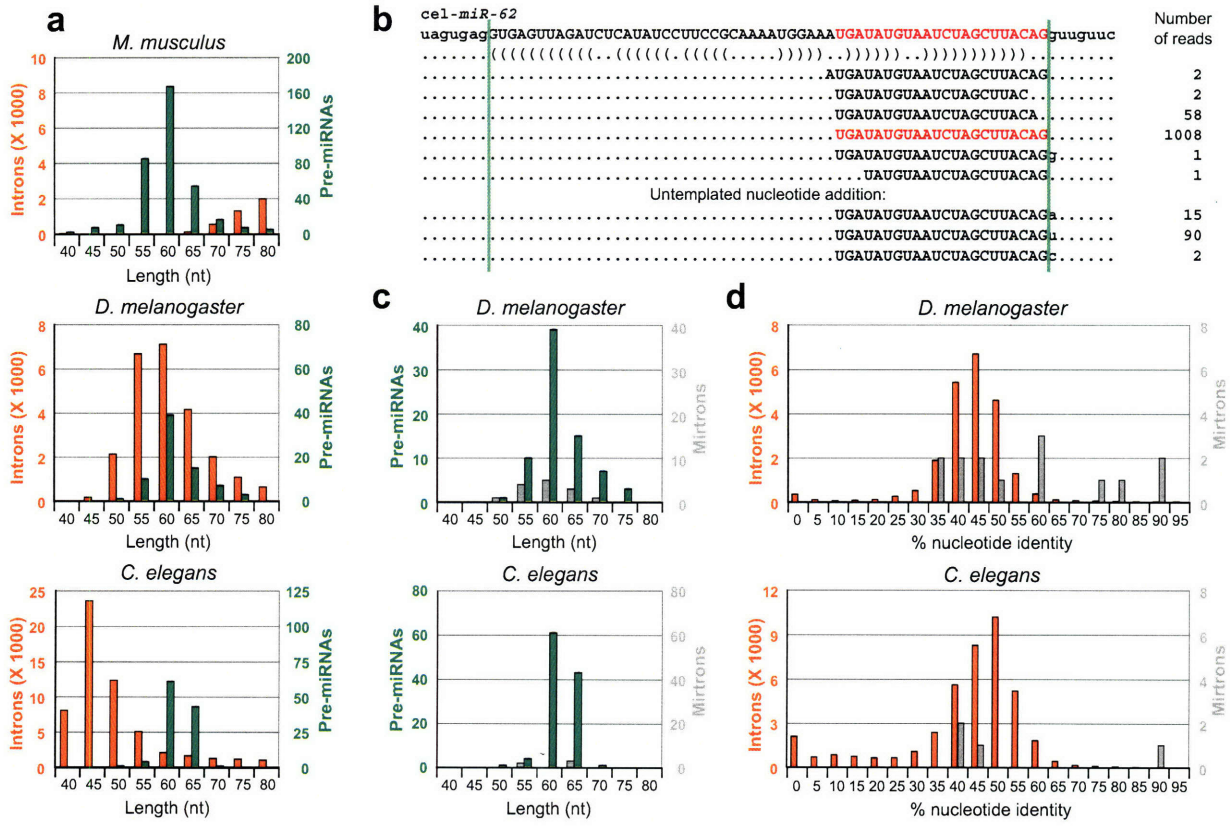

a
miR-1003
w.t.
3´ Mut
5´ Mut
b
w.t.
U1a
5´ Mut
U1a
5´ Mut
U1a-3G
1. Wild type
2. 3´ Mut
3. 5´ Mut + wt U1a
4. 5´ Mut + U1a-3G
U6
pre-miR-1003
miR-1003
c
Marker
GFP
Drosha
Dicer-1
Loquacious
Dicer-2
R2D2
Drosha + Dicer-1
DBR(1)
DBR(2)
78 nt
60 nt
24 nt
21 nt
U6
let-7 pre-miRNA
let-7 miRNA
pre-miR-1003 (probe1)
pre-miR-1003 (probe2)
miR-1003
d
pre-miR-1006 (probe1)
pre-miR-1006 (probe2)
miR-1006

Figure 3


a
M. musculus
Introns (X 1000)
Pre-miRNAs
Length (nt)
D. melanogaster
C. elegans
b
cel-miR-62
uagugagGUGAGUUAGAUCUCAUAUCCUUCCGCAAAAUGGAAAUGAUAUGUAAUCUAGCUUACAGguuguuc
Number of reads
AUGAUAUGUAAUCUAGCUUACAG 2
UGAUAUGUAAUCUAGCUUAC 2
UGAUAUGUAAUCUAGCUUACA 58
UGAUAUGUAAUCUAGCUUACAG 1008
UGAUAUGUAAUCUAGCUUACAGg 1
UAUGUAAUCUAGCUUACAG 1
Untemplated nucleotide addition:
UGAUAUGUAAUCUAGCUUACAGa 15
UGAUAUGUAAUCUAGCUUACAGu 90
UGAUAUGUAAUCUAGCUUACAGc 2
c
D. melanogaster
Pre-miRNAs
Mirtrons
Length (nt)
C. elegans
d
D. melanogaster
Introns (X 1000)
Mirtrons
% nucleotide identity
C. elegans

SUPPLEMENTAL INFORMATION

Intronic microRNA precursors that bypass Drosha processing

J. Graham Ruby*, Calvin H. Jan*, David P. Bartel

*these authors contributed equally to this work

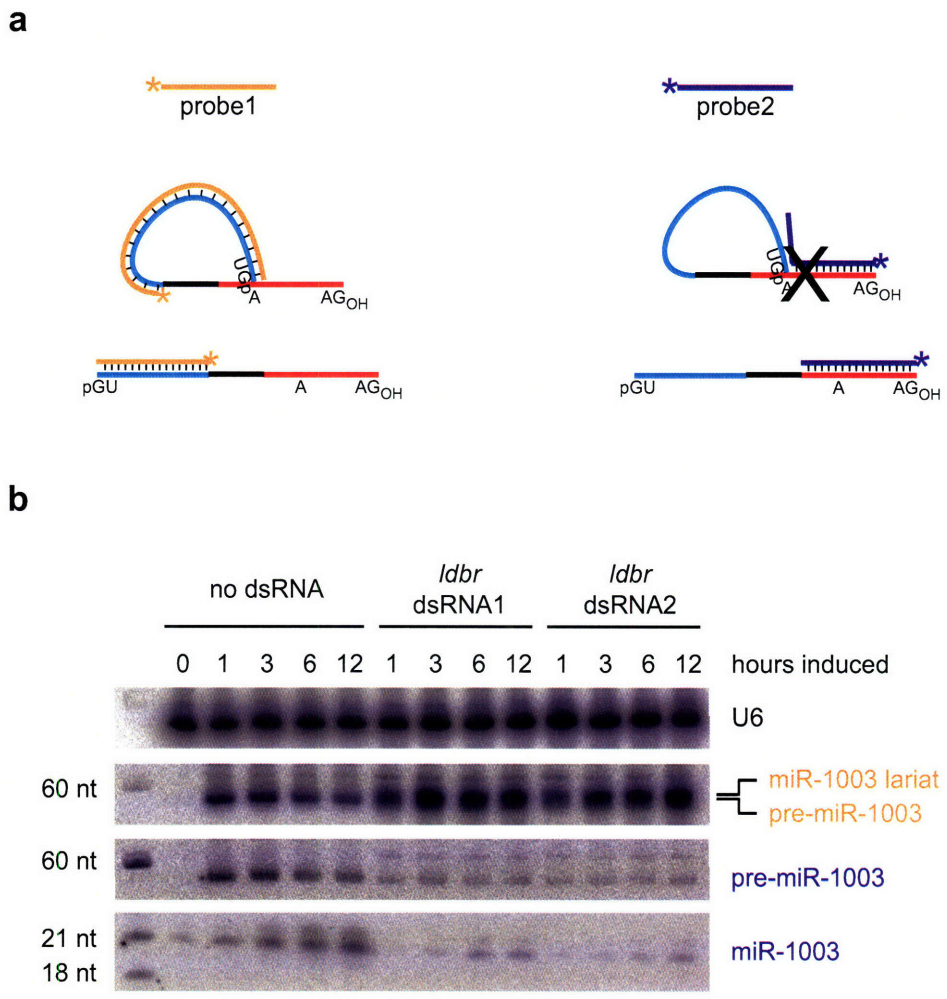


Figure S1. Mirtrons accumulate as lariats after splicing and require debranching enzyme (Ldbr) for conversion into functional pre-miRNAs. a, Left, hybridization of probe1 to miR-1003 intron lariat or linear pre-miR-1003. Right, stable hybridization of probe2 occurs only with linear pre-miR-1003, and is inhibited by the presence of the branch-point adenosine in the lariat. **b,** Northern blotting was used to analyze miR-1003 maturation in a time course after induction of mini-gene expression. Prior to induction, cells were soaked with either of two dsRNAs targeting *ldbr* (CG7942) or left untreated. RNA was resolved on a denaturing 15% acrylamide gel. Under these conditions, the lariat runs slightly above the pre-miRNA hairpin. In DBR dsRNA lanes, the major band detected by probe1 is absent when the blot is hybridized to probe2, indicating the presence of a lariat in these samples. When separated on a 17% gel, the lariat runs significantly higher (Fig. 2c). Changes in relative mobility in gels with different polyacrylamide densities are characteristic of non-linear RNA species.

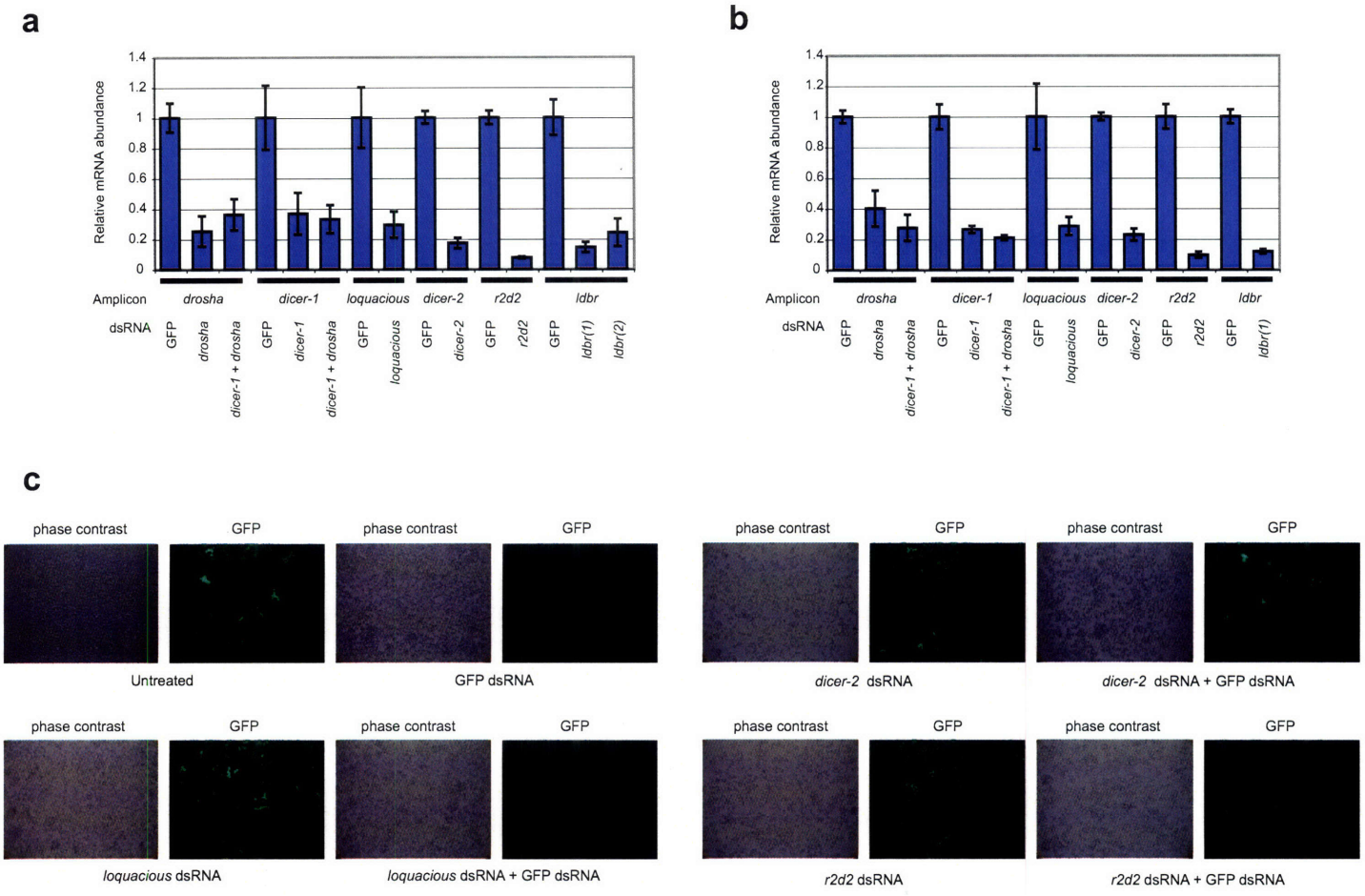


Figure S2. Confirmation of RNAi knockdowns. a, Quantitative RT-PCR analysis of samples from Fig. 2c. Relative abundance was measured using the ΔΔCt method, normalizing to *actin 5c* (ΔCt), and then to samples soaked in GFP dsRNA (ΔΔCt). Values are reported as geometric mean ± s.d. (n=3). **b,** Analysis as in (**a**), using samples from Fig. 2d. **c,** Functional analysis of *dicer-2* and *r2d2* knockdown by fluorescence microscopy. Cells stably expressing GFP were soaked in dsRNAs targeting *loquacious*, *dicer-2*, or *r2d2*. After 4 days, dsRNA targeting GFP was added. Depletion of Dicer-2 or R2D2 reduces the ability of GFP dsRNA to silence GFP. Depletion of Loquacious serves as a negative control. Functional efficacy of the other dsRNAs was assessed by northern blot analysis of miRNA or intron processing (Fig. 2).

Table S1

mir-1003

 # reads: 352
Most abundant read: UCUCACAUUUACAUAUUCACAG
Host gene: CG6695-RA, CG6695-RB
Intron coordinates: chr3R:20484326-20484382(+)

	# reads	# loci
GUGGGUAUCUGGAUGUGGUUGGCUCUGGCGGUCCUCUCACAUUUACAUAUUCACAG		
((((((((.(((((((((..((..(....)..))..)))))))))).))))))))..		
GUGGGUAUCUGGAUGUGGUUG.....	6	1
GUGGGUAUCUGGAUGUGGUUGG.....	9	1
GUGGGUAUCUGGAUGUGGUUGGC.....	1	1
.....UCUCACAUUUACAUAUUCA...	3	1
.....UCUCACAUUUACAUAUUCAC..	15	1
.....UCUCACAUUUACAUAUUCACA.	84	1
.....UCUCACAUUUACAUAUUCACAG	233	1
.....CUCACAUUUACAUAUUCACAG	1	1

```
D.melanogaster  GTGGGTATC-TGGATGTGGTTGGCTCT------GGCGGTCCTCTCACATTTACATATTCACAG
D.simulans      GTGGGTATC-TGGATGTGGTTGGCTCT------GGCGGTCCTCTCACATTTACATATTCACAG
D.yakuba        GTGGGTATC-TGGATGTGGTTGGCTCT------GGCGGTCCTCTCACATTTACATATTCACAG
D.ananassae     GTGAGTATAGTGGATGTGGGTGGCTCTTATTTGGCCGGTCCTCTCACATCTCCATATTCACAG
D.pseudoobscura GTGGGTATCGTGCTTGTGGGTGGCTCT----TGGTCAGTCCTCTCACATCTCTATATTCACAG
D.virilis       GTGAGTAAT-CAGTTGTGGGTGGCTTT----TTGAAAGCCCTCTCACATCTCTTTATTCACAG
D.mojavensis    GTGAGTAAT-CAGTTGTGGGTGGCTCT----AGTGAAGCCCTCTCACATCTCTTTATTCACAG
```

mir-1004

 # reads: 50
Most abundant read: UCUCACAUCACUUCCCUCACAG
Host gene: CG31772-RA
Intron coordinates: chr2L:3767620-3767688(+)

	# reads	# loci
GUUGGGGGACAUUGAUCUCGGAGACGGCGGUUUAACUGAUCCAUUCUCUCACAUCACUUCCCUCACAG		
..(((((((...((((.(.(((((.((((((.....)))).))..)).))))))).).)))))..)))))))....		
.....UCUCACAUCACUUCCCUCACA.	4	1
.....UCUCACAUCACUUCCCUCACAG	46	1

```
D.melanogaster  GT-TGGGGGACAT----------TGATCTCGGAG------ACGGCGGTTTAACTGATCCAT--TCTCTCACATC-ACT---TCCC---------TCACAG
D.simulans      GT-TGGGGGACAT----------TGATCTCGGAG------ACGGCGGTTTAACTGATCCAT--TCTCTCACATC-ACT---TCCC---------TCACAG
D.yakuba        GT-TGGGGGACAT----------TGATCTCGGAG------ACGGCGGTTTAACTGATCCAT--TCTCTCACATC-ACT---TCCC---------TCACAG
D.ananassae     GT-GAGATAACACGTTCTAACCAACTGATATGAAACTACGAATGTTTTTATCG------CTGTCGATTCGAACGA-------TCTCTCATATA-ACCGTACCTA---------TTACAG
D.pseudoobscura GTGTGTGGGATAC----------TGATTTTAGAGAAAAAAAAAACCATTAACGTGAGGCTTCGTTTCTCACATC-ATTTT-CCCC---------TCACAG
D.virilis       GT-TGGGGCACAT----------TGATCTTCAAGAAACTCACAGCCAGCTCATTTACTCAC--TCTCTCTCTCT---TCTCTCCC---------TCACAG
D.mojavensis    GT--------AAT----------TGATC--------ACTCcta------tctctgtctctc--tttctctttctgactctctcccactctttATTTGTAG
```

Table S1

mir-1005

 # reads: 8
Most abundant read: UCUGGAAUCUUUAAUUCGCAG
Host gene: CG2969-RA CG2969-RB
Intron coordinates: chr2L:4343695-4343756(+)

	# reads	# loci
GUGAGUUGAUCGAUUUCGAGGUUUUGGCACACGAAUAUAAUCUGGAAUCUUUAAUUCGCAG		
((((((((..((((((.(((((((((.....))).....)))))))))))).)))))))..		
.....UCUGGAAUCUUUAAUUCGCA.	1	1
.....UCUGGAAUCUUUAAUUCGCAG	7	1

```
D.melanogaster  GTGAGT-------TGATCGATTTCGAGGTTTTGGCA------------------------CACGA-------ATATAATCTGGAATCTTTAA-------TTCGCAG
D.simulans      GTGAGT-------TGATCGATTTCGAGTTTTTGGC------------------------CACAA-------ATATAATCTGGAATCTTTAA-------TTCGCAG
D.yakuba        GTGAGT-------TGATCGATTTCGAGGTTTTGGCA------------------------GCCAA-------AAATAATCTGGAATCTTTAA-------TTCGCAG
D.ananassae     GTAAGT-------ACATTGTGGATATTGTTTATAC-------------------------------------TACAGCCTCTAATCTTATACTATATTTTTGCAG
D.pseudoobscura GTAAGTG------TCCATATCCTCGAGGCTCctgcaatgcactgcaatgcactgcaatgaccgcaatgcTGACGTATTATTTATGTTCTCCGA-------TCCCCAG
D.virilis       GTAAGGGCTGA--ATTTTAAAATTG-AATTT-----------------------------AACAAGTATACAACAATATTATAATCC-------------CACACAG
D.mojavensis    GTAAGCATAGAGCAGATCAGATTTTATGATTT-----------------------------CACATATTCTCAATACTGCTTCGATCC-------------TCCACAG
```

mir-1006

 # reads: 57
Most abundant read: UAAAUUCGAUUUCUUAUUCAUAG
Host gene: CG17332-RA CG17332-RB CG17332-RD
Intron coordinates: chr2L:16720723-16720787(-)

	# reads	# loci
GUGAGUUUGAAAUUGAAAUGCGUAAAUUGUUUGGUACAAUUUAAAUUCGAUUUCUUAUUCAUAG		
((((((..(((((((((.....(((((((.....)))))))).)))))))))..)))))))..		
.....UAAAUUCGAUUUCUUAUUCA...	1	1
.....UAAAUUCGAUUUCUUAUUCAU..	12	1
.....UAAAUUCGAUUUCUUAUUCAUA.	8	1
.....UAAAUUCGAUUUCUUAUUCAUAG	35	1
.....AAAUUCGAUUUCUUAUUCAUAG	1	1

```
D.melanogaster  GTGAGTTTGAAATTGAAATGCGTAAATTGTTTGGTACAATTTAAATTCGATTTCTTATTCATAG
D.simulans      GTGAGTTTGAAATTGAAATGCGTAAATTGTTTGGTACAATTTAAATTCGATTTCTTATTCATAG
D.yakuba        GTGAGTTTGAAATTGAAATGCGTAAATTGTTTGGTACAATTTAAATTCGATTTCTTATTCATAG
D.ananassae     GTGAGTTTGAAATTGAAATGCGTAAATTGTTTGGTACAATTTAAATTCGATTTCTTATTCATAG
D.pseudoobscura GTGAGTTTGAAATTGAAATGTGTAAATTGTTTGTGCCAATTTAAATTCGATTTCTTATTCATAG
D.virilis       GTGAGTTTGAAATTGAAATATGTAAATTGTTTGGTACAATTTAAATTCGATTTCTTATTCATAG
D.mojavensis    GTGAGTTTGAAATTGAAATGTGTAAATTGTTTGGTACAATTTAAATTCGATTTCTTATTCATAG
```

Table S1

mir-1007

 # reads: 9
Most abundant read: UAAGCUCAAUUAACUGUUUGCA
Host gene: CG1718-RA
Intron coordinates: chrX:21107060-21107125(-)

```
                                                                                  # reads        # loci
GUAAGCAGUGUUUGAACUCGAUCUUGGUUCUUGGACUCUUGAUAAGCUCAAUUAACUGUUUGCAG
(((((((((..((((.((..(((..((((....))).)..))).)).))))...)))))))))..
.........................................UAAGCUCAAUUAACUGUUUGC..              2            1
.........................................UAAGCUCAAUUAACUGUUUGCA.              6            1
.........................................UAAGCUCAAUUAACUGUUUGCAG              1            1

D.melanogaster  GTAAGCAGTGTTTGAACTCGATC--TTGGTTC----TTG---GACTCT-------TGATAAGCTCAATTAACTGTTTGCAG
D.simulans      GTAAGCAGTGTTTGAACTCGATC--TTGGTTC----TTG---GACTCT-------TGATAAGCTCAATTAACTGTTTGCAG
D.yakuba        GTAAGCAGTGTTTGAACTCGATC--TAGGATC----TTG---GACTCT-------TGATAAGCTCAATTAACTGTTTGCAG
D.ananassae     GTAAGCAGTGTTTGAACTCGATC--TTGGAAT----------AGCTCC-------CGATAAGCTCAATTAACTGTTTGCAG
D.pseudoobscura GTAAGCAGCGATTGA--TCAATCaattgaatc----------gaatcgaatcgaatGATAAACTCCATTAACTGTTTGCAG
D.virilis       GTAAGCAGTGCTTGAGCTTATTC--TCTGGCTTCATTTGACCATTTTC-------TGATAAGCTCAATTAACTGTTTGCAG
D.mojavensis    GTAAGCAGTGTTTGAACTAAATC--TCTGGCT--ACTTGGCCGTATAT-------TGATAAGCTCAACTAACTGTTTGCAG
```

mir-1008

 # reads: 46
Most abundant read: UCACAGCUUUUUGUGUUUACA
Host gene: CG18004-RA CG18004-RB
Intron coordinates: chr2R:6401439-6401496(+)

```
                                                                                  # reads        # loci
GUAAAUAUCUAAAGUUGAACUUGGCCAAUGGCAAGUCACAGCUUUUUGUGUUUACAG
(((((((((..(((((((.(((((.(....).))))))..)))))))..)))))))..
GUAAAUAUCUAAAGUUGAACU.....................................              1            1
.....................................UCACAGCUUUUUGUGU......              1            1
.....................................UCACAGCUUUUUGUGUUU....              1            1
.....................................UCACAGCUUUUUGUGUUUAC..              6            1
.....................................UCACAGCUUUUUGUGUUUACA.             22            1
.....................................UCACAGCUUUUUGUGUUUACAG             14            1
.....................................CAGCUUUUUGUGUUUACAG                1            1

D.melanogaster  GTAAATAT----CTAAAGTTGAAC---TTGGCCAATGGCAAGTCACA----GCTTTTTGTGTTTACAG
D.simulans      GTAAATAT----CTAAAGTTGAAC---TTGGCCAACGGCAAGTCACA----GCTTTTTGTGTTTACAG
D.yakuba        GTAAATAT----CTAAAGTTGAAC---TTGGCCAACGGCAAGTCACA----GCTTTTTGTGTTTACAG
D.ananassae     GTAAGGAA----CTCAATTTTTAC--ATTAAACCGAAGCAATTTAAC----ACGTTTCTTATTT-CAG
D.pseudoobscura GTAAGGGATCGGCGAGAGTTTTTCCCACGGAATATCATTATATTATA----TTGTTATGTGCCTGCAG
D.virilis       GTAAGTGA----TGAT--GCGTCC---ATTGGGAATATCATTTAATT----------TGTGTTGGTAG
D.mojavensis    GTAAGTAG----TAATAGGTGTTT---GTAGACATATTCAGTTAATTTTCGCATTTGTGTATTGGCAG
```

Table S1

mir-1009

 # reads: 14
Most abundant read: UCUCAAAAAUUGUUACAUUUCAG
Host gene: CG3860-RA
Intron coordinates: chr2R:19500653-19500714(-)

	# reads	# loci
GUAAGUGUAAGACUUUCUUGAGUUACCCGCGAUGAGUAUCUCAAAAAUUGUUACAUUUCAG		
(.(((((((((.(.(((.(((((.(((.(.....).))).)))))))).).))))))))..		
.UAAGUGUAAGACUUUCUUGAGU.....	1	1
.....UCUCAAAAAUUGUUACAUUUCA.	3	1
.....UCUCAAAAAUUGUUACAUUUCAG	10	1

```
D.melanogaster  GTAAGTGTAAGACTTTCT------TGAGTT----ACCCGCGATGAGTATCTCAAAAATTGT--TACATTTCAG
D.simulans      GTAAGTGTAAGACTTTCT------TGAGTT----ACCCGCGATGAGTATCTCAAAAATTGT--TACATTTCAG
D.yakuba        GTAAGTGTAAGACTTTCT------TGAGTT----ACCCGCGAGGAGTATCTCAAAAATTAT--TACATTTCAG
D.ananassae     GTAAGTTTTGAATACTTC------TACTCT----ATCTTGGA-ATGTCTCTCATAATTGTC--CATCTTCTAG
D.pseudoobscura GTAAGTTCGAGAATCCCAAACACATAAGTT----CTTTTTTA-----ACATCCAAA-------TATTTTGTAG
D.mojavensis    GTAAGGACCAA---TTGG------TAAAGTGAGAAATTGGAAGAAATTTATTGAAATTTATACTCGTGTTTAG
```

mir-1010

 # reads: 193
Most abundant read: UUUCACCUAUCGUUCCAUUUGCAG
Host gene: CG31163-RA CG31163-RB CG31163-RC
Intron coordinates: chr3R:18118600-18118671(+)

	# reads	# loci
GUAAGUGGUGUAGAUGAAACAAAUUUACCAACAAUUUUGUUGGAUUGUUUCACCUAUCGUUCCAUUUGCAG		
(((((((.((((.(((((((((.....((((((....))))))..)))))))).)))).....))))))))..		
GUAAGUGGUGUAGAUGA.....	1	1
GUAAGUGGUGUAGAUGAAA.....	2	1
GUAAGUGGUGUAGAUGAAAC.....	3	1
GUAAGUGGUGUAGAUGAAACA.....	30	1
GUAAGUGGUGUAGAUGAAACAA.....	1	1
.....UUUCACCUAUCGUUCCAUUUG...	10	1
.....UUUCACCUAUCGUUCCAUUUGC..	39	1
.....UUUCACCUAUCGUUCCAUUUGCA.	38	1
.....UUUCACCUAUCGUUCCAUUUGCAG	64	1
.....UUCACCUAUCGUUCCAUUUGC..	3	1
.....UUCACCUAUCGUUCCAUUUGCA.	1	1
.....UUCACCUAUCGUUCCAUUUGCAG	1	1

```
D.melanogaster  GTAAGTGGTGTAGATGAAACAAATTTACCAAC-AAT----TTTGTTGGATTGTTTCACCTATCGTTCCATTTGCAG
D.simulans      GTAAGTGGTGTAGATGAAACAAATTTACCAAC-AATA-TTTTTGTTGGATTGTTTCACCTATCGTTCCATTTGCAG
D.yakuba        GTAAGTGGTGTAGATGAAACAAATTTACCAAC-AATATTTTTTGTTGGATTGTTTCACCTATCGTTCCATTTGCAG
D.ananassae     GTAAGTGGTGTAGATGAAACAAATTTACCAAC-AATA-TTTTTGTTGGATTGTTTCACCTATCGTTCCATTTGCAG
D.pseudoobscura GTAAGTGGTGTAGATGAAACAAATTAACCAAC-CTTT-ATTGTGTTGGATTGTTTCACCTATCGTTCCATTTGCAG
D.virilis       GTAAGTGGTGTAGATGAAACAAATTTCACAACAAAT--TTTTTGTTGGATTGTTTCACCTATCGTTCCATTTGCAG
D.mojavensis    GTAAGTGGTGTAGATGAAACAAATTTCACAAC-AAT--TTTTTGTTGGATTGTTTCACCTATCGTTCCATTTGCAG
```

Table S1

mir-1011

 # reads: 2
Most abundant read: UUAUUGGUUCAAAUCGCUCGCAG
Host gene: CG17274-RA CG17274-RB
Intron coordinates: chr3R:16679026-16679080(-)

	# reads	# loci
GUGAGUUUUUGAGCCAGGAAUAUAGUUCUUAUUAUUGGUUCAAAUCGCUCGCAG		
((((((.((((((((((.((((.......)))).)))))))))..))))))..		
.............................UUAUUGGUUCAAAUCGCUCGCAG	2	1

```
D.melanogaster  GTGAGTTTTTGAGCCAGG-----AATATAGTT--------CTTAT-----TAT-TGGTTCAAATCGCTCGCAG
D.simulans      GTGAGTTTTTGAGCCAGG-----AATATAGTT--------CTTAT-----TAT-TGGTTCAAATCGCTCGCAG
D.yakuba        GTGAGTTTTTGAGCCAGG-----AATATAATT--------CTTAT-----TAT-TGGTTCAAATCGCTCGCAG
D.ananassae     GTGAGTCTTTGAACCAGG-----AATATAATT--------TGTAT----ATAT-TGGTTCAAATCGCTCGTAG
D.pseudoobscura GTGAGATTTTGAATCTAATATATAATATAATC--------CGTACGTGTATATATGGTTCAAATTACTCGTAG
D.virilis       GTGAGTCATTGAACCAGG-----AATATATGTATGTAATTCTTAT----ATAT-TGGTTCAAATTTCTCGCAG
D.mojavensis    GTGAGTCTTTGAGCCAGG-----AATATATGTTCAT----CTTAT-----TAT-TGGTTCAAATCTCTCGTAG
```

mir-1012

 # reads: 101
Most abundant read: UUAGUCAAAGAUUUUCCCCAUAG
Host gene: CG31072-RA CG31072-RB
Intron coordinates: chr3R:22687070-22687129(-)

