

Computational and Experimental Identification of *C. elegans* microRNAs

Yonatan Grad,^{1,3} John Aach,^{1,3}
Gabriel D. Hayes,^{2,3} Brenda J. Reinhart,^{2,4}
George M. Church,¹ Gary Ruvkun,^{1,*}
and John Kim^{2,3}

¹The Lipper Center for Computational Genetics and
Department of Genetics
Harvard Medical School
Boston, Massachusetts 02115

²Department of Molecular Biology
Massachusetts General Hospital and
Department of Genetics
Harvard Medical School
Boston, Massachusetts 02114

Summary

MicroRNAs (miRNAs) constitute an extensive class of noncoding RNAs that are thought to regulate the expression of target genes via complementary base-pair interactions. To date, cloning has identified over 200 miRNAs from diverse eukaryotic organisms. Despite their success, such biochemical approaches are skewed toward identifying abundant miRNAs, unlike genome-wide, sequence-based computational predictions. We developed informatic methods to predict miRNAs in the *C. elegans* genome using sequence conservation and structural similarity to known miRNAs and generated 214 candidates. We confirmed the expression of four new miRNAs by Northern blotting and used a more sensitive PCR approach to verify the expression of ten additional candidates. Based on hypotheses underlying our computational methods, we estimate that the *C. elegans* genome may encode between 140 and 300 miRNAs and potentially many more.

Introduction

Diverse eukaryotic organisms harbor a class of noncoding, small RNAs, termed microRNAs (miRNAs), which are thought to function as regulators of gene expression. The role of miRNAs as potential translational regulators of target genes is based on functional studies of the *Caenorhabditis elegans* genes *lin-4* and *let-7*, the first two miRNA genes discovered (Rougvie, 2001; Pasquinelli and Ruvkun, 2002). *lin-4* and *let-7* mutations cause defects in the temporal regulation of larval stage-specific programs of cell divisions, resulting in the abnormal repetition of certain earlier patterns of cell lineage. *lin-4* expression begins at the first larval stage and persists in subsequent stages while *let-7* expression starts at the late third larval stage and continues throughout the adult life cycle (Feinbaum and Ambros, 1999; Lee et al.,

1993; Reinhart et al., 2000). These miRNAs regulate the stage-specific pattern of cell lineage by base-pairing to partially complementary sites in the 3' untranslated region (UTR) of their target mRNAs and repressing their translation (Lee et al., 1993; Wightman et al., 1993; Reinhart et al., 2000; Slack et al., 2000).

Both *lin-4* and *let-7* are processed from ~70 nucleotide (nt) precursors predicted to fold into hairpin secondary structures (Lee et al., 1993; Pasquinelli et al., 2000). The partially double-stranded stems of the hairpins are cleaved by the RNase III-like enzyme Dicer (DCR-1 in *C. elegans*) to release the ~22 nt mature miRNAs (Bernstein et al., 2001; Grishok et al., 2001; Hutvagner et al., 2001; Ketting et al., 2001). Dicer also functions in RNA interference (RNAi) to cleave introduced double-stranded RNAs (dsRNAs) into ~22 nt small interfering RNAs (siRNAs), which ultimately hybridize to homologous mRNA sequences and target them for degradation (Hannon, 2002). Significant differences between the two pathways exist. siRNAs are generated from exogenously introduced dsRNAs and hybridize with perfect complementarity to target mRNAs, marking them for destruction. In contrast, miRNAs such as *let-7* and *lin-4* are expressed endogenously, bind target mRNAs at the 3' UTR through imperfect base-pairing, and, at least in the case of *lin-4* and its target mRNA *lin-14*, regulate target mRNAs at the translational level (Ha et al., 1996; Olsen and Ambros, 1999; Wightman et al., 1993).

Recently, over 200 miRNAs have been identified by cloning from various eukaryotes including *C. elegans*, *Drosophila melanogaster*, *Mus musculus*, and *Homo sapiens* (Dostie et al., 2003; Lagos-Quintana et al., 2001, 2002, 2003; Lau et al., 2001; Lee and Ambros, 2001; Mourelatos et al., 2002; Lim et al., 2003). Studies of miRNAs in various species have hinted at their wide-ranging importance and additional possible mechanisms of miRNA-mediated regulation. Significantly, the sequence of mature ~22 nt *let-7* miRNA and the temporal regulation of its expression, as well as complementary sequences in the 3' UTR of its target, *lin-41*, are conserved among bilaterians, suggesting that *let-7*-mediated temporal control may be functionally conserved as well (Pasquinelli et al., 2000). Similar cross-species sequence conservation (but not complete identity) has also been noted for several miRNAs, indicating that they may be functionally conserved (Grosshans and Slack, 2002). In addition, some murine miRNAs show restricted expression patterns, suggesting tissue-specific functions (Lagos-Quintana et al., 2002, 2003). Recent evidence from *Arabidopsis thaliana* indicates that miRNAs are also expressed in plants (Reinhart et al., 2002). Many of these miRNAs bind with near-perfect complementarity to target mRNA coding regions and could cause target degradation similarly to RNAi (Llave et al., 2002; Rhoades et al., 2002).

Despite the bevy of miRNAs that has emerged from cloning, such screens are likely to be far from saturated, as they are biased to abundant miRNAs. In this study, we developed computational methods that predict miRNAs encoded by the *C. elegans* genome independently of

*Correspondence: ruvkun@molbio.mgh.harvard.edu

³These authors contributed equally to this work.

⁴Present address: Department of Plant Biology, Carnegie Institution of Washington, Stanford, California 94305.

abundance. Such methods take advantage of properties of known miRNAs, including their length (typically 21–24 nt), precursor hairpin structure (typically ~70–90 nt, with multiple 1–4 nt bulges and mismatches), and tendency to be found in intergenic regions. However, the short length and high degree of sequence and structure variation also limit the accuracy of computational prediction based on sequence and structure alone. Therefore, we improved prediction methods by focusing on miRNAs that appear conserved across species or within a species. Our criterion of apparent conservation is “correspondence”—the presence in two or more genomes of short, very similar sequences embedded in the same stems of predicted hairpins with otherwise variable sequence. Through systematic searches for correspondence, coupled with careful tuning of sequence and structure constraints (see the Experimental Procedures), we developed two algorithms that each generated a set of candidate *C. elegans* miRNAs and possible homologs in other species. The algorithms considered correspondences across *C. elegans*, *C. briggsae*, *D. melanogaster*, and *H. sapiens*. A third algorithm was based on apparent homology to known miRNAs from any species. Although a bioinformatic approach to identify miRNAs in *C. elegans* has been reported to be feasible (Lee and Ambros, 2001), and a systematic bioinformatic search has been performed in vertebrates (Lim et al., 2003), we present the first systematic examination of *C. elegans* miRNAs that are conserved in *D. melanogaster* and *H. sapiens* as well as provide computational methods to identify miRNA sequence family members. Based on our computational findings, we predict that the *C. elegans* genome may encode between 140 and 300 miRNAs and discuss how the *C. elegans* genome may encode even more than this upper estimate.

Results and Discussion

Computational Algorithms to Identify Conserved microRNAs in the *C. elegans* Genome

We surveyed repeat-masked intergenic regions of the *C. elegans* genome and generated a large list of sequences that have the potential to form imperfect hairpin structures of similar length to known miRNA hairpins (see the Experimental Procedures and Supplemental Experimental Procedures at <http://www.molecule.org/cgi/content/full/11/5/1253/DC1>). We filtered these to an initial set of 8713 *C. elegans* hairpins on the basis of the score from a hairpin-prediction program we developed (srnaloop; see the Experimental Procedures and Supplemental Experimental Procedures for the program), GC content, predicted structure minimum free energy, and presence of multiloops (see the Experimental Procedures and Supplemental Experimental Procedures). A total of 39 out of 61 (64%) hairpin precursors associated with cloned *C. elegans* miRNAs were present in the intergenic genome sequence we analyzed, and, of these, 29 out of the 39 (74%) passed our filters. We used this collection of 8713 hairpins to generate three sets of predictions of *C. elegans* miRNA hairpins (Figure 1 and Supplemental Table S1 at <http://arep.med.harvard.edu/miRNA/>).

