Letter

Repertoire and evolution of miRNA genes in four divergent nematode species

Elzo de Wit,[1,3] Sam E.V. Linsen,[1,3] Edwin Cuppen,[1,2,4] and Eugene Berezikov[1,2,4]

[1]*Hubrecht Institute-KNAW and University Medical Center Utrecht, Cancer Genomics Center, Utrecht 3584 CT, The Netherlands;*
[2]*InteRNA Genomics B.V., Bilthoven 3723 MB, The Netherlands*

miRNAs are ~22-nt RNA molecules that play important roles in post-transcriptional regulation. We have performed small RNA sequencing in the nematodes *Caenorhabditis elegans*, *C. briggsae*, *C. remanei*, and *Pristionchus pacificus*, which have diverged up to 400 million years ago, to establish the repertoire and evolutionary dynamics of miRNAs in these species. In addition to previously known miRNA genes from *C. elegans* and *C. briggsae* we demonstrate expression of many of their homologs in *C. remanei* and *P. pacificus*, and identified in total more than 100 novel expressed miRNA genes, the majority of which belong to *P. pacificus*. Interestingly, more than half of all identified miRNA genes are conserved at the seed level in all four nematode species, whereas only a few miRNAs appear to be species specific. In our compendium of miRNAs we observed evidence for known mechanisms of miRNA evolution including antisense transcription and arm switching, as well as miRNA family expansion through gene duplication. In addition, we identified a novel mode of miRNA evolution, termed "hairpin shifting," in which an alternative hairpin is formed with up- or downstream sequences, leading to shifting of the hairpin and creation of novel miRNA* species. Finally, we identified 21U-RNAs in all four nematodes, including *P. pacificus*, where the upstream 21U-RNA motif is more diverged. The identification and systematic analysis of small RNA repertoire in four nematode species described here provides a valuable resource for understanding the evolutionary dynamics of miRNA-mediated gene regulation.

[Supplemental material is available online at http://www.genome.org. The sequencing data from this study have been submitted to Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) under accession no. GSE15169. The novel miRNA sequences identified in this study have been submitted to miRBase under accession nos. MI0010974–MI0011250.]

miRNAs are small ~22-nucleotide noncoding RNAs that function as post-transcriptional regulators in plants and animals. They are implicated to play roles in diverse processes ranging from development and tumorigenesis to neurological and cardiac diseases (Carrington and Ambros 2003; Croce and Calin 2005; Bilen et al. 2006; Plasterk 2006; Thum et al. 2007; Stefani and Slack 2008). miRNA silencing is effectuated by Watson–Crick base pairing of the miRNA to the 3′ untranslated region (UTR) of an mRNA, thereby preventing the onset of translation, causing degradation of the messenger RNA and potential inhibition of translational elongation (Filipowicz et al. 2008; Gu et al. 2009).

miRNA genes are mainly transcribed by RNA polymerase II, producing a primary (pri)-miRNA that is capped, polyadenylated, and spliced (Lee et al. 2002, 2004). The processing of miRNA genes depends on their secondary structure, and the stable hairpin structure of the pre-miRNA is the clearest characteristic of miRNA transcripts. The majority of miRNA genes are processed in the nucleus by the microprocessor complex consisting of the proteins RNASEN (also known as Drosha) and DGCR8 (Lee et al. 2003; Gregory et al. 2004; Han et al. 2004). The 50–80-nt Drosha product (pre-miRNA) is exported to the cytoplasm and subsequently processed by the enzyme Dicer (Bernstein et al. 2001; Lee et al. 2002). A minority of the miRNAs, called mirtrons, do not require Drosha processing and instead are produced from intronic hairpins that are formed after splicing of protein-coding genes (Berezikov et al. 2007; Ruby et al. 2007a).

[3]These authors contributed equally to this work.
[4]Corresponding authors.
E-mail e.cuppen@hubrecht.eu; fax 31-30-2516554.
E-mail e.berezikov@hubrecht.eu; fax 31-30-2516554.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.093781.109.

After Dicer processing, one of the two arms of the miRNA duplex is loaded into the RNA induced silencing complex (RISC). Incorporation of either of the strands depends on the thermodynamic stability of the 5′ region of the miRNA in the duplex (Khvorova et al. 2003; Schwarz et al. 2003). The nonincorporated miRNA* species is subsequently degraded. Targeting of miRNAs to the 3′ UTR of mRNAs depends on RISC, which mediates the imperfect base pairing of miRNAs to mRNAs (Lai 2002; Filipowicz 2005). Nucleotides 2–7 of mature miRNAs, known as the "seed region," are the primary determinants of miRNA target recognition (Lewis et al. 2003; Bartel 2009), and the target sites in 3′ UTRs are often conserved between species (Lewis et al. 2003; Stark et al. 2003) and lie in thermodynamically open structures (Kertesz et al. 2007; Hammell et al. 2008).

miRNAs are highly conserved throughout the animal kingdom (Wheeler et al. 2009), with miR-100 as an extreme example being conserved among eumetazoa (Grimson et al. 2008). Computational and cloning efforts have boosted the discovery of novel miRNAs (Berezikov et al. 2006a). Since the number of predicted hairpins from genomic sequences greatly exceeds the number of true miRNAs (Pervouchine et al. 2003; Bentwich 2005), a hairpin alone is not sufficient for the identification of a novel miRNA. Identification of novel miRNA has focused mainly on two methods. First, the power of comparative genomics data has been used to identify novel miRNA genes (Berezikov et al. 2005; Xie et al. 2005; Stark et al. 2007) due to the higher level of conservation of miRNA hairpins compared to non-miRNA hairpins. Second, the advent of high-throughput sequencing methods has facilitated the rapid cloning and sequencing of large numbers of small RNAs. Combining both methods allows for the most robust identification of miRNAs.

The first miRNAs were discovered by forward genetics approaches in the nematode *Caenorhabditis elegans* (Lee et al. 1993).