	# reads	# loci
GUGGGUAGAACUUUGAUUAAUAUUGCUUGAAAAAUAUUAGUCAAAGAUUUUCCCCAUAG		
(((((.((((((((((((((((((.....)))))))))))))))).)))).))))..		
GUGGGUAGAACUUUGAUUA....................................	1	1
GUGGGUAGAACUUUGAUUAA...................................	5	1
GUGGGUAGAACUUUGAUUAAU..................................	20	1
GUGGGUAGAACUUUGAUUAAUA.................................	16	1
GUGGGUAGAACUUUGAUUAAUAU................................	1	1
..................................UUAGUCAAAGAUUUUCCCCAUA.	2	1
..................................UUAGUCAAAGAUUUUCCCCAUAG	56	1

```
D.melanogaster  GTGGGTAGAACTTTGATTAAT----------------------ATTGCTTGAAAAAT-------ATTAGTCAA----AGATTTT-C------CCCATAG
D.simulans      GTGGGTAGAACTTTGATTAAT----------------------ATTGCTTGAGAA-T-------ATTAGTCAA----AGATTTT-C------CCCATAG
D.yakuba        GTGGGTAGAACTTTGATTAAT----------------------ATTGCTTGCAAGAT-------ATTAGTCAA----AGGTTTTTC------CCCATAG
D.ananassae     GTAGGT-----TTCAACCAAA----------------------TTTCCTTTGAGAGT--------TCAGTTAACTTTATATATTT-C------TTTTTAG
D.pseudoobscura GTGGGTAGT-CTCTCATATAT----------------------AGTTATAAAAGAACGAACACCAGTGGTTAA-GCAATGCATT-T------CTTGTAG
D.virilis       GT---------ACGGATTGTTTATTTA----------------AATGCTTTATATAT--------TTATCTAT----AAGCTAT-CTTTTTGTTTGCAG
D.mojavensis    GTGTGTAAA-TATGGATTATT-ATTTATAAATTATCGAAAACTTAACTTCTAATGTTT--------TTATATTT----ATATTTT-CAACATACTCTCAG
```

Table S1

mir-1013

 # reads: 17
Most abundant read: AUAAAAGUAUGCCGAACUCG
Host gene: CG12072-RA
Intron coordinates: chr3R:26617357-26617418(-)

	# reads	# loci
GUGAGUUUCGUACACUUAAUUAAUAGGAUCGGCCGUUAAUAAAAGUAUGCCGAACUCGCAG		
((((((((.(((.((((.((((((.((.....)))))))).)))).)))).))))))))..		
.....UAAUAGGAUCGGCCGUUAAU.....	2	1
.....AUAAAAGUAUGCCGAACUCG...	4	1
.....AUAAAAGUAUGCCGAACUCGC..	4	1
.....AUAAAAGUAUGCCGAACUCGCA.	2	1
.....AUAAAAGUAUGCCGAACUCGCAG	4	1
.....UAAAAGUAUGCCGAACUCGCAG	1	1

```
D.melanogaster  GTGAGTT------TCGTACACTTAATTAATAGGATCGGCCGTTAATAAAAGTATGCC---GAACTCGCAG
D.simulans      GTGAGTT------TCGTACACTTAATTAATAGGATCGGCCGTTAATAAAAGTATGCC---GAACTCGCAG
D.yakuba        GTGAGTT------TCGTACACTTAATTAATTGGGACGGCCGTTAATAAAAGTATGCC---GAACTCGCAG
D.ananassae     GTAATCT------TTGAATAATTATCTGTGAGTTGTGGCATCTAATGATTGT-------TATCTTCCAG
D.pseudoobscura GTAAGTCCATGAATTGCATCCCCCTTTGAT-----TATTCTTTAATCTGGAAATCCCTGTGATCCCATAG
```

mir-1014

 # reads: 3
Most abundant read: AAAAUUCAUUUUCAUUUGCAG
Host gene: CG2196-RA
Intron coordinates: chr3R:27579245-27579313(-)

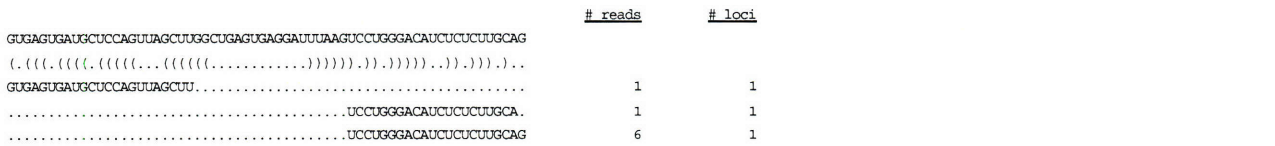
	# reads	# loci
GUAUAAUGGAAAUAGAUUUUAAUCGCAGGCGCGUCAGUGGUUGAAUUAAAAUUCAUUUUCAUUUGCAG		
(((.(((((((((..(((((((((..(((.(((....))).))).)))))))))..))))))))))))..		
.....UAAAAUUCAUUUUCAUUUGCAG	1	1
.....AAAAUUCAUUUUCAUUUGCAG	2	1

```
D.melanogaster  GTATAATGGAAATAGATTTTAATCGCAGGCGCGTCAGTGGTTGAATTAAAATTCATTTTCATTTGCAG
D.simulans      GTATAATGGAAATAGATTTTAATCGCTGGCGCGTCAGTGGTTGAATTAAAATTCATTTTCATTTGCAG
D.yakuba        GTATAATGGAAATAGATTTTAATCGCAGGCGCGTCAGTGGTTGAATTAAAATTCATTTTCATTTGCAG
D.ananassae     GTATAATGAAAATTGATTTTAATCACACGGATCGGAGTGGCAAAATTAAAATTCATTTTCATTTGCAG
D.pseudoobscura GTACAATGGAAATAGATTTTAATCGGGTTTCGTTTGGCGGTGAAATTAAAATTCATTTTCATTTACAG
```

Table S1

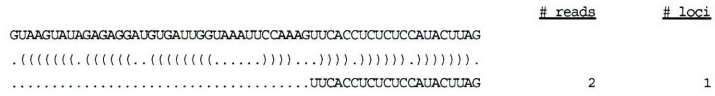
mir-1015

 # reads: 8
Most abundant read: UCCUGGGACAUCUCUCUUGCAG
Host gene: CG6432-RA
Intron coordinates: chr3R:20164953-20165017(+)



mir-1016

 # reads: 2
Most abundant read: UUCACCUCUCUCCAUACUUAG
Host gene: CG8479-RA CG8479-RB
Intron coordinates: chr2R:9747992-9748050(-)



mir-1017

 # reads: 148
Most abundant read: GAAAGCUCUACCCAAACUCAUCC
Host gene: CG6844-RA CG6844-RB
Intron coordinates: chr3R:20314333-20314502(+)

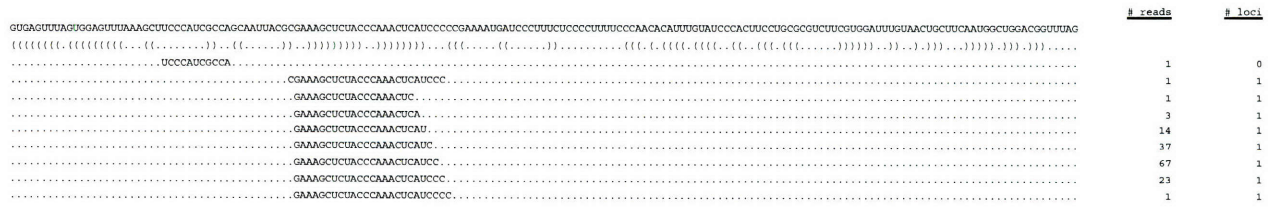


Table S2

mir-62

 # reads: 1071
Most abundant read: UGAUAUGUAAUCUAGCUUACAG
Host gene: T07C5.1b T07C5.1c
Intron coordinates: chrX:12692524-12692582(+)

	# reads	# loci
GUGAGUUAGAUCUCAUAUCCUUCCGCAAAAUGGAAAUGAUAUGUAAUCUAGCUUACAG		
((((((((((..((((((.((((((.....)))))).))))))).)))))))))))..		
.....AUGAUAUGUAAUCUAGCUUACAG	2	1
.....UGAUAUGUAAUCUAGCUUAC..	2	1
.....UGAUAUGUAAUCUAGCUUACA.	58	1
.....UGAUAUGUAAUCUAGCUUACAG	1008	1
.....UAUGUAAUCUAGCUUACAG	1	1

C.elegans GTGAGTTAGATCTCATATCCTTCCGCAAAATGGAAATGATATGTAATCTAGCTTACAG
C.briggsae GTGGGTTAGATCCCATATCCTTCCGCTTGATGGAAATGATATGTAATCTAGCTTACAG

mir-1018

 # reads: 2
Most abundant read: AGAGAGAUCAUUGGACUUACAG
Host gene: Y59E1B.1
Intron coordinates: chrX:1879451-1879507(+)

	# reads	# loci
GUAAGUUCAUGAUUUCUCCCAUAUAUUUUUCAUGAGAGAGAUCAUUGGACUUACAG		
((((((((((((((((((.(((.........)))))))))))))))))))))..		
.....AGAGAGAUCAUUGGACUUACAG	2	1

mir-1019

 # reads: 2
Most abundant read: GUGAGCAUUGUUCGAGUUUCAUUU
Host gene: M04C9.5
Intron coordinates: chrI:9369650-9369719(+)

	# reads	# loci
GUGAGCAUUGUUCGAGUUUCAUUUUUAAUAAAAUUUAUUUAAAAACUGUAAUUCCACAUUGCUUUCCAG		
..((((((.(((..((((((.(((((((((.....)))))))).))).))))))..))).))))).....		
GUGAGCAUUGUUCGAGUUUCAUUU.....	1	1
.....CUGUAAUUCCACAUUGCUUUCCAG	1	1

C.elegans GTGAGCATTGTTCG-AGTTTCatttttaataaaatttatttaaaaaCTGTAATTCCACATTGCTTTCCAG
C.briggsae GTTTGCATTCTTTGAAGTGTCACAGTT----------CTATAACGTCGCA----GTTCCAG

mir-1020

 # reads: 2
Most abundant read: GUAAGUGUUACAGAAUAAUCU
Host gene: T16G12.1
Intron coordinates: chrIII:10047630-10047700(-)

	# reads	# loci
GUAAGUGUUACAGAAUAAUCUUAGACAAAACAACUAAAAUUAAUGAAAAAUUAUUCUGUGACACUUUCAG		
(.((((((((((((((((.....)))))))))))))))))).)..		
GUAAGUGUUACAGAAUAAUCU.....	1	1
.....AUUAUUCUGUGACACUUUCAG	1	1

C.elegans GTAAGTGTTACAGAATAATCTTAGACAAAACAACTAAAATTAATGAAAAATTATTCTGTGACACTTTCAG
C.briggsae GTGAGGAACACATACAAATGTTGGAT----------TTTATTC----GAATTTTCAG

Table S3. Quantification of signals from RNA blots of Figure 2c and 2d. Signals were first normalized to that of the loading control (U6), then to that of the control dsRNA (GFP). When signal was below detection (b.d.), the upper bound of the value, based on the normalized detection limit, is shown for relevant lanes.

Fig. 2c Quantification

	dsRNA								
	GFP	*drosha*	*dicer-1*	*loquacious*	*dicer-2*	*r2d2*	*drosha* + *dicer-1*	*ldbr*(1)	*ldbr*(2)
pre-*let-7* miRNA	1.0	0.03	3.54	0.90	0.51	0.60	0.14	0.48	0.91
let-7 miRNA	1.0	0.45	1.36	1.58	1.37	2.02	0.23	1.90	3.80
pre-miR-1003 probe1	1.0	0.12	0.57	0.65	0.32	0.29	0.35	0.06	0.08
pre-miR-1003 lariat	b.d.	b.d.	b.d.	b.d.	b.d.	b.d.	b.d.	0.36	0.57
pre-miR-1003 probe2	1.0	0.10	0.51	0.68	0.36	0.32	0.31	0.03	0.03
miR-1003	1.0	0.92	0.08	0.09	0.81	0.31	0.10	b.d. (<.04)	b.d. (<.04)

Fig. 2d Quantification

	dsRNA							
	GFP	*drosha*	*dicer-1*	*loquacious*	*dicer-2*	*r2d2*	*drosha* + *dicer-1*	*ldbr*(1)
pre-*let-7* miRNA	1.0	b.d. (<.05)	4.56	2.19	1.15	1.51	0.15	1.05
let-7 miRNA	1.0	0.17	0.85	1.61	1.21	0.41	0.21	0.91
pre-miR-1006 probe1	1.0	0.36	1.37	1.33	0.92	0.73	1.15	0.46
pre-miR-1006 lariat	b.d.	b.d.	b.d.	b.d.	b.d.	b.d.	b.d.	0.18
pre-miR-1006 probe2	1.0	0.34	1.37	1.41	1.10	0.86	1.28	0.53
miR-1006	1.0	0.73	0.14	0.15	0.56	0.31	0.37	0.37

>pCJ19 (pMT-puro_miR-1006)
actagtAACAACGAATACATCCAGTCGGTGGCCCGGTGCCTGCAGATGATGCTCCGTGTGGATGAGTATCGATTTGCCTTTGTGGGAGTCGACGGAATCAGCACTCTGATCCGTATCTTGTCGAC
CCGTGTCAACTTCCAGGTGAGTTTGAAATTGAAATGCGTAAATTGTTTGGTACAATTTAAATTCGATTTCTTATTCATAGGTGCAATACCAGTTGATCTTTTGCCTGTGGGTGCTGACCTTCAAT
CCCCTGCTGGCCGCCAAGATGAATAAGTTCAGCGTGATCCCCATCCTGGCTGATATCCTCAGCGATTGTGCCAAGGAGAAGGTGACACGCATTATCCTCGCGGTCTTCCGCAATCTGATCGAGAA
GCCGGAGGATTCATCGGTGGCCAAGGACCATTGCATCGCCATGGTCCAGTGCAAGGTGCTGAAGCAGCTATCCATCCTGGAGCAGCGTCGCTTCGACGACGAGGACATTACCGCCGACGTAGAGT
ACCTGAGCGAGAAGCTCCAGAATTCGGTGCAAGACTTGAGCTCCTTTGATGAGTACGCCACAGAGGTGCGCAGCGGTCGCTTGGAATGGTCGCCTGTGCACAAGTCGGCCAAGTTCTGGCGCGAG
AATGCCCAGCGCCTAAACGAAAAGAACTACGAGTTGCTGCGCATCCTCGTCCACCTCCTGGAAACCTCAAAAGATGCCATCATCCTTTCCGTCGCCTGCTTCGACATCGGGGAGTATGTGCGCCA
CTATCCCCGCGGCAAGCAgcggccgc

>pCJ20 (pMT-puro_miR-1003)
actagtATAAAGCCGATAAGCGTGCGGAAATCGAAAAGACCGAGCGTGAACGACTGCAACAGCAGGAACGCGAGGATGAGATGCGCGAACTGGCCCTCAAGCTGCGCAGAAGGTGGGTATCTGGA
TGTGGTTGGCTCTGGCGGTCCTCTCACATTTACATATTCACAGGCGCCGTGAGCTGCGTCACAAATATGGAACGCCCTCGAGTGGAAAGCTCTCGGACAGCGATGCCGAATCTGTGGCATCGGAA
AGCCAGCTGGCAAAGTCGTCGACGCGTCGCAGTCGCAGCAGATCGGAGAGTCGTAGGCGGTCGCCACCGCCGGAAACACAGCACAGCAGACGGAGCAGCAGCCGAACGGAAAGGAGGAGACGCAC
AAGGACGgcggccgc

>pCJ24 (pMT-puro_let-7)
actagtGACAAATGGCGGAGTAAGGATAAAAAGTCGAGTCAAAAACGGATTAATACGAAATAAATAACTACTAAAATAACTAAATAGAAGATCAACAGCGATCCATTAAACAAACTATAACAATA
ATATGTAATATGAAAACCATCGATAATAAATCGTAAAACTAAATTAAACTAAATATCCAACGTCATATAAATACAACTCAAAACGTCATAATAAGAAAACTGATATGGTATAACAAACTCAAGTT
TAAGTTTGAATAACACAACAAAGTAATTTACTTAAATACTCATCTTATGTATATTTTATTTAACATTCAAATGTAACCTTCAAAAGCATTTTTAATATGATTTCTCCGTATTATTTTTCTTCTTG
TTTGCCATCATCGTTTCAACCAAAAACCGAACCAATGATATCCAGAAGATCCTTTAAATACCAAACCACCTAGCAAAAAGGACTACACCAAGGACCTTTTTCTCTCTGGCAAATTGAGGTAGTAG
GTTGTATAGTAGTAATTACACATCATACTATACAATGTGCTAGCTTTCTTTGCTTGACTACAAGCCGCATTTGATAAAAGAATCCCAATCGAACTGCACCACTTAATAAACCCAATCCCCAGCCA
TACAAAAGTTGGTGGTGCAACAAATTTTGATTGTGGACAACAAAGAAAGTGTTCTGAGCCAAACTATTGTTAAATATCATTACGAATGCCAAAGTATGTAAATGCAACCGGGCATATGTAAATAT
TGGCATTGGTGACATGTGCAAATGTTTGTATGGCTGATTCCCTGAGACCCTAACTTGTGACTTTTAATACCAGTTTCACAAGTTTTGATCTCCGGTATTGGACGCAAACTTGCTGgcggccgc

>pCJ30 (p2032_miR-1003)
ggtaccATAAAGCCGATAAGCGTGCGGAAATCGAAAAGACCGAGCGTGAACGACTGCAACAGCAGGAACGCGAGGATGAGATGCGCGAACTGGCCCTCAAGCTGCGCAGAAGGTGGGTATCTGGA
TGTGGTTGGCTCTGGCGGTCCTCTCACATTTACATATTCACAGGCGCCGTGAGCTGCGTCACAAATATGGAACGCCCTCGAGTGGAAAGCTCTCGGACAGCGATGCCGAATCTGTGGCATCGGAA
AGCCAGCTGGCAAAGTCGTCGACGCGTCGCAGTCGCAGCAGATCGGAGAGTCGTAGGCGGTCGCCACCGCCGGAAACACAGCACAGCAGACGGAGCAGCAGCCGAACGGAAAGGAGGAGACGCAC
AAGGACGgcggccgc

>pCJ31 (p2032_miR-1006)
ggtaccAACAACGAATACATCCAGTCGGTGGCCCGGTGCCTGCAGATGATGCTCCGTGTGGATGAGTATCGATTTGCCTTTGTGGGAGTCGACGGAATCAGCACTCTGATCCGTATCTTGTCGAC
CCGTGTCAACTTCCAGGTGAGTTTGAAATTGAAATGCGTAAATTGTTTGGTACAATTTAAATTCGATTTCTTATTCATAGGTGCAATACCAGTTGATCTTTTGCCTGTGGGTGCTGACCTTCAAT
CCCCTGCTGGCCGCCAAGATGAATAAGTTCAGCGTGATCCCCATCCTGGCTGATATCCTCAGCGATTGTGCCAAGGAGAAGGTGACACGCATTATCCTCGCGGTCTTCCGCAATCTGATCGAGAA
GCCGGAGGATTCATCGGTGGCCAAGGACCATTGCATCGCCATGGTCCAGTGCAAGGTGCTGAAGCAGCTATCCATCCTGGAGCAGCGTCGCTTCGACGACGAGGACATTACCGCCGACGTAGAGT
ACCTGAGCGAGAAGCTCCAGAATTCGGTGCAAGACTTGAGCTCCTTTGATGAGTACGCCACAGAGGTGCGCAGCGGTCGCTTGGAATGGTCGCCTGTGCACAAGTCGGCCAAGTTCTGGCGCGAG
AATGCCCAGCGCCTAAACGAAAAGAACTACGAGTTGCTGCGCATCCTCGTCCACCTCCTGGAAACCTCAAAAGATGCCATCATCCTTTCCGTCGCCTGCTTCGACATCGGGGAGTATGTGCGCCA
CTATCCCCGCGGCAAGCAgcggccgc

>pCJ32 (p2032_Let-7)
ggtaccGACAAATGGCGGAGTAAGGATAAAAAGTCGAGTCAAAAACGGATTAATACGAAATAAATAACTACTAAAATAACTAAATAGAAGATCAACAGCGATCCATTAAACAAACTATAACAATA
ATATGTAATATGAAAACCATCGATAATAAATCGTAAAACTAAATTAAACTAAATATCCAACGTCATATAAATACAACTCAAAACGTCATAATAAGAAAACTGATATGGTATAACAAACTCAAGTT
TAAGTTTGAATAACACAACAAAGTAATTTACTTAAATACTCATCTTATGTATATTTTATTTAACATTCAAATGTAACCTTCAAAAGCATTTTTAATATGATTTCTCCGTATTATTTTTCTTCTTG
TTTGCCATCATCGTTTCAACCAAAAACCGAACCAATGATATCCAGAAGATCCTTTAAATACCAAACCACCTAGCAAAAAGGACTACACCAAGGACCTTTTTCTCTCTGGCAAATTGAGGTAGTAG
GTTGTATAGTAGTAATTACACATCATACTATACAATGTGCTAGCTTTCTTTGCTTGACTACAAGCCGCATTTGATAAAAGAATCCCAATCGAACTGCACCACTTAATAAACCCAATCCCCAGCCA
TACAAAAGTTGGTGGTGCAACAAATTTTGATTGTGGACAACAAAGAAAGTGTTCTGAGCCAAACTATTGTTAAATATCATTACGAATGCCAAAGTATGTAAATGCAACCGGGCATATGTAAATAT
TGGCATTGGTGACATGTGCAAATGTTTGTATGGCTGATTCCCTGAGACCCTAACTTGTGACTTTTAATACCAGTTTCACAAGTTTTGATCTCCGGTATTGGACGCAAACTTGCTGgcggccgc

>Drosha dsRNA
GGTCACCAAGCCGGGTTATAAGCCGTGTTCGCTACGCGTTGACCAATTGGACAGGAATAACTCCGATTTGCCCGAGTGCGTAGATCGCGAGACTGGAATCTCACATCCAGCAATCGTGCACTTTG
GCATTTGTCATCCTCAGCTAAGCTACGCTGGAAATCCAGAGTACCAGAAGGCGTGGCGAGAGTACGTTAAGTACCGTCATCTGATGGCCAACATGTCGAAGCCCTCTTTCAAGGATAAGCGCAAG
CTAGAGGAGAAGGAGCAACGTCTTCAGGAGATGCGAACTCAGGGGCGCATGAAACGAAATATCACAGTGGCGATCAGCTCGGAGGGCTTCTATCGCACCGGCATTATGTGCGACGTTGTGCAGCA
TGCCATGTTGATTCCTGTCCTAACTGGTCACCTTCGCTTTCACAAGTCGCTGGACCTGCTAGAGGAGAGTATCGGGTACCGCTTTAAAAATCGGTACCTTCTCCAATTGGCGCTGACGCATCCCT
CATACAAGGAGAACTACGGTACCAATCCGGATCACGCCCGTAATTCGCTGACTAACTGCGGAATTCGTCAGCCGGAGTACGGAGATCGCAAGATCCATTACATGAACACACGCAAGCGGGGTATC
AACACATTAGTGAGCATTATGT

>Dicer-1 dsRNA
CTACTGGCCACCGCCTACGAGCGGAACACGATTATTTGCCTGGGCCATCGAAGTTCCAAGGAGTTTATAGCCCTCAAGCTGCTCCAGGAGCTGTCGCGTCGAGCACGCCGACATGGTCGTGTCAG
TGTCTATCTCAGTTGCGAGGTTGGCACCAGCACGGAACCATGCTCCATCTACACGATGCTCACCCACTTGACTGACCTGCGGGTGTGGCAGGAGCAGCCGGATATGCAAATTCCCTTTGATCATT
GCTGGACGGACTATCACGTTTCCATCCTACGGCCAGAGGGATTTCTTTATCTGCTCGAAACTCGCGAGCTGCTGCTGAGCAGCGTCGAACTGATCGTGCTGGAAGATTGTCATGACAGCGCCGTT
TATCAGAGGATAAGGCCTCTGTTCGAGAATCACATTATGCCAGCGCCACCGGCGGACAGGCCACGGATTCTCGGACTCGCTGGACCGCTGCACAGCGCCGGATGTGAGCTGCAGCAACTGAGCGC
CATGCTGGCCACCCTGGAGCAGAGTGTGCTTTGCCAGATCGAGACGGCCAGTGATATTGTCACCGTGTTGCGTTACTGTTCCCGACCGCACGAATACATCGTACAGTGCGCCCCCTTCGAGATGG
ACGAACTG

>Loquacious dsRNA
ATGGACCAGGAGAATTTCCACGGCTCCAGCTTGCCGCAGCAGCTACAGAACCTCCACATCCAGCCGCAGCAGGCGTCCCCCAATCCTGTCCAGACGGGATTTGCTCCACGGCGGCACTATAATAA
CCTTGTCGGCCTGGGCAATGGAAATGCCGTCAGTGGTAGTCCGGTGAAGGGTGCTCCGCTGGGGCAGCGCCATGTGAAGCTCAAGAAGGAGAAGATATCCGCCCAGGTTGCGCAGCTGTCTCAGC
CAGGTCAGCTGCAGCTGTCAGATGTTGGTGATCCTGCCTTGGCGGGCGGATCGGGCTTACAAGGTGGAGTCGGCCTTATGGGCGTAATATTGCCCAGCGACGAGGCC

>Dicer-2 dsRNA
AAAGTTCAACGCGAAGAGCAAGGCTAAGATGAAAGTTATTTTGATTCCGGAGCTATGCTTCAATTTTAACTTTCCTGGGGATTTATGGCTTAAGTTGATCTTCCTACCCAGCATTTTAAACCGCA
TGTACTTCCTTCTCCACGCAGAGGCCTTACGTAAGCGATTTAATACGTATTTAAACCTCCATCTGCTGCCTTTTAATGGAACTGATTACATGCCCAGACCACTAGAAATTGATTATTCGCTAAAG
CGGAATGTCGACCCCTTGGGCAATGTCATACCAACTGAGGATATCGAGGAGCCGAAATCCCTTTTAGAGCCAATGCCCACAAAGTCCATTGAGGCGTCCGTGGCCAATCTTGAAATAACAGAATT
CGAAAATCCCTGGCAAAAGTATATGGAGCCGGTTGATCTGTCGCGAAATCTTTTGAGTACGTATCCCGTAGAGCTGGACTACTACTATCATTTTAGCGTTGGTAATGTATGTGAGATGAATGAGA
TGGATTTTGAAGATAAGGAATACTGGGCAAAAAATCAGT

>R2D2 dsRNA
TGCTGCCCGGCATACACGGCTTGATGAAGGATTCGACTGTGGGTGATCTGGATGAGGAACTGACTAACCTCAACCGGGACATGGTGAAGGAGCTGCGTGACTACTGCGTCCGCCGCGAGATGCCA
CTGCCCTGCATTGAGGTAGTGCAGCAAAGCGGCACCCCGAGCGCCCCGGAATTCGTGGCCTGTTGCTCCGTGGCCTCCATAGTACGCTACGGAAAGTCGGACAAAAAGAAGGATGCCCGTCAGCG
AGCGGCCATTGAAATGCTGGCCTTAATCTCCAGCAATTCGGACAATTTGCGTCCGGATCAAATGCAAGTAGCGAGCACAAGCAAATTGAAAGTTGTTGATATGGAAGAATCTATGGAGGAATTGG
AGGCATTGCGCAGAAAGAAATTTACCA

>CG7942[1](Debranching enzyme) dsRNA
GTGGCGAGCTGGTAGCACCAGTGCTGACCATATTCATTGGCGGCAACCATGAGGCCTCCAATTACCTGCAGGAGCTCCCATACGGCGGTTGGGTGGCTCCAAATATTTACTACCTTGGTTATGCC
GGCGTCGTCAATGTAAACGGTGTTCGGATAGCTGGAATAAGCGGAATCTTCAAGGGTCACGACTTTTTGCGCGGCCATCACGAATTCCCTCCGTATACCGATTCCACGTGTCGCAGTGTCTACCA
TGTGCGGCAGCTAGAAGTCTTTCGGCTGAAACAAATTTCCGGGCGAGTTGATATTTTCCTGTCCCACGACTGGCCCACCGGCATCTATGAATACGGAAACAAGGCGCAACTGCTCCGCAAGAAAC
CATTTTTTGCTGCAGACATGGAAAGCGGGAAGCTGGGTAGCCAGCCACTGGAGGAGTTACTGAAAGCGGTCCAACCGGCCTACTGGTTTGCTGCCCATTTGCATTGCAAGTTTGCCGCTTTGGTG
CCGCACAATCACAGCCAGAAGCTAGGAGATGCTGAATC

>CG7942[2](Debranching enzyme)
GCAGCAGTGAAGATGAAGACGAGGAAAGGGAGAAGGTAAAGAAAGCTGCTCCTGTACCTCCACCATCCAAATCTGTTCCCGTGACCAAGTTTCTGGCTCTCGACAAATGCCTGCCACGTCGTGCT
TTCCTGCAAGTGGTAGAGGTACCCAGTGACCCCATCGAAGGCACTCCCCGCCTGGAATACGACGCAGAGTGGCTAGCCATCTTGCACAGTACAAATCACTTGATTTCAGTGAAGGAGAATTATTA
TTACCTGCCCGGAAAAAAGGCGGGAGAGTTTACAGAGCGATCAAACTTTACCCCCACTGAAGAAGAACTAGAAGCAGTGACCGCAAAGTTTCAGAAACTTCAAGTCCCCGAGAACTTTGAGCGCA
CAGTGCCAGCTTTCGATCCCGCGGAGCAGTCTGATTATAAGCACATGTTTGTGGATCAACCCAAGGTTCAACTAAACCCCCAGAGCAATACGTTCTGTGCCACTCTGGGTATAGACGATCCGCTG
TGCTTAGTTTTGTTGGCGAATGGCCTAGATCTGCCTGC

Table S4

>GFP dsRNA
GATCACATGGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCTCGGCATGGACGAGCTGTACAAGTAAAGCGGCCGCGACTCTAGATCATAATCAGCCATACCACATTTGTAGAGGTTTTA
CTTGCTTTAAAAAACCTCCCACACCTCCCCCTGAACCTG