Two of the algorithms were based on interspecies sequence and secondary structure conservation of can-

didate miRNAs. The comparative genomics approaches were motivated by the example of *let-7* in which identical ~21 nt mature miRNA sequences are phylogenetically conserved within the same stems of the ~70 nt hairpin precursor (Pasquinelli et al., 2000). As a first step, we searched the *D. melanogaster* genome for hairpins corresponding to the initial set of *C. elegans* hairpins, finding 2523 distinct *C. elegans* hairpins in correspondence with 3505 distinct *D. melanogaster* hairpins (see Correspondence Determination in the Supplemental Experimental Procedures). We compared these corresponding sets of hairpins to genome sequences of a third species. In the first algorithm, we looked for sequences among the *C. elegans* hairpins with corresponding hairpins in *D. melanogaster* for those that closely matched sequences in *C. briggsae*, a sister nematode species (see the Experimental Procedures and Supplemental Experimental Procedures). Eliminating repetitive sequences resulted in 81 candidate hairpins, termed the *C. elegans*, *D. melanogaster*, and *C. briggsae* (CDC) set (Figure 1). The CDC set includes six conserved *C. elegans* miRNAs previously cloned (*mir-1*, -34, -45, -60, -79, and *let-7*). After testing by Northern blotting (see below), we applied additional structure and repeat filtering to obtain a higher quality set of 28 predictions (CDC-f; see the Supplemental Experimental Procedures). Of these, many may represent *C. elegans* miRNAs that are conserved in arthropods and nematodes but have not been identified by previous biochemical experiments.

In the second cross-species conservation algorithm, we used the set of *D. melanogaster* hairpin sequences derived as described above to search the *H. sapiens* genome for corresponding hairpins and then applied a transitivity filter for the two sets of correspondences, between *C. elegans* and *D. melanogaster*, and between *D. melanogaster* and *H. sapiens* (see the Supplemental Experimental Procedures). Application of additional repetitive sequence and structural filters yielded a set of 40 hairpins with conservation in the stem region across *C. elegans*, *D. melanogaster*, and *H. sapiens* (CDH set; Figure 1). Six of 40 were known, conserved *C. elegans* miRNAs (*mir-1*, -2, -34, -57, -79, and *let-7*), of which four were present in the CDC set, two (*mir-2* and *mir-57*) were unique to the CDH set, eight were candidates also predicted by the CDC set, and the remaining 26 represent additional new miRNA candidates (see Supplemental Table S1 at <http://arep.med.harvard.edu/miRNA/>). The appearance of four known miRNAs in both the CDC and CDH sets suggests that these are the most evolutionarily conserved miRNAs by our methods among the 29 miRNAs analyzed.

In our third algorithm, we searched for possible *C. elegans* homologs of 164 miRNAs cloned from *C. elegans*, *D. melanogaster*, *M. musculus*, and *H. sapiens* (Figure 1; Lee and Ambros, 2001; Lau et al., 2001; Lagos-Quintana et al., 2001). After narrowing the initial set of *C. elegans* hairpins to a smaller set of 6086 through more stringent filters for intergenic sequences, repetitive sequences and exact duplicates, and hairpin duplex topology (see the Supplemental Experimental Procedures), we employed a Smith-Waterman algorithm to compare each hairpin to each of the 164 miRNA sequences. Excluding exact matches to the known miRNAs, we obtained 116 candidate homolog hairpins in *C. ele-*

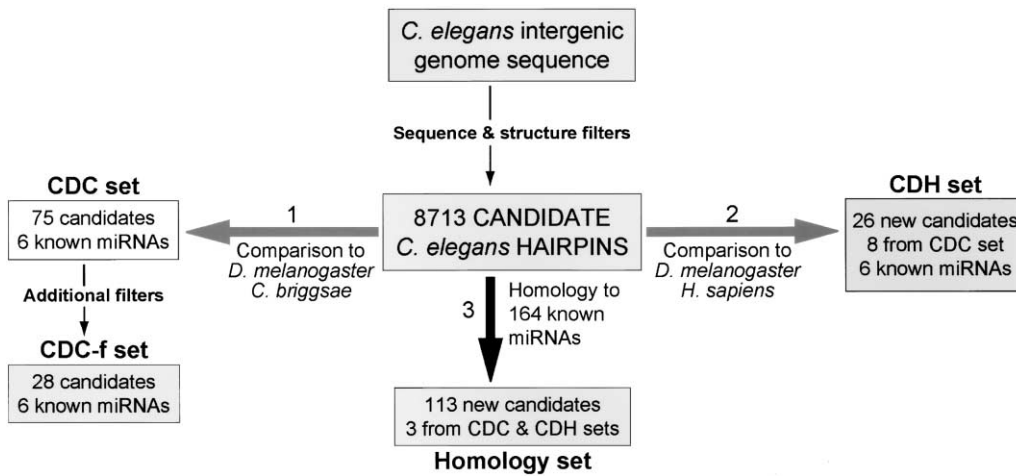


Figure 1. Schematic Representation of the Computational Algorithms Used in This Study
Details are discussed in the text.

gans (homology set), of which three were also identified by the two cross-species algorithms, and the remaining 113 represent potentially new miRNAs. The low overlap among the three sets of predictions indicates either the presence of noise in the predictions or that the true number of miRNAs in the genome is large. We show below that many of the candidates may be real but hard to detect experimentally, leaving open the latter possibility. In sum, the three algorithms employed in this study yielded a total of 214 miRNA candidates, of which 101 were identified by cross-species conservation and 113 by homology to known miRNAs (Figure 1; see Supplemental Table S1 at <http://arep.med.harvard.edu/miRNA/>). Complete details on algorithms are presented in the Experimental Procedures and Supplemental Experimental Procedures.

Experimental Verification of miRNA Candidates

The expression of candidate miRNAs was evaluated by Northern blot analysis of total RNA extracted from mixed developmental stage populations of *C. elegans*. We tested the CDC set of 81 hairpins that included six known *C. elegans* miRNAs. Because the mature form of the miRNA can arise from either the 5' or 3' arm of the stem of the hairpin, we made antisense probes to both sides of each of the candidate hairpins and asked whether either of these probes could detect a predicted ~21–24 nt mature miRNA in the total RNA from wild-type worms and a ~70–90 nt precursor in the total RNA from the *dcr-1* mutant, which is defective for processing of the precursor miRNAs. As some predicted thermodynamic characteristics of hairpins (e.g., folding energy, bulge locations, and sizes) may vary between reverse complementary sequences, in some cases hairpins from only one strand pass our filters while in other cases hairpins from both strands pass. We thus only tested hairpin probes strand-specifically based on the strand sequences that pass the filters. Based on this protocol, we verified two new miRNAs: *mir-236*, which is derived from the 3' arm of its hairpin, and *mir-228*, which is

derived from the 5' arm of its hairpin (Figure 2D). Both accumulated precursors of ~70 nt in the *dcr-1* mutant background, similar to the observation for *let-7*, indicating that they are substrates for Dicer cleavage (Figure 2A). Even though the two new miRNAs are apparently unrelated to each other by sequence, their expression is regulated similarly with a peak in expression at the L1 larval stage (Figure 2B). This is consistent with the observation made by Lee and Ambros (2001) that many *C. elegans* miRNAs have a peak of expression during early *C. elegans* larval development (Lee and Ambros, 2001). The remaining 73 candidate miRNAs from this algorithm could not be detected by conventional Northern blotting (data not shown).

