Current Biology, Vol. 13, 807–818, May 13, 2003, ©2003 Elsevier Science Ltd. All rights reserved. DOI 10.1016/S0960-9822(03)00287-2

# MicroRNAs and Other Tiny Endogenous RNAs in *C. elegans*

**Victor Ambros,[1,3,*] Rosalind C. Lee,[1,3]**
**Ann Lavanway,[1] Peter T. Williams,[1]**
**and David Jewell[2]**
[1]Dartmouth Medical School Department of Genetics
[2]Dartmouth College Research Computing
Hanover, NH 03755

## Summary

**Background:** MicroRNAs (miRNAs) are small noncoding RNAs that are processed from hairpin precursor transcripts by Dicer. miRNAs probably inhibit translation of mRNAs via imprecise antisense base-pairing. Small interfering RNAs (siRNAs) are similar in size to miRNAs, but they recognize targets by precise complementarity and elicit RNA-mediated interference (RNAi). We employed cDNA sequencing and comparative genomics to identify additional *C. elegans* small RNAs with properties similar to miRNAs and siRNAs.

**Results:** We found three broad classes of small RNAs in *C. elegans*: (1) 21 new miRNA genes (we estimate that *C. elegans* contains approximately 100 distinct miRNA genes, about 30% of which are conserved in vertebrates; (2), 33 distinct members of a class of tiny noncoding RNA (tncRNA) genes with transcripts that are similar in length to miRNAs (approximately 20–21 nt) and that are in some cases developmentally regulated but are apparently not processed from a miRNA-like hairpin precursor and are not phylogenetically conserved; (3) more than 700 distinct small antisense RNAs, about 20 nt long, that are precisely complementary to protein coding regions of more than 500 different genes and therefore seem to be endogenous siRNAs.

**Conclusions:** The presence of diverse endogenous siRNAs in normal worms suggests ongoing, genome-wide gene silencing by RNAi. miRNAs and tncRNAs are not predicted to form complete Watson-Crick hybrids with any *C. elegans* RNA target, and so they are likely to regulate the activity of other genes by non-RNAi mechanisms. These results suggest that diverse modes of small RNA-mediated gene regulation are deployed in normal worms.

## Introduction

microRNAs (miRNAs) are about 22 nt (~22 nt) in length and have been shown in some cases to function as antisense regulators of gene expression (reviewed in [1] and [2]). These small RNAs are cleaved by Dicer from a foldback hairpin RNA transcript, yielding a mature miRNA from one strand of the hairpin [3–5]. *lin-4* and *let-7*, the founding members of the miRNAs, control developmental timing in *C. elegans* and were identified based on their mutant phenotypes [6, 7]. Additional miRNA genes have been identified in animals and plants through molecular and genomics approaches [8–15]. Many of these miRNAs are phylogenetically conserved, suggesting strong evolutionary pressure to conserve their primary sequences [8–16]. Conserved miRNAs probably acquired multiple antisense targets, which would inhibit subsequent evolutionary variation in the miRNA sequence. Nonconserved miRNAs could play more species-specific roles or regulate less well-constrained target sequences. Although more that 200 different miRNAs have been identified in plants and animals, we do not yet have a comprehensive picture of the scope and spectrum of miRNAs in particular species, which miRNAs are evolutionarily conserved, or what other classes of ~22 nt noncoding RNAs function in eukaryotic cells.

Small interfering RNAs (siRNAs) are similar in length to miRNAs but generally differ from miRNAs in their origin [17]. siRNAs are produced by Dicer cleavage of a long double-stranded RNA into short duplexes of about 21–25 nt in length. Although siRNAs are initially double stranded [18], the antisense strand probably associates with the mRNA target as part of the RNA induced silencing complex (RISC) [19–21]. In contrast to siRNAs, miRNAs are produced by the processing of a small (approximately 70 nt) single-stranded hairpin precursor transcript [3–6]. Although most animal miRNAs are thought to imprecisely base-pair with their targets and thereby elicit translational repression [6, 22–25], it has been shown that if an animal miRNA encounters a target with complete complementarity, it can enter the RNAi pathway and trigger target cleavage [26, 27]. This suggests that the chief difference between miRNAs and siRNAs in animals lies in their origins (either from an endogenous small hairpin RNA or from a bimolecular duplex, respectively) and that the outcome of miRNA or siRNA action on a target RNA depends more specifically on the nature of the interaction of the small RNA with the target (imprecise versus precise complementarity).

RNAi-mediated gene silencing has been implicated in surprisingly diverse phenomena of normal cells. In plants, some miRNAs naturally contain complete complementarity to messenger RNAs and seem to function as endogenous siRNAs by directing cleavage of their mRNA targets [13, 28–30]. In the fission yeast, *Sacchromyces pombe*, the production of small endogenous RNAs from centromeric repeat regions, together with RNAi-related protein machinery, seems to be associated with the transcriptional silencing of the corresponding centromeric heterochromatin [31, 32]. In animal cells, RNAi is important for the silencing of transposons and transgenic repeats in the germ line [33–35]. These observations suggest that small RNAs of the ~22 nt class could be broadly employed in normal eukaryotic cells for diverse modes of gene regulation. A full understanding of the regulatory mechanisms and biological functions carried out by small antisense RNAs in a given organism requires a thorough characterization of the number, diversity, and function of miRNA genes and the

*Correspondence: vambros@dartmouth.edu
[3]These authors contributed equally to this work.

identification of other genes that produce similar small antisense RNAs.

To identify as many of the *C. elegans* miRNA genes as possible, and to also identify other kinds of ~22 nt endogenous RNAs expressed in *C. elegans*, we employed phylogenetic comparisons of noncoding genomic sequences and large-scale sequencing of size-selected cDNA libraries. To permit the detection of novel classes of small RNAs that may have somewhat different biogenesis than miRNAs, we generated the cDNA libraries by methods not dependent on the presence of the 5′ phosphate typical of Dicer products [8].

Among the new small RNAs detected are 21 distinct miRNA sequences that were not previously described in *C. elegans*. We estimate the size of the *C. elegans* miRNA gene family to be somewhat more than 100 distinct genes. About 80% of the *C. elegans* miRNA genes are conserved in a related nematode, *Caenorhabditis briggsae*, and about 30% have apparent homologs in insects and/or vertebrates. In addition to miRNAs, we also identified 33 distinct members of a second class of small RNA, which we call "tiny noncoding RNAs" (tncRNAs). tncRNAs are similar in length to miRNAs, but unlike miRNAs, they are apparently not processed from short hairpin precursors and therefore are not counted among the miRNAs [36]. However, like miRNAs, tncRNAs are transcribed from noncoding genomic sequence, and can exhibit developmentally regulated expression. tncRNAs also resemble miRNAs in their lack of precise complementarity to any worm mRNA. Therefore, tncRNAs could act in a fashion exemplified by the canonical miRNAs, *lin-4* and *let-7*, which regulate translation of mRNA targets through imprecise antisense base-pairing [6, 22–25]. Finally, we identified a third class of ~22 nt RNAs in *C. elegans* that are produced directly from protein coding sequences and are predicted to have precise antisense complementarity to messenger RNAs. These apparent endogenous siRNAs represent sequences from more than 500 different protein-coding genes. These results suggest that RNA-mediated gene silencing may be broadly deployed in normal worm cells through the action of diverse classes of small RNAs.

## Results and Discussion

### cDNA Cloning of Three Classes of Very Small Expressed RNA Sequences

miRNAs have been identified by computational and cDNA cloning approaches tailored to the key features of the canonical miRNAs, *lin-4* and *let-7* [8–15]. These features include a fold-back hairpin RNA precursor [6, 7], evolutionary conservation of the hairpin structure [16], and Dicer-dependent processing of an ~22 nt miRNA from one arm of the precursor [3–5]. To identify new small RNAs in the 22 nt size class, including novel miRNAs, we generated cDNA libraries by using size-fractionated RNA from total worm RNA. RNA samples were treated with phosphatase prior to cloning, and cDNA synthesis was performed without a 5′ RNA ligation step, obviating the need for a 5′ phosphate on the RNA [8]. Sequences were obtained for 2957 small cDNA sequences that ranged in length from 17 to 28 nucleotides. All of these cDNA sequences precisely matched the *C. elegans* genome at least once, as determined by Blast searches [37] to the WS83 version of the Wormbase *C. elegans* genomic sequence. Most of these sequences were represented by only one clone. Others (presumably from more abundant small RNAs) were represented by as many as 128 independent clones (see Supplemental Data). About half (1496) of the ~22 nt cDNAs corresponded to the expressed (sense) strand of loci that encode longer, previously-identified coding or noncoding RNAs, and hence they were probably generated from degradation products of these longer RNAs (Table 1).

A total of 746 other cDNA sequences precisely match genomic sequences antisense to predicted protein-coding exons. More than 95% of these cDNAs only match exons, and no other location in the genome, suggesting that these correspond to short RNAs produced by transcription of the protein coding genes themselves. (The other approximately 5% of these cDNAs matched not only coding sequence but also other sequences not annotated as protein-coding. In many of these cases, the other matches seem to be rearranged gene fragments.) We interpret these 746 antisense cDNA sequences to represent endogenous siRNAs produced in the course of ongoing RNA-mediated interference (RNAi) in the worms from which the RNA samples were obtained (see below).