>UTR insert CG11094
actagtTGATAATTTTTCATTAACTAGAGTAACGAATACTACTTTGCCCCGATATTTATTATTGTTCAGCATCACATATTAGCTTAATGCTTCGGTGAAATCGCGCGAATTTAACTTTTATAACT
TAGAGTTGAGTAACTTAGAGTTTTATGGAGCAAAACCTCTGTAAATAAATCGAATTTATCGGTAAACTAAAGCGCGACTTGGACTATCTTCAATCAACAAGCCAAATATGTCGATGTGTGACAGC
CGTTCTACGCGTCAGCTTTCTTCAATCAACATTACCCCGTGCTGAGATGTCTGGCCTCAATGTTAATAATCTCAATCTACAATCAACATTCTCTTCTCTTCAATCAACAATCCGCAAACGGATCT
AATCgcggccgc

>UTR insert CG11094-mutant
actagtTGATAATTTTTCATTAACTAGAGTAACGAATACTACTTTGCCCCGATATTTATTATTGTTCAGCATCACATATTAGCTTAATGCTTCGGTGAAATCGCGCGCAGTGAACTTTTATAACT
TAGAGTTGAGTAACTTAGAGTTTTATGGAGCAAAACCTCTGTAAATAAATCGCAGTGATCGGTAAACTAAAGCGCGACTTGGACTATCTTCAATCAACAAGCCAAATATGTCGATGTGTGACAGC
CGTTCTACGCGTCAGCTTTCTTCAATCAACATTACCCCGTGCTGAGATGTCTGGCCTCAATGTTAATAATCTCAATCTACAATCAACATTCTCTTCTCTTCAATCAACAATCCGCAAACGGATCT
AATCgcggccgc

>UTR insert CG1849
actagtCCTGGAAATCAGACTCCGGCGAAGTTTTATGCTCGGACTCATAAAATCGTGCGACGAGTTTGAATCACAGGCCCTCGATTTTCACCAGGATTTTTTACAAATCCCAGCAGAAAACACGA
AAACTCAAAAACTCAGCCCAAAAAGAAAATACCAAGAAAGCAAACTTTAGTTCAATTTCATTTCAACACAAAACAACAACAACACAATTTGTACATAGCTAACTAGTTGTAACACTCATAACTTT
TTTTTTTTGAGAACCTATTTTTTTTCGATGGATAATATGCGAATTTAGCTATTTTTAATCATTATGTTTAACTAGTCGTCTAAGCGAGAAATCAATTTTTTTGTCTAGCCATAAGTTTTAGCGCGA
AAAGAGATCTAACACAAAAATCGAATTTGAAACAAAACCAAATAAAAAACAAAAATCACACACAAAAAgcggccgc

>UTR insert CG1849-mutant
ActagtCCTGGAAATCAGACTCCGGCGAAGTTTTATGCTCGGACTCATAAAATCGTGCGACGAGTTTGAATCACAGGCCCTCGATTTTCACCAGGATTTTTTACAAATCCCAGCAGAAAACACGA
AAACTCAAAAACTCAGCCCAAAAAGAAAATACCAAGAAAGCAAACTTTAGTTCAATTTCATTTCAACACAAAACAACAACAACACAATTTGTACATAGCTAACTAGTTGTAACACTCATAACTTT
TTTTTTTTGAGAACCTATTTTTTTTCGATGGATAATATGCGCAGTGAGCTATTTTTAATCATTATGTTTAACTAGTCGTCTAAGCGAGAAATCAATTTTTTTGTCTAGCCATAAGTTTTAGCGCGA
AAAGAGATCTAACACAAAAATCGCAGTGGAAACAAAACCAAATAAAAAACAAAAATCACACACAAAAAgcggccgc

>UTR insert CG5166a
actagtGACACCAGAAACCCAAGTCATCATTCCAAGTTAGTTTTTCCACCGGCGCAAGGAAAGGGCCGCGCTTCATCCAGCATTCCGATTGTAAACTTACTTAGCATATAATGTGAACTCGGTTC
GGAAGGAGCTGATCGCTGATCGCTGATCGAAGCTGCAAGCTGGATGGAAGCTCTTTGCTTGCCCTGCGGGAAATGAAAAACGAATGTGAGATTTTAGAGAGCTTCAAATTTATTCGTTTCCTTTT
CGAAATTCGGTAGAACTAATTAATTTTTGTTTAAATTGAATTTTGTTGCCACTTCTCCGCCTCTTCTTACACATTATTCGCCAGCATTTACCAGAAATGTAATGACATCGATATATAAATGATTG
TTTTGACGTTTCTCGGAGAAATTTCCTTGCTAGCTTTACAGGCAGAAGCTAATGTGAGAGCAAGAGCTTGAGTCAGGCTTCCTTTGGGTTTTAGTGCCTCCGTTGTCTCCGAATTAATGAAAAAT
TAACAAGAACAATCCGTATTACTTCTTTGCCCGTCATAAATCGGTTTGGTTATATTTCGTATGATCTAGAAGCATCTGTTGTGGTCTGTTTTGTTTTGTAAACCTTCAAGTTTCCTTAAATGAAg
cggccgc

>UTR insert CG5166a-mutant
actagtGACACCAGAAACCCAAGTCATCATTCCAAGTTAGTTTTTCCACCGGCGCAAGGAAAGGGCCGCGCTTCATCCAGCATTCCGATTGTAAACTTACTTAGCATATAATGTGAACTCGGTTC
GGAAGGAGCTGATCGCTGATCGCTGATCGAAGCTGCAAGCTGGATGGAAGCTCTTTGCTTGCCCTGCGGGAAATGAAAAACGAATCTCACATTTTAGAGAGCTTCAAATTTATTCGTTTCCTTTT
CGAAATTCGGTAGAACTAATTAATTTTTGTTTAAATTGAATTTTGTTGCCACTTCTCCGCCTCTTCTTACACATTATTCGCCAGCATTTACCAGAAATGTAATGACATCGATATATAAATGATTG
TTTTGACGTTTCTCGGAGAAATTTCCTTGCTAGCTTTACAGGCAGAAGCTAATCTCACAGCAAGAGCTTGAGTCAGGCTTCCTTTGGGTTTTAGTGCCTCCGTTGTCTCCGAATTAATGAAAAAT
TAACAAGAACAATCCGTATTACTTCTTTGCCCGTCATAAATCGGTTTGGTTATATTTCGTATGATCTAGAAGCATCTGTTGTGGTCTGTTTTGTTTTGTAAACCTTCAAGTTTCCTTAAATGAAg
cggccgc

>UTR insert CG6551
actagtTGATATCCACCCGATTCAAACCACAGCATCAGCATCCGCATCTATATTCGCATCAGCAACAGGAAACCTCTTGCCATGCTACCCACACATCTGAGGACACTGATTTGTTAGCTCAAGAC
AACACAACTGAAATCGAAACGCATTGAATTTAGATCAAATTCGAGCTGGTATCGAATATTAACCATACAAACAAACATAAACAAAAGGCTCCCTAAATGATTTAAATATTGGTCTGGTCCCCTTA
AGATTTAAAAATATCAATTAGTTTTTATGGAAATAGTTAGTTTCAATCGTAATAGGCATTTAAAAAACATTTTACCCTAATTGAGTTTTTAAATCTCCAGAGGATTTCAACGCACCAATATTTTG
TACACAACACACATTGTTAAATTTAAATTTTCACTCGAATTTCAAGTATTCTATTTTGCAAAAATTATTTTGTGTAAATCTCGgcggccgc

>UTR insert CG6551-mutant
actagtTGATATCCACCCGATTCAAACCACAGCATCAGCATCCGCATCTATATTCGCATCAGCAACAGGAAACCTCTTGCCATGCTACCCACACATCTGAGGACACTGATTTGTTAGCTCAAGAC
AACACAACTGAAATCGAAACGCATTGAATTTAGATCAAATTCGAGCTGGTATCGAATATTAACCATACAAACAAACATAAACAAAAGGCTCCCTAAATGATTTAAATATTGGTCTGGTCCCCTTA
AGATTTAAAAATATCAATTAGTTTTTATGGAAATAGTTAGTTTCAATCGTAATAGGCATTTAAAAAACATTTTACCCTAATTGAGTTTTTAAATCTCCAGAGGATTTCAACGCACCAATATTTTG
TACACAACACACATTGTTAAATTTAAATTTTCACTCGCAGTGCAAGTATTCTATTTTGCAAAAATTATTTTGTGTAAATCTCGgcggccgc

Chapter 4

Evolution, Biogenesis, Expression, and Target Predictions of a Substantially Expanded Set of *Drosophila* MicroRNAs

J. Graham Ruby[1,2], Alexander Stark[3,4], Wendy K. Johnston[1,2], Manolis Kellis[3,4], David P. Bartel[1,2], and Eric C. Lai[5]

[1]Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA, 02142
[2]Howard Hughes Medical Institute and Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, 02139
[3]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[4]Broad Institute of MIT and Harvard, Cambridge, MA 02141, USA
[5]Department of Developmental Biology, Memorial Sloan-Kettering Cancer Center, New York, New York 10021, USA

J.G.R. performed the computational analysis excluding target predictions, which were performed by A.S. Libraries were prepared for sequencing by W.K.J. and E.C.L. J.G.R., A.S., M.K., and D.P.B. wrote the manuscript.

Electronic supplemental files are provided on the accompanying CD-ROM, under the directory "Chapter 4". All files are ASCI text; .html files are best opened with a web browser, and .xls are tab-delimited text that is best viewed with a spreadsheet application such as Microsoft Excel.

Published as:
JG Ruby, A Stark, WK Johnston, M Kellis, DP Bartel, and EC Lai. (2007) Biogenesis, expression, and target predictions for an expanded set of microRNA genes in *Drosophila*. *Genome Research*. 17:1850-64.

Abstract

MicroRNA (miRNA) genes give rise to small regulatory RNAs in a wide variety of organisms. We used computational methods to predict miRNAs conserved among *Drosophila* species and large-scale sequencing of small RNAs from *Drosophila melanogaster* to experimentally confirm and complement these predictions. In addition to validating 20 of our top 45 predictions for novel miRNA loci, the large-scale sequencing identified many miRNAs that had not been predicted. In total, 59 novel genes were identified, increasing our tally of confirmed fly miRNAs to 148. The large-scale sequencing also refined the identities of previously known miRNAs and provided insights into their biogenesis and expression. Many miRNAs were expressed in particular developmental contexts, with a large cohort of miRNAs expressed primarily in imaginal discs. Conserved miRNAs typically were expressed more broadly and robustly than were nonconserved miRNAs, and those conserved miRNAs with more restricted expression tended to have fewer predicted targets than those expressed more broadly. Predicted targets for the expanded set of microRNAs substantially increased and revised the miRNA-target relationships that appear conserved among the fly species. Insights were also provided into miRNA gene evolution, including evidence for emergent regulatory function deriving from the opposite arm of the miRNA hairpin, exemplified by *mir-10*, and even the opposite strand of the DNA, exemplified by *mir-iab-4*.

Small RNA sequences were deposited in the Gene Expresion Omnibus (GSE7448). Computational tools for miRNA prediction (MiRscan3) are available for

anonymous download at http://web.wi.mit.edu/bartel/pub/. Two HTML tables and two MS Excel tables are provided as supplementary material.

Introduction

MicroRNAs (miRNAs) are ~23-nt RNA species that direct the post-transcriptional repression of messenger RNAs (mRNAs) (Bartel 2004). They are generated from primary transcripts (pri-miRNAs) that can fold into characteristic hairpin secondary structures. In animals, those hairpins are typically first cleaved away from the rest of the primary transcript by the nuclear RNase III enzyme Drosha to generate miRNA pecursors (pre-miRNA), and are then cleaved near their loops by the cytoplasmic RNase III enzyme Dicer to generate a heteroduplex of two ~23-nt RNAs (Lee et al. 2003). The mature miRNA is preferentially packaged into the RNA-induced silencing complex (RISC), while the other species, known as the miRNA star (miRNA*) is discarded (Lau et al. 2001; Lim et al. 2003b). The decision as to which species is incorporated into the silencing complex is influenced by the difference in pairing stabilities between the two ends of the miRNA:miRNA* duplex, with preferential incorporation of the strand whose 5′ end is less stably paired (Khvorova et al. 2003; Schwarz et al. 2003).

Once incorporated into the silencing complex, metazoan miRNAs pair to the messages of their mRNA targets, primarily in 3′ untranslated regions (3′ UTRs). Complementarity between the message and a segment in the 5′ region of the miRNA known as the seed (miRNA nucleotides 2-7) appears to be the most crucial requirement of target recognition. Conserved pairing to the seed region is a feature of most genetically

identified miRNA-target interactions (Lai 2002; Lee et al. 1993; Wightman et al. 1993). Indeed, the requirement of conserved pairing to the miRNA seed enables miRNA targets to be predicted in excess of the noise of false-positive predictions (Brennecke et al. 2005; Krek et al. 2005; Lewis et al. 2005; Lewis et al. 2003). Short, 7- to 8-nt sites matching the seed region of the miRNA are not only important but sometimes can suffice for repression in reporter assays (Brennecke et al. 2005; Doench and Sharp 2004; Lai et al. 2005). Consistent with the *in vivo* sufficiency of 7mer seed-matching sites in mediating repression, many messages preferentially co-expressed with a highly expressed miRNA are depleted in 7mer sites matching that miRNA, presumably due to selective avoidance of miRNA-mediated repression during evolution (Farh et al. 2005; Stark et al. 2005). Moreover, miRNAs that share the same seed sequence but are diverse throughout the remainder of their sequences can be functionally redundant (Abbott et al. 2005; Lim et al. 2005), which justifies their grouping into members of the same miRNA 'family' (Ambros et al. 2003; Lim et al. 2003b). The seeds that define families are often conserved throughout diverse species even as the individual miRNA genes within the family vary (Ruby et al. 2006). The arms of the hairpin precursors are less conserved than the seeds, but are more conserved than either the surrounding genomic sequence or the intervening loop sequence (Lai et al. 2003; Lim et al. 2003b).

Most known miRNAs were discovered through the cloning and sequencing of small-RNA cDNAs (Griffiths-Jones 2004). However, this method can miss miRNAs expressed at low levels or in only specific cell types or conditions. One approach for identifying low-abundance miRNAs that has previously been applied in *Drosophila* is to identify candidate miRNA hairpins computationally and then validate their expressions

using more directed, and thereby potentially more sensitive, experimental methods (Lai et al. 2003). Because identification of plausible candidates is aided by comparative genomics, this approach gains efficacy as the genome sequences of additional related species become available. A second approach for identifying rare miRNAs is simply to increase the scale of small-RNA sequencing well beyond the reach of prior efforts. This high-throughput sequencing approach has not been used previously in insects, but in other lineages it has revealed miRNAs and miRNA candidates that escaped earlier detection because they are rare or not well conserved in related genomes (Berezikov et al. 2006; Fahlgren et al. 2007; Lu et al. 2006; Rajagopalan et al. 2006; Ruby et al. 2006).

Here, we use the two complementary approaches described above – computational prediction and high-throughput sequencing – to identify nearly 60 additional fly miRNAs and to refine the descriptions of about half of those that had been previously annotated. These results provided insights into miRNA evolution, biogenesis, and expression in insects. When combined with improved target prediction, which utilized information from all 12 sequenced *Drosophila* genomes (Consortium 2007a; Consortium 2007b) to increase prediction accuracy, these new and revised miRNAs substantially expanded and improved the view of miRNA-directed regulation in flies.

Results

Computational prediction of fly miRNAs

Novel miRNA genes were sought computationally as hairpins with secondary structure and conservation patterns resembling those of previously annotated miRNAs, using an approach with similarities to that described for miRscan, which had previously been

applied to nematodes and vertebrates (Lim et al. 2003a; Lim et al. 2003b).For each of six *Drosophila* genomes (*melanogaster*, *ananassae*, *pseudoobscura*, *mojavensis*, *virilis*, and *grimshawi*) (Adams et al. 2000; Consortium 2007a; Consortium 2007b; Richards et al. 2005), RNAfold (Hofacker et al. 1994) was used to identify candidate hairpins from across the entire genome. Candidate hairpins from each genome were first scored based on the relative frequencies of structural characteristics in the background candidate set versus a foreground training set of annotated miRNA hairpins. This training set comprised 37 miRNA hairpins from *D. melanogaster* that were selected randomly from the 78 previously annotated in miRBase v8.1 (Griffiths-Jones 2004); the remaining 41 miRNAs were reserved as a test set, the performance of which was not evaluated until after the completion of the prediction process. Candidates with scores far below that of the lowest foreground hairpin were removed from the background set, altering its aggregate properties. Unlike the previous method, candidates were then re-scored based on the properties of the minimized background, and the worst candidates were again eliminated. Following several rounds of eliminating candidates from each individual genome, candidates from different genomes (nodes) were paired as putative orthologs (edges) using Blast (Altschul et al. 1990) and put through the same process of iterative scoring and elimination, now simultaneously evaluating conservation.

The surviving ortholog pairs included 565 hairpin candidates from *D. melanogaster* that could form complete networks between all six genomes, including all 15 possible edges. These successful candidates were ranked by the sum of the 15 pairwise scores from their first round of pairwise scoring and elimination (Table S1). Of the 37 members of the training set, 26 survived and fell mostly within the higher scoring

tail of the distribution (Fig. 1A). Examination of the test set revealed that 35 of 41 members of the test set survived and that these 35 hairpins had similar score distributions as the training set, indicating that our prediction method did not over-fit to the training-set data (Fig. 1A).

Our concept of a candidate miRNA hairpin specified the genomic strand from which the hairpin was derived. However, the secondary structure and conservation properties of a genomic sequence frequently match those of its reverse complement. As a result, 174 of our 565 candidate hairpins were paired with a candidate locus from the opposing genomic strand and thus represented only 87 unique genomic loci. We therefore collapsed our predictions into 478 strand-independent genomic loci, which each included a prediction of which strand would give rise to the mature miRNA based on the higher score. Eliminating 151 candidates that overlapped the annotated exons of protein-coding genes (Table S1) left 327 candidate loci (Fig. 1B). The top 100 candidate loci were carried forward as our predictions. These included 55 of the previously annotated genes (24 of the 26 surviving training set genes), as well as 45 novel predictions.

Recent results from plants and worms demonstrate that for the validation of miRNA expression, large-scale sequence datasets are more reliable and sensitive compared to RNA blotting, and more reliable and roughly equally sensitive as PCR assays (Rajagopalan et al. 2006; Ruby et al. 2006). We therefore used large-scale small RNA sequence data to evaluate the quality of our predictions.

High-throughput sequencing of small RNAs

To survey the miRNAs of flies, we performed high-throughput pyrosequencing (Margulies et al. 2005; Ruby et al. 2006) on libraries of small RNAs isolated from the following ten *D. melanogaster* tissues or stages: very early embryo (0-1 hours), early embryo (2-6 hours), mid embryo (6-10 hours), late embryo (12-24 hours), larvae (first and third instars), imaginal discs, pupae (0-4 days), adult heads, adult bodies, and tissue-culture cells (S2). Pyrosequencing yielded a total of 1.14 million small RNA reads (55,761 – 174,031 reads per library) that perfectly matched the *D. melanogaster* genome.

Refinement of prior miRNA annotations

Of the 54 *D. melanogaster* miRNAs (corresponding to 60 hairpins) that had been previously cloned and sequenced (Aravin et al. 2003; Lagos-Quintana et al. 2001), all 54 were represented in our dataset of 1.14 million small RNA reads, as exemplified by miR-7 and miR-iab-4 (Fig. 2), and detailed for all the miRNAs (Table S2). For the 60 hairpins of these previously cloned miRNAs, read frequencies ranged from 60 (miR-303) to 20,049 (miR-14), with a median of 2415, and for each of these hairpins the miRNA* species was also recovered. Additional *Drosophila* miRNA genes are annotated in miRBase v.8.1 based on homology to other miRNAs or computational predictions supported by RNA blots (Aravin et al. 2003; Lai et al. 2003). Of these 18 genes for which small RNAs had not been previously cloned, 14 were represented in our dataset (Fig. 2A). The four that were missing (*mir-280, -287, -288, and -289*) had been predicted computationally and experimentally supported by RNA blots of samples from early

embryos, larvae and pupae, and adult males (Lai et al. 2003). Their absence in our libraries from these same developmental stages called their authenticity into question.

In half the cases (37 of the 74 confirmed genes), the distribution of reads across the hairpin suggested that the mature miRNA differs from the one that had been previously annotated (Table S2). Usually, the discrepancy was only at the 3′ terminus of the mature miRNA, as exemplified by miR-7 (Fig. 2B). Although proper 3′ annotation is needed for some miRNA expression profiling methods (Wang et al. 2007), re-annotation of the miRNA 3′ terminus was of little consequence because 3′ heterogeneity is a hallmark of mature miRNAs (Basyuk et al. 2003; Lau et al. 2001; Lim et al. 2003b; Ruby et al. 2006). However, in 12 cases there was discrepancy at the miRNA 5′ terminus (Table S2). The re-annotation of a miRNA 5′ terminus is far more consequential due to its role in defining the miRNA seed sequence, which in turn defines the set of targets. For example, shifting the 5′ terminus by a single nucleotide changes the identity of one or both of the two 7mers used for target prediction (Lewis et al. 2005), thereby dramatically altering the set of predicted targets, and shifting it by two or more nucleotides would have an even greater effect. Seven of these 12 cases were corrections of annotations that have been based on computational or molecular evidence not expected to identify the 5′ termini with confidence (Supplemental Text). The other five cases were more interesting because they illustrated how a single miRNA hairpin or paralogous hairpins could spawn new miRNA function.

For miR-210, there were 917 reads with the originally annotated 5′ end and 1031 reads with an extra 5′ nucleotide, all of which mapped uniquely to the genome. Combined, the abundance and equivalence of reads indicated that miR-210 was a rare

example of a single hairpin generating mature miRNAs with multiple abundant 5′ ends. As was done for miR-248 in *C. elegans* (Ruby et al. 2006), we annotated the species with an extended 5′ end as miR-210.1 and the species with the originally annotated 5′ end as miR-210.2, with the idea that both probably direct repression in the fly (Table S2).

In the case of miR-10, the dominant read was from the arm of the hairpin precursor opposite the annotated miRNA and was sevenfold more abundant (Fig. 3), a result expanding on the observation that species from both arms are easily detected (Schwarz et al. 2003). Because conservation criteria supported the function of RNAs from both arms of the hairpin, and in conjunction with a parallel study (Stark 2007b), we annotated the two major products of the *miR-10* hairpin as miR-10-5p and miR-10-3p. The seed of the original miR-10 (miR-10-5p) was conserved throughout all annotated *miR-10* genes, including those of vertebrates (Fig. 3A). The seed of the species more abundantly represented in our dataset (miR-10-3p) was not conserved in all annotated *miR-10* genes but was nonetheless conserved in at least one *miR-10* gene of each species examined (typically the *miR-10a* paralogs of vertebrates; Fig. 3A).

The *mir-281* and *mir-2* paralogs illustrated how highly related miRNA genes could have divergent function. For both sets of paralogs, miRNAs deriving from the miRNA arms of the hairpins could be mapped to multiple related hairpins. Nonetheless, the miRNA* species, which mapped uniquely to their hairpins, revealed the likely processing of each hairpin and indicated that one of the two *mir-281* hairpins and two of the five *mir-2* hairpins gave rise to miRNA species that differed from those previously annotated (Supplemental Text).

Novel miRNAs

Having revised many of the previous miRNA annotations of *Drosophila*, we next examined the overlap between our predicted miRNA loci and the small RNA reads. Of the 100 predictions, 45 had not been previously annotated as miRNA genes. Of those, 20 were supported by the reads. In all 20 cases, our prediction method identified the correct strand of the miRNA gene (as well as 21 of 24 cases from the training set and 29 of 31 cases from the test set). Given correct identification of the miRNA strand, the miRNA 5′ terminus was correctly predicted in 8 of 20 cases (plus 14 of 21 cases from the training set and 15 of 29 cases from the test set, Fig. 1C). Some the remaining 25 predictions that lacked experimental confirmation also might be authentic miRNAs. However, because most might be false-positives and because their 5′ termini were not predicted with high confidence, we did not consider any of these 25 suitable for annotation or target-prediction studies.

Looking more broadly at the small RNA reads to identify the miRNAs that were more difficult to recognize computationally increased our count of newly identified miRNA loci from 20 to 59. To confidently identify these as miRNA genes, we considered the following criteria: 1) the pairing characteristics of the hairpin; 2) the miRNA expression, as measured by the abundance of sequence reads sharing the same 5′ terminus; 3) evolutionary conservation, as evaluated by the apparent conservation of the hairpin in other fly species and grouping of the miRNA candidate into a family based on its seed sequence; 4) the absence of annotation suggesting non-miRNA biogenesis, and 5) the presence of reads corresponding to the predicted miRNA* species. The observation of both a candidate miRNA and a candidate miRNA* in a set of reads provides particularly

compelling evidence for Dicer-like processing from an RNA hairpin (Fahlgren et al. 2007; Rajagopalan et al. 2006; Ruby et al. 2006). As illustrated for miR-988 (Fig. 4A), 40 newly identified genes satisfied all five of our criteria, and nineteen others satisfied a subset of the criteria deemed sufficient for confident annotation as miRNAs (Table 1). Nine additional candidates fell within predicted miRNA-like hairpins and were sequenced more than once (Table S2). However, they were considered unlikely to be miRNAs because they did not satisfy the other criteria sufficiently and they mapped to either protein-coding transcripts (candidates 1-5) or heterochromatic DNA (candidates 6-8). Ten of the newly-identified miRNAs derived from loci that were among the top 200 predicted to form miRNA precursor hairpins in a previous effort (Lai et al. 2003). Nine of those predictions correctly identified the genomic strand from which the miRNA was derived, but prediction of the mature miRNA had not been attempted.

Two thirds of the novel miRNAs appeared to be broadly conserved in the *Drosophila* genus (Table 1). Orthologs were sought in six species spanning both the *Sophophora* subgenus (*D. simulans*, *D. yakuba*, *D. ananassae*, *D. pseudoobscura*) and the *Drosophila* subgenus (*D. mojavensis*, and *D. virilis*). Putative orthologs were found in all six species for 28 of the miRNAs and in five of six species for another 9. In 12 of the remaining cases, orthologs were found in two or fewer of the *Drosophila* species.

One newly identified locus, *mir-996*, resided 1.5 kb downstream of a related miRNA (*mir-279*) and within the transcript of CG31044, which is annotated as encoding a 140 amino acid protein (Crosby et al. 2007). We suggest that miR-996, not the putative protein, is the functional product of this gene. Consistent with this proposal, the observed miRNA was perfectly conserved across a wide scope of fly species, whereas in the ORF,

sequence polymorphisms in the closely related species *D. simulans*, *D. yakuba*, *D. ananassae*, and *D. pseudoobscura* introduced nonsense mutations at codons 73, 56, 13, and 40, respectively.

Like many known miRNA genes (Griffiths-Jones 2004), 26 of the 59 newly identified loci were clustered with other miRNA loci (Table 1, Fig. 5A), and 13 fell within annotated introns (and from the same genomic strand as the intron, Table 1). Thus, more than half (69 of 133) of the canonical *Drosophila* miRNAs were clustered (Fig. 5A), and over a quarter (36 of 133) were intronic (Table S1).

Although most of the novel miRNA genes closely resembled those previously annotated, three of the hairpin precursors were much larger than those observed previously in animals. For the vast majority of previously annotated fly genes, fewer than 30 nucleotides separated the miRNA and miRNA*, and all had fewer than 60 intervening nucleotides. The distribution of intervening sequence lengths was generally similar among the newly identified miRNAs. However, *mir-956*, *mir-989*, and *mir-997* had abnormally large intervening sequences of 82 nt, 99 nt, and 112 nt, respectively (Fig. 4C-D, Table S2). Each of these hairpins gave rise to miRNA* reads, and in each case the dominant ends of the miRNA versus the miRNA* exhibited 1- or 2-nt 3′ overhangs. In no case was there EST evidence of an intron helping to bridge the distance between the miRNA and miRNA* loci (Crosby et al. 2007; Karolchik et al. 2003). The lack of constraint on the length of the intervening sequence was illustrated by *mir-989*, whose mature miRNA sequence was perfectly conserved across all seven of the *Drosophila* species examined but whose intervening sequence length varied widely, dipping as low as 52 nucleotides in *D. pseudoobscura* (Fig. 4D).

MicroRNA biogenesis in flies

As in nematodes (Ruby et al. 2006), examining the multitude of reads arising from the previously annotated miRNA hairpins provided insights into the specificity and precision of Drosha and Dicer processing (Table S2). These RNase III enzymes preferentially leave 2-nt 3′ overhangs when cleaving perfect RNA duplexes (Basyuk et al. 2003; Lee et al. 2003), and this configuration between the ends of miRNA and miRNA* species was observed in our dataset. The 5′ ends of both the miRNA and miRNA* species were more homogenous than the 3′ ends, and the miRNA 5′ ends were more consistent than the miRNA* 5′ ends, and the miRNA* 3′ ends were more consistent than the miRNA 3′ ends, regardless of which was generated by Drosha and which by Dicer (Table S2; Supplemental Text). The heightened precision of either enzyme when it defined the miRNA seed implied that Dicer does not simply measure from the site of the Drosha cleavage and suggested that additional determinants must be employed when needed to more accurately define Dicer cleavage. Similar conclusions arise from the sequencing data in nematodes (Ruby et al. 2006).

As reported in other species (Li et al. 2005; Ruby et al. 2006), untemplated nucleotide addition also contributed to a minor fraction of 3′ heterogeneity. In *Drosophila* we observed evidence for processivity of the terminal-transferase activity, with a preference for adding adenosines (Supplemental Text).

Reads that were antisense to either the miRNA or miRNA* appeared for four previously annotated hairpins (*mir-iab-4, -124, -305, -307*). The *mir-iab-4* hairpin gave rise to nine antisense reads (Fig. 2D); each of the remaining hairpins gave rise to one

read. Antisense transcription from *mir-iab-4* has been noted previously (Cumberledge et al. 1990). Our reads from either side of the antisense hairpin paired to each other with 2-nt 3′ overhangs, indicating that the antisense transcript was processed by the miRNA biogenesis machinery and likely produced a miRNA (miR-iab4as) that enters the silencing complex.

MicroRNA expression patterns

The collection of reads from a variety of developmental stages and anatomical contexts permitted analyses of miRNA expression profiles and overall miRNA expression (Fig. 6A). MicroRNAs were clustered based on their relative expression across all ten libraries, with the expression values for a particular miRNA in a particular library set to the number of reads corresponding to that miRNA in the given library divided by the total number of reads matching miRNA hairpins in that library. For previously annotated miRNAs, this normalization scheme generated expression profiles similar to those observed previously using 2S rRNA-normalized northern blots (Fig. 6B)(Aravin et al. 2003), whereas other schemes, such as normalizing to the total reads from each library or to the number of sequenced ribosomal RNA fragments, did not generate profiles matching the published northern results (data not shown).