For the CDH set, we tested candidates by Northern blotting using antisense probes to the stem region that is conserved across the three species (*C. elegans*, *D. melanogaster*, *H. sapiens*). Of the 40 miRNAs identified by the CDH algorithm, we conducted Northern blot analysis of 20, which included *let-7* and *mir-34*, as well as the *mir-236* and *mir-228* hairpins predicted by both CDC and CDH cross-species algorithms. While *let-7*, *mir-34*, *mir-236*, and *mir-228* could be detected, the remaining 16 candidates were not detected by Northern blot analysis. Finally, to examine 39 of the 113 candidates derived from the homology algorithm, we used antisense probes to the stem containing sequence similar to a known miRNA. We detected *mir-236*, which was predicted by all three algorithms, but none of the other 38 candidates tested (data not shown).

The homology algorithm predicts potential miRNAs with close sequence homology to the *C. elegans* miRNAs, *mir-236* and *mir-228*, in both *D. melanogaster* and *H. sapiens*. The closest relative of *C. elegans mir-236* in *D. melanogaster* is *mir-8*, which was previously identified by cloning (Lagos-Quintana et al., 2001), while the other candidates in fly and human for *mir-236* and *mir-228* represent previously unreported miRNAs. By Northern blot analyses of human and fly RNA samples, we detected *mir-200b*, a novel human miRNA similar to *C. elegans mir-236*, and *mir-263*, a *D. melanogaster* miRNA

mir-228 in *C. elegans* may have distinct functional roles from *mir-263* in *D. melanogaster*.

Overall, two of the 132 (~1.5%) computationally predicted and tested candidates (excluding known miRNAs) were detected by Northern blot analysis, although many of the candidate miRNAs are indistinguishable from known miRNAs by structure and energy criteria. These findings suggest that many miRNA candidates may be expressed at levels below the threshold of experimental detection by Northern analysis, perhaps due to restricted cell-type expression or induction only in response to specific environmental cues. The Northern blotting protocol for detection of miRNAs is not particularly sensitive: the antisense probes are short, end-labeled oligonucleotides that hybridize to their 21–24 nt target miRNAs at a 1:1 ratio. Furthermore, the miRNA on the Northern blot is not enriched in any way; polyA selections are not applicable to miRNAs, and miRNAs constitute a tiny fraction of a total RNA preparation dominated by ribosomal and other RNAs. A strong correlation exists between the number of times a miRNA appears in miRNA clone libraries and its expression level: miRNAs identified in just one or a few clones are barely detectable by Northern blotting whereas those isolated many times are much more easily detected (Lau et al., 2001). Thus, miRNAs with abundance lower than this threshold would escape detection by cloning or verification by Northern blotting. We conclude that only miRNAs that are expressed at relatively high levels in the organism can be detected by this low-sensitivity method.

To increase the sensitivity of detection, we used the same biochemical procedures involved in cloning miRNAs to construct an amplified small RNA library derived from mixed-stage, wild-type worm RNA. An 18–24 nt size-selected pool of small RNAs was ligated to 5' and 3' RNA linkers. As in the miRNA cloning procedure, the 3' linker oligonucleotide is preadenylated to allow ligation to RNAs with a 3'OH, characteristic of cleavage by an RNaseIII such as Dicer, in the absence of ATP, which would allow circularization or multimerization of the small RNAs in the pool (Lau et al., 2001). Reverse-transcription and PCR using DNA oligonucleotides complementary to those linkers amplified a fraction highly enriched for miRNAs over all other cellular RNAs (see the Experimental Procedures). This PCR-amplified library of small RNAs was then used as the template in an assay that employed a second round of PCR used to detect an individual miRNA. Here, one PCR primer was complementary to the 5' linker sequence and the other primer complementary to one computationally predicted miRNA. While this protocol, like the DNA sequencing of miRNA clones, still depends on biochemical abundance for detection, we reasoned that low-abundance miRNAs could be easily sampled after this second round of amplification; a comparably deep sampling of the library for low-abundance miRNAs by sequencing would require sequencing many thousands of miRNA clones.

Because this PCR detection assay depends on successful hybridization of the primer 3' ends to the 5' end of amplified miRNA transcripts, and because our predictions of miRNA sequences do not precisely predict this 5' end, we selected candidate miRNAs for testing from the CDH and homology algorithm predictions, as these gave more information on the possible location

of a mature miRNA within a predicted miRNA hairpin than did the CDC algorithm. The CDH algorithm provides a refined sequence prediction by virtue of the overlap of two short sequence matches in predicted hairpins across three species (*C. elegans* to *D. melanogaster* and *D. melanogaster* to *H. sapiens*). The homology algorithm starts from known mature miRNA sequences, and hence candidates from this method are most likely to have proper ends. In contrast, the CDC predictions are based on only a single short sequence match (*C. elegans* to *D. melanogaster*) with homology to *C. briggsae* judged by whole-hairpin alignments.

We used our PCR assay to test 15 miRNA candidates from the CDH algorithm: *mir-236* and *mir-228*, which have expression detectable by Northern analysis (Figure 2), as well as 13 others which were not detected by conventional Northern blotting. We also examined 42 miRNA candidates from the homology algorithm, including *mir-236* and 41 other candidates, which were undetectable by Northern blotting. Except for including previously verified *mir-236* and *mir-228*, we applied no sequence, structure, or experimental criteria in picking these 56 candidates from the 134 unique candidates from the CDH and homology algorithms. As positive controls, we tested probes against a subset of previously reported, conserved miRNAs: *let-7*, *mir-1*, *mir-2*, *mir-34*, and *mir-47*. As negative controls, we designed 22 nt probes (starting at position +10 from the start ATG sequence) complementary to 11 of the 20 most abundant mRNA transcripts in mixed-stage worms as determined by serial analysis of gene expression (Jones et al., 2001). For three of the mRNA transcripts, a probe was designed for the middle of the transcript as well as the 5' region of the transcript (Table 1). We reasoned that if the small RNA library contained a significant fraction of contamination resulting from mRNA degradation products, then the PCR assay would be sensitive enough to amplify these mRNA degradation products.

Of the 54 miRNA candidates that were previously negative by Northern blotting, 1 of the 13 CDH candidates and 9 of the 41 homology candidates were positive by the PCR assay, thus constituting a ~20% (10/54) verification of expression by the PCR assay for the candidate miRNAs that could not be detected by conventional Northern blotting analysis (Figure 3A). In addition, all of the known miRNAs assayed were detected by the PCR assay, including the new miRNAs, *mir-236* and *mir-228*. None of the mRNA probes amplified a PCR product. The results of the PCR assay are summarized in Table 1. We detected no sequence or structure characteristics that might distinguish the candidates that were positive by the PCR assay from those that did not result in amplification of a product.

To assess the probability that the primers for the candidate miRNAs generated PCR products by adventitiously priming against rRNAs, tRNAs, and previously cloned miRNAs, we performed computational comparisons of all candidate predictions and PCR primers against a database of 888 noncoding *C. elegans* RNA sequences (<http://www.wormbase.org>, release WS95, date February 18, 2003). This comparison established that positive results due to adventitious priming were very unlikely by showing that short sequence matches of primer 3' ends against miRNA sequences expected

Table 1. PCR Assay Results for Candidate miRNAs, Known miRNAs, and Abundant *C. elegans* mRNAs