19:2064–2074 © 2009 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/09; www.genome.org

Since then, the availability of a reference genome in combination with experimental and computational approaches has resulted in the description of 155 miRNAs in this species (Griffiths-Jones et al. 2008). Since recently, the *C. briggsae* and *C. remanei* genomes are available as well (Kiontke and Fitch 2005), allowing for similar approaches. Within this group, *C. briggsae* and *C. remanei* are sister species and *C. elegans* represents an outgroup. The evolutionary distance between *C. elegans* and *C. briggsae* and *C. remanei* is estimated to be about 100 million years (Myr) (Stein et al. 2003). While *C. remanei* reproduces gonochoristically, both *C. briggsae* and *C. elegans* reproduce in a hermaphroditic fashion, which likely evolved independently in both species (Kiontke et al. 2004). The genome size of *C. elegans* and *C. briggsae* is ~100 Mb, whereas the genome of *C. remanei* is ~140 Mb. The natural habitat for these species is compost or garden soil, where they feed on microorganisms (Kiontke and Sudhaus 2006). A fourth species for which genomic sequence information is available is *Pristionchus pacificus*, a necromenic nematode, which had a last common ancestor with the *Caenorhabditis* species 280–430 Myr ago (Dieterich et al. 2008). Although genomic contigs are not completely assembled yet, its genome size is considerably larger than that of *C. elegans*.

In this study we have sequenced and analyzed small RNAs from these four species. We found that within the *Caenorhabditis* genus the majority of miRNAs are conserved between species, while *P. pacificus* miRNAs show a decreased level of conservation. In addition, we show that between species, a change of the dominantly expressed miRNA arm, a process known as arm switching (Liu et al. 2008), is a common mechanism. Antisense transcription of miRNA loci (Ruby et al. 2007b) is observed for three miRNAs in our data set, and this process could lead to novel miRNAs as well. Finally, we have identified a new mechanism of formation of novel miRNAs, called hairpin shifting, in which a novel up- or downstream hairpin relative to the mature miRNA is formed, leading to shifting of the miRNA from the 5′ arm to the 3′ arm, or vice versa. This process of hairpin shifting creates a novel miRNA* species, which opens a way for formation of novel miRNA sequences via the process of arm switching during subsequent evolution. Our results provide a valuable resource for understanding of miRNA evolution as well as for studying miRNA-based regulatory mechanisms.

Results

Sequencing of small RNAs from four nematode species

Recent sequencing of the genomes of several nematode species provides the basis for systematic investigation of evolutionary processes occurring over long evolutionary periods within the nematode clade. We have used an established small RNA cloning approach (Berezikov et al. 2006a) and 454 Life Sciences (Roche) sequencing technology to investigate the composition of small RNAs in three *Caenorhabditis* species (*C. elegans*, *C. briggsae*, and *C. remanei*) and in the necromenic nematode *Pristionchus pacificus*. Approximately 160,000, 41,000, 32,000, and 29,000 alignable small RNA reads were generated for the respective species (Table 1). Primary analysis of the obtained sequencing data was performed by the miR-Intess software (see Methods for details), which provides annotation of small RNA reads as well as identification of novel miRNA candidate genes. In the identification process, the software distinguishes between confident and candidate miRNA genes (see Methods for details on the classification).

Table 1. Composition of the small RNA libraries

	No. of reads			
Category	*C. elegans*	*C. briggsae*	*C. remanei*	*P. pacificus*
miRBase miRNAs	143,092	30,247[a]	24,839[a]	17,382[a]
Novel miRNAs	6	2203	2978	4781
Candidate miRNAs	59	5918	317	2119
rRNA	7335	829	0[b]	856
tRNA	1405	231	0[b]	0[b]
Other ncRNA	218	0[b]	0[b]	0[b]
siRNA	3291	402	253	19
senseRNA	1095	127	111	13
Repeats	708	192	0[b]	86
21U	1609	0[b]	0[b]	0[b]
Rest	1454	1729	3945	4637
Total	160,272	41,878	32,443	29,893

[a]Including clear homologs of *C. elegans* miRBase miRNAs.
[b]Based on available genome annotations.

In all four species the largest fraction of small RNA reads corresponds to known miRBase miRNAs (or clear miRBase homologs in the case of *C. remanei* and *P. pacificus*) and varies from 89% in *C. elegans* and 74% in *C. briggsae* to 77% in *C. remanei* and 58% in *P. pacificus*. In addition, 5%, 9%, and 16% of the small RNA reads in *C. briggsae*, *C. remanei*, and *P. pacificus*, respectively, correspond to confident novel miRNAs (see below), bringing the total fraction of miRNA reads in all libraries close to or above 80% of the reads. Beside miRNAs, small RNAs corresponding to rRNA, tRNA, snoRNA, repeats, 21U, and sense—as well as antisense—to protein-coding regions, were observed. However, the fraction of reads corresponding to these categories does not exceed several percent (max 5% for rRNA in *C. elegans*). Since we rely on genome annotations for the classification of small RNAs to various non-miRNA categories, some of the reads corresponding to these categories might not be recognized as such in species other than *C. elegans* due to incomplete genome annotations. However, there is no substantial influence on the analysis of the miRNA fraction, as only a small percentage of the reads are affected by these potential gaps in the annotations.

Identification of known and novel miRNA genes

The miRNA repertoire of *C. elegans* has been extensively studied in a number of (deep)-sequencing efforts (Lau et al. 2001; Lee and Ambros 2001; Ruby et al. 2006), and currently, 155 miRNA loci are reported for *C. elegans* in miRBase (v.12). We have found experimental support for 112 of these miRBase miRNAs in our sequencing data (Supplemental Table 1). Some of the *C. elegans* miRBase miRNAs missing in our study could be explained by the limited coverage depth of our sequencing data (~160,000 reads). However, 14 miRNAs that were not picked up in our sequencing effort (i.e., miR-256, miR-261, miR-264–miR-273, miR-356, miR-1021), were also missed by a much deeper sequencing effort (Batista et al. 2008) (GSE11734, ~20 million aligned reads) and some of them (*mir-261*, *mir-267*, *mir-271*) do not show detectable promoter activity (Martinez et al. 2008), and thus, likely do not represent genuine miRNAs. Notably, increasing sequencing coverage more than 100-fold allows identification of only 17 additional miRBase miRNAs (GSE11734, data not shown), indicating that our data is sufficient to capture the vast majority of bona fide miRNAs and gives us confidence that a comprehensive indexation of miRNAs can be made in *C. briggsae*, *C. remanei*, and *P. pacificus* based on these data.