Most miRNAs were observed across several libraries. However, several large sets of miRNAs exhibited strong preference for expression in a single context. The 33 miRNAs that exhibited the narrowest ranges of expression (>70% of their library-normalized reads deriving from a single library) were prevalent in the imaginal discs, adult heads, and to a lesser extent, adult bodies and late embryos (61%, 24%, 12%, and

3% of narrowly expressed miRNAs, respectively), and most were first sequenced in this study (88% of narrowly expressed miRNAs; Fig. 6A).

The normalization of read counts across libraries also permitted an approximate but informative assessment of relative overall expression (Fig. 6A). As reported in vertebrates, worms, and plants (Bartel 2004; Rajagopalan et al. 2006; Ruby et al. 2006), miRNA abundance correlated strongly with the extent of conservation, with those miRNAs found only within the subgenus *Sophophora* expressed significantly less than those conserved beyond that clade (Fig. 6C, Wilcoxon rank-sum test, $p < 2.7 \times 10^{-9}$). Notably, the more highly conserved miRNAs were also observed more evenly across the ten libraries examined (Wilcoxon rank-sum test, $p < 8.5 \times 10^{-7}$; Fig. 6D).

As observed previously in worms and mammals (Baskerville and Bartel 2005; Lau et al. 2001; Sempere et al. 2004), miRNAs that were clustered in the *Drosophila* genome usually had similar expression profiles (Fig. 6E). The correlation of miRNA expression patterns diminished as the distance separating miRNAs surpassed 10,000 nt. Proximally located miRNAs are thought to derive generally from common primary transcripts (Lagos-Quintana et al. 2001; Lau et al. 2001). The *mir-991-992* and *mir-310~313* clusters, separated from each other by only 1.0 kb, provided a counter example (Fig. 5C). Although these two clusters each exhibited internally consistent expression patterns, there was little correlation of expression between the two clusers (Fig. 6E), implying that the *mir-991/992* and *mir-310~313* clusters derived from independent transcripts. A more intriguing example was provided by *mir-283, -12 and -304*, all three of which map within a single intron. The expression patterns of *mir-12* and *mir-304* correlated very closely with each other (Pearson correlation coefficient = .94), but neither

correlated well with that of *mir-283* (correlation coefficients of .16 and -.05, respectively), which resided only 1.0 kb upstream of *mir-304* and 1.5 kb upstream of *mir-12*.

MicroRNA targets

In order to gain insight into the functional consequences of the known *D. melanogaster* miRNAs, including those whose annotations were established or modified here, we predicted their targets using comparative genomics of the sequenced genomes of the *Drosophila* genus. As done previously, sites were identified in annotated *D. melanogaster* 3′ UTRs that matched the seed region of each miRNA. Two types of 7mer sites were sought: the perfect Watson-Crick match to miRNA nucleotides 2-8 (Brennecke et al. 2005; Krek et al. 2005; Lewis et al. 2005; Lewis et al. 2003) and the perfect Watson-Crick match to miRNA nucleotides 2-7, supplemented with an adenosine opposite miRNA position 1 (Lewis et al. 2005).

Conservation of 7mer sites was assessed using a multi-genome alignment of 12 *Drosophila* species (Adams et al. 2000; Consortium 2007a; Consortium 2007b; Richards et al. 2005). The phylogenetic distribution of each seed-match motif was used to calculate the total branch-length, a measure of evolutionary distance, across which the motif was conserved (Stark 2007a). Requiring perfect conservation across all of the available species maximized confidence in predicted targets, defined as the fraction of sites that were conserved above chance expectation, but also substantially reduced sensitivity (Fig. 7A-C). This trend extended to arbitrary subsets of the currently available species, including subsets that have been used for other prediction efforts in flies (Brennecke et al.

2005; Enright et al. 2003; Grun et al. 2005; Stark et al. 2003). Such loss of sensitivity is partly attributed to artifacts in sequencing coverage, assembly, or alignment, whose impacts on predictions increase with the number of genomes considered (Grun et al. 2005). Discarding the traditional requirement for perfect conservation within a species set and replacing it with a branch-length requirement enabled confident predictions to be reported in spite of the absence of the motif in particular genomes (Stark 2007a). The confidence of miRNA target predictions increased with the total branch length and approached a maximum, averaged over all conserved miRNAs, of 0.64 (Fig. 7A), corresponding to a signal-to-noise ratio of 2.7:1. These improved predictions for the expanded and revised set of miRNAs are available at targetscan.org.

For comparison, we used the same procedure to predict targets for the *D. melanogaster* miRNAs as annotated in miRBase v8.1 (Griffiths-Jones 2004). By both increasing the number of annotated conserved miRNAs and correcting the identities of previously annotated miRNAs, our study increased the numbers of both miRNAs and miRNA families with significantly conserved targets (confidence ≥ 0.5) by 1.7 fold (Fig. 7D). While 9292 miRNA-target gene pairs were unaffected by the miRNA annotation additions and changes, 2484 were removed and 5475 were added, thereby changing the predicted network of miRNAs and targets in *D. melanogaster* by 68% [(2484 + 5475)/(9292 + 2484)]. Of the 3424 unique genes predicted to be conserved targets of miRNAs, 706 had conserved sites for only novel miRNAs. Conversely, 290 genes were erroneously predicted to be conserved targets due to miRNA annotations that were adjusted based on our sequencing data (Fig. 7E).

The scope of miRNA targeting varied between those miRNAs broadly expressed across many libraries compared to those expressed more narrowly, independent of the relationship between breadth of expression and conservation discussed above. Of those miRNAs conserved beyond the scope of the *Sophophora* subgenus, the narrowly expressed miRNAs tended to have fewer predicted target genes (Fig. 7F, Wilcoxon rank-sum test, $p < .0015$).

Discussion

Hairpin characteristics

The sets of miRNAs initially identified in nematodes, flies and mammals derive from short hairpins, whereas many of those identified in plants derive from longer precursors (Bartel 2004). Three somewhat longer exceptions have been noted [*Drosophila mir-31b* (Aravin et al. 2003), *C. elegans mir-229* (Ambros et al. 2003; Lim et al. 2003b), and *C. briggsae mir-72* (Ohler et al. 2004)], but the prevalence of short hairpin precursors has seemed to justify limiting the length of the sequenced folded during the initial steps of many prediction protocols (Grad et al. 2003; Lai et al. 2003; Lim et al. 2003b), including the approach described here. Several protocols even explicitly evaluate the distance between the predicted miRNA and miRNA* as a characteristic feature of miRNA hairpins (Bentwich et al. 2005), again including the approach described here. Although imposing these constraints likely boosts the specificity of miRNA prediction, our sequencing results indicated that this comes at the cost of missing some miRNAs with unusually long hairpins, particularly in *Drosophila* where we found three hairpins (*mir-*

956, *mir-989*, and *mir-997*) with at least 80 nt separating the miRNA and miRNA* strands (Fig. 4C-D, Table S2).

Our observation that metazoan miRNA precursors can be much longer than previously recognized confirmed that minimal sequence or structural requirements are imposed upon the loops of miRNA hairpins (Berezikov et al. 2005; Han et al. 2006; Lai et al. 2003) but raised the question of why long miRNA hairpins are not more frequent in animals. A large, open loop can lead to Drosha processing on the incorrect end of the hairpin by mimicking the single-stranded RNA normally present at the base (Han et al. 2006). This opportunity for dead-end side reactions implies selective pressure for the maintenance of a tight loop. Consistent with this idea, *mir-956*, *mir-989*, and *mir-997* each exhibited extensive secondary structure in the segment connecting the miRNA and miRNA* (Table S2). Deletions can tighten a loop even if they disrupt secondary structure, making them more tolerable than insertions, which must be accompanied by compensatory changes in order to maintain the tightness of the loop. Thus, miRNA hairpins might be expected to shorten rather than lengthen over evolutionary time. Another possibility is that shorter pre-miRNAs might be more suitable substrates for downstream events such as nuclear export, and longer pre-miRNAs might only rarely bypass these constraints.

The evolutionary origins of novel miRNA genes

High-throughput sequencing of miRNAs in *D. melanogaster* provided insight into the origins of novel miRNA genes and how their origins might differ from those of protein-coding genes. Generally, the first step in the emergence of a new gene is the duplication

of all or part of an ancestral gene (Ohno 1970). A redundant copy of a gene eventually faces one of three fates: the accumulation of mutations that render the copy functionless (nonfunctionalization), the accumulation of mutations that confer a novel and independently-selectable function (neofunctionalization), or, in cases where the ancestral gene had multiple functions, the accumulation of complementary degenerative mutations in both gene copies that specialize each to perform one of the parental functions (subfunctinoalization) (Force et al. 1999). Protein-coding genes provide some examples consistent with subfunctionalization and others consistent with neofunctionalization. We observed examples of miRNA genes that were consistent with each of these models, and also examples that appeared to be the products of *de novo* emergence.

The process of subfunctionalization first requires that an ancestral gene acquire multiple functions. Mechanisms that could impart multiple functions on a miRNA locus include imprecise processing that generates alternative miRNA 5′ termini, like that observed for *mir-210*, and transcription from both orientations with subsequent processing of both pri-miRNAs, as observed for *mir-iab-4*. But perhaps the most available mechanism for acquiring new functions is bringing the miRNA* into service. MicroRNA* species are initially generated at an obligate 1:1 stoichiometric ratio compared to mature miRNAs and to varying degrees are incorporated into RISC just like their complementary counterparts, albeit at a generally lower frequency. They thereby represent an easily accessible substrate for the evolution of novel functionality (Lai et al. 2004) (Fig. 8A). Examples of genes in which conservation data, read abundance, or experimental data suggested that both strands could be functional included *mir-10*, *mir-iab-4* and *mir-313*. The miR-313* seed was only conserved within the *melanogaster*

complex, which diverged from the *yakuba* and *erecta* complexes of the *melanogaster* subgroup 6-15 million years ago (Lachaise et al. 1988). Although the *melanogaster* complex species were insufficiently diverged to conclude selective maintenance of the seed, the high abundance of the miRNA* implied the capacity to affect the expressions of target messages.

If a locus with multiple miRNA products, such as one of those listed above, were to duplicate, selective pressure would diminish for each daughter copy to continue producing all miRNA species, and eventually each daughter copy might retain the ability to produce just one of the functional miRNA products. This process might be in progress for the vertebrate *miR-10* paralogs; miR-10-3p is maintained in all of the *mir-10a*, but not the *mir-10b*, hairpins (Fig. 3).

Neofunctionalization requires that gene copies find novel selectable functions after duplication and prior to loss of expressional competence. Because 5′ heterogeneity was very rare among the known miRNAs, the divergent processing of both the *mir-2* paralogs and the tandemly duplicated *mir-281* paralogs likely emerged after duplication and thus represent attractive candidates for neofunctional origins. In the case of *mir-281*, divergent processing not only shifted the seed of the ancestral miRNA, thereby potentially altering its target specificity, but also changed the miRNA:miRNA* pairing asymmetry, which significantly enhanced expression of the presumptive ancestral miRNA* species (Supplemental Text). We speculate that future drift of the two loci, with one increasingly specialized to produce a mature miRNA from the former star strand, would result in two genes with common ancestry yet no recognizable sequence identity.

Many pairs of apparently unrelated modern miRNA hairpins might have arisen from common ancestors through the processes of subfunctionalization and neofunctionalization. However, common descent is not always easy to identify, and the two mechanisms can be difficult to distinguish from each other even when descent from a common ancestor appears evident. For example, the *mir-4*, the three *mir-9*, and the *mir-79* loci appeared to have all derived from a common ancestor, as did the *mir-5* and the three *mir-6* loci (Lai et al. 2004). However, in each of these cases, the structure of the gene family tree was ambiguous.

Novel protein-coding genes derive from duplication and divergence of an ancestral gene, and the active sites of their products generally evolve within the context of the ancestral tertiary structures. For protein-coding genes, the requirements of transcription, translation, protein folding, and protein function impose a myriad of informational constraints, making the completely *de novo* evolution of novel protein-coding genes highly improbable and therefore exceedingly rare. MicroRNAs, in contrast, have much more limited requirements. They must be transcribed, and the subsequent transcript must be capable of folding into a secondary structure that is competent for Drosha/Dicer processing (Han et al. 2006). The secondary structure requirements imposed on miRNA hairpin precursors are not excessively stringent, with a wide variety of bulge distributions, hairpin lengths, and loop sizes tolerated among the miRNAs of any given organism. The minimal informational requirements for miRNA-target interactions make it likely that any expressed miRNA will have a physiological consequence, enabling a young miRNA to find a selectively advantageous physiological role. Perhaps the most difficult obstacle for the emergence of new functional miRNA genes would be

the plethora of co-expressed messages with fortuitous sites in their 3′UTRs whose expression would be dampened as the expression of the emergent miRNA became consequential.

The genomic contexts of many miRNAs, including many of the youngest (most narrowly conserved) miRNAs described here, suggested that the pliancy of the miRNA processing machinery facilitates the emergence of new miRNA genes. Most derived from introns or miRNA clusters. In either of those contexts, a miRNA gene can emerge from otherwise unconstrained portions of pre-existing transcripts with little or no effect on the other products of those transcripts, thereby circumventing the otherwise required *de novo* acquisition of transcriptional competence. The varying extents of conservation observed within the *mir-972~979* cluster, which was preferentially expressed in the imaginal discs, reflected a variety of ages for the miRNA genes of that transcript (Fig. 5B). The oldest hairpins, *mir-974/976/977*, spanned the *Drosophila* genus, indicating that they are over 30 million years old (Beverley and Wilson 1984). In contrast, the other hairpins of the same cluster appeared to have emerged after the *D. melanogaster/simulans* split, indicating that they are less than 2.5 million years old (Lachaise et al. 1988). The presence of hairpins with intermediate scopes of conservation, limited to the *melanogaster* species group (*mir-975/978*) or complex (*mir-972/973*), implied a model of functional miRNA genes emerging and presumably disappearing with some temporal regularity from the context of this transcript.

Two other miRNA genes that appeared to have emerged after the *D. melanogaster/simulans* split deserved special mention. The first, *mir-984*, expressed a miRNA whose 6-nt seed matched that of the *let-7* miRNA and thus could repress many

of the same target mRNAs (Lewis et al. 2005). Despite their seed identity, *mir-984* and *let-7* shared little sequence identity and had clearly distinct expression profiles (Fig. 6A, Table S2), suggesting that *mir-984* emerged de novo rather than as a paralog of *let-7*. The second gene was *mir-954*, notable for being the first miRNA gene to be identified on the dot chromosome of *D. melanogaster*, chromosome 4 (Fig. 5A). Portions of the euchromatic chromosome 4 exhibit some heterochromatin-like properties such as variegated expression of inserted reporter constructs, and two such sites of variegated expression flank the *mir-954* locus (Riddle and Elgin 2006).

The scope of miRNA genes and targets in flies

Our current tally of confidently identified mRNA genes in *D. melanogaster* stands at 148. These include 74 of the previously annotated genes, 59 novel genes reported in this study, and another 15 novel genes (mir-1003~1017) whose transcripts bypass Drospha processing (Ruby et al. 2007). Forty-five of our top 47 computational predictions and 75 of our 100 predictions were either previously known or newly validated miRNAs (Fig. 1, Table S1). Independent predictions from a parallel effort had even greater specificity (Stark 2007b), which might be attributed to the use of different training sets; the set used here was smaller and included miRNAs annotated in miRBase but whose authenticity is now in doubt (miR-280, and -289), as well as other miRNAs whose 5′ termini appear to have been incorrectly annotated (miR-2a-2, -33, -274, -284, and -303). The high specificity of both approaches implied that very few highly conserved miRNAs remain to be discovered in flies. However, most of the miRNAs identified by our sequencing were missed by both prediction methods because these miRNAs were insufficiently conserved.

Because the less broadly conserved fly miRNAs tended to be expressed at lower levels, it was impossible to use the cloning results to estimate a lower limit on the overall specificity of the computational gene predictions and thereby derive a meaningful upper limit on the number of miRNAs remaining to be identified in flies. Reliable upper limits on miRNA gene numbers face similar constraints in mammals, worms, and plants (Bartel 2004; Rajagopalan et al. 2006; Ruby et al. 2006).

The implication that there are many more miRNAs to be discovered in flies but almost none of them will be widely conserved, relates to the observed correlation between miRNA conservation and breadth of expression (Fig. 6D), which was likely understated here. All of the libraries from which small RNAs were sampled, with the exception of the S2 library, comprised a conglomerate of cell types, and many of the libraries surveyed thick slices of developmental time. The stronger direct correlation between conservation and total magnitude of expression that was observed here and in other systems may imply that the scarce miRNAs were actually expressed in even narrower contexts that contributed only a small fraction to their encompassing libraries. Thus, the remaining undiscovered miRNAs will inhabit niches of increasingly restricted physiological and evolutionary scopes.

Following that conclusion, another observation becomes relevant: given a consistent scope of conservation, the number of predicted targets decreased with more narrow breadth of miRNA expression (Fig. 7F). The regulatory reach of miRNAs, as indicated by the abundance of genes with conserved miRNA target sites, is likely quite vast. However, the as-yet-undiscovered miRNAs appear to have remained hidden thanks to the narrow scopes of their expression. Consequentially, the set of consequential

miRNA targets will likely grow at a diminishing rate relative to the catalogue of fly miRNAs, and our overall picture of the biological reach of miRNAs will likely not change substantially. This being said, the biology of some of the as-yet-undiscovered miRNAs is still likely to be quite interesting. As illustrated by *lsy-6* in *C. elegans* (Johnston and Hobert 2003), a single miRNA expressed in only a few cells and acting on a limited set of targets can make quite a difference to the animal.

Methods

MicroRNA gene prediction is described in Supplemental Text.

Library construction and sequencing. Total RNA was extracted from Canton S *D. melanogaster* and from S2 cells using Trizol. Embryos were collected using a population cage whose food had been changed regularly to minimize egg withholding. Staged collections of 0-1 hr, 2-6 hr, 6-10 hr and 12-24 hr embryos were obtained by culturing at 25°C. First instar larvae were obtained by aging a 0-12 hr embryo collection on a plate for 24 hours. Wandering third instar larvae were collected from vial cultures and rinsed several times in PBS to remove excess food. Total imaginal discs, brains and salivary glands were isolated from wandering instar larvae to make a pooled "disc" preparation. Separate collections of 0-1 day, 1-2 day and 2-4 day pupae were prepared and pooled to make a pupal library. Equal numbers of 1- to 5-day-old adult female and male flies were frozen at -80°C, vortexed, and sieved onto dry ice blocks to obtain adult head and body fractions. S2 cells were grown in Schneider's medium and rinsed several times in PBS prior to extraction. A cDNA library was generated from each RNA sample as described

(Lau et al. 2001) and was prepared for high-throughput pyrophosphate sequencing (Margulies et al. 2005) as described for run 4 of Ruby et al. (2006). Each library underwent a single sequencing run except for the 2-6 hr. embryo library, which underwent two sequencing runs. A total of 2,514,465 reads were generated. The processing of sequencing data is described (Supplemental Text).

Expression analysis is described in Supplemental Text.

Target Prediction and Analysis. For each miRNA we defined two 7mer motifs that corresponded to the two types of 7mers matching the seed region (the Watson-Crick match to miRNA nucleotides 2-8 and the Watson-Crick match to miRNA nucleotides 2-7 followed by an A). All occurrences of the two motifs were identified within annotated *D. melanogaster* 3′ UTRs from FlyBase Release 4.3 (Crosby et al. 2007), and the conservation of each of these sites was assessed using whole-genome alignments of *D. melanogaster* and 11 additional *Drosophila* species (Schwartz and Pachter 2007). To allow for alignment errors or gaps, sites were scored as conserved if they fell within 50 nt of the aligned site in each informant species. For each site, evolutionary conservation was evaluated as the total branch length corresponding to its species distribution as described (Stark 2007a). A site was considered conserved if this branch length representing the subset of species containing the site met the specified cut-off. To prevent double-counting of 8mer sites that contained both of the two 7mers, target-prediction results reported non-overlapping sites, obtained by first removing sites that did not meet the

specified conservation cut-off and then removing overlapping sites, such that the maximum possible number of non-overlapping sites was retained.

To estimate the conservation expected by chance, we repeated the target-prediction analyses using control motifs and compared the conservation frequencies of their sites with the conservation frequencies of sites obtained for the authentic miRNA. For each miRNA, 9 controls were generated for each of the two motifs. For the motif matching miRNA nucleotides 2-8, each control had equal nucleotide composition and a similar number of matches in *D. melanogaster* 3′ UTRs (deviation < 15%) as the authentic motif. The last six nucleotides of these controls were each extended by an A to obtain the nine controls for the other motif. Signal-to-noise ratios were calculated for each individual miRNA by dividing the frequency of conservation for the authentic sites by the average frequency of the control sites. Signal-to-noise ratios for all miRNAs combined were calculated in the same manner, aggregating all sites for all miRNAs under consideration and their controls. In each case, the number of conserved sites expected by chance was determined by multiplying the total number of sites by the control conservation frequency. Confidence was defined as the number of conserved sites above those expected by chance (i.e., above noise) divided by the total number of conserved sites. Confidence reflected the likelihood of a single conserved site being under selection.

For analysis of expression breadth versus number of predicted targets, miRNAs whose conservation did not extend beyond the *Sophophora* subgenus were not considered. A set of narrowly expressed miRNAs was defined as those with >70% of their library-normalized reads deriving from a single library, and a set of broadly expressed miRNAs as those with no more than 25% of their library-normalized reads

deriving from a single library. Each set was collapsed into families based on their 6 nt seed, resulting in 17 narrowly expressed families and 19 broadly expressed families. The number of predicted targets was determined for each family in each set by requiring targets to be conserved across 70% of the available branch length. In cases where the number of predicted target genes differed among family members because of differences at microRNA nucleotide 8 (which changes one of the two 7mer sites), the largest number of predicted target genes for the family was used.

Figure legends

Figure 1. Performance of miRNA gene prediction.

(**A**) The summed pairwise scores across all 15 two-species comparisons for each miRNA hairpin candidate. Those candidates overlapping the training, test, newly identified, and unvalidated sets of miRNA hairpins are colored as indicated in the key (right) and listed (Table S1).

(**B**) The candidate loci, following strand collapse and exon filtering, depicted as in (A). The top 100 candidates, which had scores above 698, were carried forward as the set of computational gene predictions (Table S1). Of the remaining candidates, only a few were likely to be authentic miRNAs.

(**C**) Specificity of the 100 predictions. Plotted are the number of predicted loci that were validated, the number that correctly identified the strand of the miRNA gene, and the number that correctly identified the miRNA 5′ end (Table S1), colored as in (A).

(**D**) The overlap of the 100 predicted miRNA loci with the training set, test set, and newly identified miRNA loci. Two loci from the training set and two from the test set were not validated by sequencing (red).

Figure 2. Correspondence between previously annotated miRNA hairpins and sequenced miRNAs.

(**A**) Overlap between previously annotated miRNA hairpins and the total set of 133 hairpins of canonical miRNAs supported by our high-throughput sequencing (Table S2). Mirtronic loci are described elsewhere (Ruby et al. 2007).

(**B**) Small RNAs derived from the *mir-7* hairpin. A portion of the *mir-7* transcript is shown above its bracket-notation secondary structure, mature miRNA annotation from miRBase v8.1 (Griffiths-Jones 2004) flanked by asterisks, and sequences from the current study. For each sequence, the number of reads giving rise to that sequence and the number of loci to which the sequence maps in the *D. melanogaster* genome are shown on the right. Highlighted are the most abundant sequences corresponding to the miRNA (red), miRNA* (blue), intervening loop (green), and fragment flanking the 5′ Drosha cleavage site (orange, Supplemental Text). Analogous data for all previously annotated *D. melanogaster* miRNAs is provided (Table S2).

(**C**) The predicted hairpin structure of the *mir-7* hairpin, colored as in (B). Lines indicate inferred Drosha and Dicer cleavage sites.

(**D**) Small RNAs derived from the *mir-iab-4* and *mir-iab4as* hairpins, displayed as in (B).

(**E**) The predicted secondary structure of the sense *mir-iab-4* hairpin precursor, formatted as in (C).

(**F**) The predicted secondary structure of the *mir-iab-4* reverse complement, *mir-iab4as*, formatted as in (C).

Figure 3. Expression and conservation of *mir-10*.

(**A**) The sequence and bracket-notation secondary structure of the *mir-10* hairpin, highlighting the mature miR-10-5p (blue) and the mature miR-10-3p (red), with read abundance along the length of the sequence plotted above and orthologous hairpins aligned below. Nucleotides differing from the *D. melanogaster* identities are in grey. Vertical lines indicate the edges of the 6-nt seed of each mature RNA.

(**B**) The *mir-10* hairpin predicted secondary structure, colored as in (A). Horizontal lines indicate the inferred Drosha and Dicer cleavage sites.

Figure 4. Newly identified miRNAs.

(**A**) The sequence and bracket-notation secondary structure of the *mir-988* hairpin, highlighting the miRNA (red) and the miRNA* (blue), with read abundance along the length of the sequence plotted above and orthologous hairpins aligned below (nonconserved nucleotides in gray) (Consortium 2007a; Consortium 2007b). Vertical lines indicate the inferred Drosha and Dicer cleavage sites. Analogous data for all newly identified *D. melanogaster* miRNAs is provided (Table S2).

(**B**) The predicted secondary structure of the *mir-988* hairpin, colored as in (A). Horizontal lines indicate the inferred Drosha and Dicer cleavage sites.

(**C**) The unusually large hairpin of *mir-989*, colored as in (A).

(D) The sequence and bracket-notation secondary structure of the *mir-989* hairpin, with coloring and read-abundance display as in (A). Conservation across the length of the hairpin is shown below as a histogram, with bar depth indicating for each nucleotide the number of orthologs from the organisms shown in (A) with that nucleotide conserved.

Figure 5. Genomic landscape of miRNA genes.

(A) The distribution of miRNA genes and clusters across the *D. melanogaster* genome, with newly identified miRNAs indicated (red). Euchromatic portions of the genome are drawn to scale, with (+) strand annotations marked above each chromosome and (–) strand annotations marked below. MicroRNA gene clusters, listed together (with gene numbers separated by slashes), were each defined as series of miRNA loci on the same strand of a given chromosome with no intervening gaps greater than 10 kb.

(B) Genomic arrangement and conservation of members of the *mir-972~979* cluster. Detection of an ortholog in the specified species is indicated (black box).

(C) Genomic arrangement of the *mir-310* cluster. Expression profiles among the constituent miRNAs of each labeled sub-cluster indicated that the two sub-clusters were expressed independently (Fig. 6E).

Figure 6. Expression of *D. melanogaster* miRNAs.

(A) The expression profiles of the *D. melanogaster* miRNAs across the ten libraries (left) and total level of expression (right). For each library, miRNA reads are normalized to the total reads deriving from miRNA hairpins in that library. Increasing red color intensity indicates an increasing percentage of normalized reads deriving from that library. Read

counts and normalized counts for each miRNA in each library are provided (Tables S3 and S4). The summed normalized expressions across all ten libraries are shown on the right; units are number of miRNA reads per 100,000 total miRNA hairpin reads per library. Tree and image on left were generated using publicly available software packages Cluster (Eisen et al. 1998) and MapleTree (L. Simirenko).

(**B**) The expression profiles following normalization of four miRNAs whose profiles can be compared to those determined by stage-specific northern blot (Aravin et al. 2003).

(**C**) The relationship between miRNA conservation and magnitude of total expression. MicroRNAs were separated into two groups based on whether they were conserved (Cons.) or not conserved (Not cons.) beyond the subgenus *Sophophora*. Black bars indicate the median expression for each category; red bars indicate the 25^{th} and 75^{th} percentiles. Total expression is defined as in (A).

(**D**) The relationship between conservation and breadth of expression, portrayed as in (C). The y-axis indicates the maximum percentage of expression for a given miRNA derived from a single library.

(**E**) The relationship between the genomic distances separating miRNAs and the correlation of their expressions. Each point represents a pair of miRNAs from (A), including all pairs from the same strand of the same chromosome, but excluding those that can be attributed to multiple genomic loci. The x-axis indicates the distance between the mature miRNAs in nucleotides. The y-axis indicates the Pearson correlation coefficient between the normalized expression patterns of the two miRNAs, as displayed in (A). The red dots represent miR-991 or miR-992 paired with members of the miR-

310~313 cluster, and miR-283 paired with miR-12/304. Despite their proximity, these subclusters appeared to be expressed independently.

Figure 7. MicroRNA target predictions.

(**A**) Confidence of miRNA target prediction versus phylogenic branch length over which sites were conserved in the *Drosophila* genus. Confidence increased with branch length within 12 *Drosophila* species (blue line). Confidence versus branch length values for the following fixed sets of species, strictly requiring conservation in every species, are shown as dots of the indicated colors. Green: seven species used by Grun et al (Grun et al. 2005) (*D. melanogaster, erecta, yakuba, ananassae, pseudoobscura, mojavensis, virilis*). Orange: members of the *Sophophora* subgenus (*D. melanogaster, sechellia, simulans, erecta, yakuba, ananassae, persimilis, pseudoobscura, willistoni*). Red: members of the *melanogaster* subgroup (*D. melanogaster, sechellia, simulans, erecta, yakuba, ananassae*). Purple: *D. melanogaster and pseudoobscura* only (Enright et al. 2003; Stark et al. 2003).

(**B**) Sensitivity of target prediction, shown as the average number of sites per conserved miRNA, versus confidence threshold. Colored as in (A). Note that strict conservation requirements cannot accommodate reduced confidence thresholds, as illustrated by dashed lines.

(**C**) Average number of retained target sites per miRNA for each analysis depicted in (A) and (B) at a confidence threshold of 0.5, colored as in (A).

(**D**) The number of miRNAs and miRNA families with targets above a confidence threshold of 0.5. Numbers for miRNAs from miRBase v8.1 (Griffiths-Jones 2004) are compared to those for our expanded/corrected set of miRNA annotations.
(**E**) Change to the scope of the predicted miRNA-target network (left) and set of genes predicted to be targeted by miRNAs (right) as a result of miRNA annotation additions and changes. Target-miRNA pairs and target genes identified based on miRBase v8.1 annotations (Griffiths-Jones 2004) are in blue; those based on the expanded/corrected set of miRNA annotations provided by the current study are in red.
(**F**) Specifically expressed miRNAs had fewer predicted targets than did broadly expressed miRNAs. Sets of the most broadly and narrowly expressed miRNAs were collapsed into families based on 6-nt seeds, including only miRNAs conserved beyond the *Sophophora* subgenus. The number of predicted targets for each family was set to the maximum number of predicted targets of any family member. The median (black bars) and 25^{th} and 75^{th} percentiles (red bars) of the number of targets per miRNA family are indicated for each set.