The other 715 cDNAs only match regions of the *C. elegans* genome that lie between protein coding genes or within introns (Table 1). Most of these potential small noncoding RNAs were not detected by Northern blot hybridization (see Experimental Procedures), and so they have not been given a specific designation as small RNAs (Table 1). The noncoding transcripts that were detected by Northern blot were miRNAs (Table 2) and another class of small RNA (Table 3) called "tiny noncoding RNAs" (tncRNAs). tncRNAs are similar in length to miRNAs, but they have certain properties that appear to distinguish them from miRNAs (see below).

### Identification of 12 New *C. elegans* MicroRNA Genes by cDNA Cloning

The 715 distinct intergenic sequences identified from cDNA libraries included 38 of the 58 previously identified *C. elegans* microRNAs (see complete list in Supplemental Data; [8–10]) and 12 new microRNAs (Table 2). All 12 new miRNAs correspond to single-copy genomic loci. We confirmed these miRNA genes by using a set of criteria designed to distinguish bona fide miRNAs from other small RNAs [36]. In particular, we used the mfold program [38] to test computationally for the potential of genomic sequences surrounding the cDNA sequence to form a short hairpin-like secondary structure typical of miRNA precursors (see Experimental Procedures). All the new miRNAs listed in Table 2 have such a predicted hairpin precursor sequence (see Supplemental Materials). In many cases, a presumed precursor transcript of about 65–70 nt was detected by Northern blot, in addition to the ~22 nt species. The precursor was usually a minor species compared to the ~22 nt miRNA, although in some cases the precursor predominated (see, for example, Figure 3C), suggesting relatively inefficient

Table 1. Genomic Distribution of 2957 Distinct cDNA Sequences from *C. elegans* Size-Selected RNA Samples

| Genomic Location[a] | Number of Distinct Sequences: Subtotal | Total | Length[b] |
|---|---|---|---|
| Sense strand fragments | | 1496 | |
| Sense strand of larger noncoding RNAs[c] | 1447 | | |
| Sense strand of mRNAs | 49 | | 20.55 ± 0.52[d] |
| Antisense to protein coding sequence | | 746 | 20.32 ± 0.09[e] |
| Intergenic sequences | | 715 | |
| 38 Known microRNAs[f] + 12 new microRNAs[g] | 50 | | 21.84 ± 0.47[h] |
| Other tiny noncoding RNAs (tncRNAs)[j] | 33 | | 19.97 ± 0.39[i] |
| Chromosome X cluster[k] | 41 | | |
| Not validated by Northern blot[l] | 591 | | |
| All sequences | | 2957 | |

[a] Wormbase WS94.
[b] Arithmetic mean and 95% confidence limits of the standard error.
[c] Includes rRNA, tRNA, URNA, etc.
[d] Range, 18–29; standard deviation (SD) = 1.87.
[e] Range, 18–26; SD = 1.28.
[f] [8, 9, 10]; see Supplemental Data.
[g] Table 2.
[h] range, 18–26; SD = 1.68.
[i] Range, 18–23; SD = 1.15.
[j] Table 3.
[k] See text and Supplemental Data.
[l] Includes sequences not tested by Northern blot, sequences tested but not detected, and sequences detected, but not in the ~22 size fraction.

processing. Intergenic cDNA sequences whose in vivo expression as an ~22 nt RNA was confirmed by Northern blot, but for whom no satisfactory hairpin precursor structure could be predicted with mfold, were classified as tncRNAs (see below; Table 3).

### Computational Identification of Nine New *C. elegans* MicroRNA Genes

Some miRNAs might have been missed by cDNA cloning because of their relatively low abundance or other factors affecting their representation in cDNA libraries. Another approach for identifying novel noncoding RNAs is the analysis of regions of genomic sequence alignment between different species ([8, 15, 39]; reviewed by [40]). In previous work, we employed genomic sequence comparisons between *Caenorhabditis elegans* and *Caenorhabditis briggsae* to identify novel miRNAs [8]. That computational search for miRNAs was limited to only the 10% of the *C. briggsae* genome that was then available. In this current study, we utilized the annotated, complete genomic sequence assemblies of *C. briggsae* and *C. elegans* available online from Wormbase. All the regions of alignment between *C. elegans* and *C. briggsae* genomes were obtained from Wormbase and processed computationally to identify single-copy *C. elegans* sequences that lie outside of protein coding sequences. These sequences were further processed as described in the Experimental Procedures to identify candidate miRNA sequences that are conserved between *C. elegans* and *C. briggsae*. This computational analysis yielded 6985 distinct pairs of *C. briggsae* and *C. elegans* miRNA candidates that contained at least 19 consecutive nucleotides of interspecific identity. As described in the Experimental Procedures, each *C. elegans/C. briggsae* candidate pair was assigned a "hairpin score" (Figure 1A) based on the candidates' similarity to each other and to a miRNA hairpin structure model. Among 36 known miRNAs that contain at least 19 nt of identity between *C. briggsae* and *C. elegans*, 31 were included in the 288 candidate pairs with scores of 6.5 or greater (Figure 1A; see also Supplemental Data). We tested the other candidates in this range for expression by Northern blot hybridization by using probes to each arm of the predicted *C. elegans* hairpin. Among these candidates, we confirmed the expression of six new miRNAs that were not among our cDNA clones and that had not been previously described (Table 2).

Because we used relatively long Northern hybridization probes (about 50 nt) to verify computationally predicted miRNAs, our data could not precisely indicate the positions of their 5′ and 3′ ends. However, five of the miRNAs that we predicted computationally were also identified as cDNA sequences by Lim and coworkers [15], allowing specification of their 5′ and 3′ ends (Table 2). The likely 5′ and 3′ ends of the sixth computationally predicted miRNA (*mir-124*) are specified in Table 2 because of its similarity to a previously identified miRNA [11].

A second computational approach for identifying new miRNA genes was based on the observation that many miRNAs, even those from separate taxa, have similar ~22 nt sequences [8–16, 39]. Accordingly, we used the patscan pattern-matching program [41] to search *C. elegans* genomic sequences for 22 nt blocks similar to each of the known animal miRNAs (those previously published, plus those confirmed here). Up to ten mismatches and one single-nucleotide deletion or insertion were allowed in the search pattern. All matches were tested computationally by mfold to identify candidate precursor structures, and these were scored for their similarity to a miRNA hairpin model as described above. 150 high-scoring candidates were tested by Northern blot hybridization, and the expression of three novel

Table 2. New *C. elegans* MicroRNA Genes

| Name[a] | Source[b] | Dev[c] | Cb[d] | Family[e] | Fold[f] | Sequence | Chromosome and Cosmid Address[g,h] | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| mir-124 | cg | U | 100 | mir-124 iv | 3′ 9.1 | UAAGGCACGCGGUGAAUGCCA | IV | C29E6[i] | 8382 | 8402 |
| mir-227 | cDNA1 | U | 100 | | 5′ 6.0 | AGCUUUCGACAUGAUUCUGAAC | III | K01F9 | 1038 | 1017 |
| mir-228 | cg | P | 100 | mir-182 v | 5′ 7.3 | AAUGGCACUGCAUGAAUUCACGG | IV | T12E12 | 23979 | 24001 |
| mir-229 | cDNA3 | U | 0 | | 5′ 2.0 | AAUGACACUGGUUAUCUUUUCCAUCG | III | Y48G9A | 92591 | 92616 |
| mir-230 | cg | C | 100 | | 3′ 6.5 | GUAUUAGUUGUGCGACCAGGA | X | F46G11 | 21143 | 21163 |
| mir-231 | cDNA1 | U | 95 | | 3′ 6.9 | UAAGCUCGUGAUCAACAGGC | III | R13A5 | 6384 | 6365 |
| mir-232 | cDNA2 | U | 100 | | 3′ 5.8 | UAAAUGCAUCUUAACUGCGGU | IV | F13H10 | 3209 | 3189 |
| mir-233 | cg | U | 100 | | 3′ 7.3 | UUGAGCAAUGCGCAUGUGCGG | X | W03G11[j] | 445 | 425 |
| mir-234 | cg | U | 100 | mir-137 v | 3′ 6.6 | UUAUUGCUCGAGAAUACCCUU | II | C13B4 | 1911 | 1891 |
| mir-235 | mism | nd | 85 | mir-25 i | 3′ 7.2 | UAUUGCACUCUCCCCGGCCUGA | I | T09B4 | 12677 | 12656 |
| mir-236 | cDNA1 | U | 100 | mir-8 iv | 3′ 8.3 | UAAUACUGUCAGGUAAUGACG | II | C52E12 | 37155 | 37133 |
| mir-237 | cDNA1 | Pf | 86 | lin-4 iv | 5′ 8.3 | UCCCUGAGAAUUCUCGAACAGC | X | F22F1 | 5551 | 5572 |
| mir-238 | cDNA2 | U | 90 | mir-12 i | 3′ 6.8 | UUUGUACUCCGAUGCCAUUCAG | III | K01F9 | 3054 | 3033 |
| mir-239a | mism | Ph | 90 | mir-12 i | 5′ 7.0 | UUUGUACUACACAUAGGUACUGG | X | C34E11 | 1237 | 1259 |
| mir-256 | mism | U | 0 | mir-1 iv | 3′ 5.4 | UGGAAUGCAUAGAAGACUGUA | V | T07H8 | 6363 | 6343 |
| mir-257 | cDNA2 | U | 0 | | 5′ 6.4 | GAGUAUCAGGAGUACCCAGUGA | IV | Y102A5D | 18800 | 18821 |
| mir-258 | cDNA2 | U | 0 | | 5′ 4.7 | GGUUUUGAGAGGAAUCCUUUU | X | AC8 | 11066 | 11086 |
| mir-259 | cg | U | 100 | | 5′ 7.6 | AGUAAAUCUCAUCCUAAUCUGG | V | F25D1 | 9514 | 9493 |
| mir-260 | cDNA1 | prEh | 0 | | 3′ 4.1 | UGAUGUCGAACUCUUGUAG | II | F39E9[k] | 22387 | 22405 |
| mir-261 | cDNA1 | P | 0 | | 3′ 2.5 | UAGCUUUUUAGUUUUCACG | II | B0034 | 25257 | 25275 |
| mir-262 | cDNA3 | C | 0 | | 3′ 1.7 | GUUUCUCGAUGUUUUCUGAU | V | ZK384 | 7973 | 7992 |