Table 2. miRNA component of the small RNA libraries

Category	No. of cloned loci			
	C. elegans	*C. briggsae*	*C. remanei*	*P. pacificus*
miRBase miRNAs	112	87	0	0
Homologs of miRBase miRNAs	0	28	99	33
Novel miRNAs	1	15	10	91
Candidate miRNAs	12	22	40	96
Noncloned miRBase miRNAs	45	11	0	0
No. of families[a]	21	26	26	25
No. of clusters[b]	10	14	9	30

[a]miRNAs with the same seed sequence.
[b]miRNAs located within 1 kb from each other in the genome.

Since *C. elegans* small RNAs have already been sequenced to extensive depth in a number of previous studies (Lau et al. 2001; Lee and Ambros 2001; Ruby et al. 2006), in our data set we identified only one novel confident miRNA gene and 12 miRNA candidates (Table 2). At the same time, in the less-studied worm *C. briggsae* we identified 87 miRNAs present in miRBase, 28 miRNAs not annotated in miRBase but with clear homology with *C. elegans* miRBase miRNAs, 15 confident novel miRNA genes, and 22 miRNA candidates. For *C. remanei* and *P. pacificus* there are no annotated miRNA sequences present in the current version of miRBase (v.12). Our analysis revealed 99 and 33 clear miRBase homologs that are expressed in *C. remanei* and *P. pacificus*, respectively. In addition, 10 and 91 novel miRNAs and 40 and 96 miRNA candidates were identified in these species, respectively (Table 2; Supplemental Tables 2–4). To validate the sequencing results we have performed Northern blot analysis on 19 miRNAs candidates that were selected randomly from all species (Supplemental Table 15), with the exception of *C. elegans*. We observed distinct bands around 22 nt in 17 cases (Supplemental Fig. 1), indicating that 89% could be validated by a sequencing-independent method.

miRNA conservation patterns

Generally, the most strongly conserved part of a miRNA corresponds to its seed region (nucleotides 2–7 of mature miRNA), which is the dominant determinant in recognition of miRNA targets (Bartel 2009). To identify conservation levels of all identified miRNAs regardless of their small RNA-cloning evidence in different species, we performed genome-wide searches requiring conservation of the seed sequence in addition to the presence of a hairpin with an RNAforester score above 0.3 (see Methods). Using these criteria, the largest fraction of miRNAs—115 miRBase and 17 novel miRNAs—are conserved at the seed-sequence level in all four examined species (Fig. 1), although only 15 miRBase miRNAs are cloned in all four species (Supplemental Tables 5,6). Applying the same homology-searching approach to randomly selected *C. elegans* hairpins that possess all miRNA-like properties yielded two times less conserved loci between *C. elegans* and *P. pacificus* compared with genuine miRNAs, further demonstrating the specificity of using seed conservation for identification of miRNA homologs. Plotting the average miRNA conservation profiles between *C. elegans* and other species separately for miRNAs with 5′- and 3′-arm mature sequences reveals the characteristic bimodal distribution of miRNA conservation profiles and the seed conservation tendency, which is most prominent in *C. elegans*/*P. pacificus* comparisons (Supplemental Fig. 2). While conservation of seed sequences and hairpin structures suggests that our computational approach identifies genuine miRNA homologs in other species, experimental evidence for expression of mature miRNAs from these homologs will be required for definitive miRNA classification.

The number of miRNAs that appear to be species specific is relatively small in the *Caenorhabditis* species (seven in *C. elegans*, eight in *C. briggsae*, and one in *C. remanei*), but it is more substantial in *P. pacificus* (30 miRNAs), which is in accordance with greater evolutionary divergence of this species. Species-specific miRNAs have most likely emerged in the specific lineages rather than lost in all other lineages, which allows estimation of miRNA gene fixation rate. Given the estimated 80–110 Myr divergence time between *C. elegans* and *C. briggsae* (Hillier et al. 2007) and

A

conservation	Number of reads (miRBase / Novel)				Number of loci	
	C. elegans	*C. briggsae*	*C. remanei*	*P. pacificus*	miRBase	Novel
CBRI_CREM		83/715	8/1020		2	7
CBRI_CREM_PPAC		0/7	4/629	206/388	2	8
CBRI_ONLY		73/1223			2	6
CBRI_PPAC		0/30		11/133	1	5
CEL_CBRI	32/0	28/0			3	0
CEL_CBRI_CREM	22363/0	11324/202	9814/712		29	2
CEL_CBRI_CREM_PPAC	115138/6	23826/28	19839/5	14128/1366	115	17
CEL_CBRI_PPAC	670/0	10/6		0/80	4	6
CEL_CREM	24/0				2	0
CEL_CREM_PPAC	482/0		192/0	4778/272	7	6
CEL_ONLY	5341/0				7	0
CEL_PPAC	16/0			1817/167	3	5
CREM_ONLY			0/8		0	1
CREM_PPAC			0/91	0/967	0	21
PPAC_ONLY				0/1914	0	30

B

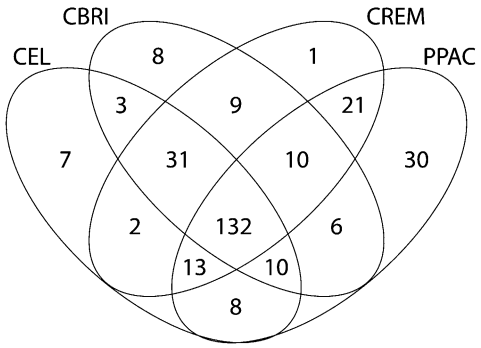

Figure 1. miRNA conservation and cloning frequencies in four nematodes. (*A*) miRNAs were divided based on whether they were conserved between species (i.e., CEL_CBRI_CREM means conserved between *C. elegans*, *C. briggsae*, and *C. remanei*, but not in *P. pacificus*). Columns show the number of reads for these loci, stratified by whether they map to known miRBase or novel miRNAs. Last two columns indicate how many loci are either known miRBase miRNAs or novel miRNAs. (*B*) Venn diagram for the number of miRNA loci in different conservation groups. (CEL) *C. elegans*; (CBRI) *C. briggsae*; (CREM) *C. remanei*; (PPAC) *P. pacificus*.

