Molecular Cell

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]
[1]The Lipper Center for Computational Genetics and
Department of Genetics
Harvard Medical School
Boston, Massachusetts 02115
[2]Department of Molecular Biology
Massachusetts General Hospital and
Department of Genetics
Harvard Medical School
Boston, Massachusetts 02114

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

[3]These authors contributed equally to this work.

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

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; Lee 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

Published Online April 23, 2003
DOI: 10.1016/S1097276503001539

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 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 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 Experimental Procedures and Supplemental Experimental Procedures at end of paper). 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 Experimental Procedures and Supplemental Experimental Procedures for the program), GC content, predicted structure minimum free energy, and presence of multiloops (see 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 candidate 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

DOI: 10.1016/S1097276503001539

the *C. elegans* hairpins with corresponding hairpins in *D. melanogaster* for those that closely matched sequences in *C. briggsae*, a sister nematode species (see 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 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 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 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. elegans* (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

DOI: 10.1016/S1097276503001539

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 we identified as a relative of *mir-228* (Figure 2C). While these two new miRNAs are closely related to their *C. elegans* counterparts and thus likely to be homologous, they may not be functionally equivalent. The developmentally regulated expression patterns of *let-7* in *C. elegans* and *D. melanogaster* are analogous to each other, suggesting that they function in a similar fashion. However, the developmental expression pattern of *mir-228* in *C. elegans* peaks at the L1 larval stage, while its sequence-related family member, *mir-263* in *D. melanogaster,* exhibits an oscillatory expression pattern during development. This observation supports the view that *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 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 in the library are not enough to generate a PCR product (see and 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

DOI: 10.1016/S1097276503001539

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 conform to the expression criteria, which are biased toward abundant miRNAs detectable via Northerns 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 Northerns 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 ~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 (~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 (~3.4 * ~2.1 = ~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, ~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

DOI: 10.1016/S1097276503001539

blot, and ten additional conserved miRNA hairpins confirmed by our PCR amplification procedures. This gives a lower estimate of 7.1 * 20 ≈ 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 Supplemental Experimental Procedures). The extent of sequence similarity within each cluster ranged from near-perfect identity to blocks of 6 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 multiple 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.

DOI: 10.1016/S1097276503001539

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 Supplemental Experimental Procedures.

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*.

srnaloop

srnaloop 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), srnaloop supports shorter word lengths and aligns complementary base pairs (including GUs). See the Supplemental Experimental Procedures for details on how srnaloop 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 srnaloop were filtered using a variety of criteria. For full details see the Supplemental Experimental Procedures. (1) Stutter filtering: srnaloop 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 srnaloop and filters of the sort described above. If the sequence

Published Online April 23, 2003
DOI: 10.1016/S1097276503001539

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 srnaloop 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 EMBOSS 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 Supplemental Experimental Procedures) and structure quality filtering were applied and resulted in a set of 116 candidate worm hairpin orthologs and paralogs of known miRNAs.

Additional Computational Methods

See Supplemental Experimental Procedures 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: April 22, 2003

DOI: 10.1016/S1097276503001539

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, 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. and Ruvkun, G. (2002). Control and developmental timing by microRNAs and their targets. Annu. Rev. Cell Dev. Biol. *18*, 495–513.

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.

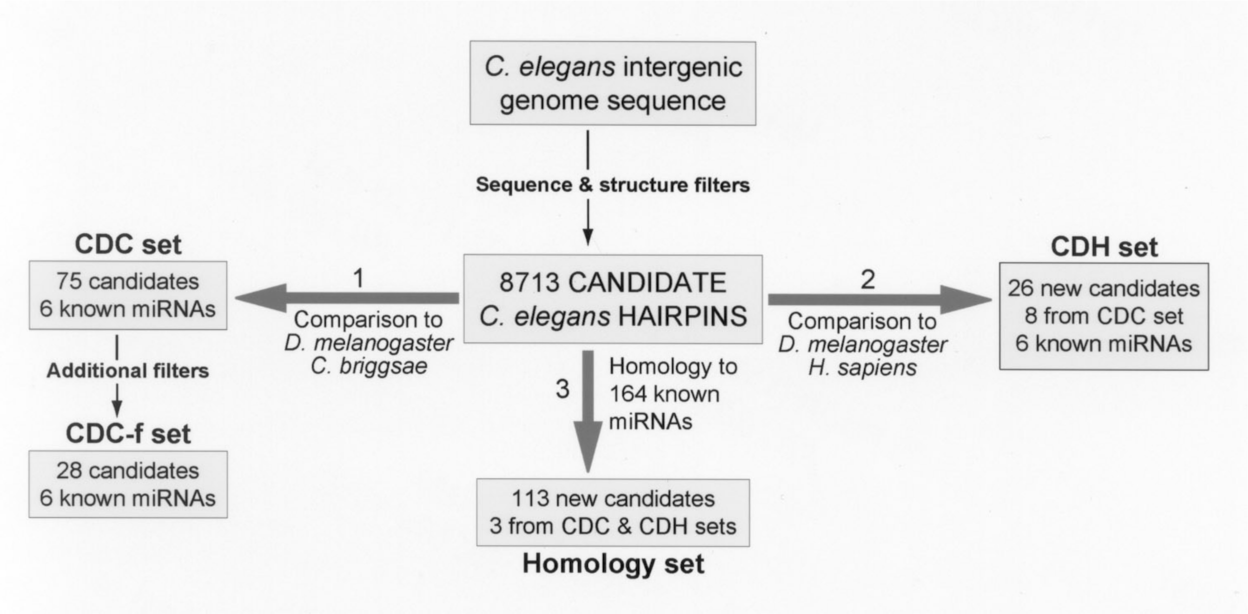
Rougvie, A. E. (2001). Control of developmental timing in animals. Nat. Rev. Genet. *2*, 690–701.

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.