[a] miRNA genes are named according to established conventions [36]; all miRNAs were confirmed by Northern blot hybridization.
[b] cDNAn, miRNA was identified by cDNA sequencing (n = number of clones); cg, miRNA was identified by comparative genomics with *C. elegans*::*C. briggsae* alignments; and mism, miRNA was identified a search of the *C. elegans* genomic sequence for sequences that mismatch each known miRNA, as described in the text.
[c] Summary of developmental Northern blot hybridization: U, uniform expression from embryo through adult; Ef, embryonic expression is shut off by feeding; Pf, postembryonic expression is activated by feeding; EfPh, embryonic expression of one hybridizing species is shut off by feeding, and postembryonic expression of another species is activated by hatching; P, postembryonic expression (food or hatching not assessed); C, partially embryonic and partially postembryonic; prE, processing primarily in embryos; and nd, developmental profile not tested.
[d] Percent nucleotide identity between the *C. elegans* ~22 nt miRNA and an apparently syntenic *C. briggsae* ortholog (Wormbase version WS94).
[e] Families of similar miRNAs are named for the first member of the group to be identified; families that include insect or vertebrate members are indicated by i and v, respectively.
[f] Predicted hairpin precursor folding topology (3′ or 5′ ~22 nt miRNA is contained within the 3′ or 5′ arm of the hairpin, respectively) and hairpin score (an average of *C. elegans* and *C. briggsae* scores is shown for conserved miRNAs; see Figure 1 and Experimental Procedures; see also Supplemental Data for secondary structures).
[g] Cosmid name, with start and end of the miRNA sequence within the cosmid.
[h] All miRNA genes are in noncoding genomic sequences; noteworthy annotations (Wormbase WS94);
[i] In an intron of C29E6.2 (sense).
[j] In an intron of W03G11.4 (sense).
[k] Antisense to an intron of F39E9.7.

miRNA genes was confirmed (Table 2). Together, our two computational approaches identified nine new miRNA genes. Therefore, the results of our cDNA and computational approaches bring the set of new miRNA genes to 21. These three methods, cDNA cloning, interspecific genomic comparisons, and miRNA similarity searches, were complementary; several miRNAs were identified uniquely by each method.

Seven of the 21 new miRNAs appear not to be conserved in *C. briggsae*. This fraction (33%) seems somewhat unusual because over all, only about 16% of the 96 known *C. elegans* miRNAs are not conserved in *C. briggsae* [8, 9, 15]. Perhaps these seven miRNAs are each relatively rare transcripts, since they were represented by only a few clones (Table 2). If the conserved miRNAs tend to be enriched among those that are more abundant, that could account for a bias for nonconserved miRNAs among these relatively rare ones.

Another apparently unusual feature of this group of seven nonconserved miRNAs is the occurrence of a 5′ G on three of them (*mir-257*, *mir-258*, and *mir-262*; Table 2). Our method of cDNA cloning utilizes a linker oligonucleotide that ordinarily contains a triple-G stretch at the linker/cDNA junction [8]. This could theoretically introduce a false 5′ G if occasionally a fourth G were to be introduced at the junction; the extra G would go unnoticed in cases such as *mir-257*, *mir-258*, and *mir-262*, where the genomic sequence contains a G at the corresponding position. However, *mir-257*, *mir-258*, and *mir-260* were each identified by more than one cDNA clone, all of which contained the 5′ G, so it is likely that in these cases, the 5′ G is authentic. Our cloning approach may be particularly efficient at capturing RNAs with a 5′ G; 85% of the tncRNAs and siRNAs we identified contained a 5′ G (Table 3; Supplemental Data).

### One Hundred or More *C. elegans* miRNA Genes

The *C. elegans* genome contains at least 96 miRNA genes if one counts the 21 new miRNA genes described here, the 58 genes described previously [8–10], and 17 additional *C. elegans* miRNAs among those recently identified by Lim and coworkers [15]. These probably represent most of the *C. elegans* miRNA genes that are relatively abundant (permitting them to be cloned as

Table 3. Other *C. elegans* Tiny Noncoding RNA Loci

| Name[a] | Hits | Dev[b] | Dicer[c] | Fold[b] | Transcript Sequence | Chromosome and Cosmid Address[d] | | | |
|---|---|---|---|---|---|---|---|---|---|
| tncR1 | 1 | P | nd | | AAAUUAGAAAAACAUGAGC | I | F31C3[e] | 12572 | 12590 |
| tncR2 | 1 | E+ | gone | | GUUGAAACUGUAAAAAAUUAAA | IV | Y59H11AR[f] | 16809 | 16788 |
| tncR3 | 2 | nd | nd | | GAAAUCUCAUUCAUUCCUGC | IV | K08F11 | 23939 | 23920 |
| tncR4 | 2 | C | - | | GUUAGUCGUGUUCCGUUGCA | V | K02E2[g] | 32314 | 32295 |
| tncR5 | 2 | E | pre | | GAAACUCUUGUAACAUCCG | III | Y39E4B[h] | 96058 | 96040 |
| tncR6 | 2 | nd | nd | | GUACAUUUUCGAUGAACUGU | II | C47G2 | 28501 | 28520 |
| tncR7-1 | 2 | U | pre | | GACAACCAUUCCGUAGGCUG | II | F39E9 | 23553 | 23572 |
| tncR7-2 | 2 | U | pre | | GACAACCAUUCCGUAGGCUG | II | F39E9 | 26072 | 26091 |
| tncR7-3 | 2 | U | pre | | GACAACCAUUCCGUAGGCUG | II | F39E9 | 21822 | 21841 |
| tncR8-1 | 2 | E | pre | | GACCUCCAUGUAAACGUACAA | II | F39E9[i] | 22304 | 22324 |
| tncR8-2 | 2 | E | pre | | GACCUCCAUGUAAACGUACAA | II | F39E9[i] | 25338 | 25358 |
| tncR9 | 2 | E | nd | | GAAACGGAAUCGCUGUCCC | II | F39E9[i] | 22554 | 22572 |
| tncR10 | 2 | E | gone | | GUAAAAAUUCUUCUCAUGUG | II | F39E9[i] | 22451 | 22470 |
| tncR11-1 | 2 | U | gone | | GUACAACAAAAUUAUUGCG | II | F39E9[i] | 25353 | 25371 |
| tncR11-2 | 2 | U | gone | | GUACAACAAAAUUAUUGCG | II | F39E9[i] | 22319 | 22337 |
| tncR12 | 2 | U | nd | | GUUUUUACAAUUACACGACU | IV | Y37E11B[j] | 1516 | 1497 |
| tncR13 | 2 | E | - | | GCAAUGUGCGGACAAAAAGG | II | F10E7[k] | 21646 | 21627 |
| tncR14 | 2 | U | nd | | GAAAAAGUAAGUUUAUUUAA | V | B0250[l] | 15927 | 15946 |
| tncR15 | 2 | U | pre | | GAAUCAAUUUAUAUGUGC | I | W04A8 | 17669 | 17652 |
| tncR16 | 2 | C | pre | | GUCGUAGGGUACCCCAUUGCC | I | Y37E3 | 91119 | 91139 |
| tncR17 | 2 | C | pre | | GUAGAUAUCAAGUAUGAGGU | III | F40G9 | 17568 | 17587 |
| tncR18 | 2 | nd | nd | | GUCUUAGUUGCACGUUGU | I | Y37E3 | 90887 | 90904 |
| tncR19 | 2 | E | nd | | GUUGAAUAUAGUGUUUUAAU | V | Y46H3C | 4052 | 4033 |
| tncR20 | 2 | Ph | nd | | GAAAAUUUGUUUCGCUGACCG | II | Y53F4B[m] | 92202 | 92222 |
| tncR21 | 2 | E | nd | | GUCAGCGUCUUCUCACAC | X | F02C12[n] | 8806 | 8789 |
| tncR22 | 1 | U | nd | | GACAAUCAGCCGAUCAUAGUC | I | Y37E3 | 90791 | 90811 |
| tncR23 | 1 | C | gone | | UUGCAGAUAACUGGUACAAG | IV | K02E2 | 35092 | 35073 |
| tncR24-1 | 1 | U | nd | | GAUUUCUGAGCUAAAAAUUGGG | X | T08D2 | 7315 | 7336 |
| tncR24-2 | 1 | U | nd | | GAUUUCUGAGCUAAAAAUUGGG | V | Y60A3A | 4502 | 4523 |
| tncR25 | 2 | U | pre | 5′ | GUAUUCCAACUUUCUGAACU | II | C06C3 | 19336 | 19355 |
| tncR26 | 2 | U | nd | 3′ | GCAAUGUGUUUAUUUUGAGG | II | F59H5 | 7896 | 7915 |
| tncR27 | 2 | U | nd | 5′ | GGAUUUUCAGUAAUGGCAC | I | Y71A12B[o] | 89248 | 89230 |
| tncR28 | 2 | U | nd | | GAAUUGUGAAGAUAUCUCU | II | F59H6 | 31097 | 31079 |
| tncR29 | 3 | U | gone | | GAAUAGAUAUCCGGUAUGA | II | K08A2[p] | 33550 | 33532 |
| tncR30 | 2 | C | gone | 3′ | UUGCAGAUAACUGGUACAAG | V | K02E2 | 35092 | 35073 |
| tncR31 | 1 | EfPh | - | | GUUUAUAUUGAGAUUUGAUG | IV | Y59H11AR | 16121 | 16102 |
| tncR32-1 | 1 | Ef | nd | | UAUCGAUAUGUUUAUUGAGAG | II | K08A2[q] | 32895 | 32875 |
| tncR32-2 | 1 | Ef | nd | | UAUCGAUAUGUUUAUUGAGAG | IV | T05E11[r] | 4220 | 4200 |
| tncR32-3 | 1 | Ef | nd | | UAUCGAUAUGUUUAUUGAGAG | IV | T05E11[s] | 3131 | 3111 |
| tncR33 | 0[t] | P | nd | 3′ | UGUCAUGGAAGGCGCCCCUUUCU | V | F02D8 | 21698 | 21676 |