280–430 Myr between *C. elegans* and *P. pacificus* (Dieterich et al. 2008), the miRNA birth rate appears to be similar in the different nematode branches and is in the range of one miRNA born per 9–16 Myr. This is about 10 times lower than observed for the distantly related *Drosophila* species (estimated one gene per 1 Myr) (Lu et al. 2008). Estimates on the rates of miRNA gene birth should be taken with some care, as miRNA discovery may still not be fully exhausted in nematodes. For example, in our data set there are 12 *C. elegans* miRNA candidates that were not taken into account for calculating the birth rate because of their limited experimental support, but many of them could be genuine miRNAs. Indeed, our reanalysis of the large *C. elegans* small RNA data set GSE11734 (Batista et al. 2008) identified more than 30 additional confident miRNAs in *C. elegans*, 20 of which appear to be species specific (data not shown), supporting the notion that there are numerous nonconserved miRNAs expressed at low levels that require large sequencing depth for their identification (Berezikov et al. 2006b). However, even when considering these additional novel nonconserved candidate miRNAs in *C. elegans* miRNAs, the miRNA birth rate in the nematodes would still appear to be three- to fourfold lower than in the *Drosophila* species.

Comparison of miRNA expression levels between species is complicated by the fact that our small RNA datasets were generated from mixed-stage populations, which may introduce substantial biases. However, we do see a good correlation between expression levels for high- and medium-expressed miRNAs, with miRNAs highly expressed in one species also expressed at high levels in other species (Supplemental Fig. 3). Similar to previous studies (Berezikov et al. 2006b), we also observe that less-conserved miRNAs tend to be expressed at lower levels.

Extensive analysis of miRNA expression patterns in *C. elegans* based on promoter reporters has become available recently (Martinez et al. 2008), allowing comparison of mRNA conservation levels with their tissue specificity. We have used information on miRNA conservation between *C. elegans* and *P. pacificus* to test whether there are any biases toward preferential expression of conserved or nonconserved miRNAs in certain *C. elegans* tissues, but have not found any statistically significant differences (data not shown).

3′ Arm bias in nematode miRNAs

After hairpin processing by Dicer, a small RNA from either the 5′ or the 3′ arm of the hairpin is loaded into RISC. Which of the two arms ultimately supplies the mature miRNA depends on the thermodynamic properties of the intermediate RNA duplex. The strand with the lowest 5′ stability in the duplex is incorporated into RISC (Khvorova et al. 2003; Schwarz et al. 2003). We found that in all four investigated nematode species there are more miRNA genes with the mature miRNA located in the 3′ arm of the hairpin (2.1–3.3-fold increase, Fig. 2A).

We extended the analysis of arm bias beyond the nematode clade by calculating the arm bias for species that have more than 50 entries in miRBase (v.12.0). We found that in *Drosophila melanogaster* and *D. pseudoobscura*, and in plants, there is also a bias for the mature miRNA originating from the 3′ arm (Supplemental Fig. 4). In vertebrates, however, the majority of the species show 5′ arm bias. These results suggest the trend in arm bias is similar between related species but can vary considerably in more distant species, which could be indicative of subtle changes in miRNA processing.

To further investigate the differences between 5′ dominant and 3′ dominant miRNA hairpins we separately calculated thermodynamic profiles for miRNAs originating from different arms of