DOI: 10.1016/S1097276503001539

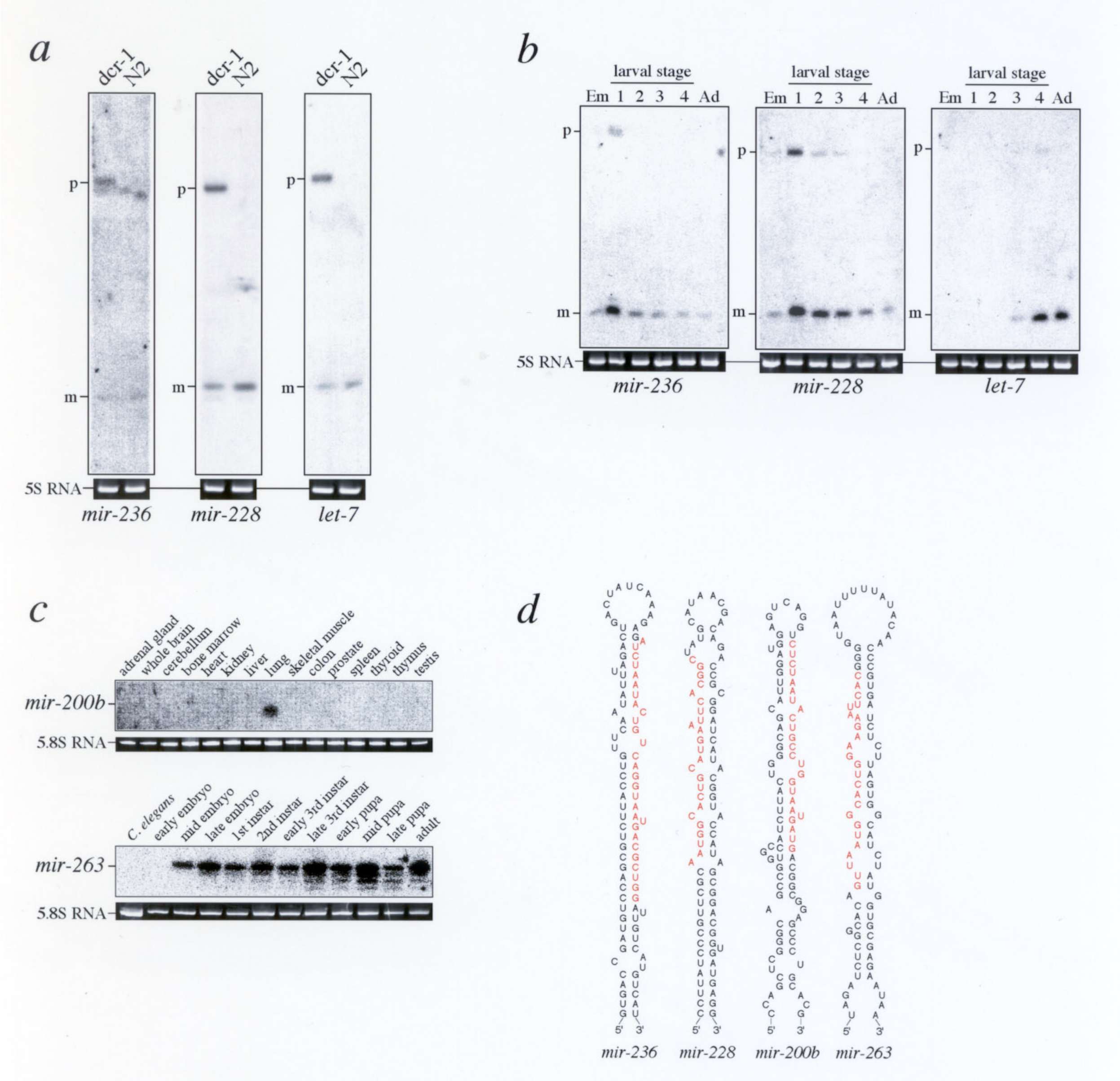
Figure 1. Schematic Representation of the Computational Algorithms Used in This Study


C. elegans intergenic genome sequence
Sequence & structure filters
8713 CANDIDATE *C. elegans* HAIRPINS
1
Comparison to *D. melanogaster* *C. briggsae*
CDC set
75 candidates
6 known miRNAs
Additional filters
CDC-f set
28 candidates
6 known miRNAs
2
Comparison to *D. melanogaster* *H. sapiens*
CDH set
26 new candidates
8 from CDC set
6 known miRNAs
3
Homology to 164 known miRNAs
113 new candidates
3 from CDC & CDH sets
Homology set

Details are discussed in the text.

Published Online April 23, 2003
DOI: 10.1016/S1097276503001539

Figure 2. Northern Blot Analysis of Novel miRNAs Identified by Computational Analysis



(A) Two novel *C. elegans* miRNAs, *mir-236* and *mir-228,* show *dcr-1* dependent accumulation of the precursor miRNA form. Total RNA from wild-type and *dcr-1* (*ok-247*) mutant strains was subjected to Northern analysis using antisense probes against the mature form of *let-7* and the stem regions of the two *C. elegans* miRNA candidates. The mature (m) and precursor (p) forms of the miRNAs are indicated.

(B) The expression of *mir-236* and *mir-228* is developmentally regulated. Northern blot of total RNA from wild-type *C. elegans* developmental stages probed as in (A) indicates a peak in expression at L1 larval stage. The mature (m) and precursor (p) forms of the miRNAs are indicated.

(C) Expression of *mir-200b,* a human homolog of *mir-236,* and *mir-263,* a fly homolog of *mir-228*. *mir-200b* was detected on a Northern blot of total RNAs from human tissues (top panel) and *mir-263* was detected on a Northern blot of total RNAs from fly developmental stages with antisense probes as described in (A).