[a] RNAs are named according to the tncR designation (tiny noncoding RNA), as described in the text.

[b] The notation for expression (as determined by Northern blot) is the same as in Table 2; E+ indicates expression detected in embryos and in early larval stages; the notation for putative precursor folds is the same as in Table 2; only the indicated four tncRNAs have predicted hairpin folds (see Figure 1 and Supplemental Data).

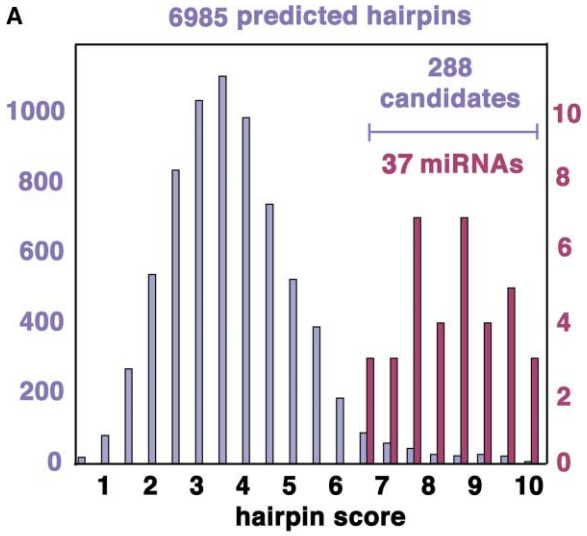
[c] Expression of ~21 nt tncRNA in total RNA of Dicer mutant animals, as determined by Northern blot; -, expression normal; gone, ~21 nt RNA absent; pre, larger precursor accumulates; nd, not tested.

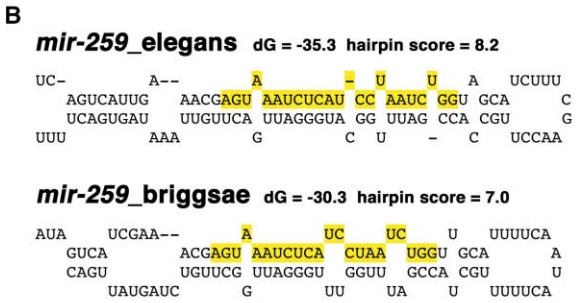
[d] All tncRNA genes are located in noncoding sequences; noteworthy annotations (Wormbase version WS94) are as follows: [e] antisense to intergenic linker of operon CEOP1776; [f] in intron of Y59H11AR.4 (sense); [g] antisense to 3′ UTR of K02E2.6; [h] antisense to EST OSTR047C4_1, which contains sequence from an intron (and/or perhaps an unknown exon) of Y39E4B.14; [i] antisense to intron of F39E9.7; [j] antisense to intergenic linker of operon CEOP4096; [k] possibly antisense to long 3′ UTR of F10E7.4; [l] antisense to intron of B0250.8; [m] in intron of Y53F4B.15 (sense) and antisense to multiple apparently noncoding ESTs; [n] antisense to apparently noncoding EST yk521h3.5, which defines apparent 5′ UTR of F02C12.3; [o] antisense to intron of Y71A12B.10; [p] antisense to multiple ESTs that define the 5′ UTR of K08A2.1; [q] antisense to sequences upstream of 5′ UTR of K08A2.1 (see *tncR29*), in a region with multiple bidirectional, apparently noncoding, ESTs; [r] in 5′ end of EST yk550a12.5, which is antisense to numerous apparently noncoding ESTs; [s] antisense to numerous apparently noncoding ESTs (same as for *tncR32-3*); and [t] identified by similarity to *mir-47* (18 of 22 nt).

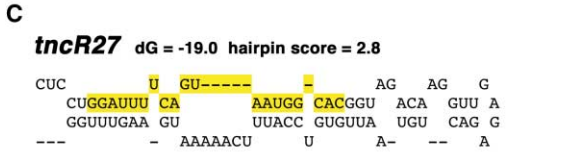
cDNAs and to be detected by Northern blots) and/or well conserved in sequence (permitting them to be identified by comparative genomics). However, there could be another class of miRNA that are less abundant and that are not well conserved in sequence and secondary structure between *C. briggsae* and *C. elegans*. These could lie still undiscovered among the noncoding cDNAs or computationally predicted candidates that we tested by Northern blot but could not detect (Table 1). They could also lie among the many hundreds of hairpin candidates that did not score high enough to merit validation (Figure 1A). Confirmation of very low abundance miRNAs awaits the application of detection methods more sensitive than Northern blots; also, higher-throughput assays for miRNA expression would allow for validation of greater numbers of computationally predicted candidates.

### Families of Similar miRNAs

Some microRNA genes are similar to each other in the sequence of their ~22 nt transcripts and in their precursor secondary structures [8–16, 39]. Of 79 *C. elegans*



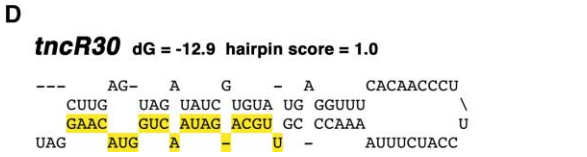




Figure 1. Secondary-Structure Prediction of Putative miRNA and tncRNA Precursors

(A) MicroRNA candidates identified by comparative genomic analysis of *C. elegans* and *C. briggsae* sequences. *C. elegans/C. briggsae* genomic sequence alignments were obtained from Wormbase and processed computationally as described in the Experimental Procedures. A total of 6985 *C. elegans/C. briggsae* aligned sequence pairs were scored via a secondary-structure model based on the hairpin precursors of published miRNAs. The score range from 6.5 to 10 contained 31 miRNAs that had been previously published or identified as cDNA sequences in this study. Six additional novel miRNA sequences were confirmed by Northern blot from among the 288 best-scoring candidates.

(B) Examples of miRNA candidates predicted computationally based on *C. elegans/C. briggsae* sequence and structural alignment.

(C and D) Predicted putative foldback structures for two tncRNAs. These structures do not meet minimal criteria for annotation as miRNAs [36], based on excessive bulges (C) and fewer than 16 base pairs involving the small RNA (D).

miRNAs, (58 from references [8 and 9] and 21 from this work), 66 (84%) are conserved in *C. briggsae* to at least 75% sequence identity, and 48 (61%) to at least 90% identity. Furthermore, 27 of these 79 worm miRNAs (34%) have apparent homologs in insects or vertebrates (Figure 2). Phylogenetic conservation of a miRNA sequence has been proposed to reflect conservation of its complementary target sites within mRNAs of multiple genes [16]. If homologs perform similar roles in distinct species, then perhaps almost one-third of the *C. elegans* miRNAs may be valid experimental models for understanding the roles of their counterparts in higher animals.