Figure 8. Three models for the genesis of miRNA genes. Blue bars represent ancestral miRNAs; orange bars represent novel miRNAs. (**A**) An example of subfunctionalization: a miRNA* acquires function; following gene duplication, one daughter copy maintains the function of the original miRNA while the other maintains the function of the former miRNA*. Another example of subfunctionalization begins with heterologous 5′ processing. (**B**) Neofunctionalization: a miRNA gene duplicates; one daughter copy maintains the function of the original miRNA while the other accumulates mutations that

confer novel functionality to either the former miRNA or miRNA*. (C) *De novo* gene emergence: an unselected portion of a pre-existing transcript, such as an intron or part of a pri-miRNA, acquires the capacity to fold into a hairpin that can be processed into a mature miRNA. That product is selectively maintained due to the fortuitous benefit of gene silencing guided by its seed.

Acknowledgements

We thank Ann Hammonds, Michael Axtell, and Gerald Rubin for assistance and support during library construction, Joseph Rodriguez, Robin Ge, Katherine Gurdziel, George Bell, and Fran Lewitter for constructing the TargetScanFly database of predicted targets (targetscan.org), and Ramya Rajagopalan for comments on the manuscript. Supported by a grant from the NIH (D.P.B.). D.P.B is a HHMI Investigator. E.C.L. was supported by grants from the Burroughs Wellcome Foundation, the Leukemia and Lymphoma Society, and the V Foundation for Cancer Research.

References

Abbott, A.L., E. Alvarez-Saavedra, E.A. Miska, N.C. Lau, D.P. Bartel, H.R. Horvitz, and V. Ambros. 2005. The *let-7* MicroRNA family members *mir-48*, *mir-84*, and *mir-241* function together to regulate developmental timing in *Caenorhabditis elegans*. *Dev Cell* **9:** 403-414.

Adams, M.D. S.E. Celniker R.A. Holt C.A. Evans J.D. Gocayne P.G. Amanatides S.E. Scherer P.W. Li R.A. Hoskins R.F. Galle R.A. George S.E. Lewis S. Richards M. Ashburner S.N. Henderson G.G. Sutton J.R. Wortman M.D. Yandell Q. Zhang L.X. Chen R.C. Brandon Y.H. Rogers R.G. Blazej M. Champe B.D. Pfeiffer K.H. Wan C. Doyle E.G. Baxter G. Helt C.R. Nelson G.L. Gabor J.F. Abril A. Agbayani H.J. An C. Andrews-Pfannkoch D. Baldwin R.M. Ballew A. Basu J. Baxendale L. Bayraktaroglu E.M. Beasley K.Y. Beeson P.V. Benos B.P. Berman D. Bhandari S. Bolshakov D. Borkova M.R. Botchan J. Bouck P. Brokstein P. Brottier K.C. Burtis D.A. Busam H. Butler E. Cadieu A. Center I. Chandra J.M. Cherry S. Cawley C. Dahlke L.B. Davenport P. Davies B. de Pablos A. Delcher

Z. Deng A.D. Mays I. Dew S.M. Dietz K. Dodson L.E. Doup M. Downes S. Dugan-Rocha B.C. Dunkov P. Dunn K.J. Durbin C.C. Evangelista C. Ferraz S. Ferriera W. Fleischmann C. Fosler A.E. Gabrielian N.S. Garg W.M. Gelbart K. Glasser A. Glodek F. Gong J.H. Gorrell Z. Gu P. Guan M. Harris N.L. Harris D. Harvey T.J. Heiman J.R. Hernandez J. Houck D. Hostin K.A. Houston T.J. Howland M.H. Wei C. Ibegwam M. Jalali F. Kalush G.H. Karpen Z. Ke J.A. Kennison K.A. Ketchum B.E. Kimmel C.D. Kodira C. Kraft S. Kravitz D. Kulp Z. Lai P. Lasko Y. Lei A.A. Levitsky J. Li Z. Li Y. Liang X. Lin X. Liu B. Mattei T.C. McIntosh M.P. McLeod D. McPherson G. Merkulov N.V. Milshina C. Mobarry J. Morris A. Moshrefi S.M. Mount M. Moy B. Murphy L. Murphy D.M. Muzny D.L. Nelson D.R. Nelson K.A. Nelson K. Nixon D.R. Nusskern J.M. Pacleb M. Palazzolo G.S. Pittman S. Pan J. Pollard V. Puri M.G. Reese K. Reinert K. Remington R.D. Saunders F. Scheeler H. Shen B.C. Shue I. Siden-Kiamos M. Simpson M.P. Skupski T. Smith E. Spier A.C. Spradling M. Stapleton R. Strong E. Sun R. Svirskas C. Tector R. Turner E. Venter A.H. Wang X. Wang Z.Y. Wang D.A. Wassarman G.M. Weinstock J. Weissenbach S.M. Williams WoodageT K.C. Worley D. Wu S. Yang Q.A. Yao J. Ye R.F. Yeh J.S. Zaveri M. Zhan G. Zhang Q. Zhao L. Zheng X.H. Zheng F.N. Zhong W. Zhong X. Zhou S. Zhu X. Zhu H.O. Smith R.A. Gibbs E.W. Myers G.M. Rubin and J.C. Venter. 2000. The genome sequence of *Drosophila melanogaster*. *Science* **287:** 2185-2195.

Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. 1990. Basic local alignment search tool. *J Mol Biol* **215:** 403-410.

Ambros, V., R.C. Lee, A. Lavanway, P.T. Williams, and D. Jewell. 2003. MicroRNAs and other tiny endogenous RNAs in *C. elegans*. *Curr Biol* **13:** 807-818.

Aravin, A.A., M. Lagos-Quintana, A. Yalcin, M. Zavolan, D. Marks, B. Snyder, T. Gaasterland, J. Meyer, and T. Tuschl. 2003. The small RNA profile during *Drosophila melanogaster* development. *Dev Cell* **5:** 337-350.

Bartel, D.P. 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116:** 281-297.

Baskerville, S. and D.P. Bartel. 2005. Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA* **11:** 241-247.

Basyuk, E., F. Suavet, A. Doglio, R. Bordonne, and E. Bertrand. 2003. Human *let-7* stem-loop precursors harbor features of RNase III cleavage products. *Nucleic Acids Res* **31:** 6593-6597.

Bentwich, I., A. Avniel, Y. Karov, R. Aharonov, S. Gilad, O. Barad, A. Barzilai, P. Einat, U. Einav, E. Meiri, E. Sharon, Y. Spector, and Z. Bentwich. 2005. Identification of hundreds of conserved and nonconserved human microRNAs. *Nat Genet* **37:** 766-770.

Berezikov, E., V. Guryev, J. van de Belt, E. Wienholds, R.H. Plasterk, and E. Cuppen. 2005. Phylogenetic shadowing and computational identification of human microRNA genes. *Cell* **120:** 21-24.

Berezikov, E., F. Thuemmler, L.W. van Laake, I. Kondova, R. Bontrop, E. Cuppen, and R.H. Plasterk. 2006. Diversity of microRNAs in human and chimpanzee brain. *Nat Genet* **38:** 1375-1377.

Beverley, S.M. and A.C. Wilson. 1984. Molecular evolution in *Drosophila* and the higher Diptera II. A time scale for fly evolution. *J Mol Evol* **21:** 1-13.

Brennecke, J., A. Stark, R.B. Russell, and S.M. Cohen. 2005. Principles of microRNA-target recognition. *PLoS Biol* **3:** e85.

Consortium, D.C.G.S.a.A. 2007a. Evolution of Genes and Genomes in the Genus *Drosophila. Nature* **In preparation.**

Consortium, D.C.G.S.a.A. 2007b. Initial comparative genomics analysis of 12 *Drosophila*
genomes. *Nature* **In preparation.**

Crosby, M.A., J.L. Goodman, V.B. Strelets, P. Zhang, and W.M. Gelbart. 2007. FlyBase: genomes by the dozen. *Nucleic Acids Res* **35:** D486-491.

Cumberledge, S., A. Zaratzian, and S. Sakonju. 1990. Characterization of two RNAs transcribed from the cis-regulatory region of the abd-A domain within the *Drosophila* bithorax complex. *Proc Natl Acad Sci U S A* **87:** 3259-3263.

Doench, J.G. and P.A. Sharp. 2004. Specificity of microRNA target selection in translational repression. *Genes Dev* **18:** 504-511.

Eisen, M.B., P.T. Spellman, P.O. Brown, and D. Botstein. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* **95:** 14863-14868.

Enright, A.J., B. John, U. Gaul, T. Tuschl, C. Sander, and D.S. Marks. 2003. MicroRNA targets in *Drosophila. Genome Biol* **5:** R1.

Fahlgren, N., M.D. Howell, K.D. Kasschau, E.J. Chapman, C.M. Sullivan, J.S. Cumbie, S.A. Givan, T.F. Law, S.R. Grant, J.L. Dangl, and J.C. Carrington. 2007. High-Throughput Sequencing of *Arabidopsis* microRNAs: Evidence for Frequent Birth and Death of MIRNA Genes. *PLoS ONE* **2:** e219.

Farh, K.K., A. Grimson, C. Jan, B.P. Lewis, W.K. Johnston, L.P. Lim, C.B. Burge, and D.P. Bartel. 2005. The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. *Science* **310:** 1817-1821.

Force, A., M. Lynch, F.B. Pickett, A. Amores, Y.L. Yan, and J. Postlethwait. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151:** 1531-1545.

Grad, Y., J. Aach, G.D. Hayes, B.J. Reinhart, G.M. Church, G. Ruvkun, and J. Kim. 2003. Computational and experimental identification of *C. elegans* microRNAs. *Mol Cell* **11:** 1253-1263.

Griffiths-Jones, S. 2004. The microRNA Registry. *Nucleic Acids Res* **32:** D109-111.

Grun, D., Y.L. Wang, D. Langenberger, K.C. Gunsalus, and N. Rajewsky. 2005. microRNA target predictions across seven *Drosophila* species and comparison to mammalian targets. *PLoS Comput Biol* **1:** e13.

Han, J., Y. Lee, K.H. Yeom, J.W. Nam, I. Heo, J.K. Rhee, S.Y. Sohn, Y. Cho, B.T. Zhang, and V.N. Kim. 2006. Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. *Cell* **125:** 887-901.

Hofacker, I.L., W. Fontana, P.F. Stadler, L.S. Bonhoeffer, M. Tacker, and P. Schuster. 1994. Fast folding and comparison of RNA secondary structures. *Monatshefte Fur Chemie* **125:** 167-188.

Johnston, R.J. and O. Hobert. 2003. A microRNA controlling left/right neuronal asymmetry in *Caenorhabditis elegans. Nature* **426:** 845-849.

Karolchik, D., R. Baertsch, M. Diekhans, T.S. Furey, A. Hinrichs, Y.T. Lu, K.M. Roskin, M. Schwartz, C.W. Sugnet, D.J. Thomas, R.J. Weber, D. Haussler, and W.J. Kent. 2003. The UCSC Genome Browser Database. *Nucleic Acids Res* **31:** 51-54.

Khvorova, A., A. Reynolds, and S.D. Jayasena. 2003. Functional siRNAs and miRNAs exhibit strand bias. *Cell* **115:** 209-216.

Krek, A., D. Grun, M.N. Poy, R. Wolf, L. Rosenberg, E.J. Epstein, P. MacMenamin, I. da Piedade, K.C. Gunsalus, M. Stoffel, and N. Rajewsky. 2005. Combinatorial microRNA target predictions. *Nat Genet* **37:** 495-500.

Lachaise, D., M.L. Cariou, J.R. David, F. Lemeunier, L. Tsacas, and M. Ashburner. 1988. Historical Biogeography of the *Drosophila melanogaster* Species Subgroup. *Evolutionary Biology* **22:** 159-225.

Lagos-Quintana, M., R. Rauhut, W. Lendeckel, and T. Tuschl. 2001. Identification of novel genes coding for small expressed RNAs. *Science* **294:** 853-858.

Lai, E.C. 2002. Micro RNAs are complementary to 3' UTR sequence motifs that mediate negative post-transcriptional regulation. *Nat Genet* **30:** 363-364.

Lai, E.C., B. Tam, and G.M. Rubin. 2005. Pervasive regulation of *Drosophila* Notch target genes by GY-box-, Brd-box-, and K-box-class microRNAs. *Genes Dev* **19:** 1067-1080.

Lai, E.C., P. Tomancak, R.W. Williams, and G.M. Rubin. 2003. Computational identification of *Drosophila* microRNA genes. *Genome Biol* **4:** R42.

Lai, E.C., C. Wiel, and G.M. Rubin. 2004. Complementary miRNA pairs suggest a regulatory role for miRNA:miRNA duplexes. *RNA* **10:** 171-175.

Lau, N.C., L.P. Lim, E.G. Weinstein, and D.P. Bartel. 2001. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans. Science* **294:** 858-862.

Lee, R.C., R.L. Feinbaum, and V. Ambros. 1993. The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14. Cell* **75:** 843-854.

Lee, Y., C. Ahn, J. Han, H. Choi, J. Kim, J. Yim, J. Lee, P. Provost, O. Radmark, S. Kim, and V.N. Kim. 2003. The nuclear RNase III Drosha initiates microRNA processing. *Nature* **425:** 415-419.

Lewis, B.P., C.B. Burge, and D.P. Bartel. 2005. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120:** 15-20.

Lewis, B.P., I.H. Shih, M.W. Jones-Rhoades, D.P. Bartel, and C.B. Burge. 2003. Prediction of mammalian microRNA targets. *Cell* **115:** 787-798.

Li, J., Z. Yang, B. Yu, J. Liu, and X. Chen. 2005. Methylation protects miRNAs and siRNAs from a 3'-end uridylation activity in *Arabidopsis. Curr Biol* **15:** 1501-1507.

Lim, L.P., M.E. Glasner, S. Yekta, C.B. Burge, and D.P. Bartel. 2003a. Vertebrate microRNA genes. *Science* **299:** 1540.

Lim, L.P., N.C. Lau, P. Garrett-Engele, A. Grimson, J.M. Schelter, J. Castle, D.P. Bartel, P.S. Linsley, and J.M. Johnson. 2005. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* **433:** 769-773.

Lim, L.P., N.C. Lau, E.G. Weinstein, A. Abdelhakim, S. Yekta, M.W. Rhoades, C.B. Burge, and D.P. Bartel. 2003b. The microRNAs of *Caenorhabditis elegans*. *Genes Dev* **17:** 991-1008.

Lu, C., K. Kulkarni, F.F. Souret, R. MuthuValliappan, S.S. Tej, R.S. Poethig, I.R. Henderson, S.E. Jacobsen, W. Wang, P.J. Green, and B.C. Meyers. 2006. MicroRNAs and other small RNAs enriched in the *Arabidopsis* RNA-dependent RNA polymerase-2 mutant. *Genome Res* **16:** 1276-1288.

Margulies, M., M. Egholm, W.E. Altman, S. Attiya, J.S. Bader, L.A. Bemben, J. Berka, M.S. Braverman, Y.J. Chen, Z. Chen, S.B. Dewell, L. Du, J.M. Fierro, X.V. Gomes, B.C. Godwin, W. He, S. Helgesen, C.H. Ho, G.P. Irzyk, S.C. Jando, M.L. Alenquer, T.P. Jarvie, K.B. Jirage, J.B. Kim, J.R. Knight, J.R. Lanza, J.H. Leamon, S.M. Lefkowitz, M. Lei, J. Li, K.L. Lohman, H. Lu, V.B. Makhijani, K.E. McDade, M.P. McKenna, E.W. Myers, E. Nickerson, J.R. Nobile, R. Plant, B.P. Puc, M.T. Ronan, G.T. Roth, G.J. Sarkis, J.F. Simons, J.W. Simpson, M. Srinivasan, K.R. Tartaro, A. Tomasz, K.A. Vogt, G.A. Volkmer, S.H. Wang, Y. Wang, M.P. Weiner, P. Yu, R.F. Begley, and J.M. Rothberg. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437:** 376-380.

Ohler, U., S. Yekta, L.P. Lim, D.P. Bartel, and C.B. Burge. 2004. Patterns of flanking sequence conservation and a characteristic upstream motif for microRNA gene identification. *Rna* **10:** 1309-1322.

Ohno, S. 1970. *Evolution by gene duplication*. Springer-Verlag, Berlin, New York,.

Rajagopalan, R., H. Vaucheret, J. Trejo, and D.P. Bartel. 2006. A diverse and evolutionarily fluid set of microRNAs in *Arabidopsis thaliana*. *Genes Dev* **20:** 3407-3425.

Richards, S., Y. Liu, B.R. Bettencourt, P. Hradecky, S. Letovsky, R. Nielsen, K. Thornton, M.J. Hubisz, R. Chen, R.P. Meisel, O. Couronne, S. Hua, M.A. Smith, P. Zhang, J. Liu, H.J. Bussemaker, M.F. van Batenburg, S.L. Howells, S.E. Scherer, E. Sodergren, B.B. Matthews, M.A. Crosby, A.J. Schroeder, D. Ortiz-Barrientos, C.M. Rives, M.L. Metzker, D.M. Muzny, G. Scott, D. Steffen, D.A. Wheeler, K.C. Worley, P. Havlak, K.J. Durbin, A. Egan, R. Gill, J. Hume, M.B. Morgan, G. Miner, C. Hamilton, Y. Huang, L. Waldron, D. Verduzco, K.P. Clerc-Blankenburg, I. Dubchak, M.A. Noor, W. Anderson, K.P. White, A.G. Clark, S.W. Schaeffer, W. Gelbart, G.M. Weinstock, and R.A. Gibbs. 2005. Comparative genome sequencing of *Drosophila pseudoobscura*: chromosomal, gene, and cis-element evolution. *Genome Res* **15:** 1-18.

Riddle, N.C. and S.C. Elgin. 2006. The dot chromosome of *Drosophila*: insights into chromatin states and their change over evolutionary time. *Chromosome Res* **14:** 405-416.

Ruby, J.G., C. Jan, C. Player, M.J. Axtell, W. Lee, C. Nusbaum, H. Ge, and D.P. Bartel. 2006. Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. *Cell* **127:** 1193-1207.

Ruby, J.G., C.H. Jan, and D.P. Bartel. 2007. Intronic microRNA precursors that bypass Drosha processing. *Nature* **448:** 83-86.

Schwartz, A.S. and L. Pachter. 2007. Multiple alignment by sequence annealing. *Bioinformatics* **23:** e24-29.

Schwarz, D.S., G. Hutvagner, T. Du, Z. Xu, N. Aronin, and P.D. Zamore. 2003. Asymmetry in the assembly of the RNAi enzyme complex. *Cell* **115:** 199-208.

Sempere, L.F., S. Freemantle, I. Pitha-Rowe, E. Moss, E. Dmitrovsky, and V. Ambros. 2004. Expression profiling of mammalian microRNAs uncovers a subset of brain-expressed microRNAs with possible roles in murine and human neuronal differentiation. *Genome Biol* **5:** R13.

Stark, A. 2007a. Genome-wide identification of fly regulatory motifs and their functional roles. *in preparation.*

Stark, A. 2007b. MicroRNA gene prediction. *submitted.*

Stark, A., J. Brennecke, N. Bushati, R.B. Russell, and S.M. Cohen. 2005. Animal microRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution. *Cell* **123:** 1133-1146.

Stark, A., J. Brennecke, R.B. Russell, and S.M. Cohen. 2003. Identification of *Drosophila* microRNA targets. *PLoS Biol* **1:** E60.

Wang, H., R.A. Ach, and B. Curry. 2007. Direct and sensitive miRNA profiling from low-input total RNA. *Rna* **13:** 151-159.

Wightman, B., I. Ha, and G. Ruvkun. 1993. Posttranscriptional regulation of the heterochronic gene *lin-14* by *lin-4* mediates temporal pattern formation in *C. elegans*. *Cell* **75:** 855-862.

Table 1. Newly identified miRNAs in *D. melanogaster*

		Reads				Conserved In						Other family members		
miRNA	**Sequence**	**miRNA**	**miRNA***	**Clusterd**	**Intronic**	***dsi***	***dya***	***dan***	***dps***	***dmo***	***dvi***	***dme***	***cel***	***vert.***
miR-137	UAUUGCUUGAGAAUACACGUAG	48	7			Y	Y	Y	Y	Y	Y			miR-137
miR-190	AGAUAUGUUUGAUAUUCUUGGUUG	513	25		Y	Y	Y	Y	Y	Y	Y		miR-50	miR-190
miR-193	UACUGGCCUACUAAGUCCCAAC	755	44			Y	Y	Y	Y	Y	Y		miR-240	miR-193
miR-252	CUAAGUACUAGUGCCGCAGGAG	7271	145		Y	Y	Y	Y	Y	Y	Y	miR-1002	miR-252	
miR-375	UUUGUUCGUUUGGCUUAAGUUA	339	20			Y	Y	Y	Y	Y	Y			miR-375
miR-927	UUUAGAAUUCCUACGCUUUACC	389	14			Y	Y	Y	Y	Y	Y			
miR-929	CUCCCUAACGGAGUCAGAUUG	119	14		Y	Y	Y	Y	Y	Y	Y			
miR-932	UCAAUUCCGUAGUGCAUUGCAG	616	13		Y	Y	Y	Y	Y	Y	Y			
miR-954	UCUGGGUGUUGCGUUGUGUGU	31	10											
miR-955	CAUCGUGCAGAGGUUUGAGUGUC	14				Y	Y	Y		Y	Y			
miR-956	UUUCGAGACCACUCUAAUCCAUU	109	1			Y	Y	Y	Y	Y	Y			
miR-957	UGAAACCGUCCAAAACUGAGGC	137				Y	Y	Y	Y	Y	Y			
miR-958	UGAGAUUCUUCUAUUCUACUUU	1721	110			Y	Y	Y	Y	Y	Y			
miR-959	UUGUCAUCGGGGGUAUUAUGAA	61	18	Y		Y	Y	Y	Y					
miR-960	UGAGUAUUCCAGAUUGCAUAGC	54	14	Y		Y	Y	Y	Y	Y		miR-12		
miR-961	UUUGAUCACCAGUAACUGAGAU	5	4	Y		Y	Y	Y	Y					
miR-962	AUAAGGUAGAGAAAUUGAUGCUGU	50	9	Y		Y	Y	Y	Y	Y	Y			
miR-963	ACAAGGUAAAUAUCAGGUUGUUUC	92	2	Y		Y	Y	Y		Y	Y			
miR-964	UUAGAAUAGGGGAGCUUAACUU	87	1	Y		Y	Y	Y		Y	Y			
miR-965	UAAGCGUAUAGCUUUUCCCCUU	137	63		Y	Y	Y	Y	Y	Y	Y			
miR-966	UGUGGGUUGUGGGCUGUGUGG	10	2		Y									
miR-967	AGAGAUACCUCUGGAGAAGCG	5	1		Y	Y						miR-977		
miR-968	UAAGUAGUAUCCAUUAAAGGGUUG	84	63	Y		Y	Y	Y	Y		Y		miR-252	miR-562
miR-969	GAGUUCCACUAAGCAAGUUUU	10		Y		Y	Y	Y	Y	Y	Y			
miR-970	UCAUAAGACACACGCGGCUAU	487	22		Y	Y	Y	Y	Y	Y	Y			
miR-971	UUGGUGUUACUUCUUACAGUGA	52	1				Y	Y	Y	Y	Y			miR-333
miR-972	UGUACAAUACGAAUAUUUAGGC	11		Y		Y								
miR-973	UGGUUGGUGGUUGAACUUCGAUUU	21	3	Y		Y								
miR-974	AAGCGAGCAAAGAAGUAGUAUU	4	1	Y		Y		Y			Y			
miR-975	UAAACACUUCCUACAUCCUGUAU	66	2	Y		Y	Y	Y						
miR-976	UUGGAUUAGUUAUCAUCAAUGC	31	1	Y		Y	Y	Y		Y	Y			
miR-977	UGAGAUAUUCACGUUGUCUAA	251	8	Y		Y	Y	Y		Y	Y	miR-967		
miR-978	UGUCCAGUGCCGUAAAUUGCAG	51	6	Y		Y	Y	Y						miR-198
miR-979	UUCUUCCCGAACUCAGGCUAA	1	1	Y										
miR-980	UAGCUGCCUUGUGAAGGGCUUA	197	13			Y	Y	Y	Y	Y	Y			miR-22
miR-981	UUCGUUGUCGACGAAACCUGCA	1744	3			Y	Y	Y	Y	Y	Y		miR-76	
miR-982	UCCUGGACAAAUAUGAAGUAAAU	29	3	Y										
miR-983-1	AUAAUACGUUUCGAACUAAUGA	29	39	Y										miR-655
miR-983-2	AUAAUACGUUUCGAACUAAUGA	29	39	Y										miR-655
miR-984	UGAGGUAAAUACGGUUGGAAUUU	173	6	Y								let-7	let-7	let-7
miR-985	CAAAUGUUCCAAUGGUCGGGCA	14	3			Y								
miR-986	UCUCGAAUAGCGUUGUGACUGA	41	1		Y	Y	Y	Y	Y	Y	Y			
miR-987	UAAAGUAAAUAGUCUGGAUUGAUG	875	4			Y	Y	Y	Y	Y	Y			miR-559
miR-988	CCCCUUGUUGCAAACCUCACGC	1908	46		Y	Y	Y	Y	Y	Y	Y			
miR-989	UGUGAUGUGACGUAGUGGAAC	196	5			Y	Y	Y	Y	Y	Y			
miR-990	AUUCACCGUUCUGAGUUGGCC	13	1		Y	Y	Y							
miR-991	UUAAAGUUGUAGUUUGGAAAGU	28		Y		Y								
miR-992	AGUACACGUUUCUGGUACUAAG	148	2	Y			Y							
miR-993	GAAGCUCGUCUCUACAGGUAUCU	287	7			Y	Y	Y	Y	Y	Y		miR-231	
miR-994	CUAAGGAAAUAGUAGCCGUGAU	233	14	Y		Y	Y	Y	Y	Y	Y			
miR-995	UAGCACCACAUGAUUCGGCUU	1326	88		Y	Y	Y	Y	Y	Y	Y	miR-285	miR-49	miR-29
miR-996	UGACUAGAUUUCAUGCUCGUCU	4509	322	Y		Y	Y	Y	Y	Y	Y	miR-279	miR-44	
miR-997	CCCAAACUCGAAGGAGUUUCA	10	2			Y								
miR-998	UAGCACCAUGAGAUUCAGCUC	519	190	Y	Y	Y	Y	Y	Y	Y	Y	miR-285	miR-49	miR-29
miR-999	UGUUAACUGUAAGACUGUGUCU	420	16		Y	Y	Y	Y	Y	Y	Y			
miR-1000	AUAUUGUCCUGUCACAGCAGU	331	31			Y	Y	Y	Y	Y	Y			
miR-1001	UGGGUAAACUCCCAAGGAUCA	35	2		Y	Y								miR-555
miR-1002	UUAAGUAGUGGAUACAAAGGGCGA	73	55	Y		Y	Y	Y	Y		Y	miR-252	miR-251	

Figure 1

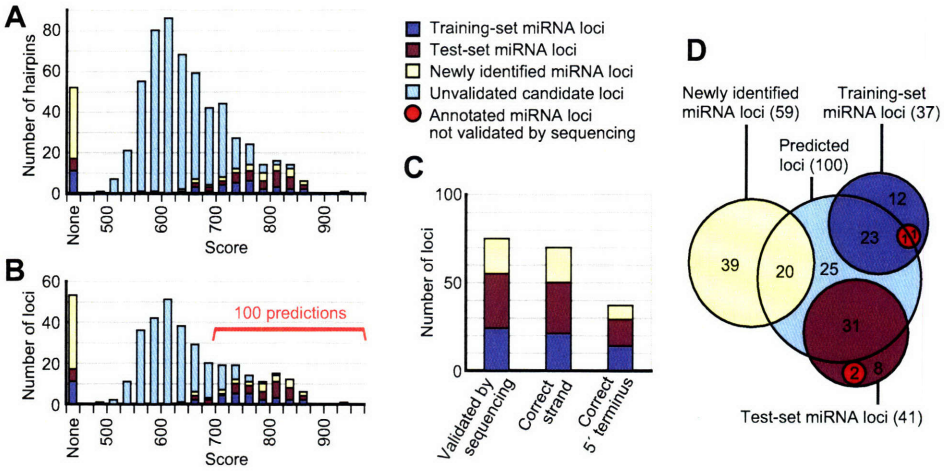

A
Number of hairpins
80
60
40
20
0
None
500
600
700
800
900
Score
B
Number of loci
60
40
20
0
100 predictions
None
500
600
700
800
900
Score
Training-set miRNA loci
Test-set miRNA loci
Newly identified miRNA loci
Unvalidated candidate loci
Annotated miRNA loci not validated by sequencing
C
Number of loci
100
50
0
Validated by sequencing
Correct strand
Correct 5′ terminus
D
Newly identified miRNA loci (59)
Training-set miRNA loci (37)
Predicted loci (100)
12
23
1
39
20
25
31
2
8
Test-set miRNA loci (41)