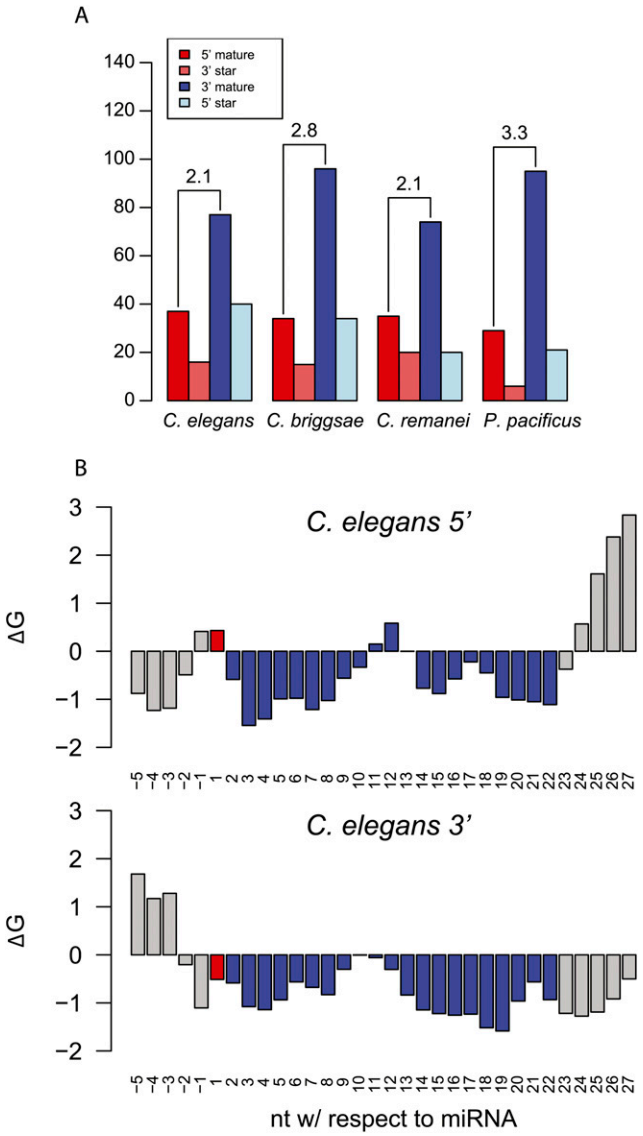

Figure 2. Arm bias and thermodynamic profiles. (*A*) Barplot showing the arm of origin for the mature and miRNA* sequences for all four species. Grouping of 5′ mature miRNA and 3′ miRNA* species and vice versa, reflects that they originate from the same precursor. Numbers *above* the bars indicate the fold enrichment of 3′ mature over 5′ mature. (*B*) Thermodynamic profiles for mature miRNAs originating from either the 3′ or 5′ arm. Secondary structures were calculated with mfold (Zuker 2003) and ΔG values extracted for single nucleotides. Positive ΔG values denote instability in the miRNA hairpin. The red bar identifies the starting nucleotide of the miRNA, the blue bars the rest of the canonical 22-nt miRNA.

the hairpins using mfold (Zuker 2003). We averaged the profiles of 5′ dominant and 3′ dominant miRNAs for each species. In Figure 2B we show the thermodynamic profiles for *C. elegans* miRNA precursors, which is representative for all species (Supplemental Fig. 5). Both 5′ dominant and 3′ dominant miRNA hairpins show high stability at the 3′ end of the mature miRNA and have a region of lower stability overlapping with the middle of the miRNA. Overall, miRNAs originating from the 5′ arm have a more strongly defined thermodynamic profile, and the first nucleotide in the miRNA overlaps with a very unstable region in the hairpin, in concordance with the mechanism put forward by Khvorova et al.

(2003). Surprisingly, miRNAs from the 3′ arm do not show a strong instability at the first nucleotide, suggesting that additional, yet undetermined factors, might contribute to the selection of mature miRNA strand.

miRNA evolution mechanisms

It has recently been demonstrated that miRNA* species can be physiologically relevant. In *Drosophila* a substantial number of miRNA* species are well conserved, can be loaded into RISC, and can regulate expression of the target genes (Okamura et al. 2008). It has been suggested that acquisition of biological function by miRNA* species is a way of evolutionary diversification of miRNAs, and that it preferentially acts on duplicated miRNA genes. A number of miRNA families that have undergone transition from miRNA* to mature sequences (arm switching) or are still in the process of switching, have been identified (Liu et al. 2008). In our nematode small RNA data set there are 172 miRNAs for which both mature and star sequences have been cloned, allowing investigation of miRNA:miRNA* ratios. We have found several examples that support the arm-switching model of miRNA evolution (Supplemental Table 7). Specifically, for *mir-246* the 3′ arm is dominant in *C. elegans* (5′/3′ read ratio: 3/22) and *C. briggsae* (5′/3′ read ratio: 2/10), but in *C. remanei* there are more reads cloned from the 5′ arm (Fig. 3; 5′/3′ read ratio: 19/4). For these miRNA loci, the miRNAs from the 5′ arm belong to one family and the miRNAs from the 3′ arm belong to one family. Notably, there are two copies of the *mir-246* gene present in the current *C. remanei* genome assembly, whereas only one copy is present in *C. elegans* and *C. briggsae* genomes. This observation supports the hypothesis that miRNA gene duplication might be a prerequisite for arm switching.

Expansion of the miRNA repertoire through gene duplication is a known mechanism, e.g., the *mir430* family has more than 70 members in zebrafish, and the *MIRLET7* (*let-7*) family consists of 11 members in human (Griffiths-Jones et al. 2008). One of the most extreme examples of such expansion events observed in our nematode small RNA data set is the miRNA family defined by the seed sequence [UA]AUGACA. In the *Caenorhabditis* genus this family has only five copies (*mir-63–mir-66* and *mir-229* in *C. elegans*), but in *P. pacificus* it has expanded to encompass 16 members in multiple clusters (Supplemental Fig. 6). This once more consolidates the importance of miRNA gene duplication in the evolution of miRNAs.

The above-described mechanisms of miRNA evolution (duplication and arm switching) can result in the generation of novel miRNAs, with the retained miRNA always residing on the original arm. Intriguingly, however, upon studying miRNA families, we identified family members that originated from different arms of the hairpins (Fig. 4). We therefore studied all mature miRNAs that differ only in 3 or 4 nucleotides of sequence. We found 15 families in which the member miRNAs do not originate from the same arm of the hairpin (Supplemental Table 8). This would suggest a mechanism whereby an up- or downstream genomic region of the original hairpin mutates so that a novel hairpin can be formed (Fig. 4). This process of "hairpin shifting" would lead to the mature miRNA being expressed from the other arm of the hairpin and, more importantly, the emergence of a novel miRNA* sequence, which could ultimately give rise to a novel miRNA via the mechanisms discussed above. We found cloning evidence for formation of such novel miRNA* sequences in eight of the 15 families (Supplemental Table 8). Although a novel miRNA could spuriously gain the same seed as an already known miRNA family, the high degree of conservation between certain hairpin shifted miRNAs and the sheer number of miRNAs that show hairpin shifting argues against a coincidental event. Rather, it suggests a novel mechanism of miRNA evolution.