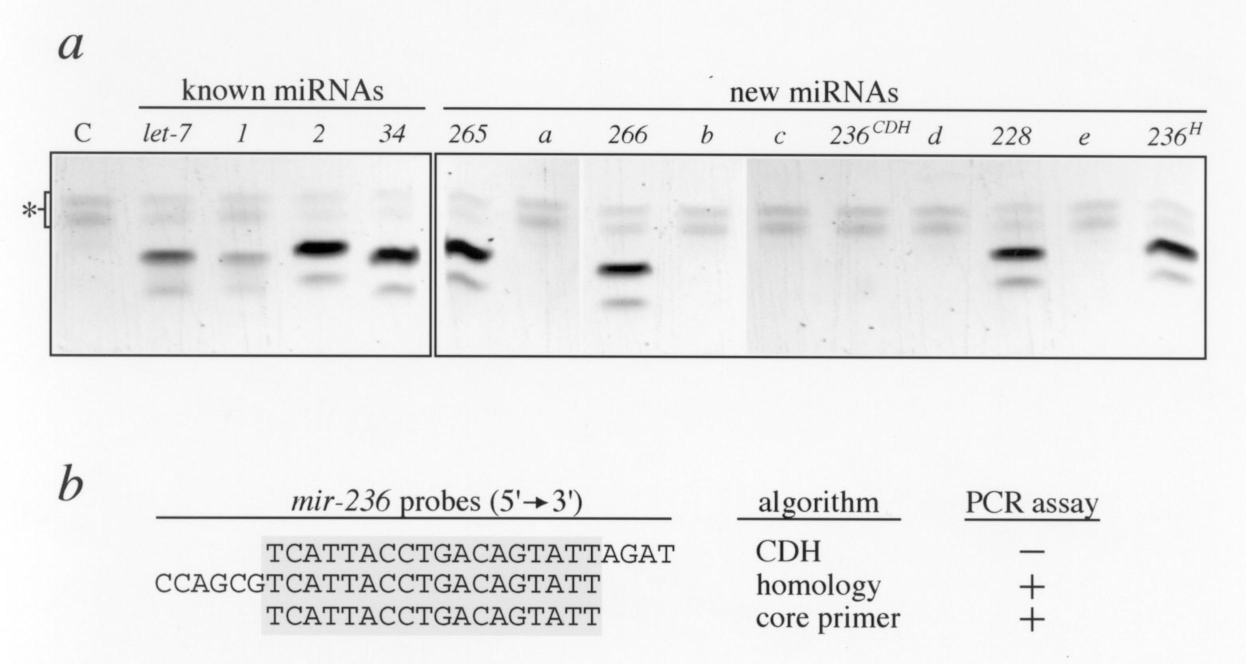
Published Online April 23, 2003

DOI: 10.1016/S1097276503001539

(D) Hairpin secondary structures for the sequence regions around which novel miRNAs are predicted to be encoded. The sequence boundaries of the precursor miRNAs remain to be determined. The stem regions to which the antisense probes were designed are indicated in red type; in the case of *mir-236* and *mir-263,* predicted mature miRNA sequences were from the shortest probe which was positive by Northern blot. In cases where overlapping short probes were used, the full sequence including the nonoverlap region was used to designate the mature sequence in subsequent computational analyses.

Published Online April 23, 2003
DOI: 10.1016/S1097276503001539

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

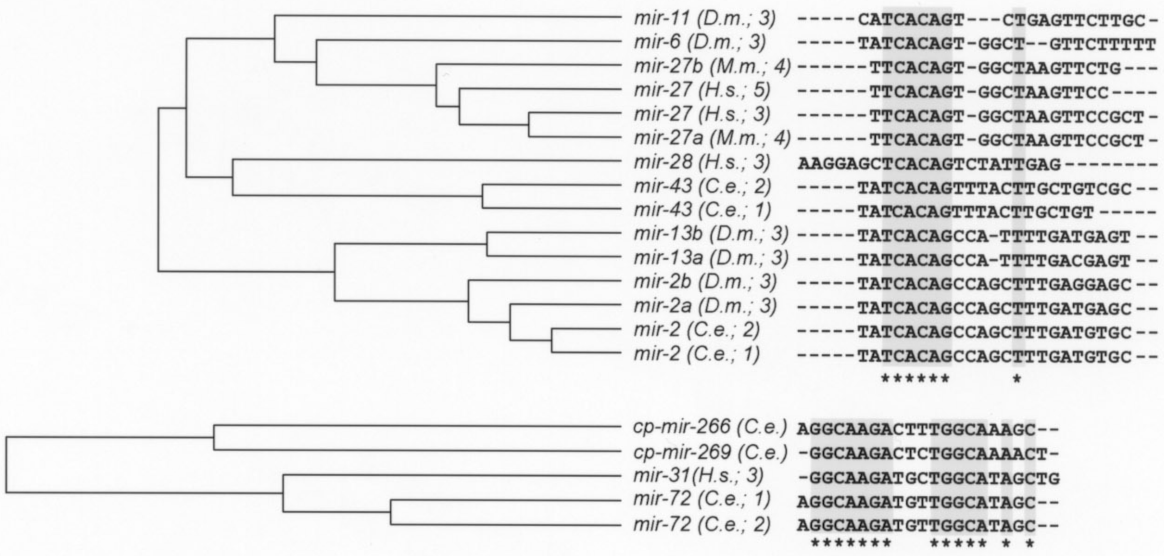


An amplified cDNA library of the small RNAs was generated by RT-PCR using oligonucleotides corresponding to the linker sequences. A second round of PCR amplification was performed using a 5′ primer to the 5′ linker sequence and a 3′ primer specific for the candidate miRNA (see Experimental Procedures).

(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.

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


mir-11 (D.m.; 3) -----CATCACAGT---CTGAGTTCTTGC-
mir-6 (D.m.; 3) -----TATCACAGT-GGCT--GTTCTTTTT
mir-27b (M.m.; 4) ------TTCACAGT-GGCTAAGTTCTG---
mir-27 (H.s.; 5) ------TTCACAGT-GGCTAAGTTCC----
mir-27 (H.s.; 3) ------TTCACAGT-GGCTAAGTTCCGCT-
mir-27a (M.m.; 4) ------TTCACAGT-GGCTAAGTTCCGCT-
mir-28 (H.s.; 3) AAGGAGCTCACAGTCTATTGAG--------
mir-43 (C.e.; 2) -----TATCACAGTTTACTTGCTGTCGC--
mir-43 (C.e.; 1) -----TATCACAGTTTACTTGCTGT-----
mir-13b (D.m.; 3) -----TATCACAGCCA-TTTTGATGAGT--
mir-13a (D.m.; 3) -----TATCACAGCCA-TTTTGACGAGT--
mir-2b (D.m.; 3) -----TATCACAGCCAGCTTTGAGGAGC--
mir-2a (D.m.; 3) -----TATCACAGCCAGCTTTGATGAGC--
mir-2 (C.e.; 2) -----TATCACAGCCAGCTTTGATGTGC--
mir-2 (C.e.; 1) -----TATCACAGCCAGCTTTGATGTGC--
******* *
cp-mir-266 (C.e.) AGGCAAGACTTTGGCAAAGC--
cp-mir-269 (C.e.) -GGCAAGACTCTGGCAAAACT-
mir-31(H.s.; 3) -GGCAAGATGCTGGCATAGCTG
mir-72 (C.e.; 1) AGGCAAGATGTTGGCATAGC--
mir-72 (C.e.; 2) AGGCAAGATGTTGGCATAGC--
******** ***** * *