In some miRNA families (exemplified by the *lin-4* and *let-7* families), sequence similarity is greatest in the 5′ portion of the miRNA sequence (Figure 2), consistent with target recognition primarily via these 5′ sequences [6, 22, 42]. For other miRNA families (such as the *mir-8*, *mir-30*, or *mir-182* families), similarity is more uniformly distributed along the miRNA sequence (Figure 2), suggesting that in these cases, target recognition may more uniformly involve the whole length of the miRNA.

miRNAs of similar sequence could recognize a consensus target sequence and hence might act on common target mRNAs. This could allow redundant control of target gene expression by multiple miRNA genes and/or the deployment of different miRNAs for the regulation of a target in different developmental contexts. For example, *mir-237* is similar in sequence to *lin-4* and is predicted to recognize some of the same sites in the UTRs of heterochronic genes as *lin-4*. *mir-237* RNA accumulates during the L3 and L4 stages (our unpublished data) and so could contribute to the repression of heterochronic gene translation during later larval stages, perhaps acting in combination with *let-7* on targets that contain predicted complementary elements for both *let-7* and *lin-4* [7, 42].

## tncRNAs Share Some Properties with MicroRNAs but Are Not Processed from MicroRNA-like Hairpin Precursors

Some cDNA sequences were similar to miRNAs in that they were from genomic locations outside of the protein coding sequence and were detectable as ~20–22 nt species by Northern blots. However, because these RNAs did not meet a key secondary-structure criterion for classification as miRNAs [36], they are referred to by a separate designation, "tiny noncoding RNAs," or tncRNAs (Table 3). tncRNAs are not predicted to form the sort of small foldback hairpin precursors characteristic of miRNAs. Also, none of the tncRNAs are well conserved in sequence outside *C. elegans*, and tncRNAs do not seem to fall into families of related sequences (our unpublished data), further distinguishing them from miRNAs.

Although none of the tncRNAs are predicted to come from miRNA hairpin precursors, five of the tncRNAs are predicted to form potential foldback structures somewhat reminiscent of miRNAs (Table 3; Figures 1C and 1D; see also Supplemental Data). However, these putative tncRNA precursor structures deviate significantly from the miRNA hairpins in key characteristics [36]; for example, they

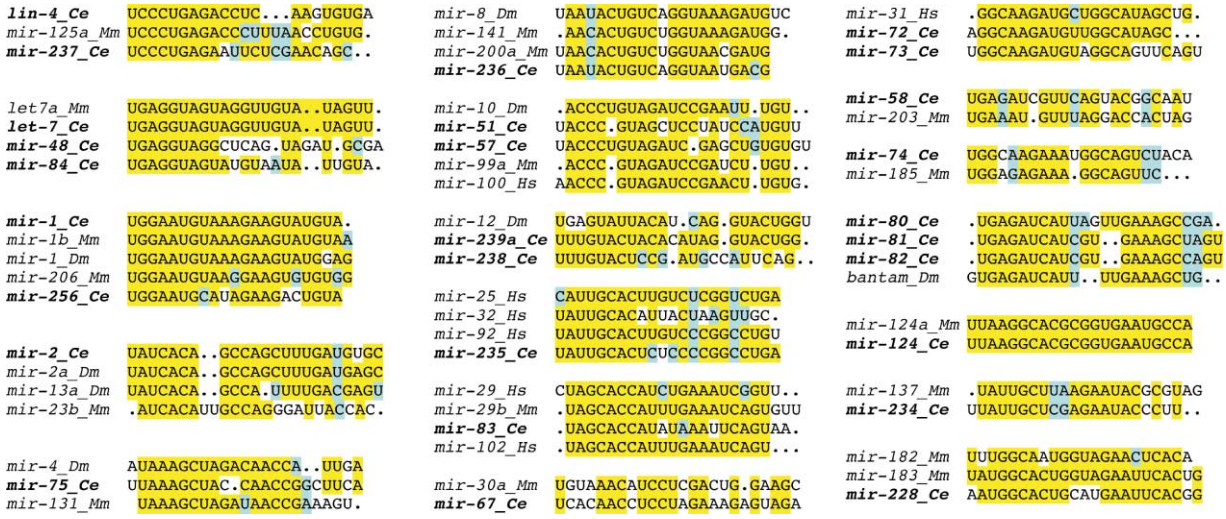



Figure 2. Families of Similar MicroRNAs

The threshold for inclusion was 80% sequence identity. The only families shown are those that include members from *C. elegans* [8, 9], insects [10, 45], and/or vertebrates [10–12]. Identical nucleotides are shaded yellow; conservation of purines or pyrimidines is indicated by blue shading. Sequences were aligned with ClustalW [48] and adjusted by hand.

may exhibit excessive numbers of bulged nucleotides in the stem (Figure 1C) or have fewer than 16 base pairs involving the small RNA (Figure 1D). We tested 16 of the tncRNAs for whether their accumulation in vivo depends on Dicer activity, and in 13 cases, the tncRNA was absent, and/or a larger precursor accumulated, in RNA from Dicer animals (Table 3). This suggests the involvement of some sort of duplex structure in the biogenesis of these tncRNAs—perhaps a single-stranded foldback or a bimolecular duplex. Five of the 33 tncRNA sequences overlap previously identified longer, noncoding EST sequences, as annotated in Wormbase (Table 3). Significantly, in all five cases, the tncRNA sequence is antisense to one or more of the overlapping EST sequences, suggesting that these particular tncRNAs may be siRNA-like molecules; in one case, (*tncR32*), ESTs are reported from both strands, consistent with a duplex precursor for *tncR32*.

## Developmental Regulation of Small-RNA Gene Expression

A characteristic of *C. elegans* miRNAs is that they can exhibit temporal or tissue-specific patterns of gene expression [8, 9, 15, 16, 43]. Although many of the miRNAs and tncRNAs we tested displayed continuous expression during development (Figures 3A and 4D), seven of the 21 new worm miRNAs (Table 2), and 18 of the 33 tncRNAs (Table 3) display distinct temporal changes in abundance. This behavior is consistent with possible regulatory roles in the control of developmental timing, as is the case for the miRNAs *lin-4* and *let-7* [6, 7]. We have not assessed the anatomical patterns of miRNA or tncRNA gene expression, and so our data do not address whether they could be involved in spatial patterning of cell fates.

The observed temporal regulation of miRNAs and tncRNAs can be roughly classified into three categories: (1) primarily embryonic (Figures 3C, 3E, and 3G), (2) primarily postembryonic (Figures 3B, 3D, 3F, and 4B), or (3) complex (embryonic and partially postembryonic; for example, Figures 4A and 4C). Our data do not address whether miRNA or tncRNA accumulation is regulated transcriptionally or posttranscriptionally. However,

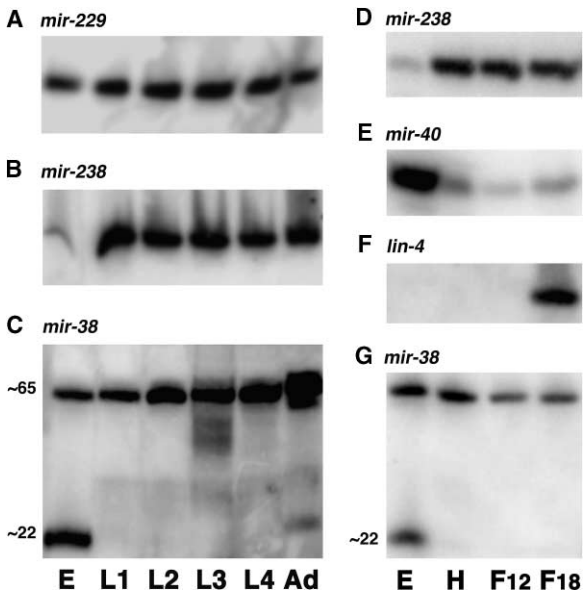



Figure 3. Examples of Developmental Patterns of Expression of Worm MicroRNAs Detected by Northern Blot Hybridization

Panels (A), (B), and (D–F) show the ~22 nt (mature miRNA) size range, and panels (C) and (G) also include the ~65 nt (miRNA precursor) range. In panels (A)–(C), the lanes contain total RNA samples from (left to right): lane 1, embryos; lane 2, L1 larvae; lane 3, L2 larvae; lane 4, L3 larvae; lane 5, L4 larvae; and lane 6, adults. In panels (D)–(G), the lanes contain total RNA samples from (left to right): lane 1, embryos; lane 2, L1 larvae held in developmental arrest for 24 hours by starvation; lane 3, L1 larvae 12 hours after feeding; lane 4, L1 larvae 18 hours after feeding.