Another further way of evolving new miRNA function is so-called seed shifting, in which the mature sequence of a given miRNA is moved one or several nucleotides relative to its original position. Seed shifting has been recently observed when

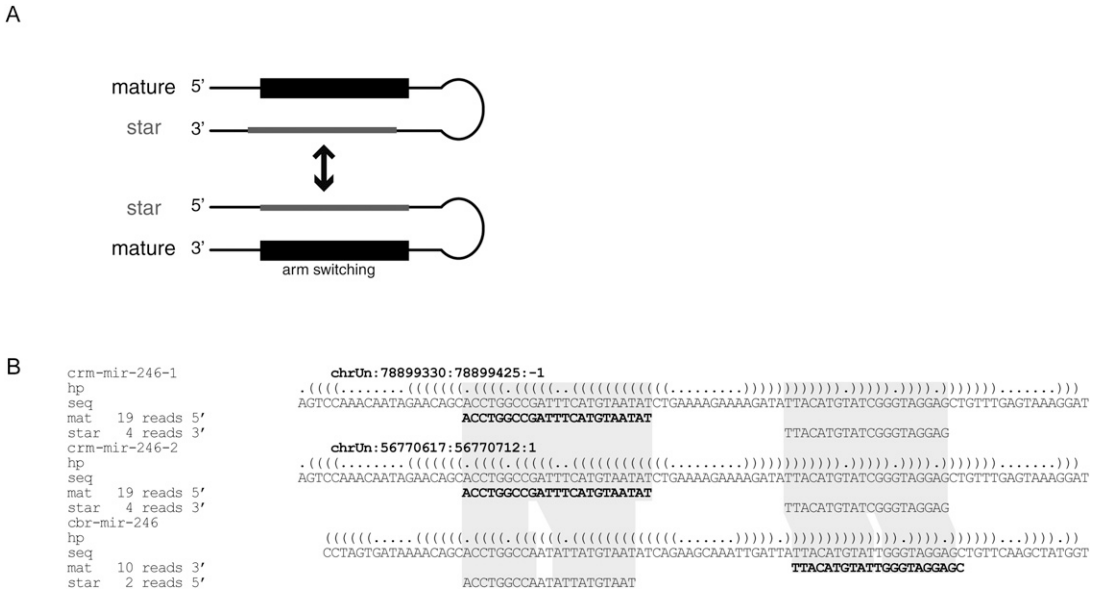

Figure 3. Arm-switching model of miRNA evolution. (*A*) Schematic representation of arm switching. In one miRNA hairpin, the 3′ arm provides the mature miRNA. In a slightly modified hairpin, the mature miRNA is the one that comes from the 5′ arm. (*B*) An example of arm switching. According to our cloning data, the duplicated loci of *mir-246* in *C. remanei* express the 5′ arm as mature miRNAs, whereas this locus in *C. briggsae* expresses the 3′ arm as mature miRNA.

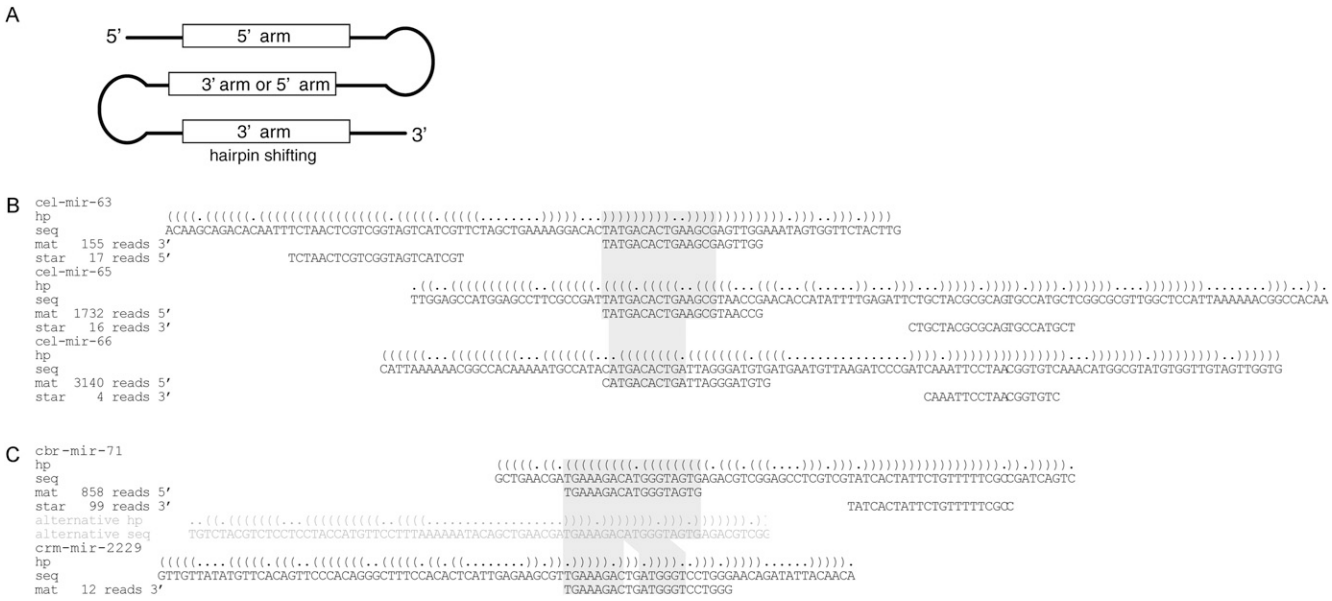

Figure 4. Hairpin shifting model of miRNA evolution. (*A*) Schematic representation of hairpin shifting. If the hairpin shifts in the 3′ direction, the 3′ arm of the existing hairpin becomes the 5′ arm of the new hairpin. With this new assembly, the original 5′ arm is lost and potentially a new 3′ arm miRNA is born. A similar mechanism shifts the hairpin in the 5′ direction. (*B*) In *C. elegans*, the first 5′ 15-nt miR-63 and miR-65 are identical; however, miR-63 originates from the 3′ arm, whereas miR-65 originates from the 5′ arm. Further divergence may have led to the birth of *mir-66* and miRNA stars that are also detected from the *mir-65/66* loci. Gray planes highlight identical sequence. (C) Intermediates of shifted hairpins are present alongside miRNA loci. *C. briggsae cbr-mir-71* and *C. remanei crm-mir-2229* are highly similar, but arise from different arms. Their common ancestry is confirmed by the observation that the 5′ flank as well as the 3′ flank of *C. briggsae mir-71* can fold into a stable hairpin, as indicated by "alternative seq" and "alternative hp."

comparing miRNAs from greatly diverged species (Grimson et al. 2008; Wheeler et al. 2009). Many miRNA genes in our data set have reads with varying 5′ start positions, as is customary for small RNA deep-sequencing data. However, when cloned orthologous miRNAs were compared with known miRNAs from *C. elegans*, the most abundant read sequence had the same 5′ start sites in all species except for one case. However, there are eight *C. elegans* miRBase miRNAs that can be grouped with the most abundant read sequence from our data set based on shifted seeds (nucleotides 2–8) (Supplemental Table 13), indicating that duplication precedes seed divergence during evolution of miRNA genes.