A set of 305 miRNA sequences, including redundant miRNAs published in multiple articles (Lagos-Quintana et al., 2001, 2002; Lau et al., 2001; Lee and Ambros, 2001; 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 Experimental Procedures and Supplemental Experimental Procedures). A dendrogram was generated from the resulting clustering (see Supplemental Figure S1 at http://arep.med.harvard.edu/miRNA/ for 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 Supplemental Experimental Procedures). Several examples of such clusters are presented here, with conserved sequences highlighted in gray.

DOI: 10.1016/S1097276503001539

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)	*C. elegans* mRNA	na	GCGACCAAAAGAATTAGGATGG	–
R09B5.3 (m)	*C. elegans* mRNA	na	ACCTCCCATCATTCCTGGGTAT	–
R09B5.9 (f)	*C. elegans* mRNA	na	GCGACAAGAAGGACGAGAATGT	–
R09B5.9 (m)	*C. elegans* mRNA	na	TCCACCACCATACCCGCCATAT	–
Y2H9A.3 (f)	*C. elegans* mRNA	na	CTAGCGCCGTATACAAGAGTCT	–
Y2H9A.3 (m)	*C. elegans* mRNA	na	ACCATTCCATGCCAGGCACCAC	–
F53F1.4 (f)	*C. elegans* mRNA	na	ATAGCAGCGAAGAAAATGACAA	–
F57H12.3 (f)	*C. elegans* mRNA	na	GCAGCTGAGATTTTAAAGATAT	–
C45B2.1 (f)	*C. elegans* mRNA	na	GCAAACGCCAGGAAAACGGCAG	–
T23G11.3 (f)	*C. elegans* mRNA	na	ACACCGTAAGTTGGAGTGGTGC	–
F28D1.5 (f)	*C. elegans* mRNA	na	AGGAGAGCGAGAGTGAGCTTGA	–
T23G11.2 (f)	*C. elegans* mRNA	na	AGGATTTCATCGAAAATTGAAA	–
W05F2.3 (f)	*C. elegans* mRNA	na	ATCGTGAGAACGGCGAGTGTCA	–
C33G8.2 (f)	*C. elegans* mRNA	na	ATGGCAAGGATAAGTATAGGTA	–
let-7	known microRNA	UGAGGUAGUAGGUUGUAUAGU	ACTATACAACCTACTACCTCA	+
mir-1	known microRNA	UGGAAUGUAAAGAAGUAUGUA	TACATACTTCTTTACATTCCA	+
mir-2	known microRNA	UAUCACAGCCAGCUUUGAUGUGC	GCACATCAAAGCTGGCTGTGATA	+
mir-34	known microRNA	AGGCAGUGUGGUUAGCUGGU	ACCAGCTAACCACACTGCCT	+
mir-47	known microRNA	UGUCAUGGAGGCGCUCUCUUCA	TGAAGAGAGCGCCTCCATGACA	+
mir-56-2	known microRNA	UACCCGUAAUGUUUCCGCUGAG	CTCAGCGGAAACATTACGGGTA	+
mir-79	known microRNA	AUAAAGCUAGGUUACCAAAGCU	AGCTTTGGTAACCTAGCTTTAT	+
mir-236	CDH candidate	AUCUAAUACUGUCAGGUAAUGA	TCATTACCTGACAGTATTAGAT	–
mir-228	CDH candidate	AAUGGCACUGCAUGAAUUCACGG	CCGTGAATTCATGCAGTGCCATT	+
mir-264	CDH candidate	GGCGGGUGGUUGUUGUUAUG	CATAACAACAACCACCCGCC	+
candidate-31	CDH candidate	UUCUGGAGCAGGAACUGCAGCUG	CAGCTGCAGTTCCTGCTCCAGAA	–
candidate-42	CDH candidate	AGUGGCAGUGGACAUUUGACGG	CCGTCAAATGTCCACTGCCACT	–
candidate-52	CDH candidate	AAAGUUGCUAAAGUUGGUGGA	TCCACCAACTTTAGCAACTTT	–
candidate-72	CDH candidate	CUCGUCUACCCUGUAGAUCGA	TCGATCTACAGGGTAGACGAG	–
candidate-77	CDH candidate	AAGUUGUGAGCCGCCACUGCGACG	CGTCGCAGTGGCGGCTCACAACTT	–
candidate-100	CDH candidate	ACGAUUUGGCAUUGGAUGUGG	CCACATCCAATGCCAAATCGT	–
candidate-108	CDH candidate	UAUGCGAGUGUGUGGGGCUCC	GGAGCCCCACACACTCGCATA	–
candidate-127	CDH candidate	CGUUUUCUUCUGUCGUUCCCC	GGGGAACGACAGAAGAAAACG	–
candidate-135	CDH candidate	CAUUACCUGACAGUAUUAGAU	ATCTAATACTGTCAGGTAATG	–
candidate-169	CDH candidate	GCUUCUGCGGUGCGUGCGUGGG	CCCACGCACGCACCGCAGAAGC	–
candidate-169	CDH candidate	UUCUAUUUGUGCUUGUGCGCA	TGCGCACAAGCACAAATAGAA	–
candidate-187	CDH candidate	AGUUGAGCCUGCUGGUGGGUUU	AAACCCACCAGCAGGCTCAACT	–
mir-236	homology candidate	AAUACUGUCAGGUAAUGACGCUGG	CCAGCGTCATTACCTGACAGTATT	+
cp-mir-265	homology candidate	UGAGGGAGGAAGGGUGGUAU	ATACCACCCTTCCTCCCTCA	+
cp-mir-266	homology candidate	AGGCAAGACUUUGGCAAAGC	GCTTTGCCAAAGTCTTGCCT	+
cp-mir-267	homology candidate	CCCGUGAAGUGUCUGCUGCA	TGCAGCAGACACTTCACGGG	+
cp-mir-268	homology candidate	GGCAAGAAUUAGAAGCAGUUUGGU	ACCAAACTGCTTCTAATTCTTGCC	+