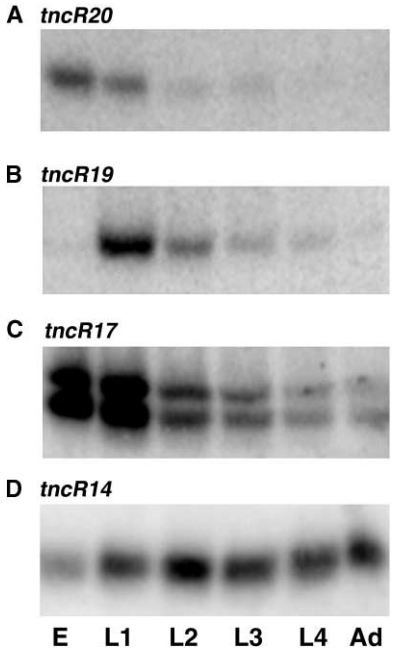



Figure 4. Examples of Developmental Patterns of Expression of Worm tncRNAs Detected by Northern Blot Hybridization

The lanes contain total RNA samples from (left to right): lane 1, embryos; lane 2, L1 larvae; lane 3, L2 larvae; lane 4, L3 larvae; lane 5, L4 larvae; and lane 6, adults. We did not test whether the two species in panel (C) represent isoforms of *tncR17* or whether one of them is a cross-hybridizing RNA from another locus.

in the case of *mir-38*, the ~22 nt mature miRNA is detected only in the embryo, whereas its precursor is more uniformly expressed (Figures 3C and 3G), suggesting that the processing of a *mir-38* precursor may be temporally regulated.

Some miRNAs and tncRNAs exhibit postembryonic downregulation after robust embryonic expression, whereas others appear to be sharply upregulated postembryonically. The embryo-to-larva transition in *C. elegans* involves at least three steps: (1) a developmental arrest at the end of embryogenesis; (2) hatching from the eggshell; and (3) a food-dependent coordinate initiation of postembryonic cell cycles and developmental programs. In the absence of food, hatched larvae remain in a developmentally arrested state, with cell cycles held in G0 and developmental pathways suspended. To characterize the developmental signals influencing embryo-to-larva changes in small-RNA gene expression, we tested whether certain miRNAs or tncRNAs respond to a signal associated with hatching or with feeding. Accordingly, we performed Northern blot analysis of RNA samples from embryos, larvae hatched and held for 24 hours in the absence of food, L1 larvae fed on E. coli for 12 hours, and larvae fed for 18 hours (Figures 3D–3G; Tables 2 and 3). Two distinct kinds of response were observed: (1) temporal regulation by hatching only (for example, Figures 3D, 3E, and 3G) or (2) temporal regulation that requires postembryonic feeding (Figure 3F and reference [43]). These results suggest that miRNA or tncRNA gene expression can be regulated by at least two physiological signals in L1 larvae: a signal associated with developmental and cell cycle arrest or a signal triggered by feeding and/or growth. These differences in response to physiological signals could reflect functional distinctions; for example, regulation of the miRNA level by hatching could signify roles for the miRNA in cell cycle arrest, whereas feeding-dependent expression may reflect roles in the progression of larval development.

### No Precisely Matched Antisense Targets for *C. elegans* MicroRNAs or tncRNAs

Computational prediction of precisely complementary RNA targets for short antisense RNAs is essentially straightforward. Potential targets for plant miRNAs have been identified that contain full-length complementarity such that the miRNA is predicted to form a complete ~22 nt helix with the target messenger RNA [29]. A precise match would be expected to trigger target degradation through RNA interference (RNAi), and indeed, certain plant miRNAs have been shown to function as natural, endogenous siRNAs [28, 30]. In contrast to the plant situation, none of the 79 worm miRNAs or 33 tncRNAs that we analyzed (see Experimental Procedures) show precise Watson-Crick complementarity to any predicted or confirmed worm mRNA sequence. Allowing for the occurrence of non-Watson-Crick G::U base pairs, we find that only three of 79 known *C. elegans* miRNAs and only three of 33 tncRNAs are predicted to form full-length duplexes with worm mRNAs (see Supplemental Data). Among these potential ~22 nt duplexes, only a handful are relatively free of centrally located G::U base pairs. Thus, although it is conceivable that endogenous *C. elegans* miRNAs and tncRNAs have the *potential* to enter the RNAi pathway (were they to be artificially presented with a target containing a precise sequence match), it seems that most of them do not normally engage in RNAi of other RNAs, but probably bind to their targets via incomplete base-pairing, as do *lin-4* and *let-7* [6, 7, 22, 23]. Based on an incomplete base-pairing model for target recognition exemplified by *lin-4* and *let-7*, computational prediction of targets for a given miRNA can easily yield more candidate target genes than can be tested (our unpublished data). To distinguish bona fide targets from irrelevant sequence matches will require additional functional data about the small RNA and its candidate targets.

### Endogenous Small Antisense RNAs Corresponding to Diverse *C. elegans* Protein-Coding Genes

It is striking that among the short cDNA sequences that correspond to protein-coding sequences, only 49 were from the sense strand of an mRNA, whereas 746 distinct cDNA sequences are antisense to the coding strand of mRNAs (Table 1; see also Supplemental Data). The siRNAs that accumulate in *C. elegans* in response to exogenous dsRNA trigger are also primarily antisense to the targeted mRNA [44], suggesting that the antisense sequences we identified represent endogenous siRNAs. A similar result was obtained for the plant *Arabidopsis thaliana*, where a significant fraction of ~22 nt cDNAs that were analyzed corresponded to protein coding se-

Table 4. Classes of Protein-Coding Genes with the Greatest Number of Distinct ~22 nt Antisense cDNA Sequences

| Gene Class[a] | cDNAs Matching Genes in the Class[b] | Canonical Gene[c] | Canonical cDNA Sequences[d] | |
|---|---|---|---|---|
| Retrotransposon | 32 | K02E2.6 | Ct0598 | GAAATTGGCAAGTAACTGAT |
| F box | 28 | Y56A3A.14 | B4C1C | GGAAATCAAAATGTCGTTTAGC |
| Zn finger | 21 | F54F2.2b | E7247 | GTCTCATTTGAACCAAGACC |
| Kinase | 20 | F39F10.2 | E4AC5 | GATGGCATGAATATCTTGCA |
| Transcription factors | 19 | ZK856.13 | A91D2 | TACCCAGTCATCTGGTGG |
| DNA metabolism | 15 | Y57A10A.15 | B567A | GCAATTCCTTGTATTTTCG |
| Transposase | 14 | F16D3.5 | BA5A1 | GATTGTCATCGTGGCAAAGTA |
| RING | 13 | B0564.7 | B55ED | GTTTTTTGTAGAGATACAG |

[a] A total of 746 distinct cDNA sequences were complementary to exons in 551 distinct protein-coding genes. Wormbase WS94 annotations.
[b] Only distinct, nonoverlapping antisense matches to exon sequences are counted.
[c] Chosen arbitrarily from among the genes hit by cDNAs.
[d] Chosen arbitrarily from among the cDNAs that hit the indicated canonical gene.

quences [13]. In the *Arabidopsis* case, cDNA sequences were obtained for either the sense or antisense orientation relative to the open reading frame, which is consistent with the duplex nature of plant siRNAs [18].

The 746 distinct antisense cDNA sequences that we obtained identified 551 different *C. elegans* genes and are distributed uniformly across all six chromosomes (our unpublished data). This suggests that RNAi phenomena are widely deployed among genomic loci in normal worms. Transposon sequences, including transposases and retrotransposon sequences, were among the classes of genes frequently represented by antisense cDNA sequences (Table 4), consistent with previous results implicating RNAi in the silencing of transposons in the *C. elegans* germ line [33, 34].

Some *C. elegans* genes were represented by as many as nine distinct antisense cDNA sequences, suggesting that the corresponding mRNAs may be relatively more abundant and/or more avidly targeted by endogenous RNAi (Table 5). The antisense cDNA sequences from these genes also tended to include sequences that were identified in multiple clones, consistent with the idea that these represent particularly abundant siRNAs. Indeed, some of the endogenous siRNAs were detected as discrete ~22 nt species by northern blot hybridization (our unpublished data), indicating that they accumulate in vivo.

A peculiar locus on Chromosome X ("X cluster") extends about 2000 bp upstream of F47E1.1 and contains 41 distinct (although in some cases partially overlapping) cDNA sequences (Table 1), all oriented in the same direction on the chromosome (see Supplemental Data). Although it is possible that this dense cluster of short cDNA sequences may represent antisense siRNAs involved in gene silencing in the F47E1.1 region, there are no previously reported cDNA sequences that overlap this cluster, nor is either strand of the DNA predicted to encode a protein. A few of the small RNAs in the X cluster seem to be contained within predicted hairpin structures characteristic of miRNAs, but because of their close association with so many distinctly non-miRNA species, none of the X cluster sequences are counted among the miRNAs. The locus could correspond to an exceptionally dense association of tncRNAs, but because this degree of clustering does not seem to be typical of tncRNAs in general, the X cluster sequences are unclassified for the time being.