Figure 2

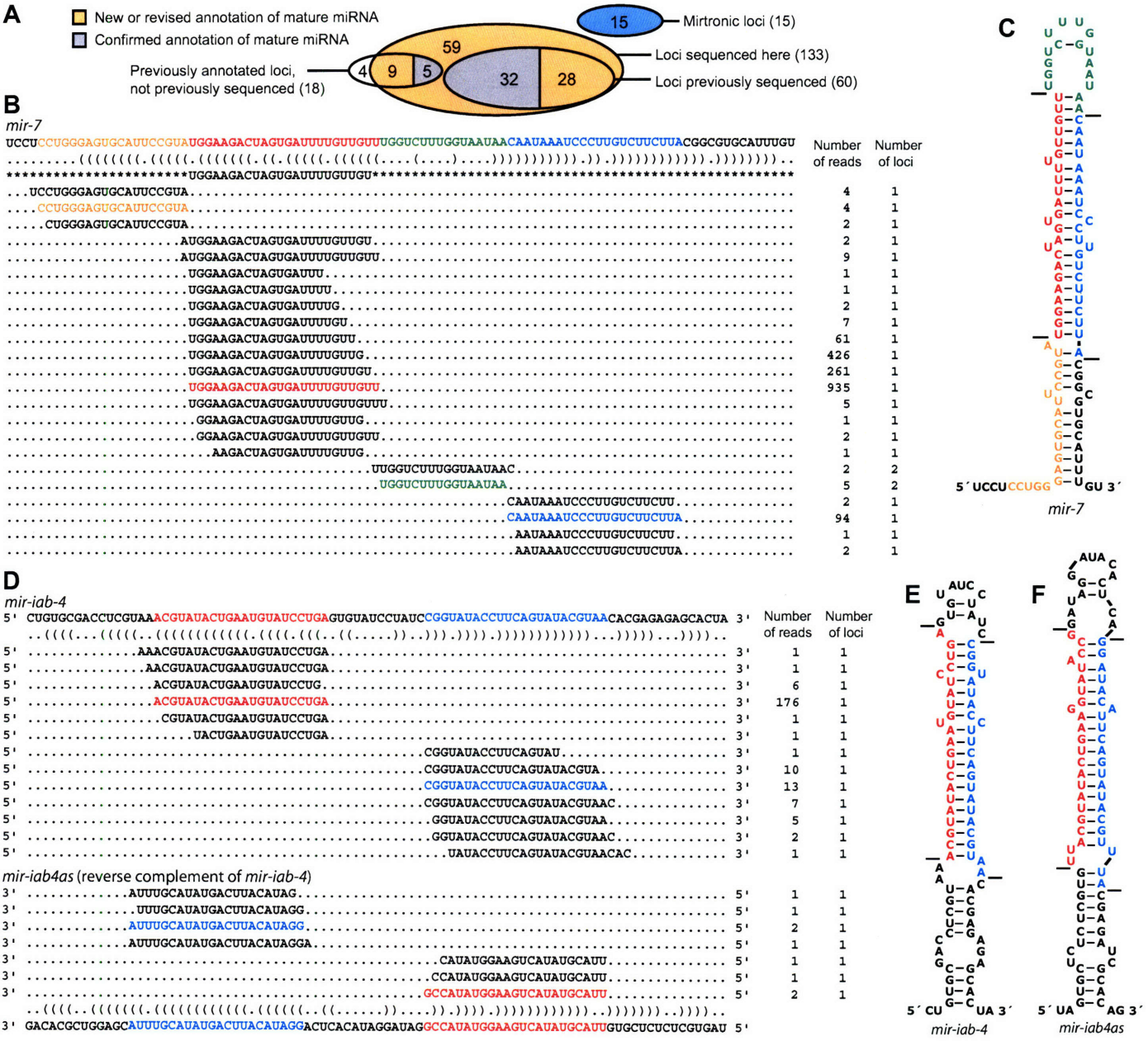

A
New or revised annotation of mature miRNA
Confirmed annotation of mature miRNA
Previously annotated loci, not previously sequenced (18)
4
9
5
59
32
28
15
Mirtronic loci (15)
Loci sequenced here (133)
Loci previously sequenced (60)
B
mir-7
Number of reads
Number of loci
C
5´ UCCUCCUGG G U GU 3´
mir-7
D
mir-iab-4
Number of reads
Number of loci
mir-iab4as (reverse complement of mir-iab-4)
E
5´ CU UA 3´
mir-iab-4
F
5´ UA AG 3´
mir-iab4as

Figure 3

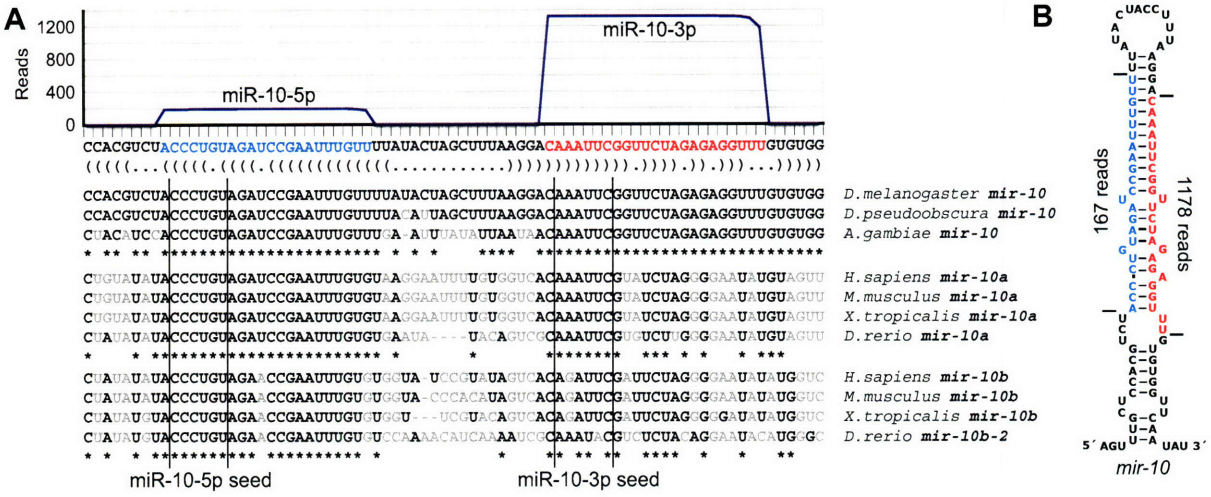

A
Reads
1200
800
400
0
miR-10-5p
miR-10-3p
CCACGUCUACCCUGUAGAUCCGAAUUUGUUUUAUACUAGCUUUAAGGACAAAUUCGGUUCUAGAGAGGUUUGUGUGG
((((((...((((((.((((.(((((((((((((((.....)))))))))))))))).)))))).)).)))...))))))
CCACGUCUACCCUGUAGAUCCGAAUUUGUUUUAUACUAGCUUUAAGGACAAAUUCGGUUCUAGAGAGGUUUGUGUGG D.melanogaster mir-10
CCACGUCUACCCUGUAGAUCCGAAUUUGUUUUACAUUAGCUUUAAGGACAAAUUCGGUUCUAGAGAGGUUUGUGUGG D.pseudoobscura mir-10
CUACAUCCACCCUGUAGAUCCGAAUUUGUUUGAAUUUAUAUUAACACAAAUUCGGUUCUAGAGAGGUUUGUGUGG A.gambiae mir-10
CUGUAUAUACCCUGUAGAUCCGAAUUUGUGUAAGGAAUUUUGUGGUCACAAAUUCGUAUCUAGGGAAUAUGUAGUU H.sapiens mir-10a
CUGUAUAUACCCUGUAGAUCCGAAUUUGUGUAAGGAAUUUUGUGGUCACAAAUUCGUAUCUAGGGAAUAUGUAGUU M.musculus mir-10a
CUGUAUAUACCCUGUAGAUCCGAAUUUGUGUAAGGAAUUUUGUGGUCACAAAUUCGUAUCUAGGGAAUAUGUAGUU X.tropicalis mir-10a
CUAUAUAUACCCUGUAGAUCCGAAUUUGUGUGAAUA----UACAGUCGCAAAUUCGUGUCUUGGGAAUAUGUAGUU D.rerio mir-10a
CUAUAUAUACCCUGUAGAACCGAAUUUGUGUGGUA UCCGUAUAGUCACAGAUUCGAUUCUAGGGAAUAUAUGGUC H.sapiens mir-10b
CUAUAUAUACCCUGUAGAACCGAAUUUGUGUGGUA-CCCACAUAGUCACAGAUUCGAUUCUAGGGAAUAUAUGGUC M.musculus mir-10b
CUAUAUGUACCCUGUAGAACCGAAUUUGUGUGGU---UCGUACAGUCACAGAUUCGAUUCUAGGGGAUAUAUGGUC X.tropicalis mir-10b
CUAUAUGUACCCUGUAGAACCGAAUUUGUGUCCAAAACAUCAAAAUCGCAAAUACGUCUCUACAGGAAUACAUGGGC D.rerio mir-10b-2
miR-10-5p seed
miR-10-3p seed
B
167 reads
1178 reads
5´ AGU
UAU 3´
mir-10

Figure 4

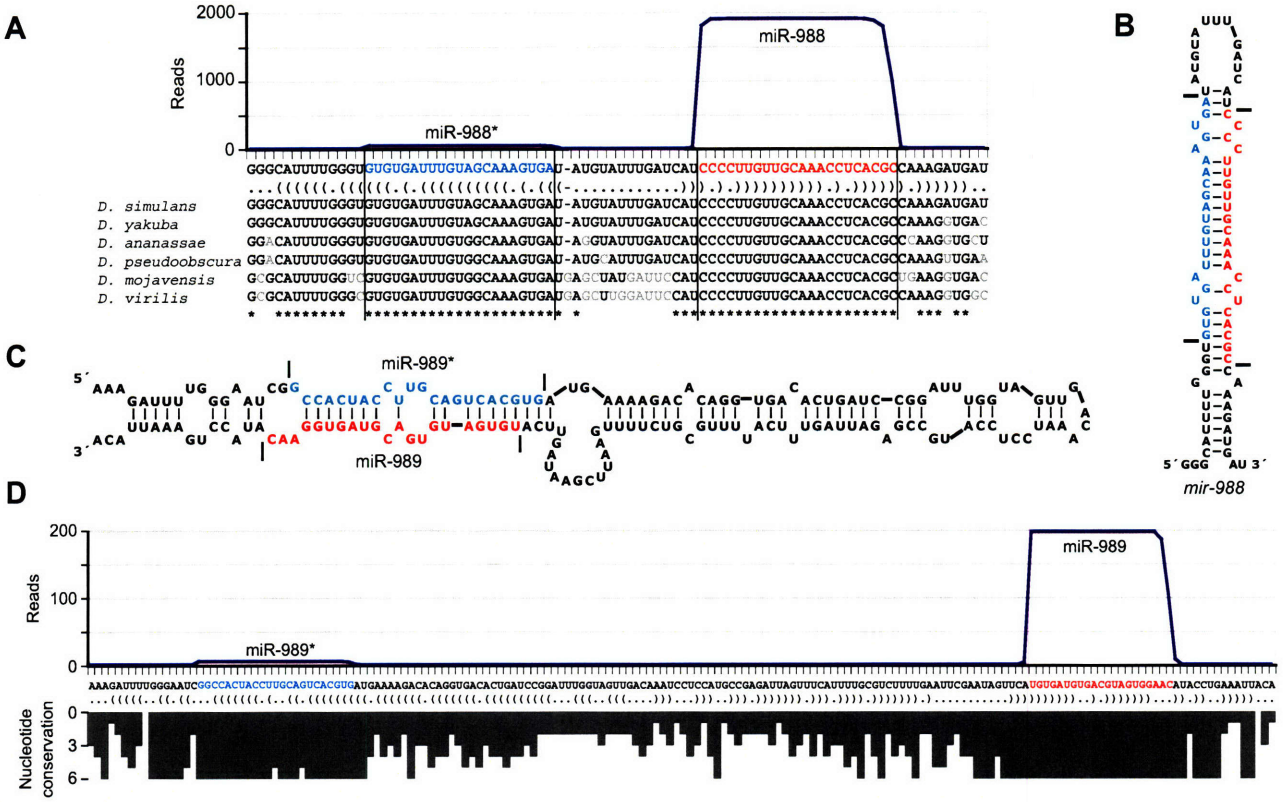

A
2000
1000
0
Reads
miR-988
miR-988*
D. simulans
D. yakuba
D. ananassae
D. pseudoobscura
D. mojavensis
D. virilis
B
mir-988
C
miR-989*
miR-989
D
200
100
0
Reads
miR-989
miR-989*
Nucleotide conservation
0
3
6

Figure 5

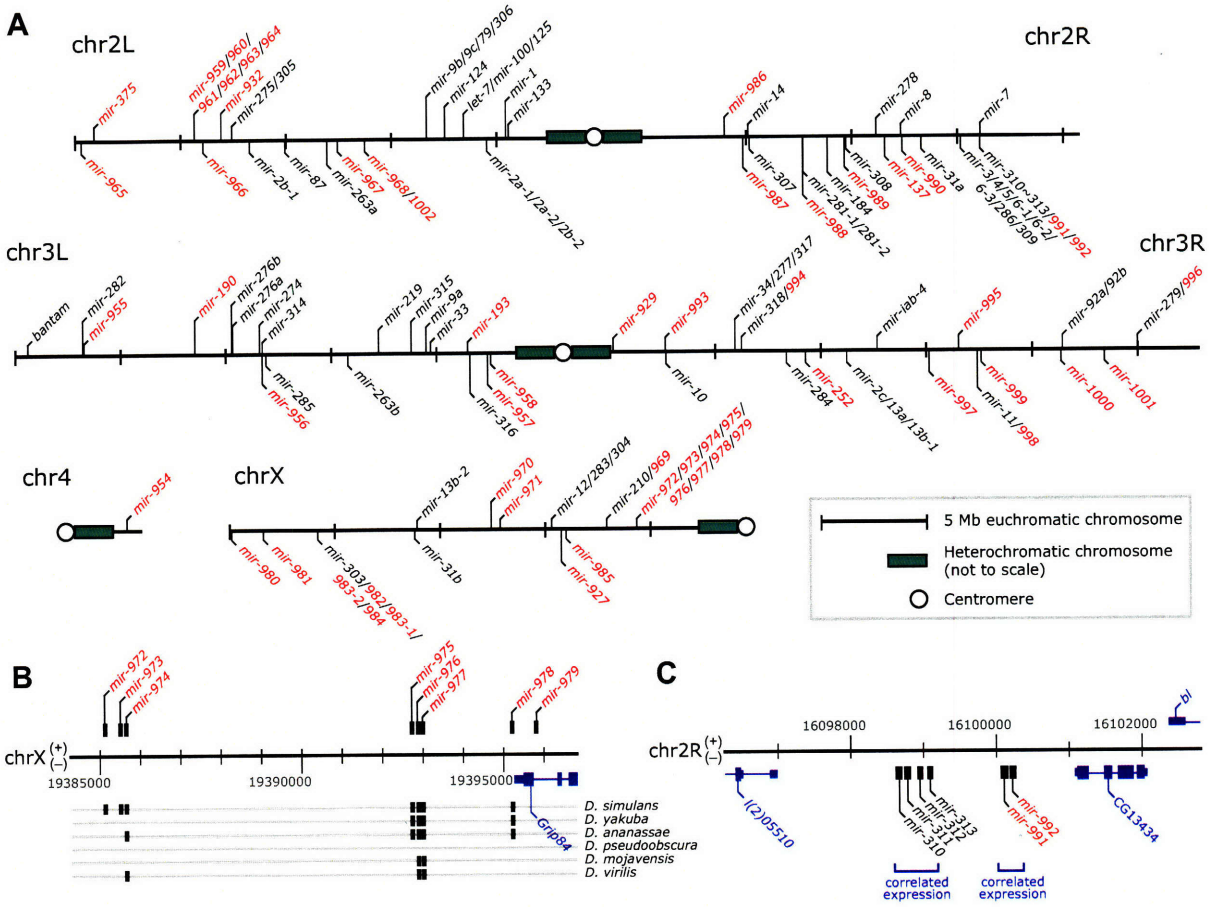

A
chr2L
chr2R
chr3L
chr3R
chr4
chrX
mir-375
mir-959/960/961/962/963/964
mir-932
mir-275/305
mir-9b/9c/79/306
mir-124
let-7/mir-100/125
mir-1
mir-133
mir-986
mir-14
mir-278
mir-8
mir-7
mir-965
mir-966
mir-2b-1
mir-87
mir-263a
mir-967
mir-968/1002
mir-2a-1/2a-2/2b-2
mir-987
mir-307
mir-988
mir-281-1/281-2
mir-989
mir-308
mir-137
mir-990
mir-31a
mir-310-313/991/992
mir-314/5/6-116-2/6-3/256/309
bantam
mir-282
mir-955
mir-190
mir-276b
mir-276a
mir-274
mir-314
mir-285
mir-956
mir-263b
mir-219
mir-315
mir-9a
mir-33
mir-193
mir-958
mir-957
mir-316
mir-929
mir-10
mir-993
mir-34/277/317
mir-318/994
mir-252
mir-284
mir-iab-4
mir-2c/13a/13b-1
mir-995
mir-997
mir-999
mir-11/998
mir-92a/92b
mir-1000
mir-279/996
mir-1001
mir-954
mir-980
mir-981
mir-303/992/983-1/983-2/984
mir-13b-2
mir-31b
mir-970
mir-971
mir-12/283/304
mir-210/969
mir-972/973/974/975/976/977/978/979
mir-985
mir-927
5 Mb euchromatic chromosome
Heterochromatic chromosome (not to scale)
Centromere
B
chrX (+) (–)
19385000
19390000
19395000
mir-972
mir-973
mir-974
mir-975
mir-976
mir-977
mir-978
mir-979
Grip84
D. simulans
D. yakuba
D. ananassae
D. pseudoobscura
D. mojavensis
D. virilis
C
chr2R (+) (–)
16098000
16100000
16102000
l(2)05510
mir-313
mir-312
mir-311
mir-310
correlated expression
mir-992
mir-991
correlated expression
CG13434

Figure 6

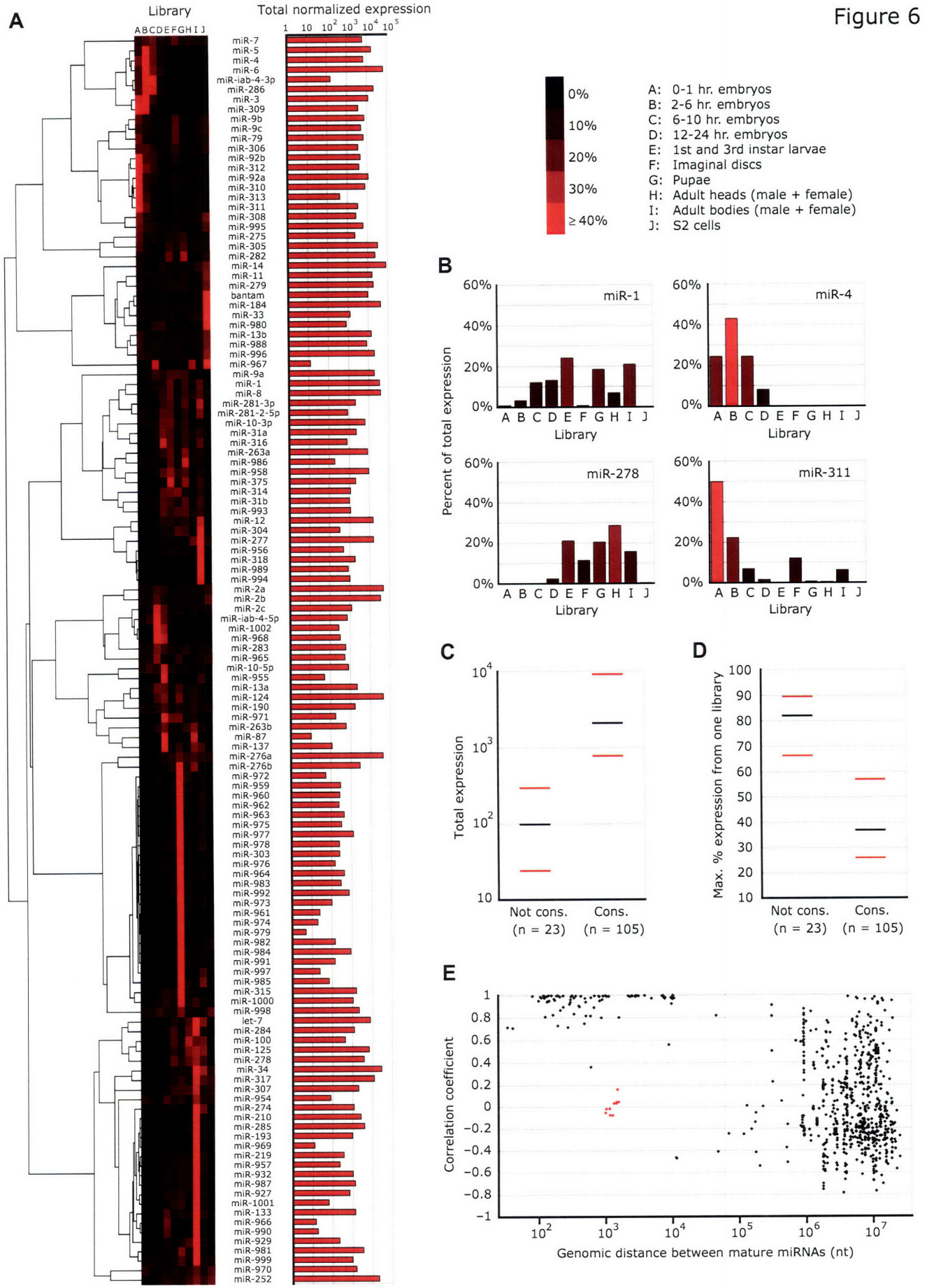

A
Library
A B C D E F G H I J
Total normalized expression
1 10 10^2 10^3 10^4 10^5
miR-7
miR-5
miR-4
miR-6
miR-iab-4-3p
miR-286
miR-3
miR-309
miR-9b
miR-9c
miR-79
miR-306
miR-92b
miR-312
miR-92a
miR-310
miR-313
miR-311
miR-308
miR-995
miR-275
miR-305
miR-282
miR-14
miR-11
miR-279
bantam
miR-184
miR-33
miR-980
miR-13b
miR-988
miR-996
miR-967
miR-9a
miR-1
miR-8
miR-281-3p
miR-281-2-5p
miR-10-3p
miR-31a
miR-316
miR-263a
miR-986
miR-958
miR-375
miR-314
miR-31b
miR-993
miR-12
miR-304
miR-277
miR-956
miR-318
miR-989
miR-994
miR-2a
miR-2b
miR-2c
miR-iab-4-5p
miR-1002
miR-968
miR-283
miR-965
miR-10-5p
miR-955
miR-13a
miR-124
miR-190
miR-971
miR-263b
miR-87
miR-137
miR-276a
miR-276b
miR-972
miR-959
miR-960
miR-962
miR-963
miR-975
miR-977
miR-978
miR-303
miR-976
miR-964
miR-983
miR-992
miR-973
miR-961
miR-974
miR-979
miR-982
miR-984
miR-991
miR-997
miR-985
miR-315
miR-1000
miR-998
let-7
miR-284
miR-100
miR-125
miR-278
miR-34
miR-317
miR-307
miR-954
miR-274
miR-210
miR-285
miR-193
miR-969
miR-219
miR-957
miR-932
miR-987
miR-927
miR-1001
miR-133
miR-966
miR-990
miR-929
miR-981
miR-999
miR-970
miR-252
0%
10%
20%
30%
≥ 40%
A: 0-1 hr. embryos
B: 2-6 hr. embryos
C: 6-10 hr. embryos
D: 12-24 hr. embryos
E: 1st and 3rd instar larvae
F: Imaginal discs
G: Pupae
H: Adult heads (male + female)
I: Adult bodies (male + female)
J: S2 cells
B
Percent of total expression
miR-1
miR-4
miR-278
miR-311
60%
40%
20%
0%
A B C D E F G H I J
Library
C
Total expression
10^4
10^3
10^2
10
Not cons. (n = 23)
Cons. (n = 105)
D
Max. % expression from one library
100
90
80
70
60
50
40
30
20
10
Not cons. (n = 23)
Cons. (n = 105)
E
Correlation coefficient
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1
10^2
10^3
10^4
10^5
10^6
10^7
Genomic distance between mature miRNAs (nt)

Figure 7

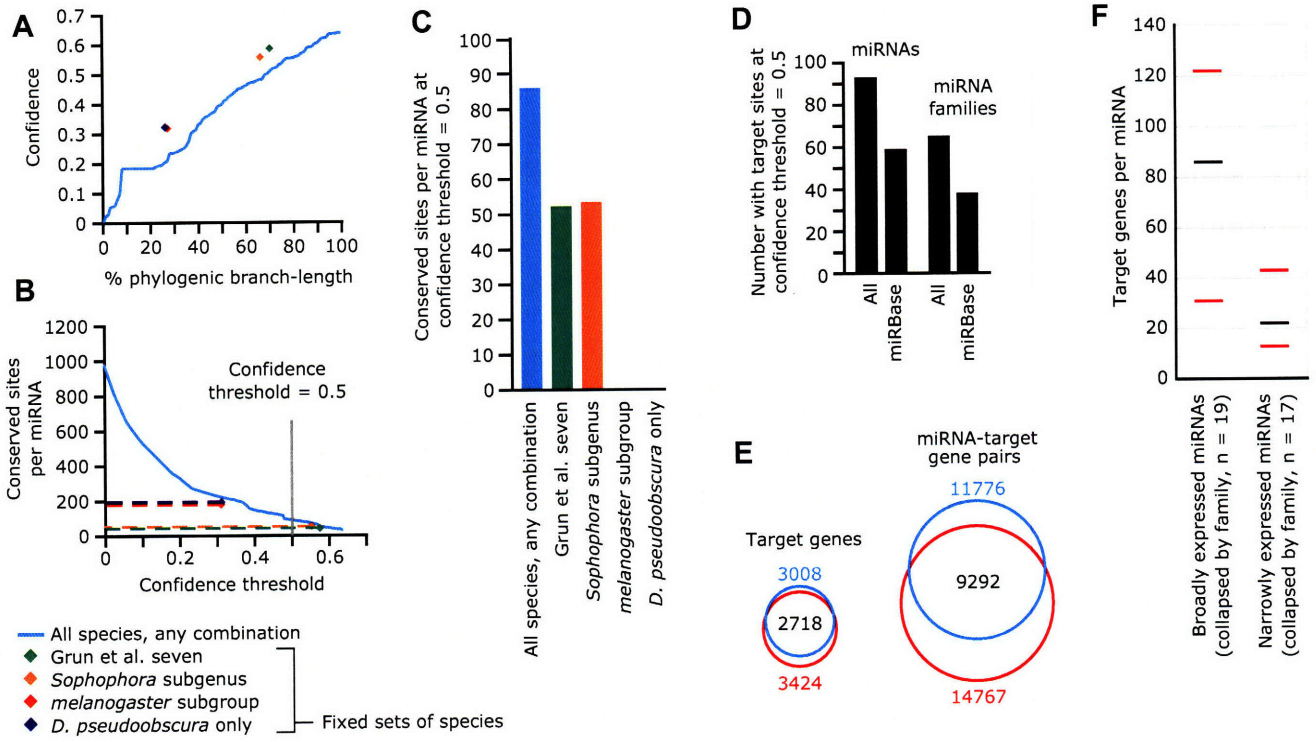
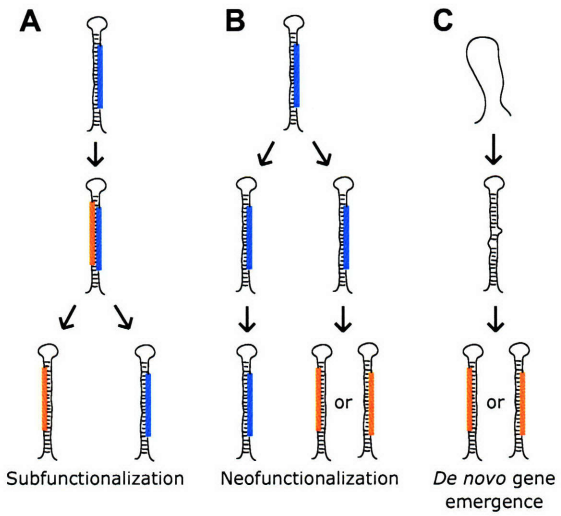

A
Confidence
% phylogenic branch-length
B
Conserved sites per miRNA
Confidence threshold = 0.5
Confidence threshold
All species, any combination
Grun et al. seven
Sophophora subgenus
melanogaster subgroup
D. pseudoobscura only
Fixed sets of species
C
Conserved sites per miRNA at confidence threshold = 0.5
All species, any combination
Grun et al. seven
Sophophora subgenus
melanogaster subgroup
D. pseudoobscura only
D
Number with target sites at confidence threshold = 0.5
miRNAs
miRNA families
All
miRBase
All
miRBase
E
miRNA-target gene pairs
Target genes
3008
2718
3424
11776
9292
14767
F
Target genes per miRNA
Broadly expressed miRNAs (collapsed by family, n = 19)
Narrowly expressed miRNAs (collapsed by family, n = 17)

Figure 8


A
B
C
or
or
Subfunctionalization
Neofunctionalization
De novo gene emergence

Supplemental Text

Refinement of prior miRNA annotations

For 37 of the 74 miRNA genes that were previously annotated in *Drosophila melanogaster* (Griffiths-Jones 2004) and validated by our reads, the distribution of reads across the hairpin suggested refinements to the annotation of the mature miRNA (Table S2). In 12 cases there was discrepancy at the miRNA 5′ terminus. For six of these mature species (miR-33,-263a, -274, -282, -283, -284), the annotated 5′ termini were predictions based on conservation (Aravin et al. 2003; Grad et al. 2003; Lai et al. 2003). In each of these cases, the annotated miRNA 5′ end was 1-4 nucleotides upstream of the observed 5′ end (Table S2). These six miRNAs have been experimentally supported by RNA blots, but there has been no previous attempt to map their 5′ termini (Aravin et al. 2003; Grad et al. 2003; Lai et al. 2003), and thus discrepancies between the annotated and actual sequences were expected. The 5′ end of miR-303 has been based on a single read (Aravin et al. 2003); our larger set of reads contained ~50-fold more reads with a 5′ terminus offset by 2 nt in the 3′ direction. In the case of miR-87, the total number of reads from our dataset matching the hairpin was insufficient to challenge the current 5′ terminus annotation (Table S2).

As with *mir-10* (Main Text), most reads matching the *mir-313* hairpin were from the arm opposing the previously annotated miRNA. The original annotation for miR-313 shared a 5′ end with 74 reads from our dataset, whereas the star species gave rise to 392 reads sharing a common 5′ end. In contrast to miR-10, however, conservation criteria did not support function of the most abundantly sequenced species. It was not conserved even

in the closely related *D. yakuba* or *D. ananassae*, whereas the miRNA was highly conserved and shared a seed with five other conserved miRNAs. This scenario resembled that observed for *Arabidopsis* miR403 (Rajagopalan et al. 2006) and supported either the idea that the read frequency does not always indicate the more abundant species, or the idea that the less abundant species is occasionally the functional one.

For miR-210, an additional case with discrepancy at the 5′ terminus, the reads that matched the hairpin suggested that two mature miRNAs are generated from the same arm of the hairpin (miR-210.1 and -210.2, Main Text).

Other miRNAs with revised 5′ termini included those in the miR-2 group. *D. melanogaster* has five annotated *mir-2* hairpins: *mir-2a-1, mir-2a-2, mir-2b-1, mir-2b-2,* and *mir-2c*. The mature miR-2a and miR-2b miRNAs were identified by cloning and sequencing, whereas miR-2c had been predicted based on similarity to the other two (Aravin et al. 2003; Lagos-Quintana et al. 2001; Lai et al. 2003). Our reads confirmed expression from all five hairpin precursors. Due to the high similarity of these five hairpins, reads from the miRNA arm of the hairpin whose sequences could be attributed uniquely to a single hairpin were scarce, but when considering all the data together, including the dominant miRNA* species from each hairpin, we concluded that each of these hairpins generates a single preferred miRNA:miRNA* duplex, and that the 5′ ends of miR-2a-2 and miR-2c are offset from their prior annotations by two nucleotides in the 3′ direction (Table S2).