Published Online April 23, 2003
DOI: 10.1016/S1097276503001539

cp-mir-269	homology candidate	GGCAAGACUCUGGCAAAACU	AGTTTTGCCAGAGTCTTGCC	+
cp-mir-270	homology candidate	GGCAUGAUGUAGCAGUGGAG	CTCCACTGCTACATCATGCC	+
cp-mir-271	homology candidate	UCGCCGGGUGGGAAAGCAUU	AATGCTTTCCCACCCGGCGA	+
cp-mir-272	homology candidate	UGUAGGCAUGGGUGUUUG	CAAACACCCATGCCTACA	+
cp-mir-273	homology candidate	UGCCCGUACUGUGUCGGCUG	CAGCCGACACAGTACGGGCA	+
candidate-11	homology candidate	AGCCGCACAGCACUGGUUGACA	TGTCAACCAGTGCTGTGCGGCT	–
candidate-15	homology candidate	CGGAUCGUUAAAACCAGGAAGAUG	CATCTTCCTGGTTTTAACGATCCG	–
candidate-33	homology candidate	AGAAUUAAAAAUUCUAGACC	GGTCTAGAATTTTTAATTCT	–
candidate-48	homology candidate	CUAGGCCACCAACUUUAAACGGUU	AACCGTTTAAAGTTGGTGGCCTAG	–
candidate-87	homology candidate	GAGUACGGUAGAUCUGGUACUG	CAGTACCAGATCTACCGTACTC	–
candidate-93	homology candidate	UUACAGCCGUACCUACCUGCUU	AAGCAGGTAGGTACGGCTGTAA	–
candidate-96	homology candidate	UGGCAGGCACGUAGGUAUUGG	CCAATACCTACGTGCCTGCCA	–
candidate-103	homology candidate	UUCAGUUGGAGAUGUGUGCAUC	GATGCACACATCTCCAACTGAA	–
candidate-107	homology candidate	UCUUUCACGUCGGGCCACUUG	CAAGTGGCCCGACGTGAAAGA	–
candidate-111	homology candidate	GCAAGUCUUUGGCAAAACU	AGTTTTGCCAAAGACTTGC	–
candidate-112	homology candidate	UGAGGCUAAGAAAUUGUGUAGUU	AACTACACAATTTCTTAGCCTCA	–
candidate-117	homology candidate	UAGCAACCAUUUGAAGUUGUU	AACAACTTCAAATGGTTGCTA	–
candidate-120	homology candidate	CAGUCUACCACAUGGUCGU	ACGACCATGTGGTAGACTG	–
candidate-130	homology candidate	AUUUGGAAUUUUCUAGAUCA	TGATCTAGAAAATTCCAAAT	–
candidate-133	homology candidate	GAUCUGAUCCUUCAGAGCUU	AAGCTCTGAAGGATCAGATC	–
candidate-133	homology candidate	UGUGACUGGUGAGCAAGCGA	TCGCTTGCTCACCAGTCACA	–
candidate-134	homology candidate	UGAGGCUAGAAAAUUGUGUAGUU	AACTACACAATTTTCTAGCCTCA	–
candidate-136	homology candidate	AUUAUUGAUACUGUUGCUACGGG	CCCGTAGCAACAGTATCAATAAT	–
candidate-141	homology candidate	UGAGGCUCAUAGAUUUUGUAGUU	AACTACAAAATCTATGAGCCTCA	–
candidate-142	homology candidate	GAGAUUGUAGUUUGUAGUGUA	TACACTACAAACTACAATCTC	–
candidate-152	homology candidate	UAACCGAUAGGUUUCUGCCGAG	CTCGGCAGAAACCTATCGGTTA	–
candidate-162	homology candidate	UAAGGUGCAUUUAAGGCCGAUA	TATCGGCCTTAAATGCACCTTA	–
candidate-167	homology candidate	UGGACUCCUCGUUGUUUGCC	GGCAAACAACGAGGAGTCCA	–
candidate-178	homology candidate	UAGGUGGAGUCUGAUUUUCCACAGU	ACTGTGGAAAATCAGACTCCACCTA	–
candidate-179	homology candidate	AGGAGCACGAAUGGUUCGUG	CACGAACCATTCGTGCTCCT	–
candidate-180	homology candidate	AUAGGCUAGAUAGGUUGCCUAG	CTAGGCAACCTATCTAGCCTAT	–
candidate-213	homology candidate	AAUAGUGUCUGAAAGUUGUC	GACAACTTTCAGACACTATT	–
candidate-214	homology candidate	UGGGCAAAACUUUGGCAAAACU	AGTTTTGCCAAAGTTTTGCCCA	–
candidate-215	homology candidate	GUGGAUGAGGACAUGCUUCU	AGAAGCATGTCCTCATCCAC	–
candidate-216	homology candidate	UAGCUUAGGCUUAGGCUUAUGUUUA	TAAACATAAGCCTAAGCCTAAGCTA	–
candidate-217	homology candidate	UAGGAACUUCAAAGCGUUUCCGAA	TTCGGAAACGCTTTGAAGTTCCTA	–
candidate-220	homology candidate	UCUGAUCCUUCAGAGCUUAA	TTAAGCTCTGAAGGATCAGA	–

Published Online April 23, 2003

DOI: 10.1016/S1097276503001539

Supplemental Experimental Procedures

Computational Methods

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, personal communication). 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*.

srnaloop

srnaloop 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), srnaloop supports shorter word lengths and aligns complementary base pairs (including GUs). See our website (http://arep.med.harvard.edu/miRNA/) for additional information and the software itself. Because we wanted to accommodate the possibility that there might be miRNA precursor hairpins longer than the ~70 nt associated with currently known miRNAs, we generally directed srnaloop to look for hairpins with lengths ≤95 nt (–l 95). Score thresholds (–t) were generally either 23.5 or 23. srnaloop scores are aggregate match, mismatch, and gap scores accumulated over the duplex region of the hairpin and are not penalized for hairpin loop size or normalized for hairpin length (which is controlled by –l). Therefore, srnaloop scores are correlated with many other parameters (see Parameter Correlations below).