Annotation	Description	miRNA Sequence	Probe Sequence	PCR
R09B5.3 (f)	<i>C. elegans</i> mRNA	na	GCGACCAAAAGAATTAGGATGG	-
R09B5.3 (m)	<i>C. elegans</i> mRNA	na	ACCTCCCATCATTCTGGGTAT	-
R09B5.9 (f)	<i>C. elegans</i> mRNA	na	GCGACAAGAAGGACGAGAATGT	-
R09B5.9 (m)	<i>C. elegans</i> mRNA	na	TCCACCACCATACCCGCCATAT	-
Y2H9A.3 (f)	<i>C. elegans</i> mRNA	na	CTAGCGCCGTATACAAGAGTCT	-
Y2H9A.3 (m)	<i>C. elegans</i> mRNA	na	ACCATTCCATGCCAGGCACCAC	-
F53F1.4 (f)	<i>C. elegans</i> mRNA	na	ATAGCAGCGAAGAAAATGACAA	-
F57H12.3 (f)	<i>C. elegans</i> mRNA	na	GCAGCTGAGATTTTAAAGATAT	-
C45B2.1 (f)	<i>C. elegans</i> mRNA	na	GCAAACGCCAGGAAAACGGCAG	-
T23G11.3 (f)	<i>C. elegans</i> mRNA	na	ACACCGTAAGTTGGAGTGGTC	-
F28D1.5 (f)	<i>C. elegans</i> mRNA	na	AGGAGAGCGAGAGTGAGCTTGA	-
T23G11.2 (f)	<i>C. elegans</i> mRNA	na	AGGATTTTCATCGAAAATTGAAA	-
W05F2.3 (f)	<i>C. elegans</i> mRNA	na	ATCGTGAAACGGCGAGTGTCA	-
C33G8.2 (f)	<i>C. elegans</i> mRNA	na	ATGGCAAGATAAGTATAGTA	-
<i>let-7</i>	known microRNA	UGAGGUAGUAGGUUUGUAUAGU	ACTATACAACCTACTACCTCA	+
<i>mir-1</i>	known microRNA	UGGAAUGUAAAGAAGUAUGUA	TACATACTTCTTTACATTCCA	+
<i>mir-2</i>	known microRNA	UAUCACAGCCAGCUUUGAUGUGC	GCACATCAAAGCTGGCTGTGATA	+
<i>mir-34</i>	known microRNA	AGGCAGUGUGUUAGCUGGU	ACCAGTAACCACACTGCCT	+
<i>mir-47</i>	known microRNA	UGUCAUGGAGCGCUCUCUUA	TGAAGAGAGCGCCTCCATGACA	+
<i>mir-56-2</i>	known microRNA	UACCCGUAAUGUUUCCGCGAG	CTCAGCGGAAACATTACGGGTA	+
<i>mir-79</i>	known microRNA	AUAAAGCUAGGUUACCAAAGCU	AGCTTTGGTAACCTAGCTTTAT	+
<i>mir-236</i>	CDH candidate	AUCUAAUCUGUCAGGUAUUGA	TCATTACCTGACAGTATTAGAT	-
<i>mir-228</i>	CDH candidate	AAUGGCACUGCAUAAUUCACGG	CCGTGAATTATGCAGTGCCATT	+
<i>cp-mir-264</i>	CDH candidate	GGCGGGUGGUUUGUUUAUG	CATAACAACAACCACCCGCC	+
candidate-31	CDH candidate	UUCUGGAGCAGGAACUGCAGCUG	CAGCTGCAGTCTCTGCTCCAGAA	-
candidate-42	CDH candidate	AGUGGCAGUGGACAUUUGACGG	CCGTCAAATGTCCACTGCCACT	-
candidate-52	CDH candidate	AAAGUUGCUAAAGUUGGUGGA	TCCACCAACTTAGCAACTTT	-
candidate-72	CDH candidate	CUCGUCUACCCUGUAGAUCGA	TCGATCTACAGGGTAGACGAG	-
candidate-77	CDH candidate	AAGUUGUGAGCCCGCACUGCGACG	CGTCGCAGTGGCGGCTCACAACT	-
candidate-100	CDH candidate	ACGAUUUGGCAUUGGUAUGUGG	CCACATCCAATGCCAAATCGT	-
candidate-108	CDH candidate	UAUGCGAGUGUGUGGGCUC	GGAGCCCCACACTCGCATA	-
candidate-127	CDH candidate	CGUUUUCUUCUGUCGUUCCCC	GGGGAACGACAGAAAGAAACG	-
candidate-135	CDH candidate	CAUUAACUGACAGUAUUAGAU	ATCTAATACTGTCAGGTAATG	-
candidate-169	CDH candidate	GCUUCUGCGGUGCGUGCGUGGG	CCCACGCACGCACCGCGAAGG	-
candidate-169	CDH candidate	UUCUAUUUGUCUUGUGCGCA	TGCGCACAAGCACAAATAGAA	-
candidate-187	CDH candidate	AGUUGAGCCUGCUGGUGGGUUU	AAACCACCAGCAGGCTCAACT	-
<i>mir-236</i>	homology candidate	AAUACUGUCAGGUAUUGACGCUGG	CCAGCGTCATTACCTGACAGTATT	+
<i>cp-mir-265</i>	homology candidate	UGAGGGAGGAAGGGUGGUU	ATACCACCCTTCTCCCTCA	+
<i>cp-mir-266</i>	homology candidate	AGGCAAGACUUUGCAAAGC	GCTTTGCCAAAGTCTTGCT	+
<i>cp-mir-267</i>	homology candidate	CCCGUAGAGUGUCUGCUGCA	TGCAGCAGACACTTACGGG	+
<i>cp-mir-268</i>	homology candidate	GGCAAGAAUJAGAAGCAGUUUGGU	ACCAAAGTCTTAATTTTGCC	+
<i>cp-mir-269</i>	homology candidate	GGCAAGACUCUGGCAAAACU	AGTTTTGCCAGAGTCTTGCC	+
<i>cp-mir-270</i>	homology candidate	GGCAUGAUGGAGCAGUGGAG	CTCCACTGTACATCATGCC	+
<i>cp-mir-271</i>	homology candidate	UCGCCGGUGGGAAGCAUU	AATGCTTTCCACCCGGCGA	+
<i>cp-mir-272</i>	homology candidate	UGUAGGCAUGGGUGUUUG	CAAACCCCATGCCTACA	+
<i>cp-mir-273</i>	homology candidate	UGCCCGUACUGUGCGGCGUG	CAGCCGACACAGTACGGGCA	+
candidate-11	homology candidate	AGCCGCACAGCACUGGUUGACA	TGTCAACCAGTGTGTGCGGCT	-
candidate-15	homology candidate	CGGAUCGUUAAAACAGGAAGU	CATTTTCCGTTTTAACGATCCG	-
candidate-33	homology candidate	AGAAUUAAAAUUCUAGACC	GGTCTAGAATTTTTAATTCT	-
candidate-48	homology candidate	CUAGGCCACCAACUUAAACGGUU	AACCGTTTAAAGTTGGTGCCTAG	-
candidate-87	homology candidate	GAGUACGGUAGAUCUGGUACUG	CAGTACCAGATCTACCGTACTC	-
candidate-93	homology candidate	UUACAGCCGUACCUACCGCUU	AAGCAGGTAGGTACGGCTGTA	-
candidate-96	homology candidate	UGGCAGGCACGUAGGUUUUGG	CCAATACCTACGTGCCTGCCA	-
candidate-103	homology candidate	UUUCAGUUGGAGAUUGUGCAUC	GATGCACACATCTCCAAGTAAA	-
candidate-107	homology candidate	UCUUUCACGUCGGGCCACUUG	CAAGTGGCCCGACGTGAAAGA	-
candidate-111	homology candidate	GCAAGUCUUUGGCAAAACU	AGTTTTGCCAAAGACTTGC	-
candidate-112	homology candidate	UGAGGCUAAGAAUUGUGUAGUU	AACTACACAATTTCTAGCCTCA	-
candidate-117	homology candidate	UAGCAACCAUUUGAAGUUGUU	AACAACCTCAAATGGTTGCTA	-
candidate-120	homology candidate	CAGUCUACCACAUUGGUCGU	ACGACCATGTGGTAGACTG	-
candidate-130	homology candidate	AUUUGGAAUUUUCUAGAUA	TGATCTAGAAAATTTCAAAT	-
candidate-133	homology candidate	GAUCUGAUCCUUCAGAGCUU	AAGCTCTGAAGGATCAGATC	-
candidate-133	homology candidate	UGUGACUGGUGAGCAAGCGA	TCGTTGTCTACCAGTCAACA	-
candidate-134	homology candidate	UGAGGCUAGAAAAUUGUGUAGUU	AACTACACAATTTCTAGCCTCA	-
candidate-136	homology candidate	AUUUUGAAUACUGUUGCUACGGG	CCCGTAGCAACAGTATCAATAAT	-
candidate-141	homology candidate	UGAGGCUCAUAGAUUUUGUAGUU	AACTACAAAATCTATGAGCCTCA	-
candidate-142	homology candidate	GAGAUUGUAGUUUGUAGUGUA	TACTACTACAACTACAATCTC	-