It is possible that with further analysis of the siRNAs, tncRNAs, and the X cluster small RNAs, differences and similarities in their origins and functions will come into sharper focus. It will be particularly informative to determine what components of the RNA interference machinery are required for their accumulation. siRNAs and tncRNAs have intriguing structural similarities that could reflect common origins and/or functions; about 85% of the tncRNAs (Table 3) and 85% the siRNAs (Supplemental Materials) begin with a 5′ G; also, the siRNA and tncRNA sequences average about 20 nt in length, signifi-

Table 5. Genes Containing Precise Antisense Matches to Four or More Distinct ~22 nt cDNA Sequences

| Gene[a] | Gene Class[b] | Matching cDNAs[c] | Canonical cDNA Sequence[c] | |
|---|---|---|---|---|
| T01A4.3 | MADS Box | 9 | EDD84 | GAAGCAATGGTAGAGTTTTCCC |
| T12G3.1 | Zn finger, ZZ type | 6 | E6840 | GAAGATCAGCAACCATCGTCC |
| B0047.1 | MATH domain | 5 | C5168 | ATGTCAGAGATTGTGACTTTCG |
| B0250.8 | | 5 | C4EE4 | GACGAAACTTTGATACAAGG |
| C04F12.9 | RNAse H | 5 | EDC06 | GAACGTGAACTCCATAGCCTCC |
| E01G4.5 | serpin | 5 | A59E4 | GAGGTGGTTATCATAACACA |
| T23B12.10 | | 5 | AA627 | GAGTGGAGTTGGCTGGCTGTTGCT |
| Y105E8A.20 | tRNA synthetase | 5 | EE33F | ATCATCCAATTCTTCAGTT |
| C46C2.3 | | 4 | Ct1158 | GAATAACAAAAACAACAAT |

[a] A total of 746 distinct cDNA sequences were complementary to exons in 551 distinct protein-coding genes.
[b] Wormbase WS94 annotations.
[c] Only distinct, nonoverlapping antisense matches to exon sequences are counted; the canonical cDNA shown is an example of one of these distinct sequences.

cantly less than the 22 nt average length of miRNAs (Table 1).

## Conclusions

### C. elegans *MicroRNA Diversity and Function*

The results described here and elsewhere [8–10, 15] bring the number of known *C. elegans* miRNA genes to 96. Estimates place the total number of worm miRNA genes at about 100–120 [15] and the number of vertebrate miRNA genes at about 200–250 [39], corresponding to similar fractions of the total genes for each of these animals (about 0.2%–0.5%). This suggests that miRNAs may carry out a diverse spectrum of gene regulatory activities in all animals. About 30% of the known *C. elegans* miRNAs are close in sequence to one or more insect and/or vertebrate miRNA, further suggesting that a significant fraction of miRNAs could play evolutionarily conserved developmental or physiological roles in diverse species.

Certain miRNAs have been shown to play significant roles in the control of gene expression during development in plants [28, 30] and animals [6, 7, 23, 45] (reviewed in [2]). Until recently, the developmental timing regulators *lin-4* and *let-7* of *C. elegans* had been the only miRNA genes for which genetic mutations, and hence functional data, was available in any organism [6, 7]. Conservation of the temporal profile of *let-7* in diverse animals suggested conservation of its timing role [16]. However, in principle, miRNAs could function in processes other than developmental timing, a supposition that is supported by the observation that many of the worm miRNAs are apparently not temporally regulated during worm development (Table 2; see also [15]). Genetic screens for mutants defective in *Drosophila* larval growth identified the *bantam* locus, which encodes a miRNA that functions to regulate apoptosis and cell proliferation in the developing fly [45]. *bantam* is therefore the first miRNA to be shown by genetic criteria to function in a process other than developmental timing and in an animal other than *C. elegans*. *bantam* is related to *mir-80-82* miRNAs of *C. elegans* (Figure 2), raising the possibility that the *mir-80* family of miRNAs might control developmental cell death and/or cell proliferation in the worm.

### *Other Tiny Noncoding RNAs and Widespread Gene Silencing*

In addition to new miRNAs, we also identified, in the ~22 nt size class, two classes of small expressed RNAs that appear not to be miRNAs. These are the tncRNAs, which come from noncoding regions of the worm genome, and apparent siRNAs, which correspond to the antisense strand of diverse protein coding genes. tncRNAs and abundant endogenous siRNAs have not been reported previously. If these were unadulterated Dicer products, one would have expected to identify them by cDNA cloning approaches designed to select RNAs with a 5′ phosphate [9, 10]. However, if tncRNAs and endogenous siRNAs have capped or modified 5′ ends, they would be more efficiently recovered by the procedure employed here, which is relatively insensitive to 5′ end structure. We did not report tncRNAs or siRNAs previously [8] because their significance was not apparent to us until we performed the larger-scale survey described here.

The tncRNAs could represent a heterogeneous class of small RNAs with diverse origins. Some tncRNAs might not be processed from a larger transcript but could be very short primary transcripts. However, our finding that many of the tncRNAs are sensitive to Dicer activity (Table 3), and that at least some of them overlap longer ESTs, supports the idea that some tncRNAs are processed from longer single-stranded or double-stranded RNAs. Some tncRNAs might be generated in the course of RNAi involving noncoding transcripts. Others may be processed from the long-range secondary structure of a single-stranded precursor.

The tncRNAs with predicted foldback precursors that are too poor in structure to qualify as miRNAs could represent a subclass of tncRNAs that are related to miRNAs. By the same token, some of the miRNAs whose predicted precursors are satisfactory in structure yet contain unusual features compared to the typical miRNAs (for example, *mir-229*, *mir-256*, and *mir-262*; see Supplemental Data) might be related to some of the tncRNAs. As we learn more about the origins and functions of miRNAs and tncRNAs, some of them may require reclassification.

Although the lack of phylogenetic conservation of tncRNAs may cast some doubt on the potential importance of these RNAs as regulatory molecules, many of the tncRNAs exhibit potentially interesting temporal patterns of expression (Figure 4), suggesting that they could function in developmental pathways. Like miRNAs, tncRNAs are not precisely complementary to the RNA products of other genes, so they are not likely to function as siRNAs. Some of them could act similarly to miRNAs, as antisense regulators of the expression of genes with imprecise complementarity. Resolution of these possibilities awaits mutational analysis of tncRNA genomic loci.

Our finding that total RNA from normal *C. elegans* contains endogenous siRNAs from more than 500 different genes suggests that RNA-mediated gene silencing could be a global property of the worm genome and could thus affect a variety of protein coding genes. These siRNAs could arise from a spectrum of possible RNA-silencing mechanisms, including pathways that target RNA directly or that target chromatin. Further work is required to determine how this silencing may be significant to the activity and function of particular genes and to normal worm physiology and development.

## Experimental Procedures

### RNA Preparation, Northern Blots, and cDNA Library Construction

*C. elegans* strain N2 was cultured as described [46]. Extraction of total RNA from worms, construction of cDNA libraries, and Northern blot analysis were conducted as described previously [8]. Each cDNA clone contained a single, directional insert of a ~22 nt cDNA produced by reverse transcription of size-selected (~22 nt) worm RNA.

### Sequence Databases and Genomic Annotation of cDNA Sequences

Annotated *C. elegans* and *C. briggsae* sequence data sets were obtained from Wormbase (http://wormbase.org). All cDNA se-

quences were used as queries in Blast searches [37] of the complete *C. elegans* genomic sequence and of the Wormpep database of *C. elegans* protein coding sequences. Sequence reads that did not match the *C. elegans* genome precisely were discarded. Wormbase version WS94 sequence annotations (GFF files) were used to identify cDNA sequences that overlapped known noncoding RNA genes, introns or protein-coding exons. Sequences annotated as protein-coding in genomic DNA were confirmed by a match to Wormpep.

### RNA Structure Prediction and Comparative Genomics

Potential miRNA precursors were identified with mfold [38]. The methods for scoring predicted hairpin structures and for identifying similar predicted miRNAs in *C. elegans* and *C. briggsae* via computational approaches are described in the Supplemental Data.

### Prediction of Potential miRNA or tncRNA Targets

Patscan [41] was used to search a database of predicted and confirmed *C. elegans* protein coding sequences (Wormpep) and 3′ UTR sequences (S. Stricklin and S. Eddy, personal communication) for precise complementarity with each of 79 confirmed *C. elegans* miRNAs and 33 tncRNAs. G::U base pairs were permitted.

### Supplemental Data

Tables S1–S3, Figures S1 and S2, and Supplemental Experimental Procedures are available with this article online at http://images.cellpress.com/supmat/supmatin.htm. Table S4 is available at http://chronic.dartmouth.edu/VRA/Ambros_etal_2003_sup/Ambros_etal_SuplTabl_4.html.

### Acknowledgments

We thank Wormbase and the *C. elegans* Sequencing Consortium at the Sanger Center and Washington University for sequence data and database resources. We also thank Lee Lim, Nelson Lau, and David Bartel for sharing data prior to publication and Shawn Strickland and Sean Eddy for the database of predicted *C. elegans* 3′ UTR sequences. miRNA sequences are deposited at the Rfam miRNA registry (http://www.sanger.ac.uk/Software/Rfam/mirna/). This work was supported by National Institutes of Health grant GM34028 to V.R.A.