Stutter Filtering

In searching a sequence for hairpins of a certain length, srnaloop may find two or more hairpins on the same strand that overlap for a considerable percentage of their lengths, a phenomenon we called "stuttering." Stutter filtering consists of iteratively cycling through predicted hairpins on a strand-by-strand basis, detecting overlaps whose length exceeds a threshold fraction of the smaller of the two overlapping hairpin lengths, and eliminating the hairpins with the smaller srnaloop score.

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.

Parameter Correlations

Many parameters used to identify and filter candidate miRNA hairpin sequences are strongly correlated. For example, using the hairpin sequences associated with the miRNA test set (above), we found that srnaloop score and RNAfold-computed folding energy were correlated at −0.57, hairpin sequence length and srnaloop score were correlated at 0.77, hairpin sequence length and RNAfold-computed folding energy were correlated at −0.57, and GC content and RNAfold-computed folding energy at −0.49 (Pearson correlation coefficients).

Correspondence Determination

To find correspondences between hairpins in a predicted miRNA hairpin set S_A from species A in an assembled, repeat-masked genome sequence for species B, each sequence in S_A is BLASTed against the genome sequence for B using –W 8 and –e 100 (Altschul et al., 1990). BLAST hits less than 20 nt in length are ignored if their location in a query sequence in S_A is not entirely on one side of the sequence midpoint or the other, a heuristic filter whose purpose is to ensure that BLAST hits represent possible mature miRNA sequences in hairpin stems and not hairpin loop sequences. Repeat-masked sequence regions surrounding target BLAST hits in genome B are then extracted so that the target BLAST hit is in the same location in the extracted B sequence as the query hit is in the S_A sequence, plus up to 10 nt padding on either end. Overlapping B extracts were merged. Extracted B sequence is then analyzed for hairpins with srnaloop, and reverified for the presence of a BLAST target hit on the same side of the computed B hairpin sequence midpoint as the BLAST query hit in the A hairpin for the BLAST hit that generated the extract. Srnaloop parameters for extracted B sequence may be different from those that generated the initial S_A hairpins and always specify a single-stranded search. Hairpins from B that pass this consistency check are then filtered for GC content and folding energy and structure and comprise the set of B hairpins that “correspond” to S_A. Sets of corresponding hairpins may contain multiple instances of a given sequence if that sequence is duplicated in the A or B genome.

Transitivity Filter

In cases where a set of hairpins S_A from species A is used to find corresponding sequences S_B in species B, and then S_B used to find corresponding sequences S_C in species C, there is both a BLAST hit that establishes the correspondence between a hairpin sequence H_A in S_A and hairpin sequence H_B in S_B, and another BLAST hit that establishes the correspondence between H_B in S_B to hairpin sequence H_C in S_C. However, it may be the case that the target site in H_B for the $H_A \rightarrow H_B$ correspondence does not overlap the query site in H_B for the $H_B \rightarrow H_C$ correspondence. The transitivity filter looks for the subset of corresponding hairpins for which the target site in H_B for $H_A \rightarrow H_B$ overlaps the query site in H_B for $H_B \rightarrow H_C$. For the CDH set (see below), all *C. elegans*, *D. melanogaster*, and human hairpins were BLASTed against each other, and those hairpins with BLAST matches exceeding 14 nt and overlapping in sequence position were further assembled into groups, where at most 12 nt of nonoverlap was allowed between the *D. melanogaster* sequence that is the *C. elegans* hairpin target and that which is the human hairpin query.

Short Repeat Filtering

Although all assembled genomic sequences analyzed were repeat masked, many derived sets of hairpins contained mononucleotide strings or approximate tandem repeats of short words. In some cases we therefore filtered out hairpins containing 10 nt long mononucleotide strings, or tandem consecutive repeats of 2–4 bases up to lengths 12, 15, 16, respectively, allowing in each case one single base mismatch, deletion, insertion, and a possible insertion between each repeated block. None of the available cloned miRNA sequences in our set is rejected by this filter.

Structure Quality Filtering

As a final structure filtering step in generating predicted sets of miRNA hairpins as indicated below, we regenerated hairpin structures using mfold software (http://www.bioinfo.rpi.edu/~zukerm/rna/; Matthews et al., 1999; Zuker et al., 1999) and retained only those sequences with predicted hairpins containing no multiloops. These filters also ensured that the BLAST hit sequence in the hairpin establishing the correspondence or the hairpin sequence that matches the query mature miRNA sequence is entirely within a duplexed region characterized by the following topology limitations: no bulges of greater than 3 nt, no more than a 5 nt bulge on the opposite stem, and absence from the loop region.