(continued)

Table 1. Continued

Annotation	Description	miRNA Sequence	Probe Sequence	PCR
candidate-152	homology candidate	UAACCGAUAGGUUUCUGCCGAG	CTCGGCAGAAACCTATCGGTTA	-
candidate-162	homology candidate	UAAGGUGCAUUUAAGGCCGAUA	TATCGGCCTTAAATGCACCTTA	-
candidate-167	homology candidate	UGGACUCCUGUUGUUUGCC	GGCAAACACGAGGAGTCCA	-
candidate-178	homology candidate	UAGGUGGAGUCUGAUUUUCCACAGU	ACTGTGAAAATCAGACTCCACCTA	-
candidate-179	homology candidate	AGGAGCACGAAUGGUUCGUG	CACGAACCATTCGTGCTCCT	-
candidate-180	homology candidate	AUAGGCUAGAUAGGUUGCCUAG	CTAGGCAACCTATCTAGCCTAT	-
candidate-213	homology candidate	AAUAGUGUCUGAAAGUUGUC	GACAACCTTCAGACACTATT	-
candidate-214	homology candidate	UGGGCAAACUUUGCAAACU	AGTTTTGCCAAAGTTTTGCCCA	-
candidate-215	homology candidate	GUGGAUGAGGACAUGCUUCU	AGAAGCATGCTCATCCAC	-
candidate-216	homology candidate	UAGCUUAGGCUUAGGCUUAUGUUUA	TAAACATAAGCCTAAGCCTAAGCTA	-
candidate-217	homology candidate	UAGGAACUCAAAGCGUUUCCGAA	TTCGGAACGCTTTGAAGTTCCTA	-
candidate-220	homology candidate	UCUGAUCCUUCAGAGCUUAA	TTAAGCTCTGAAGGATCAGA	-

in the library are not enough to generate a PCR product (see the Supplemental Experimental Procedures for details). This check was particularly important because non-miRNA noncoding RNA sequences have frequently been reported in libraries prepared for miRNA cloning (Lagos-Quintana et al., 2001) and because some sequences predicted by the homology algorithm are very similar to previously cloned miRNAs.

The level of specificity of the PCR assay was also illustrated by the detection of *mir-236*: a conserved region of *mir-236* was predicted by both the CDH and homology algorithms. However, the sequence overlap was not complete. An 18 nt core sequence was common to both predictions but an additional 6 nt segment was identified 5' of the core sequence by the homology algorithm, while an additional 4 nt segment was predicted 3' of the core sequence by the CDH algorithm (Figure 3B). Interestingly, an amplification product was detected only with the predicted mature *mir-236* sequence from the homology algorithm (*mir-236^h*), containing the additional 6 nt 5' segment, and not with the predicted mature *mir-236* sequence from the CDH algorithm (*mir-236^{CDH}*) that lacked the 6 nt 5' sequence (Figure 3A). This finding

illustrates the importance of correct prediction of the 5' region of the mature miRNA, as noted above. Interestingly, both *mir-236^{CDH}* and *mir-236^h* hybridize to the endogenous mature *mir-236* transcript by Northern blots, suggesting that the PCR assay is more stringent than conventional Northern blot analyses.

Taken together, these results suggest that many candidate miRNAs are indeed real but expressed at levels below the threshold for detection by Northern blots. The PCR assay provides strong but not unequivocal evidence confirming the expression of our candidates, so we annotate the candidates that were positive by the PCR assay as computationally predicted, PCR assay-supported miRNAs, or *cp-mir-264* to *cp-mir-273*. All of the new miRNAs were submitted to the miRNA Registry website at <http://www.sanger.ac.uk/Software/Rfam/mirna/> for official annotation. In a recent letter, a consortium of groups in the miRNA field has agreed upon a set of guidelines for identifying and annotating new miRNAs using both expression and biogenesis criteria (Ambros et al., 2003). Our candidate miRNAs meet the biogenesis criteria established by this letter. While the detection of the candidate miRNAs by our PCR method does not

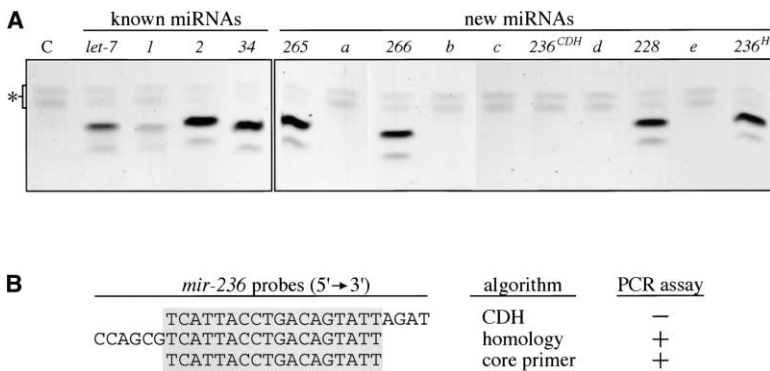


Figure 3. PCR-Based Detection of Candidate miRNAs from an Amplified Library of Small RNAs

An amplified cDNA library of the small RNAs was generated by RT-PCR using oligonucleotides corresponding to the linker sequences. A second round of PCR amplification was performed using a 5' primer to the 5' linker sequence and a 3' primer specific for the candidate miRNA (see the Experimental Procedures).

(A) Representative composite figure of the PCR products obtained after PCR amplification with the candidate miRNA-specific 3' primers. The control primer (C) is complementary to W05F2.3, an abundant mRNA

from mixed-stage *C. elegans* (Jones et al., 2001). All of the known, conserved miRNAs tested, including *let-7*, *mir-1*, *mir-2*, and *mir-34*, amplify a PCR product. New miRNAs identified by this assay include *mir-265*, which was identified by the CDH algorithm, and *mir-266*, *mir-228*, and *mir-236^h*, which were identified by the homology algorithm (see text for details). Representative miRNA candidates for which PCR products were not detected are as follows: candidate-127 (a), candidate-180 (b), candidate-77 (c), *mir-236^{CDH}* (see below), candidate-187 (d), and candidate-169 (e). See Supplemental Table S1 for complete candidate sequence information. An asterisk indicates background bands that also appear in PCR reactions with the 3' primer omitted from the amplification reaction. A summary of all of the miRNA candidates and control primers tested by the PCR assay appears in Table 1.

(B) Sequences and summary of the PCR assay for *mir-236* predicted by the CDH (*mir-236^{CDH}*) and homology (*mir-236^h*) algorithms as well as an 18 nt sequence common to both algorithms (core primer). Only the *mir-236^h* and the core primers for the sequence common to both algorithms amplify a PCR product, suggesting that the correct prediction of the 5' region of a miRNA is important for PCR amplification. See text for details.

conform to the expression criteria, which are biased toward abundant miRNAs detectable via Northern blots or cloning, we believe that our method offers a valid approach in determining the expression of rare miRNAs for which no other method currently exists.

The miRNA candidates not detected by the PCR strategy may either be expressed at even lower levels or may be expressed under specific environmental conditions such that they are not represented in the original library constructed from mixed-stage worm RNA. In addition, as demonstrated by the detection of *mir-236* from the homology algorithm but not the CDH algorithm, the correct prediction of the mature miRNA sequence at the 5' region may be critical for its hybridization, amplification, and detection by this PCR assay. Nevertheless, because all of the candidates appear structurally similar to conserved and easily detected miRNAs, we leave open the possibility that miRNA candidates unconfirmed by expression studies may also encode bona fide miRNAs.