The processing of *mir-281-1* versus *mir-281-2* appeared similarly diverged. These two hairpins, likely the result of a tandem duplication, are within 200 bp of each other (Aravin et al. 2003; Lai et al. 2003). All of the reads from the annotated miRNA strands

matched both hairpins. Two populations of reads supported two dominant 5′ termini. Analysis of the 3′ overhangs left by reads from the miRNA* arm of the hairpins indicated that the reads with the original 5′ annotation should mostly be attributed to *mir-281-1*, whereas the shifted reads should be attributed mostly to *mir-281-2*. Moreover, in this scenario, the miR-281-2* reads outnumbered the miRNA reads attributed to this hairpin (177 to 90), a result consistent with the asymmetry guidelines for loading of the silencing complex (Schwarz et al. 2003). Nonetheless, the numbers were not far from each other, and both arms of the *mir-281-2* hairpin were highly conserved, suggesting that either or both might have conserved function. Accordingly, we annotated the RNAs from this hairpin as miR-281-2-5p and miR-281-2-3p, respectively.

MicroRNA biogenesis in flies

As in nematodes (Ruby et al. 2006), examining the multitude of reads arising from the previously annotated miRNA hairpins provided insights into the specificity and precision of Drosha and Dicer processing (Table S2). Of the 117 miRNA hairpins with reads from the miRNA* strand and unambiguously defined miRNAs (*mir-2, -210, -281* hairpins excluded for reasons discussed above), 96 of the most abundant miRNA* 3′ ends had precise 2-nt overhangs relative to the miRNA 5′ ends. When considering all the reads from the miRNA* strand of the 117 hairpins, 83% exhibited 2-nt 3′ overhangs relative to the miRNA 5′ ends. Of the 21 hairpins whose most abundant miRNA* 3′ termini did not overhang by exactly 2 nt, 13 had 10 or fewer miRNA* reads, raising the possibility of insufficient sampling as an explanation for their apparent offsets.

For all ends on both arms of the hairpin, heterogeneity that extended rather than shortened the mature product comprised a sizable fraction (35.8%) of all heterogeneity. These reads whose ends extended further than the most abundant ends revealed misprocessing by Drosha or Dicer, whereas those that shortened the mature product could be attributed to either misprocessing or degradation. As reported in other species (Li et al. 2005; Ruby et al. 2006), untemplated nucleotide addition also contributed to a minor fraction of 3′ heterogeneity, and was observed for 108 of the 131 mature miRNA species. Whereas mature miRNAs were extended by a single nucleotide with an efficiency of only 3.2%, the efficiency of single nucleotide extension increased to 14.3%, 15.6%, and 16.7% for addition of a second, third, and fourth nucleotides, respectively, suggesting weak but detectable processivity for the untemplated terminal-transferase activity. The preferred untemplated base was adenine, representing 60.0%, 85.5%, 83.6%, and 74.3% of the identities in each of the first four positions, respectively.

The 5′ ends of both the miRNA and miRNA* species were more homogenous than the 3′ ends; also, the 5′ end of the miRNA was more consistent than the 5′ end of the miRNA*. Excluding the hairpins with reads of ambiguous origin described above (*mir-2, -210, -281*), the 5′ ends were 98.7% identical for miRNAs (98.1% and 98.9% identical for miRNAs deriving from the 5′ and 3′ arms of the hairpin, respectively) and 94.6% identical for miRNA*s (92.8% and 95.2% identical on the 5′ and 3′ arms, respectively). The 3′ ends were 70.2% identical for miRNAs (58.7% and 74.6% identical on the 5′ and 3′ arms) and 84.0% identical for miRNA*s (79.0% and 85.7% identical on the 5′ and 3′ arms). Thus, the miRNA 5′ ends were more consistent than the miRNA* 5′ ends, and the miRNA* 3′ ends were more consistent than the miRNA 3′ ends, regardless

of which enzyme generated them. This trend held when examining only the heterogeneity that lengthened the small RNAs (variants with extra 5′ nucleotides were 4.5-fold more abundant for the miRNA* than for the miRNA when each was normalized to the number of reads with the most common 5′ end; 3.8-fold and 5.2-fold for miRNAs on the 5′and 3′ arms, respectively), and thus was not attributable purely to degredation. The heightened precision of either enzyme when it defined the miRNA seed implied that Dicer does not simply measure from the site of the Drosha cleavage and suggested that additional determinants must be employed when needed to more accurately define Dicer cleavage. Similar conclusions arise from the sequencing data in nematodes (Ruby et al. 2006).

In addition to the miRNA and miRNA* species, other short RNA byproducts of miRNA processing with 5′ monophosphates and 3′ hydroxyl groups (required by our library construction protocol) were sometimes observed. Hairpin loop-containing sequences that connect the miRNA and miRNA* species prior to Dicer cleavage were sometimes observed at a low frequency (Fig. 2, Table S2). In addition, thirteen of the previously annotated miRNA hairpins (*mir-2a-1, -2b-2, -3, -4, -5, -6-2, -7, -9a, -13b-2, -277, -279, -283*), as well as seven of the novel miRNA hairpins (*mir-190, -964, -974, -976, -977, -988, -997*) gave rise to reads flanking the 5′ end of the pre-miRNA (Table S2). These 170 reads were more homogenous at their 3′ ends than at their 5′ ends, with the majority of 3′ ends defined by the inferred Drosha cut. Their presence implied a role for 5′→3′ exonuclease in the degradation of pri-miRNA processing byproducts. Accordingly, the eukaryotic cytoplasmic 5′→3′ exonuclease Xrn1p and its nuclear isozyme Rat1p are 5′ monophosphate-dependent (Johnson 1997; Stevens 1980). Only

three reads were observed flanking pre-miRNA 3′ ends (*mir-6-2, -11, -314*), and the 5′ ends of two of those matched the dominant inferred Drosha cleavage site.

Supplemental methods

MicroRNA gene prediction. For each of six *Drosophila* genomes (*melanogaster*, *ananassae*, *pseudoobscura*, *mojavensis*, *virilis*, and *grimshawi*) (Adams et al. 2000; Consortium 2007a; Consortium 2007b; Richards et al. 2005), RNAfold (Hofacker et al. 1994) was used to predict candidate hairpins from across the entire genome by folding 110-nt windows, advanced by 10-nt increments along each strand. Windows with a total calculated folding energy ≤ –18 kcal/mol were split into individual hairpins, and those with at least 20 predicted base pairs were retained as candidates. Forked hairpins were permitted provided that the longest forked segment contained no more than five base pairs. In the cases of *D. melanogaster* and *D. pseudoobscura*, genomic segments annotated as repetitive and >60 nt long (Grumbling and Strelets 2006) were not folded, but folding windows were allowed overlap with such segments by up to 30 nt.

Candidate hairpins were evaluated based on 35 features ($x_1 - x_{35}$), each of which could have any from a set of pre-determined values associated with it. Features were treated as either 'quantitative' or 'dimensionless'. Quantitative features were properties whose values were numerical and could be placed in bins of arbitrary size. For each hairpin evaluated, the feature value associated with that hairpin was the appropriate bin. Bins were continuous, and in cases where values could extend beyond the scope of the defined bins, the counts were assigned to the edge bins. Dimensionless features had values that were either non-quantitative or strictly Boolean. These features required that

no hairpin could return either an intermediate value or a value outside the scope of the pre-determined possibilities. For each feature x_i, for each possible value x_{iv}, a single pseudocount was added for each value in accord with Laplace's rule, and the true counts were the number of hairpins in the given set with the given value for the given feature. For quantitative features, counts were also smoothed across bins by distributing 12.5% of a bin's counts to each adjacent bin, twice. The process of training determined log-odds scores for each value of each feature by comparing the value's foreground (training set) frequency to its background (candidate set) frequency:

$$S_{iv} = \log_2(p(x_{iv} \mid \text{foreground}) / p(x_{iv} \mid \text{background}))$$

where S_{iv} was the score assigned to feature value x_{iv}, and $p(x_{iv} \mid set)$ was the probability (frequency) of feature value x_{iv} in a given set of hairpins.

The *D. melanogaster* training set comprised 37 randomly-selected miRNA hairpins from the miRBase v8.1 annotations (Griffiths-Jones 2004). The *D. pseudoobscura* training set comprised the miRBase-annotated orthologs of the *D. melanogaster* training set. The training sets for the other species were determined through manual inspection of top Blast results using the *D. melanogaster* training set as queries and the total set of candidate hairpins from the target organism as the database (Altschul et al. 1990). Foreground values were determined in reference to the annotated miRNA 5′ terminus. Background values were determined in reference to a randomly selected 5′ end from anywhere on the hairpin. A randomly selected set of 10,000 hairpins was used to generate background frequencies in each instance of training.

Features $x_{1\text{-}3}$ were quantitative. x_1 described the number of nucleotides separating a candidate miRNA from its presumed miRNA* in single-nucleotide bins ranging from 0

to 35. x_2 described the asymmetry of bulges and internal loops. At each unpaired nucleotide or segment within the portion of the hairpin bounded by the candidate miRNA, the absolute value of the difference between the numbers of nucleotides on either side of the bulge or internal loop was added to a sum. Bins were integers ranging from 0 to 9. x_3 described the sequence complexity of the full candidate hairpin. The hairpin sequence was divided into words as described (Lempel and Ziv 1976), and the complexity measure was defined as $\log_4(4^A / (B - 2A)^2)$, where A is the length of the longest word and B is the length of the entire sequence. This system was an ad-hoc attempt to arrive at a length-independent complexity measurement, which differed from the definition of information content described (Lempel and Ziv 1976). Bins were integers ranging from –2 to 5.

Features $x_{4\text{-}35}$ were dimensionless. x_4 and x_5 described the nucleotide identities at miRNA positions 1 and 9, respectively. These positions are both outside of the targeting-relevant 'seed', and both have been observed to show preference for U (Lau et al. 2001; Lewis et al. 2005). Values included each of the four nucleotides plus 'N', which indicates an uncalled base and is used in genome assemblies to connect mapped but non-overlapping contigs. $x_{6\text{-}25}$ described the base pairing of mature miRNA nucleotides at positions 1-20; these pairings will persist following liberation of the miRNA/miRNA* duplex from the hairpin precursor. Bins were Boolean indications of a base pair predicted at the given position. $x_{26\text{-}35}$ described base pairing for nucleotides outside of the miRNA/miRNA* core duplex, and were numbered based on position in the hairpin relative to the miRNA/miRNA*. Two positions were evaluated towards the loop and

eight towards the base of the stem. All of the structure-based features made use of the predicted minimal free-energy structure generated by RNAfold (Hofacker et al. 1994).

For each candidate hairpin, a score was determined for each possible position of a 22-nt miRNA by summing the scores for features $x_{1\text{-}35}$, and the hairpin was assigned the score of the maximal scoring 22-nt candidate. All training and candidate set hairpins were scored to generate foreground and background score distributions, respectively. The standard deviation of the foreground distribution was calculated, and a cut-off score was picked half a standard deviation below the minimum foreground score. Candidate hairpins scoring below the cutoff were eliminated, and the process was repeated with the more restricted background set. *D. melanogaster* and *D. pseudoobscura* candidates were each put through five rounds of such elimination; candidates from the remaining genomes were each put through three rounds.

At this point, the definition of a candidate changed from a hairpin to a pair of putatively orthologous hairpins. For each pair of organisms, candidate pairs were identified by assigning the best Blast hit (window size, 11 nt; mismatch penalty, –1) to each query hairpin. Blast searches were performed in both directions, with each set of candidates serving once as the query set and once as the database. A pairwise alignment was generated for each candidate and used to arrive at a set of candidate miRNA 5′ coordinates. A pair of coordinates, corresponding to positions in each of the hairpin sequences, was generated for each position in the alignment. The predicted 5′ end was the pair of coordinates that produced the maximum score for the candidate.

Two-species candidate sets were used for training and scoring as described above, except with 25 additional features (features $x_{36\text{-}58}$). $x_{36\text{-}55}$ were dimensionless, and

described the conservation of each miRNA nucleotide as a Boolean. Positions 1 and 9 were not considered here because their identities were being scored. Positions to be compared were determined relative to the miRNA 5′ ends rather than based on position in the pairwise alignment. x_{56} was quantitative, and described the relative conservation of the miRNA versus the sequence connecting the miRNA and miRNA*. The arithmetic difference between the fraction of the miRNA conserved minus the fraction of the connecting sequence conserved was binned over intervals of .05 extending from –1 to 1. The 5′ nucleotide identities for hairpins from each organism were considered separately; while $x_{4\text{-}5}$ described identities in the first hairpin of each pair, $x_{57\text{-}58}$ considered identities in the second hairpin. All of the other features considered for hairpin candidates were also considered for ortholog pair candidates. $x_{6\text{-}35}$ returned a Boolean True only if the given position formed a base pair in the same direction in both hairpins (i.e., only if the "(" or ")" in the bracket notation were the same). The quantitative features $x_{1\text{-}3}$ used the average of the values returned by the two hairpins.

Each set of candidates was put through four rounds of scoring and elimination as described above. Surviving candidate hairpins were defined as those hairpins for which a complete network of orthologs could be constructed across all six species examined. For an example using three species, if hairpins A, B, and C were from *D. melanogaster*, *D. pseudoobscura*, and *D. virilis*, respectively, then the requirements for A to be a surviving *D. melanogaster* candidate were that A and B had survived as an ortholog pair, A and C had survived as an ortholog pair, and B and C had survived as an ortholog pair. It would not have been sufficient for hairpins B and C to have each survived the *D.pseudoobscura*/*D.virilis* eliminations paired with other hairpins. Those *D.*

melanogaster hairpins with complete networks were ranked according to the sums of their networks' first-round pairwise scores. Candidates with multiple complete networks were assigned the maximum score from those networks. Some candidate overlapped; in cases with more than 80% overlap, the overlapping candidate with the lower score was eliminated. Inspection of the surviving candidates revealed 62 of obviously low sequence complexity, indicating a failure of the complexity score to adequately eliminate simple repeats. These 62 were removed manually to generate the final list of 565 candidate hairpins (Table S1 and Fig. 1A). Consolidating hairpins from opposite DNA strands and filtering out those overlapping with annotated exons yielded the final list of 327 candidate loci, the top 100 of which where carried forward as the computational predictions (Table S1 and Fig. 1B-C).

Library construction and sequencing. Libraries of cDNAs derived from small RNAs were prepared and sequenced as described in the methods of the Main Text. As previously described (Ruby et al. 2006), sequence reads were processed in four steps. First, perfect matches to the 9-nt segments of each linker that immediately flanked the small RNA-derived sequence were found in 2,075,098 reads; the remaining reads were discarded. Second, each sequence was compared to annotated *D. melanogaster* miRNA hairpins (miRBase 8.1) (Griffiths-Jones 2004), and those sequences ≥10 nt and with perfect matches over their entire length were set aside (224,398 reads). Third, each sequence was compared to the *D. melanogaster* 18S, 5.8S, 2S, and 28S rRNA genes (Grumbling and Strelets 2006), and those with perfect matches over their entire length were set aside (511,088 reads). Sequences with ambiguous calls (Ns) were also discarded

here. Fourth, the remaining sequences were compared to the *D. melanogaster* genome (Adams et al. 2000) using Blast (Altschul et al. 1990), and those with perfect matches across their entire length were retained (409,195 reads). Up to 50 perfect-match loci were recorded for each query. The remaining sequences that did not perfectly match the genome were queried later for examples of untemplated nucleotide addition. Sequences that were found to match miRNA hairpins were also compared to the rest of the genome so that the uniqueness of those matches could be assessed.

Expression analysis. For each library, the total number of miRNA hairpin-matching reads was calculated as a normalization factor. For each unique sequence, the number of perfect matches to miRNA hairpins was divided by the number of perfect matches to the *D. melanogaster* genome and multiplied by the number of reads that gave rise to that sequence. The number of reads matching a particular mature miRNA was calculated similarly, but only sequence matches that overlapped the center of the dominantly abundant mature miRNA sequence contributed to the miRNA tally. Each miRNA tally from each library was normalized to the total number of miRNA hairpin-matching reads for that library, and those normalized tallies were used for both relative and total expression analysis..

Relative expression analysis sought to determine the expression preferences of individual miRNAs across the biological contexts represented by our cDNA libraries. Here, the normalized tally of a particular miRNA in a particular library was divided by the sum of normalized tallies for that miRNA across all libraries. The result was an expression profile centered at 0.1, with values ranging from 0 to 1. The application

Cluster was used for hierarchical clustering of miRNAs using average linkage correlation (Eisen et al. 1998).

Total expression analysis sought to evaluate and compare the magnitudes of expression for individual miRNAs. The normalized tallies of each miRNA across all libraries were summed and multiplied by 10^6. The resulting value corresponded to the number of reads for a given miRNA per million reads matching miRNA hairpins, assuming an equal contribution by all ten libraries. It was only a rough measure of total expression because the various libraries did not proportionally represent all the tissues and stages of the flies. Comparisons of miRNA expression within a given library were more accurate, but with the exception of the S2 cells and early embryos, comparisons within a library were still confounded by a complex mixture of tissues.

References

Adams, M.D. S.E. Celniker R.A. Holt C.A. Evans J.D. Gocayne P.G. Amanatides S.E. Scherer P.W. Li R.A. Hoskins R.F. Galle R.A. George S.E. Lewis S. Richards M. Ashburner S.N. Henderson G.G. Sutton J.R. Wortman M.D. Yandell Q. Zhang L.X. Chen R.C. Brandon Y.H. Rogers R.G. Blazej M. Champe B.D. Pfeiffer K.H. Wan C. Doyle E.G. Baxter G. Helt C.R. Nelson G.L. Gabor J.F. Abril A. Agbayani H.J. An C. Andrews-Pfannkoch D. Baldwin R.M. Ballew A. Basu J. Baxendale L. Bayraktaroglu E.M. Beasley K.Y. Beeson P.V. Benos B.P. Berman D. Bhandari S. Bolshakov D. Borkova M.R. Botchan J. Bouck P. Brokstein P. Brottier K.C. Burtis D.A. Busam H. Butler E. Cadieu A. Center I. Chandra J.M. Cherry S. Cawley C. Dahlke L.B. Davenport P. Davies B. de Pablos A. Delcher Z. Deng A.D. Mays I. Dew S.M. Dietz K. Dodson L.E. Doup M. Downes S. Dugan-Rocha B.C. Dunkov P. Dunn K.J. Durbin C.C. Evangelista C. Ferraz S. Ferriera W. Fleischmann C. Fosler A.E. Gabrielian N.S. Garg W.M. Gelbart K. Glasser A. Glodek F. Gong J.H. Gorrell Z. Gu P. Guan M. Harris N.L. Harris D. Harvey T.J. Heiman J.R. Hernandez J. Houck D. Hostin K.A. Houston T.J. Howland M.H. Wei C. Ibegwam M. Jalali F. Kalush G.H. Karpen Z. Ke J.A. Kennison K.A. Ketchum B.E. Kimmel C.D. Kodira C. Kraft S. Kravitz D. Kulp Z. Lai P. Lasko Y. Lei A.A. Levitsky J. Li Z. Li Y. Liang X. Lin X. Liu B. Mattei T.C. McIntosh M.P. McLeod D. McPherson G. Merkulov N.V. Milshina C. Mobarry J. Morris A. Moshrefi S.M. Mount M. Moy B. Murphy L. Murphy D.M. Muzny D.L. Nelson D.R. Nelson K.A. Nelson K. Nixon D.R. Nusskern J.M.

Pacleb M. Palazzolo G.S. Pittman S. Pan J. Pollard V. Puri M.G. Reese K. Reinert K. Remington R.D. Saunders F. Scheeler H. Shen B.C. Shue I. Siden-Kiamos M. Simpson M.P. Skupski T. Smith E. Spier A.C. Spradling M. Stapleton R. Strong E. Sun R. Svirskas C. Tector R. Turner E. Venter A.H. Wang X. Wang Z.Y. Wang D.A. Wassarman G.M. Weinstock J. Weissenbach S.M. Williams WoodageT K.C. Worley D. Wu S. Yang Q.A. Yao J. Ye R.F. Yeh J.S. Zaveri M. Zhan G. Zhang Q. Zhao L. Zheng X.H. Zheng F.N. Zhong W. Zhong X. Zhou S. Zhu X. Zhu H.O. Smith R.A. Gibbs E.W. Myers G.M. Rubin and J.C. Venter. 2000. The genome sequence of *Drosophila melanogaster. Science* **287:** 2185-2195.

Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. 1990. Basic local alignment search tool. *J Mol Biol* **215:** 403-410.

Aravin, A.A., M. Lagos-Quintana, A. Yalcin, M. Zavolan, D. Marks, B. Snyder, T. Gaasterland, J. Meyer, and T. Tuschl. 2003. The small RNA profile during *Drosophila melanogaster* development. *Dev Cell* **5:** 337-350.

Consortium, D.C.G.S.a.A. 2007a. Evolution of Genes and Genomes in the Genus *Drosophila. Nature* **In preparation.**

Consortium, D.C.G.S.a.A. 2007b. Initial comparative genomics analysis of 12 *Drosophila*
genomes. *Nature* **In preparation.**

Eisen, M.B., P.T. Spellman, P.O. Brown, and D. Botstein. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* **95:** 14863-14868.

Grad, Y., J. Aach, G.D. Hayes, B.J. Reinhart, G.M. Church, G. Ruvkun, and J. Kim. 2003. Computational and experimental identification of *C. elegans* microRNAs. *Mol Cell* **11:** 1253-1263.

Griffiths-Jones, S. 2004. The microRNA Registry. *Nucleic Acids Res* **32:** D109-111.

Grumbling, G. and V. Strelets. 2006. FlyBase: anatomical data, images and queries. *Nucleic Acids Res* **34:** D484-488.

Hofacker, I.L., W. Fontana, P.F. Stadler, L.S. Bonhoeffer, M. Tacker, and P. Schuster. 1994. Fast folding and comparison of RNA secondary structures. *Monatshefte Fur Chemie* **125:** 167-188.

Johnson, A.W. 1997. Rat1p and Xrn1p are functionally interchangeable exoribonucleases that are restricted to and required in the nucleus and cytoplasm, respectively. *Mol Cell Biol* **17:** 6122-6130.

Lagos-Quintana, M., R. Rauhut, W. Lendeckel, and T. Tuschl. 2001. Identification of novel genes coding for small expressed RNAs. *Science* **294:** 853-858.

Lai, E.C., P. Tomancak, R.W. Williams, and G.M. Rubin. 2003. Computational identification of *Drosophila* microRNA genes. *Genome Biol* **4:** R42.

Lau, N.C., L.P. Lim, E.G. Weinstein, and D.P. Bartel. 2001. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans. Science* **294:** 858-862.

Lempel, A. and J. Ziv. 1976. Complexity of Finite Sequences. *Ieee Transactions on Information Theory* **22:** 75-81.

Lewis, B.P., C.B. Burge, and D.P. Bartel. 2005. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120:** 15-20.

Li, J., Z. Yang, B. Yu, J. Liu, and X. Chen. 2005. Methylation protects miRNAs and siRNAs from a 3'-end uridylation activity in *Arabidopsis*. *Curr Biol* **15:** 1501-1507.

Rajagopalan, R., H. Vaucheret, J. Trejo, and D.P. Bartel. 2006. A diverse and evolutionarily fluid set of microRNAs in *Arabidopsis thaliana*. *Genes Dev* **20:** 3407-3425.

Richards, S., Y. Liu, B.R. Bettencourt, P. Hradecky, S. Letovsky, R. Nielsen, K. Thornton, M.J. Hubisz, R. Chen, R.P. Meisel, O. Couronne, S. Hua, M.A. Smith, P. Zhang, J. Liu, H.J. Bussemaker, M.F. van Batenburg, S.L. Howells, S.E. Scherer, E. Sodergren, B.B. Matthews, M.A. Crosby, A.J. Schroeder, D. Ortiz-Barrientos, C.M. Rives, M.L. Metzker, D.M. Muzny, G. Scott, D. Steffen, D.A. Wheeler, K.C. Worley, P. Havlak, K.J. Durbin, A. Egan, R. Gill, J. Hume, M.B. Morgan, G. Miner, C. Hamilton, Y. Huang, L. Waldron, D. Verduzco, K.P. Clerc-Blankenburg, I. Dubchak, M.A. Noor, W. Anderson, K.P. White, A.G. Clark, S.W. Schaeffer, W. Gelbart, G.M. Weinstock, and R.A. Gibbs. 2005. Comparative genome sequencing of *Drosophila pseudoobscura*: chromosomal, gene, and cis-element evolution. *Genome Res* **15:** 1-18.

Ruby, J.G., C. Jan, C. Player, M.J. Axtell, W. Lee, C. Nusbaum, H. Ge, and D.P. Bartel. 2006. Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. *Cell* **127:** 1193-1207.

Schwarz, D.S., G. Hutvagner, T. Du, Z. Xu, N. Aronin, and P.D. Zamore. 2003. Asymmetry in the assembly of the RNAi enzyme complex. *Cell* **115:** 199-208.

Stevens, A. 1980. Purification and characterization of a *Saccharomyces cerevisiae* exoribonuclease which yields 5'-mononucleotides by a 5' leads to 3' mode of hydrolysis. *J Biol Chem* **255:** 3080-3085.

Chapter 5

Future directions

The function of 21U-RNAs

The upstream motif associated with 21U-RNAs provides some preliminary and speculative insight into their method of biogenesis, but the function of 21U-RNAs remains unknown. The lack of primary sequence conservation within the 21U-RNAs themselves, in contrast with the upstream motifs, places two constraints on models for their function. First, it makes a model in which the 21U-RNAs themselves are targeting specific genes for repression in a miRNA- or siRNA-like manner unrealistic. Second, the presence of a common upstream motif that is highly predictive in identifying expressed 21U-RNAs, together with the observation that the magnitude of 21U-RNA expression is only substantial when the 21U-RNAs are considered together as a group, suggests co-expression and cooperative function between the members of this class. The abundance and heterogeneity of 21U-RNA genes also places an important constraint on experimental approaches to their study. Because they derive from many thousands of genomic loci that are interdigitated and overlapping with a variety of protein-coding genes, they are not amenable to genetic manipulation. The identification of a nodal component of their biogenesis or action is crucial to defining their role in nematode biology.

Some insight into the function of 21U-RNAs could be gained through a better understanding of their temporal expression pattern. To this end, we have sequenced

small RNAs from each of the major developmental stages of *C. elegans* (embryo, L1, L2, L3, L4, and adult) using Solexa sequencing technology (Seo et al. 2004) (see Acknowledgements for a list of collaborators). 21U-RNA concentrations are elevated in conjunction with the proliferation of the germ line during the L4 stage, remain high in adult worms, and their abundance carries over to embryos. Their concentrations are then depleted during the L1, L2, and L3 larval stages. The developmental profile suggested germ line expression, and indeed, 21U-RNAs were severely depleted in germ line-deficient *glp-4* mutant worms.

Localization of 21U-RNA expression to the time and place of germ line proliferation suggested that 21U-RNAs could play a role in the maturation of the germ line. Accordingly, the representative 21U-RNAs were severely depleted in *prg-1* mutants (Pedro Batista, personal communication), and deep sequencing of small RNAs from the *prg-1* mutant background confirmed depletion across the entire 21U-RNA class. The PRG-1 protein is a member of the Piwi subfamily of Argonautes whose mutation leads to defects in germ line development and decreased brood size (Cox et al. 1998; Yigit et al. 2006). The 21U-RNAs also co-immunoprecipitate with the PRG-1, demonstrating a direct molecular interaction between the two (Pedro Batista, personal communication). Thus, PRG-1 overcomes the crucial constraint on investigation of 21U-RNA function by providing a single component whose activity is coupled with that of the 21U-RNA population as a whole. As a result, approaches such as a genetic screen for suppressors of the *prg-1* mutant phenotype could be interpreted in part in terms of genetic interactions with the 21U-RNA class of small RNAs as well as with the *prg-1* gene itself.

MicroRNA expression profiling in *C. elegans*

While deep sequencing of mixed-stage *C. elegans* small RNAs was useful for the identification and classification of small RNAs, the developmental time course of deep sequencing analyzed from *Drosophila* achieved an additional dimension of information about gene expression. The expression patterns of many *C. elegans* microRNAs have already been determined by northern blot, but that method becomes increasingly unreliable and non-specific as the miRNAs being analyzed become less abundant. A comprehensive set of miRNA expression profiles from *C. elegans* would undoubtedly be of great use to those who study miRNAs in nematodes, and deep-sequence datasets across *C. elegans* development have already been generated (see above). The microRNA expression profiles across these datasets are currently being analyzed.

While northern blots (Ambros et al. 2003; Lau et al. 2001; Lee and Ambros 2001; Lim et al. 2003; Sempere et al. 2004) and the sequence datasets described above provide information on the temporal expression patterns of the *C. elegans* miRNAs, almost nothing is known about their spatial expression patterns. This is not true of miRNAs in other systems; blotting, microarray hybridization, or sequencing from dissected tissues has provided spatial expression patterns for mammalian miRNAs (Baskerville and Bartel 2005; Farh et al. 2005; Kim et al. 2004; Lagos-Quintana et al. 2002; Landgraf et al. 2007), transgenic sensors have been used to detect spatial expression patterns in *Drosophila* and mouse (Brennecke et al. 2003; Mansfield et al. 2004), and *in situ* hybridization has provided similar data for zebrafish and mouse miRNAs (Kloosterman et al. 2006; Wienholds et al. 2005). The expression of reporter genes fused to miRNA gene promoters has also been used in those systems and in *C. elegans* to determine the

spatial expression patterns of miRNA primary transcripts (Johnston and Hobert 2003). Most of those approaches are labor-intensive and require an independent set of analyses for each miRNA examined. The more efficient and multiplex method of tissue isolation followed by sequencing of small RNAs is stymied in *C. elegans* by the intractability of nematode dissection.

The solution to spatial expression profiling that has been applied to mRNAs in *C. elegans* is to express an epitope-tagged poly-A binding protein as a transgene under a tissue-specific promoter and then to co-immunoprecipitate all of the mRNAs that are expressed in the relevant tissue (Roy et al. 2002). The required precedent for this approach was the identification of a factor that would interact with all expressed mRNAs in a relatively unbiased manner. Characterization of *C. elegans* miRISC has identified many such factors for miRNAs (Zhang et al. 2007). Expression of an epitope-tagged copy of any one of these factors under a series of tissue-specific promoters, followed by pull-down of associated small RNAs and high-throughput sequencing, would permit the rapid development of a comprehensive miRNA spatial expression atlas.

Vertebrate mirtrons

MicroRNAs expressed through the mirtron biogenesis pathway were identified here in both *C. elegans* and in *Drosophila*. The requirements of the mirtron biogenesis pathway have been confirmed in *Drosophila* (Okamura et al. 2007), and potential mammalian mirtrons have also been identified (Berezikov et al. 2007). However, many of those mirtrons identified in mammals are verified with scant sequence data, and most form predicted secondary structures that do not resemble those of canonical pre-miRNAs.