Received: February 17, 2003
Revised: March 31, 2003
Accepted: April 1, 2003
Published online: April 22, 2003

### References

1. Ambros, V. (2001). microRNAs: tiny regulators with great potential. Cell *107*, 823–826.
2. Pasquinelli, A.E., and Ruvkun, G. (2002). Control of developmental timing by microRNAs and their targets. Annu. Rev. Cell Dev. Biol. *18*, 495–513.
3. Grishok, A., Pasquinelli, A.E., Conte, D., Li, N., Parrish, S., Ha, I., Baillie, D.L., Fire, A., Ruvkun, G., and Mello, C.C. (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.
4. Hutvagner, G., McLachlan, J., Pasquinelli, A.E., Balint, E., Tuschl, T., and Zamore, P.D. (2001). A cellular function for the RNA-interference enzyme Dicer in the maturation of the *let-7* small temporal RNA. Science *293*, 834–838.
5. Ketting, R.F., Fischer, S.E., Bernstein, E., Sijen, T., Hannon, G.J., and Plasterk, R.H. (2001). Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in *C. elegans*. Genes Dev. *15*, 2654–2659.
6. Lee, R.C., Feinbaum, R.L., and Ambros, V. (1993). The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. Cell *75*, 843–854.
7. Reinhart, B.J., Slack, F.J., Basson, M., Pasquinelli, A.E., Bettinger, J.C., Rougvie, A.E., Horvitz, H.R., and Ruvkun, G. (2000). The 21-nucleotide *let-7* RNA regulates developmental timing in *Caenorhabditis elegans*. Nature *403*, 901–906.
8. Lee, R.C., and Ambros, V. (2001). An extensive class of small RNAs in *Caenorhabditis elegans*. Science *294*, 862–864.
9. Lau, N.C., Lim, L.P., Weinstein, E.G., and Bartel, D.P. (2001). An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. Science *294*, 858–862.
10. Lagos-Quintana, M., Rauhut, R., Lendeckel, W., and Tuschl, T. (2001). Identification of novel genes coding for small expressed RNAs. Science *294*, 853–858.
11. Lagos-Quintana, M., Rauhut, R., Yalcin, A., Meyer, J., Lendeckel, W., and Tuschl, T. (2002). Identification of tissue-specific microRNAs from mouse. Curr. Biol. *12*, 735–739.
12. Lagos-Quintana, M., Rauhut, R., Meyer, J., Borkhardt, A., and Tuschl, T. (2003). New microRNAs from mouse and human. RNA *9*, 175–179.
13. Llave, C., Kasschau, K.D., Rector, M.A., and Carrington, J.C. (2002a). Endogenous and silencing-associated small RNAs in plants. Plant Cell *14*, 1605–1619.
14. Reinhart, B.J., Weinstein, E.G., Rhoades, M.W., Bartel, B., and Bartel, D.P. (2002). MicroRNAs in plants. Genes Dev. *16*, 1616–1626.
15. Lim, L.P., Lau, N.C., Weinstein, E.G., Abdelhakim, A., Yekta, S., Rhoades, M.W., Burge, C.B., and Bartel, D.B. (2003). The microRNAs of *Caenorhabditis elegans*. Genes Dev. *17*, 991–1008.
16. Pasquinelli, A.E., Reinhart, B.J., Slack, F., Martindale, M.Q., Kuroda, M.I., Maller, B., Hayward, D.C., Ball, E.E., Degnan, B., Muller, P., et al. (2000). Conservation of the sequence and temporal expression of *let-7* heterochronic regulatory RNA. Nature *408*, 86–89.
17. Zamore, P.D. (2002). Ancient pathways programmed by small RNAs. Science *296*, 1265–1269.
18. Hamilton, A.J., and Baulcombe, D.C. (1999). A species of small antisense RNA in posttranscriptional gene silencing in plants. Science *286*, 950–952.
19. Elbashir, S.M., Lendeckel, W., and Tuschl, T. (2001). RNA interference is mediated by 21- and 22-nucleotide RNAs. Genes Dev. *15*, 188–200.
20. Martinez, J., Patkaniowska, A., Urlaub, H., Luhrmann, R., and Tuschl, T. (2002). Single-stranded antisense siRNAs guide target RNA cleavage in RNAi. Cell *110*, 563–574.
21. Hamilton, A., Voinnet, O., Chappell, L., and Baulcombe, D. (2002). Two classes of short interfering RNA in RNA silencing. EMBO J. *21*, 4671–4679.
22. Wightman, B., Ha, I., and Ruvkun, G. (1993). Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. Cell *75*, 855–862.
23. Moss, E.G., Lee, R.C., and Ambros, V. (1997). The cold shock domain protein LIN-28 controls developmental timing in *C. elegans* and is regulated by the *lin-4* RNA. Cell *88*, 637–646.
24. Olsen, P.H., and Ambros, V. (1999). The *lin-4* regulatory RNA controls developmental timing in *C. elegans* by blocking LIN-14 protein synthesis after the initiation of translation. Dev. Biol. *216*, 671–680.
25. Seggerson, K., Tang, L., and Moss, E.G. (2002). Two genetic circuits repress the Caenorhabditis elegans heterochronic gene *lin-28* after translation initiation. Dev. Biol. *243*, 215–225.
26. Hutvagner, G., and Zamore, P.D. (2002). A microRNA in a multiple-turnover RNAi enzyme complex. Science *297*, 2056–2060.
27. Zeng, Y., Wagner, E.J., and Cullen, B.R. (2002). Both natural and designed microRNAs can inhibit the expression of cognate mRNAs when expressed in human cells. Mol. Cell *9*, 1327–1333.
28. Llave, C., Xie, Z., Kasschau, K.D., and Carrington, J.C. (2002b). Cleavage of Scarecrow-like mRNA targets directed by a class of Arabidopsis miRNA. Science *297*, 2053–2056.
29. Rhoades, M.W., Reinhart, B.J., Lim, L.P., Burge, C.B., Bartel, B., and Bartel, D.P. (2002). Prediction of plant microRNA targets. Cell *110*, 513–520.
30. Tang, G., Reinhart, B.J., Bartel, D.P., and Zamore, P.D. (2003). A biochemical framework for RNA silencing in plants. Genes Dev. *17*, 49–63.
31. Reinhart, B.J., and Bartel, D.P. (2002). Small RNAs correspond to centromere heterochromatic repeats. Science *297*, 1831.
32. Volpe, T.A., Kidner, C., Hall, I.M., Teng, G., Grewal, S.I.S., and Martienssen, R.A. (2002). Regulation of heterochromatic silenc-

ing and histone H3 lysine-9 methylation by RNAi. Science *297*, 1833–1837.
33. Ketting, R.F., Haverkamp, T.H., van Luenen, H.G., and Plasterk, R.H. (1999). *mut-7* of *C. elegans*, required for transposon silencing and RNA interference, is a homolog of Werner syndrome helicase and RNaseD. Cell *99*, 133–141.
34. Tabara, H., Sarkissian, M., Kelly, W.G., Fleenor, J., Grishok, A., Timmons, L., Fire, A., and Mello, C.C. (1999). The *rde-1* gene, RNA interference, and transposon silencing in *C. elegans*. Cell *99*, 123–132.
35. Dernburg, A.F., Zalevsky, J., Colaiacovo, M.P., and Villeneuve, A.M. (2000). Transgene-mediated cosuppression in the *C. elegans* germ line. Genes Dev. *14*, 1578–1583.
36. Ambros, V., Bartel, B., Bartel, D.P., Burge, C.B., Carrington, J.C., Chen, X., Dreyfuss, G., Eddy, S.R., Griffiths-Jones, S., Marshall, M., et al. (2003). A uniform system for microRNA annotation. RNA *9*, 277–279.
37. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.
38. Mathews, D.H., Sabina, J., Zuker, M., and Turner, D.H. (1999). Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J. Mol. Biol. *288*, 911–940.
39. Lim, L.P., Glasner, M.E., Yekta, S., Burge, C.B., and Bartel, D.P. (2003). Vertebrate microRNA genes. Science *299*, 1540.
40. Eddy, S.R. (2002). Computational genomics of noncoding RNA genes. Cell *109*, 137–140.
41. Dsouza, M., Larsen, N., and Overbeek, R. (1997). Searching for patterns in genomic data. Trends Genet. *13*, 497–498.
42. Slack, F.J., Basson, M., Liu, Z., Ambros, V., Horvitz, H.R., and Ruvkun, G. (2000). The *lin-41* RBCC gene acts in the *C. elegans* heterochronic pathway between the *let-7* regulatory RNA and the LIN-29 transcription factor. Mol. Cell *5*, 659–669.
43. Feinbaum, R., and Ambros, V. (1999). The timing of *lin-4* RNA accumulation controls the timing of postembryonic developmental events in Caenorhabditis elegans. Dev. Biol. *210*, 87–95.
44. Tijsterman, M., Ketting, R.F., Okihara, K.L., Sijen, T., and Plasterk, R.H. (2002). RNA helicase MUT-14-dependent gene silencing triggered in *C. elegans* by short antisense RNAs. Science *295*, 694–697.
45. Brennecke, J., Hipfner, D.R., Stark, A., Russell, R.B., and Cohen, S.M. (2003). *bantam* encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene *hid* in *Drosophila*. Cell, *113*, 25–36.
46. Wood, W.B., ed. (1988). The nematode *Caenorhabditis elegans*. (Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press).
47. Kent, W.J., and Zahler, A.M. (2000). Conservation, regulation, synteny, and introns in a large-scale *C. briggsae-C. elegans* genomic alignment. Genome Res. *10*, 1115–1125.
48. Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. *22*, 4673–4680.