In summary, we detected, via Northern blot, four novel computationally predicted miRNAs (two worm, one fly, and one human), indicating that these novel miRNAs are expressed at high enough levels to be detected by conventional hybridization methods. Furthermore, 10 of 54 *C. elegans* miRNA candidates that were negative by Northern blots were detected by PCR assay in a library of enriched small RNAs. These findings validate the informatic methods for identifying apparently conserved miRNAs in *C. elegans* and suggest that similar computational approaches can be adopted to survey systematically other eukaryotic genomes for potential miRNAs.

Estimation of the Number of *C. elegans* miRNAs

We estimate that the *C. elegans* genome encodes between 140 to 300 miRNAs and potentially many more (see below). This estimate suggests that $\sim 1\%$ of the *C. elegans* genome encodes miRNAs, consistent with a recent study indicating that approximately 1% of the human genome encodes miRNAs (Lim et al., 2003). To estimate the number of *C. elegans* miRNAs, two adjustment factors were incorporated into the calculations. Of the 61 hairpins associated with cloned *C. elegans* miRNAs, nine have been identified as having homologs in fly, mouse, or human sequence (Lagos-Quintana et al., 2002; Lau et al., 2001; Lee and Ambros, 2001). Our independent analysis suggests that an additional nine are conserved, and thus a total of 18 of the 61 cloned *C. elegans* miRNAs ($\sim 30\%$) represent conserved *C. elegans* miRNAs. From this finding, we derive the first factor of ~ 3.4 (61 cloned *C. elegans* miRNAs / 18 conserved miRNAs) to allow estimation of the total number of *C. elegans* miRNA hairpins from the number of conserved miRNAs. Twenty-nine of the sixty-one cloned *C. elegans* miRNA hairpins were identified by our algorithms. Accounting for the fraction of miRNAs in the genome that are either in sequence we did not analyze (most significantly, large introns and the portion of the genome sequence still unassembled at the time of this analysis) or do not pass our filters results in the second factor of ~ 2.1 (61 cloned miRNAs / 29 known miRNAs that pass the filters). We then arrive at a composite adjustment factor of ~ 7.1 ($\sim 3.4 * \sim 2.1 = \sim 7.1$) required to estimate the total number of miRNAs in the *C. elegans* genome

based on the predicted number of conserved miRNAs from our algorithms.

By our PCR assay, $\sim 20\%$ of the candidates were detected. If we apply this positive rate to the 214 candidates predicted by our three algorithms, then we conclude that ~ 43 candidates are bona fide miRNAs. Using the composite adjustment factor (~ 7.1), we arrive at an estimate of ~ 300 miRNAs ($7.1 * 43$) in the *C. elegans* genome. Alternatively, a more conservative assessment considers only those 20 predictions confirmed experimentally to date, which comprise the eight previously cloned miRNAs, the two new conserved miRNAs confirmed by Northern blot, and ten additional conserved miRNA hairpins confirmed by our PCR amplification procedures. This gives a lower estimate of $7.1 * 20 \approx 140$ miRNA hairpins encoded by the *C. elegans* genome. In fact, the number of *C. elegans* microRNA genes may exceed the estimate of 300. If a higher percentage of the candidate miRNAs that were not detected by PCR encode miRNAs that are expressed only at particular times or in particular cells, or at levels too low for the detection schemes we have used so far, or if many of the predicted miRNAs have 5' ends that are mispredicted by a few nucleotides such that the PCR primers used in the analysis would fail, more than 20% of the predicted miRNAs may be bona fide. Because these predicted miRNAs are related to other predicted or experimentally verified miRNA genes, and because they are structurally similar to verified miRNA genes, they are excellent candidates for encoding real miRNAs. These estimates depend on the thresholds and definitions of conservation used by our algorithms, assume that the 61 *C. elegans* miRNA hairpins identified by cloning constitute an unbiased sample with respect to hairpin sequence and structure characteristics, and are also sensitive to the high variance associated with the small sample size of 61 miRNA hairpins. With full optimization of our amplification and detection procedures and further testing of our predicted miRNAs, estimates of the number of *C. elegans* hairpins will improve. We note that, to the extent that our predictions overestimate actual miRNAs, the incorporation into our algorithms of any forthcoming knowledge about sequence and structure determinants of miRNAs, including sequence signatures for transcription initiation, features of hairpin structures preferentially recognized and cleaved by RNase III-like enzymes, and characterization of key functional nucleotides within mature miRNAs, will likely aid in refining the algorithms to predict miRNAs in the genome.

Clustering of microRNAs into Conserved Sequence Families

Computational analysis of known miRNAs reveals that subsets of miRNAs share common sequence elements. We performed pairwise Smith-Waterman alignments of all published miRNAs and the four computationally predicted miRNAs verified by Northern blot analysis and generated a complete linkage hierarchical cluster tree based on sequence similarity scores (Figure 4; see also Supplemental Figure S1 at <http://arep.med.harvard.edu/miRNA/> and the Supplemental Experimental Procedures). The extent of sequence similarity within each cluster ranged from near-perfect identity to blocks of 6

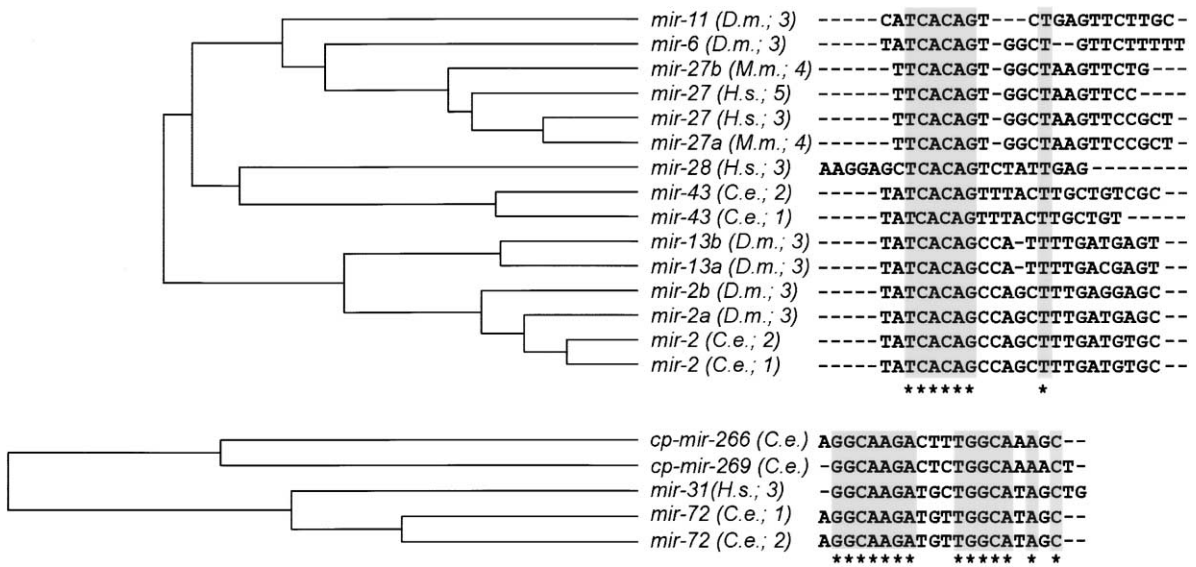


Figure 4. Sample Clusters of Published Metazoan miRNAs Based on Sequence Similarity

A set of 305 miRNA sequences, including redundant miRNAs published in multiple articles (Lee and Ambros, 2001; Lau et al., 2001; Lagos-Quintana et al., 2001, 2002; Mourelatos et al., 2002), were aligned against each other by a Smith-Waterman algorithm (EMBOSS v2.3.1; Rice et al., 2000). From the scores of the pairwise comparisons, a dissimilarity matrix was constructed and used in performing complete hierarchical clustering (see the Experimental Procedures and Supplemental Experimental Procedures at <http://www.molecule.org/cgi/content/full/11/5/1253/DC1>). A dendrogram was generated from the resulting clustering (see Supplemental Figure S1 at <http://arep.med.harvard.edu/miRNA/> for the full dendrogram). The dendrogram was then cut to yield a set of clusters, including the *let-7* variant cluster and a cluster akin to the set of miRNAs reported in Lai (2002). Alignments and membership of clusters were then adjusted by hand to improve grouping of miRNAs that share subsequences with common locations (see the Supplemental Experimental Procedures). Several examples of such clusters are presented here, with conserved sequences highlighted in gray.

to 8 nucleotide conservation. We then grouped miRNAs that share relative location of identical sequence blocks and derived ~40 miRNA “families” (Figure 4 and Supplemental Table S2 at <http://arep.med.harvard.edu/miRNA/>).