The secondary structure differences between the annotated mammalian mirtrons and canonical mammalian pre-miRNAs cast doubt on the prescribed biogenesis pathway of those intron-boundary reads that no quantity of sequencing could overcome. In the work described here, the biogenesis requirements of representative candidate mirtrons were evaluated by expression of those candidates as mini-genes and knockdown of potential biogenesis factors by RNAi. Similar experiments should be performed in mammalian cells. Mouse ES cells with a dicer knockout have already been used to evaluate the Dicer-dependence of small RNAs in those cells (Calabrese et al. 2007). Mouse ES cells with the dgcr8 gene knocked out have also been generated (Wang et al. 2007). The expression of mirtrons would be eliminated in the dicer KO background and unaffected in the dgcr8 KO background. Deep sequencing of small RNAs from wild type mouse ES cells and each of these two mutant backgrounds could thus be used to identify mirtrons that are expressed endogenously in mouse ES cells. Such analyses are underway. The same cell lines could also be used to evaluate the processing requirements of candidate mirtrons.

Evolutionary scope of microRNAs

The availability of genome sequences for many species within the *Drosophila* genus facilitated analysis of the conservation of miRNA genes. The conclusion of this analysis was that many of the rarely-sequenced miRNAs have arisen recently in the evolution of this clade, for no apparent orthologs could be found. However, the features of miRNA genes that must be selectively maintained in order to preserve function do not impose great limitations on the primary sequence of the gene. Most of the sequence

requirements are those imposed by the hairpin structure of the precursor, and primary sequence changes do not affect RNA secondary structures as long as they are complemented by compensatory mutations. To further complicate conservation analysis, the hairpin structures of miRNA precursors are quite heterogeneous, implying that many single mutations to a miRNA hairpin that modify the secondary structure would be tolerated by natural selection. Finally, the functional portion of the miRNA, the seed, is only 7 nt long, leaving a very narrow window of reliably conserved sequence to identify in a related genome. In some cases, putative orthologs have been identified whose conservation approaches the described minimum (chapter 2, figure 1D). However, the low levels of observed conservation in these cases and expected conservation in the cases of miRNAs for which no orthologs have been identified leave open the possibility of false positive and false negative errors, respectively.

The appropriate test of these evolutionary models would be to deeply sequence small RNAs from closely related species. At least two of the species whose genome sequences were used for the evolutionary analyses described here, *Caenorhabditis briggsae* and *Drosophila pseudoobscura*, have been used extensively as laboratory model organisms (Baird and Chamberlin 2006; Dobzhansky 1937). Comparison of sequencing results from *C. elegans* to similarly-generated datasets from *C. briggsae*, or of results from *D. melanogaster* to datasets from *D. pseudoobscura*, would provide several types of useful data. First, the same standards that are applied here to miRNA annotation could be applied to ortholog identification: the observation of mature miRNAs deriving from the putative miRNA locus. For the minimally conserved orthologs identified here, the ortholog model that would be tested specifies a precise 5′ end for the putatively

orthologous miRNA gene based on the conserved seed sequence. Observation of conserved expression would address the concern about spurious false-positive ortholog annotations. Second, independent annotation of miRNA genes from a related organism would reduce the search space for orthologs ~10^6 fold. This estimation assumes ~10^2 miRNA genes in a genome with ~10^8 base pairs, both of which are accurate for species within the *Caenorhabditis* and *Drosophila* genuses. Instead of looking for seed conservation anywhere in the genome, one would only need to examine the expressed miRNAs, virtually eliminating the signal-to-noise problem otherwise associated with miRNA ortholog identification.

The approach to ortholog identification described above would fail if the pattern or magnitude of expression for a miRNA had changed substantially over evolutionary timescales, but this potential source of failure also introduces a third arena of discovery using near-species analyses: the evolution of the expression patterns of conserved miRNA genes. Just like nucleotide sequences, gene expression profiles accumulate changes as a function of evolutionary divergence (Rifkin et al. 2003). Although natural selection can stabilize or drive the divergence of mRNA expression patterns (Nuzhdin et al. 2004; Rifkin et al. 2005), it has been shown that strong conservation of gene expression profiles generally reflects conservation of physiological roles (Liao and Zhang 2006). Gene expression profiles have been used to study the evolution of flies (Nuzhdin et al. 2004; Rifkin et al. 2005; Rifkin et al. 2003), nematodes (Denver et al. 2005), primates (Enard et al. 2002), fishes (Whitehead and Crawford 2006), and yeasts (Fay et al. 2004). In array-based profiling studies, array noise and the inconsistent behavior of probes in different phylogenetic contexts pose significant challenges to analysis, but the

replacement of array-based quantification of expression with high-throughput sequencing has been proposed as a solution to both of these sources of error (Khaitovich et al. 2006; Liao and Zhang 2006).

The apparent frequency of miRNA gene birth and death has already been proposed as an engine of phenotype evolution (chapter 4). Further, the established relationship between miRNA expression patterns and those of their potential targets, in which miRNA targeting is either selectively maintained or avoided by co-expressed mRNAs (Farh et al. 2005; Stark et al. 2005), implies that changes in miRNA expression patterns would generally have phenotypic consequences significant enough to be the subject of natural selection.

Endogenous siRNAs in *Drosophila*

In *C. elegans*, the endogenous siRNAs that we observed had similar properties to the secondary siRNAs that are generated during RNAi as a consequence of RdRP activity (Ambros et al. 2003; Pak and Fire 2007). In contrast to *C. elegans*, no RdRP or amplification/spreading of RNAi signal has been identified in *Drosophila*, indicating that the physiological niche occupied by endogenous siRNAs in nematodes is not maintained in insects. Nonetheless, *Drosophila* expresses two Dicers and two Argonautes, and one copy of each gene plays a specialized role in either RNAi or miRNA biogenesis. The dedication of an siRNA biogenesis pathway to the processing of dsRNA implies that there are as-yet-unidentified endogenous dsRNA triggers of RNAi in *Drosophila*.

During analysis of the sequence data from chapter 4, I identified two potential sources of endogenous RNAi triggers. The first source is antisense transcription. Large-

scale small RNA sequence data generally captures some arbitrary mRNA degradation products. These products are highly heterogeneous in terms of their length distribution. In *C. elegans*, RdRP activity generates endogenous siRNAs antisense to mRNAs that outnumber reads in the sense orientation. In contrast, mRNA sense reads greatly outnumber antisense reads in *Drosophila.* However, the antisense reads in Drosophila have a non-random length distribution with a prominent peak at 21nt in length (figure 1A), matching the length of Dicer-generated siRNAs in that organism (Bernstein et al. 2001; Zamore et al. 2000). Antisense transcripts are regulators of gene expression even in organisms without Dicer-mediated RNAi (Hongay et al. 2006). It remains to be demonstrated that the antisense RNAs described here are genuine Dicer products, and further, it has not been shown that the presumably Dicer-generated products observed here play a role in gene silencing or are simply the superfluous byproducts of gene regulation by an antisense transcription mechanism. Finally, it remains to be examined what biological processes such antisense transcription and/or Dicer processing might regulate.

The second source of endogenous RNAi triggers that I identified in *Drosophila* is endogenous long RNA hairpins (hpRNA) genes. Transgenes encoding hpRNAs are used to introduce double-stranded RNAs to the cells of tissues or whole organisms in contexts where injection, transfection, and feeding would be inappropriate or ineffective. Such hpRNA transgenes have been used for gene silencing in nematodes (Tavernarakis et al. 2000), insects (Kennerdell and Carthew 2000), plants (Chuang and Meyerowitz 2000), and mammals (Svoboda et al. 2001), though few mammalian cell types allow the use of such constructs without triggering of the pro-apoptotic interferon response (Yang et al.

2001). Among the small RNAs from *Drosophila* analyzed in chapters 3 and 4, two examples were observed of hairpins whose processing better resembled that of an hpRNA than a miRNA. The first, hpRNA-1, gave rise to only five reads (figure 1B). The two reads from the 5′ arm of the hairpin share four overlapping nucleotides, indicating that the hairpin may be processed in multiple registers. However, two nucleotides from the 3′ end of the 5′ read base paired with two nucleotides from the 3′ end of two reads from the 3′ arm of the hairpin, consistent with the sequential generation of RNA duplexes with 2nt 3′ overhangs.

Notably, the 5′ arm of hpRNA-1 almost perfectly complemented the ATP synthase beta subunit mRNA (figure 1C). Small RNAs generated from this hairpin would be expected to repress ATP synthase just as the processed products of designed hpRNAs silence the genes that they complement. This possible interaction between hpRNA-1 and ATP synthesis has yet to be investigated. Oxygen starvation induces a number of changes in gene expression and metabolic activity, including the repression of oxidative phosphorylation and the upregulation of glycolytic enzymes, thereby increasing ATP generation through the oxygen-independent mechanism of glycolysis (Semenza 1999). The potential for hpRNA-1 to target a gene that is critical for oxidative phosphorylation suggests that it could be induced in response to oxygen starvation. Notably, all five of the reads mapping to hpRNA-1 were isolated from the pupal stage of development, during which anatomical transformation places high energy demands on the fly and during which the body of the fly is inactive and compact, limiting both gas exchange and the ability of the open circulatory system to distribute oxygenated fluid

through the coelomic cavity. The relationship between hpRNA-1 expression and oxygen starvation also has yet to be investigated.

The second hpRNA, referred to here as hpRNA-2, derived from a tandem repeat on the X chromosome. The genomic sequence between annotated genes *CG6903* and *CG4068* contains 20 copies of a repeat unit ~280 basepairs in length. The ~170 nt core of each repeat unit that is best conserved across all 20 units also shares identity with its own reverse complement sequence (average: 76%), permitting those portions of the repeat to either fold into individual hairpins or basepair with adjacent repeats (figure 1D,E). Over a thousand small RNA reads mapped only to this chromosome X repeat region, though the vast majority could be mapped to almost any of the 20 repeats within the region. Those reads derive from one strand of the genome almost exclusively (1318 reads from one strand versus 5 from the other). The majority of the reads form a tandem array of 21-22mers that spans the length of the core repeat. The most abundant 5′ end accounts for 466 reads. A 21 nt periodicity extends into and across the loop of the hairpin that is formed when a single repeat unit folds back on itself, indicating that these phased reads derive from stepwise processing of a duplex formed by two repeat units rather than a single-unit hairpin. Notably, knockdown of genes in S2 cells revealed that the biogenesis of siRNAs from hpRNA-2 depended on Dcr-2 and Ago-2, components of the RNAi pathway, but not on Dcr-1, Ago-1, or Drosha, components of the miRNA biogenesis pathway (Huili Guo, personal communication).

Argonaute-associated small RNAs

The interaction between the *C. elegans* Piwi protein PRG-1 and the 21U-RNAs was identified genetically with the observation that 21U-RNAs vanish on a northern blot in *prg-1* mutants. However, this approach could only be expected to work reliably if the RNAs in question are both abundant enough to reliably detect by northern blot and can only be stabilized by a single Argonaute protein. Also, that approach left open the possibility that 21U-RNA expression is somehow downstream of prg-1 expression, and that 21U-RNAs never directly interact with PRG-1. This possibility was eliminated through the co-immunoprecipitation of 21U-RNAs with PRG-1, as described above.

The sequencing of small RNAs that co-immunoprecipitate with defined members of the Argonaute family will play a crucial role in understanding the biological roles of both the Argonautes and their associated small RNAs. The sequencing of small RNAs that co-immunoprecipitate with RNA binding proteins has already been useful in conclusively demonstrating the miRNA specificity of the ALG-1 and ALG-2 Argonaute proteins from *C. elegans* (Zhang et al. 2007) and the lack of endogenously-expressed siRNA in mouse ES cells (Calabrese and Sharp 2006). The role of PRG-1 would be better understood if the complete variety of associated small RNAs were defined by high-throughput sequencing from co-immunoprecipitate. Such efforts are underway. The plethora of uncharacterized Argonaute proteins would also be better understood if the populations of small RNAs that interact with each were known.

The raising of an antibody to specifically recognize an individual member of a large protein family like the Argonautes carries with it no guarantee of success. However, such a reagent is of general utility for study of the cell biology or substrate

specificity of such a protein. In addition, such reagents may become necessary for analysis of functionally distinct but otherwise not obviously distinguishable classes of small RNAs. For instance, the identification of the upstream sequence motif associated with 21U-RNAs was assisted greatly by their non-random genomic distribution. The length distribution of endogenous siRNAs allowed some subdivision of that class, separating the 26mers from the 21-22mers. However, the observed 21-22mers may combine the primary and secondary classes of siRNA, both of which could be expressed endogenously. In cases such as these, knowledge of which small RNAs interact with which Argonaute proteins could provide crucial insight into both the commonalities and diversity within a functional class of small RNAs.

Figure legends

Figure 1. Endogenous siRNAs and hpRNA hairpins. (A) The length and 5′ nucleotide distribution of reads that overlap annotated mRNAs in the sense (green) or antisense (blue) orientation. Read counts are normalized across libraries as described in chapter 4. (B) The hpRNA-1 hairpin, with matching reads indicated by black bars. (C) The 5′ arm of the hpRNA-1 hairpin (orange) base paired with a portion of the ATP synthase beta subunit mRNA (black). (D) The hpRNA-2 hairpin formed by two adjacent repeat units. Phased reads are represented as in (A). (E) A dot plot of the hpRNA-2 genomic region. Black dots represent sense strand matches of 25nt, 0 mismatch windows. Red dots represent antisense strand matches of 15nt, ≤1 mismatch windows.

Acknowledgements

I would like to thank these collaborators on the sequencing of small RNAs from staged *C. elegans*: H. Rosaria Chiang, Shujun Luo, Gary Schroth, David P. Bartel. I would like to thank these additional collaborators on the analysis of 21U-RNAs: Pedro J. Batista, Julie Claycomb, Noah Fahlgren, Kristin D. Kasschau, Shenghua Duan, Darryl Conte Jr., Weifeng Gu, Jessica Vasale, James Carrington, Daniel Chavez, and Craig C. Mello. I would like to thank my collaborators on the sequencing of small RNAs from wild type, *dgcr8(-)*, and *dicer(-)* mouse ES cells: Joshua Babiarz and Robert Blelloch. I would like to thank Huili Guo for communicating the processing requirements of hpRNA-2.

References

Ambros, V., R.C. Lee, A. Lavanway, P.T. Williams, and D. Jewell. 2003. MicroRNAs and other tiny endogenous RNAs in *C. elegans*. *Curr Biol* **13:** 807-818.

Baird, S.E. and H.M. Chamberlin. 2006. Caenorhabditis briggsae methods. *WormBook*: 1-9.

Baskerville, S. and D.P. Bartel. 2005. Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA* **11:** 241-247.

Berezikov, E., W.J. Chung, J. Willis, E. Cuppen, and E.C. Lai. 2007. Mammalian mirtron genes. *Mol Cell* **28:** 328-336.

Bernstein, E., A.A. Caudy, S.M. Hammond, and G.J. Hannon. 2001. Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature* **409:** 363-366.

Brennecke, J., D.R. Hipfner, A. Stark, R.B. Russell, and S.M. Cohen. 2003. bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene hid in Drosophila. *Cell* **113:** 25-36.

Calabrese, J.M., A.C. Seila, G.W. Yeo, and P.A. Sharp. 2007. RNA sequence analysis defines Dicer's role in mouse embryonic stem cells. *Proc Natl Acad Sci U S A* **104:** 18097-18102.

Calabrese, J.M. and P.A. Sharp. 2006. Characterization of the short RNAs bound by the P19 suppressor of RNA silencing in mouse embryonic stem cells. *Rna* **12:** 2092-2102.

Chuang, C.F. and E.M. Meyerowitz. 2000. Specific and heritable genetic interference by double-stranded RNA in Arabidopsis thaliana. *Proc Natl Acad Sci U S A* **97:** 4985-4990.

Cox, D.N., A. Chao, J. Baker, L. Chang, D. Qiao, and H. Lin. 1998. A novel class of evolutionarily conserved genes defined by piwi are essential for stem cell self-renewal. *Genes Dev* **12:** 3715-3727.

Denver, D.R., K. Morris, J.T. Streelman, S.K. Kim, M. Lynch, and W.K. Thomas. 2005. The transcriptional consequences of mutation and natural selection in Caenorhabditis elegans. *Nat Genet* **37:** 544-548.

Dobzhansky, T.G. 1937. *Genetics and the origin of species*. Columbia Univ. Press, New York,.

Enard, W., P. Khaitovich, J. Klose, S. Zollner, F. Heissig, P. Giavalisco, K. Nieselt-Struwe, E. Muchmore, A. Varki, R. Ravid, G.M. Doxiadis, R.E. Bontrop, and S. Paabo. 2002. Intra- and interspecific variation in primate gene expression patterns. *Science* **296:** 340-343.

Farh, K.K., A. Grimson, C. Jan, B.P. Lewis, W.K. Johnston, L.P. Lim, C.B. Burge, and D.P. Bartel. 2005. The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. *Science* **310:** 1817-1821.

Fay, J.C., H.L. McCullough, P.D. Sniegowski, and M.B. Eisen. 2004. Population genetic variation in gene expression is associated with phenotypic variation in Saccharomyces cerevisiae. *Genome Biol* **5:** R26.

Hongay, C.F., P.L. Grisafi, T. Galitski, and G.R. Fink. 2006. Antisense transcription controls cell fate in Saccharomyces cerevisiae. *Cell* **127:** 735-745.

Johnston, R.J. and O. Hobert. 2003. A microRNA controlling left/right neuronal asymmetry in *Caenorhabditis elegans*. *Nature* **426:** 845-849.

Kennerdell, J.R. and R.W. Carthew. 2000. Heritable gene silencing in Drosophila using double-stranded RNA. *Nat Biotechnol* **18:** 896-898.

Khaitovich, P., W. Enard, M. Lachmann, and S. Paabo. 2006. Evolution of primate gene expression. *Nat Rev Genet* **7:** 693-702.

Kim, J., A. Krichevsky, Y. Grad, G.D. Hayes, K.S. Kosik, G.M. Church, and G. Ruvkun. 2004. Identification of many microRNAs that copurify with polyribosomes in mammalian neurons. *Proc Natl Acad Sci U S A* **101:** 360-365.

Kloosterman, W.P., E. Wienholds, E. de Bruijn, S. Kauppinen, and R.H. Plasterk. 2006. In situ detection of miRNAs in animal embryos using LNA-modified oligonucleotide probes. *Nat Methods* **3:** 27-29.

Lagos-Quintana, M., R. Rauhut, A. Yalcin, J. Meyer, W. Lendeckel, and T. Tuschl. 2002. Identification of tissue-specific microRNAs from mouse. *Curr Biol* **12:** 735-739.

Landgraf, P., M. Rusu, R. Sheridan, A. Sewer, N. Iovino, A. Aravin, S. Pfeffer, A. Rice, A.O. Kamphorst, M. Landthaler, C. Lin, N.D. Socci, L. Hermida, V. Fulci, S. Chiaretti, R. Foa, J. Schliwka, U. Fuchs, A. Novosel, R.U. Muller, B. Schermer, U. Bissels, J. Inman, Q. Phan, M. Chien, D.B. Weir, R. Choksi, G. De Vita, D. Frezzetti, H.I. Trompeter, V. Hornung, G. Teng, G. Hartmann, M. Palkovits, R. Di Lauro, P. Wernet, G. Macino, C.E. Rogler, J.W. Nagle, J. Ju, F.N. Papavasiliou, T. Benzing, P. Lichter, W. Tam, M.J. Brownstein, A. Bosio, A. Borkhardt, J.J. Russo, C. Sander, M. Zavolan, and T. Tuschl. 2007. A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell* **129:** 1401-1414.

Lau, N.C., L.P. Lim, E.G. Weinstein, and D.P. Bartel. 2001. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science* **294:** 858-862.

Lee, R.C. and V. Ambros. 2001. An extensive class of small RNAs in *Caenorhabditis elegans*. *Science* **294:** 862-864.

Liao, B.Y. and J. Zhang. 2006. Evolutionary conservation of expression profiles between human and mouse orthologous genes. *Mol Biol Evol* **23:** 530-540.

Lim, L.P., N.C. Lau, E.G. Weinstein, A. Abdelhakim, S. Yekta, M.W. Rhoades, C.B. Burge, and D.P. Bartel. 2003. The microRNAs of *Caenorhabditis elegans*. *Genes Dev* **17:** 991-1008.

Mansfield, J.H., B.D. Harfe, R. Nissen, J. Obenauer, J. Srineel, A. Chaudhuri, R. Farzan-Kashani, M. Zuker, A.E. Pasquinelli, G. Ruvkun, P.A. Sharp, C.J. Tabin, and M.T. McManus. 2004. MicroRNA-responsive 'sensor' transgenes uncover Hox-like and other developmentally regulated patterns of vertebrate microRNA expression. *Nat Genet* **36:** 1079-1083.

Nuzhdin, S.V., M.L. Wayne, K.L. Harmon, and L.M. McIntyre. 2004. Common pattern of evolution of gene expression level and protein sequence in Drosophila. *Mol Biol Evol* **21:** 1308-1317.

Okamura, K., J.W. Hagen, H. Duan, D.M. Tyler, and E.C. Lai. 2007. The mirtron pathway generates microRNA-class regulatory RNAs in Drosophila. *Cell* **130:** 89-100.

Pak, J. and A. Fire. 2007. Distinct populations of primary and secondary effectors during RNAi in C. elegans. *Science* **315:** 241-244.

Rifkin, S.A., D. Houle, J. Kim, and K.P. White. 2005. A mutation accumulation assay reveals a broad capacity for rapid evolution of gene expression. *Nature* **438:** 220-223.

Rifkin, S.A., J. Kim, and K.P. White. 2003. Evolution of gene expression in the Drosophila melanogaster subgroup. *Nat Genet* **33:** 138-144.

Roy, P.J., J.M. Stuart, J. Lund, and S.K. Kim. 2002. Chromosomal clustering of muscle-expressed genes in Caenorhabditis elegans. *Nature* **418:** 975-979.

Semenza, G.L. 1999. Perspectives on oxygen sensing. *Cell* **98:** 281-284.

Sempere, L.F., S. Freemantle, I. Pitha-Rowe, E. Moss, E. Dmitrovsky, and V. Ambros. 2004. Expression profiling of mammalian microRNAs uncovers a subset of brain-expressed microRNAs with possible roles in murine and human neuronal differentiation. *Genome Biol* **5:** R13.

Seo, T.S., X. Bai, H. Ruparel, Z. Li, N.J. Turro, and J. Ju. 2004. Photocleavable fluorescent nucleotides for DNA sequencing on a chip constructed by site-specific coupling chemistry. *Proc Natl Acad Sci U S A* **101:** 5488-5493.

Stark, A., J. Brennecke, N. Bushati, R.B. Russell, and S.M. Cohen. 2005. Animal microRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution. *Cell* **123:** 1133-1146.

Svoboda, P., P. Stein, and R.M. Schultz. 2001. RNAi in mouse oocytes and preimplantation embryos: effectiveness of hairpin dsRNA. *Biochem Biophys Res Commun* **287:** 1099-1104.

Tavernarakis, N., S.L. Wang, M. Dorovkov, A. Ryazanov, and M. Driscoll. 2000. Heritable and inducible genetic interference by double-stranded RNA encoded by transgenes. *Nat Genet* **24:** 180-183.

Wang, Y., R. Medvid, C. Melton, R. Jaenisch, and R. Blelloch. 2007. DGCR8 is essential for microRNA biogenesis and silencing of embryonic stem cell self-renewal. *Nat Genet* **39:** 380-385.

Whitehead, A. and D.L. Crawford. 2006. Neutral and adaptive variation in gene expression. *Proc Natl Acad Sci U S A* **103:** 5425-5430.

Wienholds, E., W.P. Kloosterman, E. Miska, E. Alvarez-Saavedra, E. Berezikov, E. de Bruijn, H.R. Horvitz, S. Kauppinen, and R.H. Plasterk. 2005. MicroRNA expression in zebrafish embryonic development. *Science* **309:** 310-311.

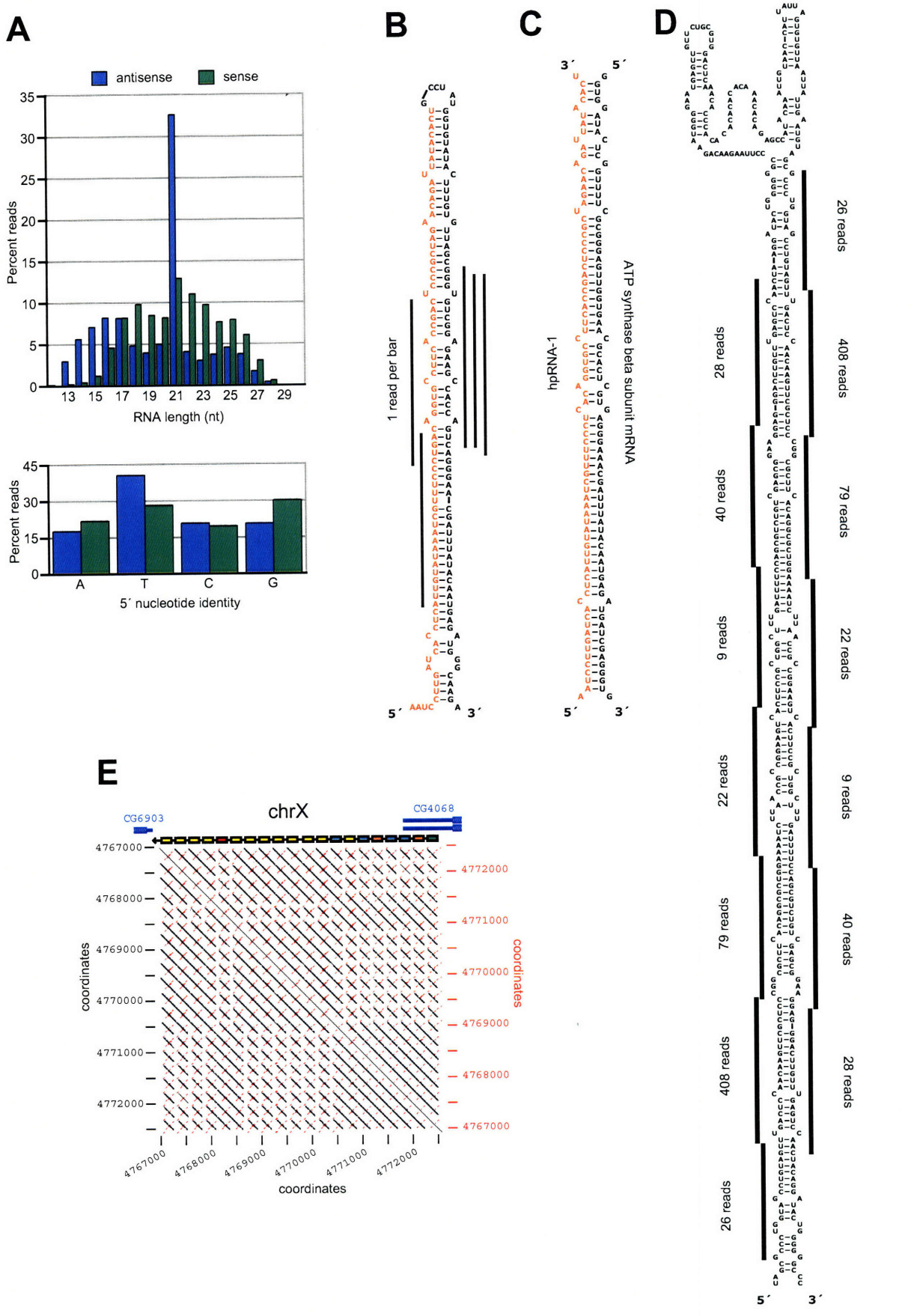
Yang, S., S. Tutton, E. Pierce, and K. Yoon. 2001. Specific double-stranded RNA interference in undifferentiated mouse embryonic stem cells. *Mol Cell Biol* **21:** 7807-7816.

Yigit, E., P.J. Batista, Y. Bei, K.M. Pang, C.C. Chen, N.H. Tolia, L. Joshua-Tor, S. Mitani, M.J. Simard, and C.C. Mello. 2006. Analysis of the C. elegans Argonaute family reveals that distinct Argonautes act sequentially during RNAi. *Cell* **127:** 747-757.

Zamore, P.D., T. Tuschl, P.A. Sharp, and D.P. Bartel. 2000. RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. *Cell* **101:** 25-33.

Zhang, L., L. Ding, T.H. Cheung, M.Q. Dong, J. Chen, A.K. Sewell, X. Liu, J.R. Yates, 3rd, and M. Han. 2007. Systematic identification of C. elegans miRISC proteins, miRNAs, and mRNA targets by their interactions with GW182 proteins AIN-1 and AIN-2. *Mol Cell* **28:** 598-613.

Figure 1


A
antisense
sense
Percent reads
RNA length (nt)
5′ nucleotide identity
B
1 read per bar
5′
3′
C
hpRNA-1
ATP synthase beta subunit mRNA
D
26 reads
408 reads
79 reads
22 reads
9 reads
40 reads
28 reads
E
chrX
CG6903
CG4068
coordinates

Curriculum Vitae

J. Graham Ruby

Education:

PhD Candidate, Massachusetts Institute of Technology (MIT), 2001-2008.
Dept. of Biology. Advisor: David Bartel

BA, Northwestern University, 1997-2001
Major: biology

Research Experience:

MIT, Dept. of Biology — Cambridge, MA
Advisor: David Bartel
Catalogued small RNA populations from *Caenorhabditis elegans*, *Drosophila melanogaster*, *Mus musculus*, and *Homo sapiens* as determined by 454 and Solexa high-throughput sequencing techniques; developed a method for predicting microRNA genes given genome sequences and applied it to *Drosophila*.

Northwestern University, Dept. of Biochemistry, Molecular Biology and Cell Biology — Evanston, IL
Advisor: Richard Morimoto
Used yeast two-hybrid system to find potential binding partners for a protein of unknown function (1999-2001)

Teaching Experience:

Instructor, MIT Dept. of Biology – Foundations of Computational and Systems Biology (2006)
Teaching assistant, MIT Dept. of Biology – Foundations of Computational and Systems Biology (2005)
Teaching assistant, MIT Dept. of Biology – Introductory Biology (2002)

Fellowships and Awards:

Gene Brown-Merck Teaching Assistant Award, MIT, 2006
Irving Klotz Award for Undergraduate Research, Northwestern University, 2001
Northwestern University Summer Research Grant, 2000

Publications:

K Okamura, WJ Chung, JG Ruby, H Guo, DP Bartel, EC Lai. (2008) The Drosophila hairpin RNA pathway generates endogenous short interfering RNAs. *Nature*. Epub ahead of print. PMID: 18463630.

JG Ruby, A Stark, WK Johnston, M Kellis, DP Bartel, and EC Lai. (2007) Biogenesis, expression, and target predictions for an expanded set of microRNA genes in *Drosophila*. *Genome Research*. 17:1850-64.

A Stark et al. (2007) Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. *Nature*. 450:219-32.

JG Ruby*, CH Jan*, and DP Bartel. (2007) Intronic microRNA precursors that bypass Drosha processing. *Nature*. 448:83-86.

JG Ruby, C Jan, C Player, MJ Axtell, W Lee, C Nusbaum, H Ge, and DP Bartel. (2006) Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. *Cell*. 127:1193-1207.

* these authors contributed equally to this work.

Skills:

Python, Java, Perl, Javascript, Lisp (Scheme), HTML