Initial C. elegans *Hairpin Set (Set of 8713 Hairpins)*

A set of 16,216 intergenic regions was extracted from the repeat masked *C. elegans* genome and both strands analyzed by srnaloop using parameters including −w 4 −dw 1 − ~2w −t 23.5 −l 95 and −sm 0. In cases where multiple transcripts for a gene were annotated, intergenic regions were based on the smallest transcript. The 494,319 resulting hairpin sequences were successively filtered to meet the following criteria: GC content ≥32.8% and ≤62.5%, stutter filtration at 66% length overlaps, RNAfold-determined minimum free energy ≤−32.5 kcal/mol and no multiloops, to yield the 8713 set of *C. elegans* hairpins. To preserve genome locations of all hairpins, duplicate hairpin sequences were maintained in the 8713 set. As our test set of miRNA sequences (above) did not contain *C. elegans* miRNAs, we could not directly verify the presence of test set sequences in the 8713 set of hairpins except for *let-7* (*lin-4* could not be used to test the 8713 set because, being found within an intron [Lee et al., 1993], it was excluded by our use of only *C. elegans* intergenic sequences). However, similar parameters (but using a lower srnaloop score threshold) passed >54% of the hairpin sequences associated with the miRNA test set. We subsequently reassessed our filters as cloned *C. elegans* miRNAs were reported: in the current set of 61 hairpins for *C. elegans* miRNAs (see below), 39 (63.9%) are present in our repeat-masked *C. elegans* intergenic sequence and 29 (47.5%) are present among the 8713, so that 29/39 (74.4%) of all hairpins available in the sequence analyzed passed our filters. The 22 hairpins not present in our analyzed sequence are either in genic sequence (including introns) or in clone sequence not found in the *C. elegans* assembly. Of the 39 miRNA hairpins present in our sequence, 37 of these 39 passed the srnaloop score threshold, all 37 of these 37 passed GC content and stutter filtering, and 29 of these 37 passed the predicted minimum free energy and multiloop threshold. We repeated this assessment for the complete set of 61 hairpins where sequence for the 22 hairpins not in our analyzed sequence was taken from *C. elegans* clones and found that 46/61 (75.4%) of all hairpins identified for cloned *C. elegans* miRNAs would have passed our filters. Of the complete set of 61, 58 of these 61 passed our srnaloop score threshold, 54 of these 58 passed our GC content threshold, and 46 of these 54 passed the predicted minimum free energy and multiloop threshold.

Refined C. elegans *Hairpin Set (Set of 6086 Hairpins)*

For the homology-based miRNA hairpin prediction set, we further filtered the 8713 *C. elegans* hairpin set by removing 145 additional sequences that potentially overlapped coding regions (based on longer transcripts for genes with multiple annotated transcripts than first used to extract intergenic regions for the 8713 set), removing duplicates, applying short repeat filtering, and applying structure quality filtering. The result was a set of 6086 *C. elegans* hairpins.

Drosophila *Correspondences to the Initial* C. elegans *Set*

Correspondences in the repeat-masked *D. melanogaster* genome sequence were determined as described above based on 900,127 BLAST hits. *D. melanogaster* genomic sequence was extracted with 10 nt padding on each end. Srnaloop applied to extracted *D. melanogaster* genomic sequence used the –t 23 parameter instead of –t 23.5, resulting in 85,999 fly hairpin sequences. These were refiltered for srnaloop score of ≥24, filtered for GC content ≥32.8%, and ≤63.5%, "stutter"-filtered at 66% overlap, and filtered for RNAfold-determined minimum free energy of ≤–30 kcal/mol and no multiloops, resulting in 4778 *D. melanogaster* hairpins. Correspondence mapping with the original 8713 query *C. elegans* hairpins resulted in 3514 *D. melanogaster* and 3019 *C. elegans* hairpins, with both sets of hairpins containing duplicate sequences. Removing duplicates led to a set of 3505 distinct *D. melanogaster* and 2523 distinct *C. elegans* hairpins. Nine miRNAs from the test set were found in the predicted fly miRNAs, including *let-7*, *mir-1*, *mir-2a*, *mir-2b*, *mir-8*, *mir-9*, *mir-10*, *mir-13a*, and *mir-13b*.

*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}, and yielded a set of 95 hairpins. Thirteen of these hairpins had between 14 and 253 hits in the *C. briggsae* genome and were removed because of the possibility that they represented repeat sequences. Two of the remaining 82 hairpins were identical except for the presence of a single "N" in one hairpin sequence; they were deemed equivalent. Thus, the set of hairpins that showed correspondence between *C. elegans* and *D. melanogaster*, and had high homology between *C. elegans* and *C. briggsae* consisted of 81 distinct hairpins; this set of hairpins included six of the eight total known miRNAs that have been identified as conserved between *C. elegans* and *D. melanogaster* (Lagos-Quintana et al., 2001; Lau et al., 2001; Lee and Ambros, 2001). All 81 were tested by Northern blot. Subsequently, we selected 28 higher quality predictions (CDC-f set) on the basis of structure and sequence. The criteria used were later codified as short repeat and structure quality filtering.

*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 as described above and based on a set of 1,328,689 BLAST hits. Sequence extraction, srnaloop hairpin identification, and stutter-filtering, GC filtering, srnaloop score refiltering, and RNAfold-computed folding energy were performed using the same parameters as for the *D. melanogaster* correspondences on the 8713 *C. elegans* set except that the folding energy criterion originally used for the 8713 *C. elegans* hairpins was used (energy ≤ –32.52 kcal/mol). After BLAST hit correspondences with the 3505 *D. melanogaster* hairpin sequences were established and duplicates were eliminated, this left 6630 distinct human and 2246 distinct *D. melanogaster* hairpin sequences, which corresponded to 1729 distinct sequences from the original 8713 *C. elegans* hairpin set. Ten test miRNA sequences from human and fly, including five *let-7* variants, were found in the combined corresponding fly and human hairpin set. At this point, we applied transitivity filtering. Groups of hairpins corresponding transitively across all three species with greater than five *D. melanogaster* hairpins and/or greater than 12 human hairpins were considered to contain possible repeat sequence, and were removed. The groups were then filtered for *C.*

elegans hairpins whose *D. melanogaster* BLAST match region was duplexed according to the structure quality filtering described above, leaving 162 *C. elegans* hairpins. Corresponding *D. melanogaster* hairpins were then subjected to the same structure filter over the BLAST match region, yielding a set of 96 *C. elegans* hairpins where the matching region between the *C. elegans* and *D. melanogaster* hairpins adheres to the structural criteria in both species. Short repeat filtering, removal of possible coding sequence, and a further selection resulted in a set of 40 hairpins, six of which are published miRNAs (Lau et al., 2001; Lee and Ambros, 2001).

C. elegans *miRNA Homolog Set*

We used matcher, a pure Smith-Waterman algorithm, from the EMBOSS 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, using default settings except for gappenalty = 10 and gaplength = 1 (parameters which are permissive for gaps). Using matcher in place of BLAST allowed us to overcome the BLAST requirement for a matching word of at least 7 nt between sequences, a limitation when applied to very short miRNA sequences. Hairpins with perfect matches or matches where gaps account for 15% or more of the sequence were removed. The resulting matches with matcher algorithm scores greater than 60 were subjected to structure quality filtering (see above), yielding a set of 190 hairpins. A round of filtering to remove hairpins with homolog matches at the extreme ends of hairpins left a total of 116 candidate worm hairpin orthologs and paralogs of known miRNAs, of which three were identified by the CDC and CDH algorithms.