These cluster alignments may facilitate computational prediction of miRNA targets. Target prediction is difficult partly because there are few well-characterized examples from which to generalize. The best described targets comprise two imperfect duplexes between *let-7* and the 3' UTR of *lin-41*, and seven imperfect duplexes between *lin-4* and the 3'-UTR of *lin-14* (Lee et al., 1993; Reinhart et al., 2000; Wightman et al., 1993; Slack et al., 2000; Pasquinelli et al., 2000). While evidence from these examples suggests that specific patterns of bulges, mismatches, and stretches of perfect duplex formation are important determinants of function, it is unclear whether these specifics can be generalized to other miRNA-target interactions. Consequently, searches for potential targets of any given miRNA lack a priori restrictions on duplex variability and thus are highly degenerate and nonspecific. However, sequence-based alignment of apparently homologous miRNAs reveals patterning that can be used to restrict the range of variability considered in target searches, under the assumption that the miRNA-target duplex structure is also conserved. These suggestions are supported by and extend a recent observation concerning a cluster alignment of six similar miRNAs from *D. melanogaster* that are complementary to the 3' UTR elements of genes known to be regulated posttranscriptionally in this organism (Lai, 2002).

In this study, we predicted miRNAs by searching mul-

iple genomes for similar, short sequences contained in hairpins satisfying energy, sequence, and structure constraints. This computational approach overcomes the difficulty of biochemically discovering low-abundance miRNAs. We have also shown that sequence similarities with known miRNAs may be exploited toward the discovery of other new miRNAs and may also be a tool for prediction of miRNA targets. While the discovery of miRNAs that are conserved across animal phylogeny implies their biological importance in gene regulation, little is known about the genetic pathways and the target genes that they regulate. Therefore, one of the major challenges will be to identify the targets of miRNA regulation, thus allowing us to place specific miRNAs in their genetic pathways and biological contexts. This endeavor will also likely require a multifaceted approach including biochemical, genetic, and computational strategies.

Experimental Procedures

RNA Analysis

Northern Blots

Total RNA isolation and Northern blot procedures have been described previously (Ausubel et al., 1995; Reinhart et al., 2000). The candidate miRNA and *let-7* sequences are presented in Supplemental Table S1; antisense probes were designed to these sequences and used for the Northern analysis and for the PCR assay. The total RNAs from human tissues were purchased from Clontech. The total RNAs from *D. melanogaster* developmental stages were kind gifts from M. Kuroda (Baylor College of Medicine, Houston, TX) and N. Perrimon (Harvard Medical School, Boston, MA). The *dcr-1 (ok-247)* strain was grown as previously described (Dent et al., 1997).

PCR Assay of an Amplified Small RNA Library

To construct a library of enriched miRNAs, endogenous 18 to 24 nt RNAs were size selected from total RNA from the N2 wild-type *C. elegans* strain, ligated with 5' and 3' RNA oligonucleotide linkers, and amplified by RT-PCR using antisense DNA oligonucleotides complementary to the linker sequences as described previously (Lau et al., 2001). To PCR-amplify candidate miRNAs from this amplified small RNA library, an oligonucleotide complementary to the 5' linker region was used with a 3' oligonucleotide complementary to the particular candidate miRNA. A list of all candidates tested, as well as known miRNAs and negative controls, appears in Table 1.

Computational Methods

We provide a brief summary of computational methods here. For full details see the Supplemental Experimental Procedures at <http://www.molecule.org/cgi/content/full/11/5/1253/DC1>.

Sequences and Annotations

We used genome assemblies as follows: *C. elegans* (produced by the *C. elegans* Sequencing Group at the Sanger Institute and Genome Sequencing Center at Washington University) downloaded from <http://www.sanger.ac.uk> on 7 March, 2001, *D. melanogaster* (release 2) downloaded from <http://www.fruitfly.org> on 9 May, 2001 (Adams et al., 2000), repeat-masked human genome sequence downloaded from <http://www.ncbi.nlm.nih.gov> on 13 August, 2001 (Lander et al., 2001). Annotations downloaded with the *C. elegans* and *D. melanogaster* genome sequences were used to identify intragenic regions. The *C. elegans* and *D. melanogaster* genome sequences were repeat masked using the RepeatMasker version dated 19 June, 2001 (A.F.A. Smit and P. Green, RepeatMasker at <http://ftp.genome.washington.edu/RM/RepeatMasker.html>). We downloaded unassembled genomic reads of *C. briggsae* (~6× coverage) from the Washington University Sequencing Center at <http://www.genome.wustl.edu/> on 21 September, 2001.

miRNA Test Set

A test set of 53 miRNAs made available to us thanks to Thomas Tuschl and later published in Lagos-Quintana, et al. (2001) was used to test the effect of combinations of parameter settings used in generating predicted miRNA hairpin sets. The test set included several variants of *let-7*.

smalooop

smalooop is a BLAST-like algorithm that looks for short complementary words within a specified distance and uses dynamic programming to determine a complete alignment. Compared to BLAST (Altschul et al., 1990), smalooop supports shorter word lengths and aligns complementary base pairs (including GUs). See the Supplemental Experimental Procedures for details on how smalooop was used for generation of candidate miRNA hairpins and our web site (<http://arep.med.harvard.edu/miRNA/>) for additional information and the software itself.

Candidate miRNA Hairpin Selection

Candidate miRNA hairpins generated by smalooop were filtered using a variety of criteria. For full details see the Supplemental Experimental Procedures. (1) Stutter filtering: smalooop may find hairpins on the same strand of a given sequence that overlap for a considerable fraction of their lengths, a phenomenon we refer to as "stuttering." Stutter filtering refers to the selection of a single hairpin out of a set of such overlaps. (2) GC content filtering: Candidate hairpins are eliminated if the GC content is outside of bounds found to apply to our miRNA test set of sequences. (3) Folding energy and structure filters: Sets of predicted hairpins were processed by RNAfold (Hofacker et al., 1994) using the $-d0$ option to compute minimum free energies of folding and structure characteristics such as numbers of multiloops which were then used to refine selections of candidate hairpins. (4) Correspondence determination: Candidate miRNA hairpins from one species A are BLASTed against the genome sequence of another species B. In the vast majority of cases, only short BLAST matches of 20 nt or less are found. Additional sequence around the BLAST match in B is extracted and examined for hairpins using smalooop and filters of the sort described above. If the sequence around the BLAST match in B forms a hairpin satisfying these criteria, and the BLAST target in the B hairpin is on the same stem of the hairpin as the BLAST source of the candidate miRNA hairpin in A, then the B hairpin is considered to correspond to the A hairpin.