In *Drosophila* it has been shown that miRNA loci can be transcribed from both strands (Ruby et al. 2007b), giving rise to two independent miRNAs. In our small RNA set we found that the *mir-66* locus in *C. elegans* produced unique miRNAs from both strands. In addition, we find that the homolog of *cbr-mir-356* in *C. remanei* is cloned from the antisense strand and not from its sense strand. These two examples show the existence of antisense transcription of miRNA genes in nematodes and represent an additional mechanism that may contribute to the genesis and evolution of miRNAs.

Evolution of 21U-RNAs

21U-RNAs are a class of small RNAs (Ruby et al. 2006) that represent piRNAs in *C. elegans* (Batista et al. 2008; Das et al. 2008; Wang and Reinke 2008). The vast majority of these small RNAs starts with a uracil and are 21-nt long. In *C. elegans* the 21U-RNA genes are predominantly located on chromosome IV and are characterized by a specific upstream motif (Ruby et al. 2006). We used our small RNA data to independently derive this motif and applied it for annotation of 21U-RNAs in our small RNA reads (see Methods). We have identified 705 21U-RNAs in *C. elegans* (Supplemental Table 9), 699 of which were described previously (Ruby et al. 2006; Batista et al. 2008). By applying the same methodology on the small RNA datasets from the other nematodes, we found that the motif is also present upstream of a number of small RNAs in these species, confirming earlier predictions of the presence of 21U-RNAs in other nematode species. In *C. briggsae*, *C. remanei*, and *P. pacificus* we have identified 250, 1314, and 1123 21U-RNAs, respectively (Supplemental Tables 10–12). In agreement with genomic comparisons described by Ruby et al. (2006), none of these small RNAs were conserved across these four species. The 21U-RNA motif shows only subtle differences between the *Caenorhabditis* species, suggesting a high level of conservation (Fig. 5A). However, in *P. pacificus* the motif is more strongly diverged, with only conservation in the GTTTCA core motif and the upstream location of the motif (data not shown), hinting at mechanistic conservation.

A striking feature of *C. elegans* 21U-RNA genes is their genomic distribution in two megabase-sized clusters on chromosome IV. When we study the 21U-RNA motif distribution in *C. briggsae* we also found these clusters on chromosome IV. In addition, however, we also found large clusters on the telomeric side of chromosome III (0–0.3 Mb) and on chromosome I (7.8–9.5 Mb, Supplemental Fig. 7). In *C. remanei* we found 46 contigs that are significantly enriched for 21U-RNA motifs. To study genomic conservation of 21U-RNA clusters, we determined whether regions on large *C. remanei* contigs (>100 kb) that are enriched for 21U-RNA motifs were syntenic with 21U-RNA clusters in *C. elegans* and *C. briggsae*. Although we found considerable rearrangement within the 21U-RNA clusters, when we compared *C. remanei* and *C. elegans*, we found that for 85% of the contigs that are enriched for 21U-RNA motifs the amount of sequence that is syntenic with a 21U-RNA cluster exceeds the background frequency. Between *C.*