Clustering and Multiple Alignments of miRNAs

We used the matcher program to align each of 233 mature metazoan miRNA sequences (Lagos-Quintana et al., 2001, 2002; Lau et al., 2001; Lee and Ambros, 2001; Mourelatos et al., 2002) and our four novel miRNAs that were confirmed by Northern blotting against each other, using the same parameters as in the homology algorithm (default, except gappenalty = 10 and gaplength = 1). We generated a dissimilarity matrix from pairwise matcher alignment scores and performed hierarchical complete clustering. From the clustering, we drew a dendrogram in Matlab (Supplemental Figure S1 at http://arep.med.harvard.edu/miRNA/). We then cut the dendrogram to yield clusters that include the large *let-7* variant cluster and additional clusters we viewed as promising (Figure 4 and Supplemental Table S2 at http://arep.med.harvard.edu/miRNA/). The resulting clusters containing multiple miRNAs were then individually aligned using CLUSTAL W (Thompson et al., 1994) using an alignment gappenalty of 8 (permissive towards gaps). Each cluster was evaluated for whether the aligned subsequences were from similar regions of the constituent miRNAs. We adjusted by hand the composition of clusters to emphasize long contiguous blocks of aligned bases in common locations. This resulted in a set of ~40 clusters (Supplemental Table S2 at http://arep.med.harvard.edu/miRNA/).

Enumeration of Cloned C. elegans *miRNAs*

We compiled cloned *C. elegans* miRNA information (Lau et al., 2001; Lee and Ambros, 2001), identified duplicates, and checked source *C. elegans* genome sequence data. We determined that there were a total of 62 miRNAs and 61 distinct hairpins. We could not confirm a stable hairpin secondary structure for one reported miRNA (*mir-89*) and hence excluded it from these counts.

Analysis of Conservation of Predicted C. elegans *miRNAs*

We analyzed conservation of *C. elegans* miRNAs in two contexts. To determine which of the 61 hairpins for cloned *C. elegans* miRNAs (above) were conserved, we first found that nine were reported in the literature as having potential homologs in the *D. melanogaster*, *M. musculus*, or *H. sapiens* genomes. We then applied the method described above for finding *D. melanogaster* correspondences to the 61 *C. elegans* miRNAs, which is more sensitive than ordinary BLASTing, to find nine additional apparently conserved sequences. Therefore we count 18 miRNA hairpins of the 61 (~30%) as conserved. We also identified conserved sequences from our own predictions: We counted all sequences in the CDC-f and CDH sets as conserved by dint of correspondences with *D. melanogaster* and/or *H. sapiens*, but only the 63 out of 116 sequences from our homology set that were found to correspond to a *D. melanogaster*, *M. musculus*, or *H. sapiens* miRNA. After removing duplicates we were left with 119 conserved predicted miRNA hairpins.

Screening of Candidate Sequences and PCR Primers against Noncoding RNA Sequence

Our first concern was to assure that our candidate miRNAs were not contained within or significantly overlapped other kinds of noncoding RNA sequences. To eliminate this possibility, we performed a BLAST with default parameters of 222 candidate miRNA sequences against the 888 noncoding RNA sequences and considered those same sense matches which were either ≥16 nt in

DOI: 10.1016/S1097276503001539

length or which overlapped an end of the candidate sequence. Of 29 such matches, nine were full-length matches between miRNAs that were in both files. Of the remaining 20 matches involving nine candidate sequences, only one candidate miRNA had matches meeting both criteria: chr_V-CIG13352.819.2 had a 51 nt overlap including its 3′ end with two putative tRNA sequences. Therefore, at most one of our candidate miRNAs potentially represents a known or suspected noncoding RNA sequence of another type.

The PCR assay for detection of candidate miRNAs uses a primer that is complementary to the entire sequence, or a large subsequence, of the predicted miRNA. We wanted to control for the possibility that such a primer might generate a spurious PCR product by partially priming against another type of noncoding RNA, or previously cloned and similar miRNA. We therefore BLASTed all primers in Table 1 against the 888 noncoding RNA sequences with parameters –W 7 and –e 100, and considered those antisense matches that contained the 3′ end of the primer. We concluded that our candidates whose primers generated a PCR product were unlikely to have generated this product adventitiously based on the following observations:

(1) Nine primers that did not generate PCR products, including four that were designed against mRNA sequence and tested as negative controls, had matches against miRNA sequences of 7 or 8 nt that included the primer 3′ end. We conclude from this that a 7–8 nt match of a primer 3′ end is insufficient to generate an adventitious PCR product against sequences that are known to be present in libraries prepared for miRNA cloning. In one of these primers, the 8 nt exact match containing the primer 3′ end was part of a 14 nt imperfect match against an miRNA sequence (mir-88), suggesting that longer exact 3′ end matches may also be insufficient to generate a PCR product. We found a similar case in a comparison of the sequences of primers that generated PCR products in our amplified library against those that did not generate PCR products: The primer for predicted miRNA candidate-111 was identical for 19 nt (including its 3′ end) with that for *cp-mir-270* except for two mismatches at positions 6 and 9 relative to its 3′ end.

(2) Examining primers for candidates that generated PCR products and excluding all the primers corresponding to previously cloned miRNAs that were included as positive controls, we found 11 candidate miRNA sequences that did not have primer 3′ end matches of longer than 9 nt with any sequence in the noncoding RNA file. By (1) we concluded that these primers were unlikely to have generated products due to adventitious priming. These candidates have been designated *mir-236*, *mir-228*, and *cp-mir-264* to *cp-mir-273*. We note that of these sequences, only *cp-mir-267* had a 9 nt match at the 3′ end, and the matching noncoding RNA sequence was identified as F09E10.10 (“probable noncoding RNA”).

Published Online April 23, 2003

DOI: 10.1016/S1097276503001539