(5) Transitivity filter: Where a hairpin in species A is found to correspond to a hairpin in species B and that hairpin in B is additionally found to correspond to a hairpin in a third species C, this filter assures that the BLAST hit that established the correspondence between A and B overlaps with the BLAST hit that established the correspondence between B and C. (6) Short repeat filtering: This filter removes candidate miRNA hairpins that contained mononucleotide sequences or short tandem repeats. Although all genomic sequences used for miRNA hairpin analysis were RepeatMasked, many short sequence repeats of this type were still found. (7) Structure quality filtering: Candidate miRNA hairpins are eliminated based on a detailed examination of the number, sizes, and positions of bulges in the predicted structure.

Hairpin Sets

(1) An initial set of 8713 *C. elegans* hairpins was generated by running smalooop and applying stutter, GC content, and RNA structure filtering. Details on the parameters used to generate this set and on the numbers of known miRNA hairpins it contained are in the text and in the Supplemental Experimental Procedures. (2) Refined set of 6086 *C. elegans* hairpins: For our homology-based miRNA predictions, the initial set of 8713 *C. elegans* hairpins was refined by application of more stringent filters for coding sequence, short repeat filtering, and structure quality filtering. (3) *D. melanogaster* correspondences to the initial set of 8713 *C. elegans* hairpins: Correspondences between the initial set of *C. elegans* hairpins and *D. melanogaster* genomic sequence were determined as described above, resulting in a set of 3514 *D. melanogaster* and 3019 *C. elegans* hairpins. Removing duplicates led to a set of 3505 distinct *D. melanogaster* and 2523 distinct *C. elegans* hairpins. (4) CDC set (*C. elegans*→*D. melanogaster*→*C. briggsae*): The set of 2523 distinct *C. elegans* hairpins was BLASTed into the *C. briggsae* genome sequence reads using an e-value cutoff of 10^{-14} , resulting in a set of 95 hairpins. Fourteen of these sequences comprising apparent repetitive sequence and a near duplicate were eliminated, resulting in a set of 81 distinct hairpins. Subsequently, a higher quality subset of 28 sequences (CDC-f) was selected on the basis of structure quality. These criteria became the basis of structure quality filtering. (5) CDH set (*C. elegans*→*D. melanogaster*→*human*): Correspondences between the 3505 distinct *D. melanogaster* hairpins found to correspond to the 8713 *C. elegans* hairpin set and human genomic sequence were determined and then subjected to GC content, stutter, RNA folding energy and structure, transitivity, short repeat, and structure quality filtering as described above, as well as additional filtering for possible coding sequence. This resulted in a set of 40 hairpins. (6) *C. elegans* miRNA homolog set: We used matcher, a pure Smith-Waterman algorithm, from the EMBOS v2.3.1 software package (Rice et al., 2000) to align each of 164 miRNA sequences against the *C. elegans* set of 6086 hairpins described above. Filters based on matcher-generated alignments (see the Supplemental Experimental Procedures) and structure quality filtering were applied and left a set of 116 candidate worm hairpin orthologs and paralogs of known miRNAs.

Additional Computational Methods

See the Supplemental Experimental Procedures at <http://www.molecule.org/cgi/content/full/11/5/1253/DC1> for further information on clustering and multiple alignments of miRNAs, enumeration of cloned *C. elegans* miRNAs, analysis of conservation of predicted *C. elegans* miRNAs, and screening of candidate sequences and PCR primers against noncoding RNA sequence.

Acknowledgments

The authors thank M. Kuroda and N. Perrimon for RNA samples from *D. melanogaster* developmental stages. We also thank N. Lau and D. Bartel for providing us a protocol for making the amplified library of small RNAs and A. Grishok for technical advice. We are grateful to T. Tuschl and D. Bartel for sharing data prior to publication. We thank members of the Ruvkun and Church Labs for helpful comments on the studies and the manuscript. We gratefully acknowledge the insightful comments of the anonymous reviewers. This research was supported by a Department of Energy HGP grant and the Lipper Foundation to G.M.C. and a National Institutes of Health GM44619 grant to G.R.

Received: October 10, 2002
Revised: March 21, 2003
Accepted: March 31, 2003
Published online: April 23, 2003

References

- Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., et al. (2000). The genome sequence of *Drosophila melanogaster*. *Science* 287, 2185–2195.
- 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.
- 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.
- Ausubel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Seidman, J.G., Smith, J.A., and Struhl, K. (1995). *Current Protocols in Molecular Biology* (New York: John Wiley and Sons).
- Bernstein, E., Caudy, A.A., Hammond, S.M., and Hannon, G.J. (2001). Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature* 409, 363–366.
- Dent, J.A., Davis, M.W., and Avery, L. (1997). *avr-15* encodes a chloride channel subunit that mediates inhibitory glutamatergic neurotransmission and ivermectin sensitivity in *Caenorhabditis elegans*. *EMBO J.* 16, 5867–5879.
- Dostie, J., Mourelatos, Z., Yang, M., Sharma, A., and Dreyfuss, G. (2003). Numerous microRNPs in neuronal cells containing novel microRNAs. *RNA* 9, 180–186.
- 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.
- 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.
- Grosshans, H., and Slack, F.J. (2002). Micro-RNAs: small is plentiful. *J. Cell Biol.* 156, 17–21.
- Ha, I., Wightman, B., and Ruvkun, G. (1996). A bulged *lin-4/lin-14* RNA duplex is sufficient for *Caenorhabditis elegans* *lin-14* temporal gradient formation. *Genes Dev.* 10, 3041–3050.
- Hannon, G.J. (2002). RNA interference. *Nature* 418, 244–251.
- Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, S., Tacker, M., and Schuster, P. (1994). Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.* 125, 167–188.
- 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.
- Jones, S.J., Riddle, D.L., Pouzyrev, A.T., Velculescu, V.E., Hillier, L., Eddy, S.R., Stricklin, S.L., Baillie, D.L., Waterston, R., and Marra, M.A. (2001). Changes in gene expression associated with developmental arrest and longevity in *Caenorhabditis elegans*. *Genome Res.* 11, 1346–1352.
- 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.
- Lagos-Quintana, M., Rauhut, R., Lendeckel, W., and Tuschl, T. (2001). Identification of novel genes coding for small expressed RNAs. *Science* 294, 853–858.
- 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.
- Lagos-Quintana, M., Rauhut, R., Meyer, J., Borkhardt, A., and Tuschl, T. (2003). New microRNAs from mouse and human. *RNA* 9, 175–179.
- Lai, E.C. (2002). Micro RNAs are complementary to 3' UTR sequence motifs that mediate negative post-transcriptional regulation. *Nat. Genet.* 30, 363–364.
- Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860–921.
- 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.
- Lee, R.C., and Ambros, V. (2001). An extensive class of small RNAs in *Caenorhabditis elegans*. *Science* 294, 862–864.
- 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.
- Lim, L.P., Glasner, M.E., Yekta, S., Burge, C.B., and Bartel, D.P. (2003). Vertebrate microRNA genes. *Science* 299, 1540.
- Llave, C., Xie, Z., Kasschau, K.D., and Carrington, J.C. (2002). Cleavage of *Scarecrow*-like mRNA targets directed by a class of Arabidopsis miRNA. *Science* 297, 2053–2056.
- Mourelatos, Z., Dostie, J., Paushkin, S., Sharma, A., Charroux, B., Abel, L., Rappsilber, J., Mann, M., and Dreyfuss, G. (2002). miRNPs: a novel class of ribonucleoproteins containing numerous microRNAs. *Genes Dev.* 16, 720–728.
- Olsen, P.H., and Ambros, V. (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.
- 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.
- 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.
- 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.
- Reinhart, B.J., Weinstein, E.G., Rhoades, M.W., Bartel, B., and Bartel, D.P. (2002). MicroRNAs in plants. *Genes Dev.* 16, 1616–1626.
- Rhoades, M., Reinhart, B., Lim, L., Burge, C., Bartel, B., and Bartel, D. (2002). Prediction of plant microRNA targets. *Cell* 110, 513–520.
- Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* 16, 276–277.
- Rougvie, A.E. (2001). Control of developmental timing in animals. *Nat. Rev. Genet.* 2, 690–701.
- 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.
- 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.