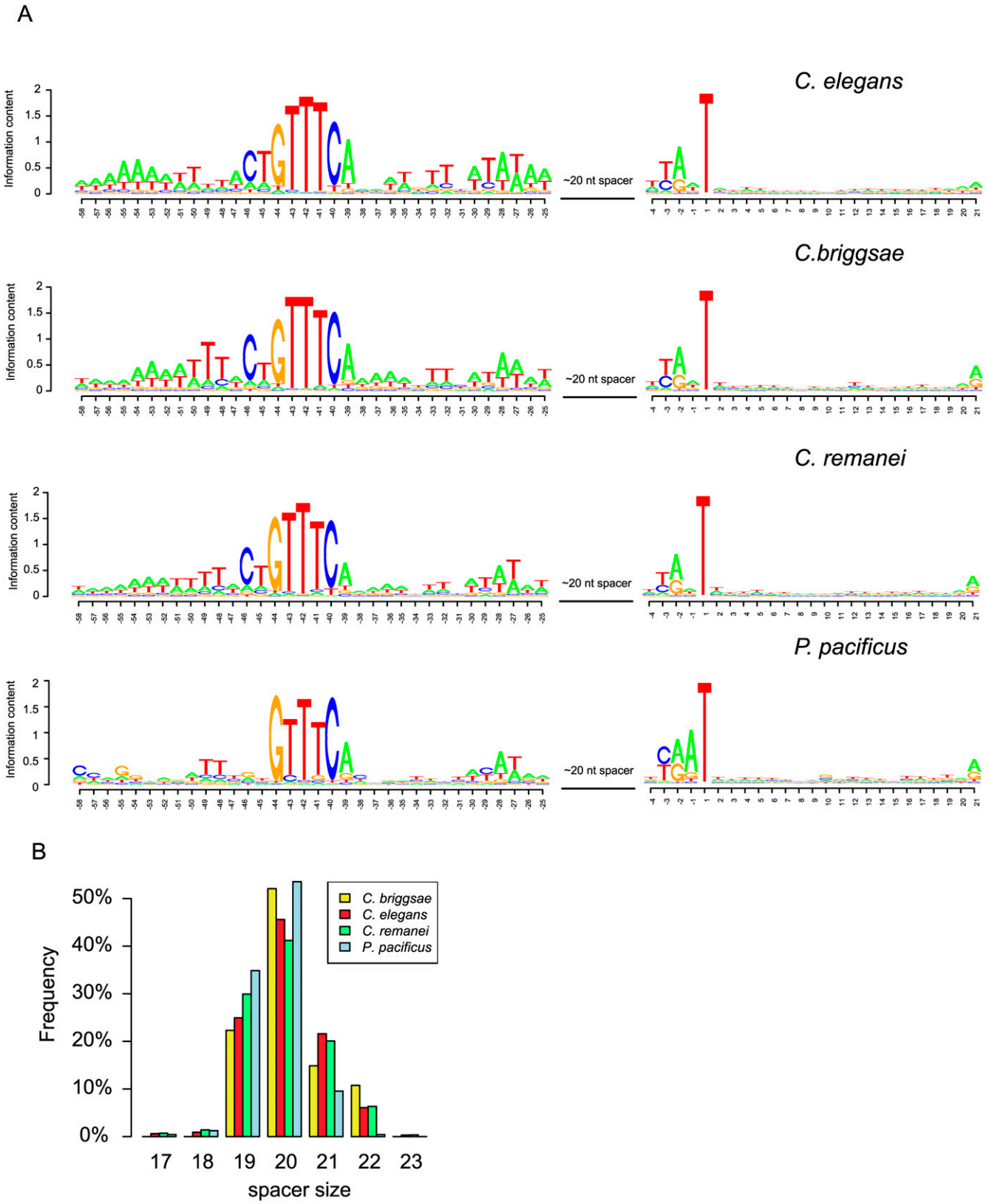

Figure 5. Upstream motifs of 21-U RNAs in different nematode species. (*A*) Consensus sequences for the upstream regions of the 21U-RNAs are displayed as sequence logos (Schneider and Stephens 1990) for all four nematode species. The consensus sequence is divided up into two regions: a small 4-nt motif directly upstream of the 21U-RNA and a larger motif ~25 nt upstream of the 21U-RNA. Spacers of variable size separate these motifs. (*B*) The distribution of the spacer size is depicted for all species.

remanei and *C. briggsae*, 89% of the contigs that are enriched for 21U-RNA motifs are syntenic with 21U-RNA clusters. As noted previously (Ruby et al. 2006), our results show that within the *Caenorhabditis* clade the 21U-RNA clusters are highly conserved.

The distribution of 21U-RNA motifs in *P. pacificus* is remarkably different from the *Caenorhabditis* species (Supplemental Fig. 7D). Although predicted 21U-RNA motifs are also clustered in *P. pacificus*, these clusters are much smaller and more widespread than the *Caenorhabditis* clusters. This is reflected in the total amount of contig sequence that is significantly enriched for 21U-RNA motifs, which amounts to 55 Mb, or nearly one-third of the genome. This is in stark contrast to the 19 Mb of enriched contig sequence for *C. remanei* or one-fifth of the genome. The conservation of 21U-RNA biology and its mechanisms over 400 Myr of evolution suggests its importance in physiological processes. At the same time, our data also shows that the genomic distribution of 21U-RNA genes has substantially diverged since the last common ancestor.

Discussion

In this study we describe cloning, sequencing, and analysis of small RNAs in four species of nematodes. For *C. briggsae*, *C. remanei*, and *P. pacificus,* no small RNA sequencing data have been previously documented. In all species the most abundant type of small RNA in our sequencing libraries is the miRNA, although it should be noted that our cloning strategy relies on the presence of 5′ monophosphates in RNA molecules and small RNAs with other 5′ modifications are not efficiently cloned by our approach. In addition to confirming the majority of *C. elegans* miRNAs from miRBase, we also experimentally identified 160 miRBase homologs that were not yet described in other species. Furthermore, we discovered 117 novel miRNA genes with high confidence, 91 of which were found in *P. pacificus*.

The resulting miRNA inventory was used to study the principles behind miRNA evolution in the nematode lineage. Our data have provided examples of all previously described mechanisms of emergence of novel miRNAs, including gene duplication, antisense transcription, arm switching, and seed shifting. Furthermore, we describe a novel emergence mechanism that we name hairpin shifting. In arm switching the once dominant arm of the miRNA is gradually replaced by the small RNA that originates from the other miRNA* arm. We found numerous such examples where the miRNA* has become the dominant species of miRNA. In many cases miRNA* sequences are also conserved at the seed level, suggesting their functional load. Our observations are in agreement with recent experiments in *Drosophila*, which provided experimental support for the importance of the miRNA* sequence (Okamura et al. 2008).

Antisense transcription is the second mechanism by which miRNAs can diverge. It has been described previously that miRNA loci can be transcribed from both strands to form mature miRNAs (Ruby et al. 2007a; Tyler et al. 2008). We found one locus that produced putative miRNAs from both directions and one locus that only produced an antisense miRNA to a known miRNA.

The third and novel mechanism of miRNA evolution that we describe is hairpin shifting. We suggest this process to explain the observation that within a set of highly similar mature miRNAs both the 5′ arm and the 3′ arm of the hairpin have been observed as the arm of origin for the mature miRNA. All of these cases were experimentally supported by the cloning of different miRNA* sequences from the same locus in different species. Since expansion of a miRNA family is generally mediated via gene duplication, the arm of origin for a miRNA is not expected to change, which is indeed the case for most miRNA families. Our results, however, indicate that for some families duplicated genes evolve independently, most likely by mutational events in the up- or downstream sequences, for 5′ or 3′ arm miRNAs, respectively. This process could give rise to a novel hairpin, and shifting will occur when this hairpin is thermodynamically more favorable than the original hairpin. The importance of this mechanism in light of miRNA evolution is that it leads to the generation of a new miRNA* sequence. Combined with arm switching, this might ultimately lead to the generation of new mature miRNAs.

We propose that the processes of duplication, arm switching, hairpin shifting, and antisense transcription can act sequentially in various permutations, making miRNA evolution a highly dynamic process. It is generally assumed that given the large number of hairpins in eukaryotic genomes (Pervouchine et al. 2003; Bentwich 2005), new miRNAs arise from the spurious expression of one of these hairpins. When the regulation by such a novel miRNA provides a selective advantage, this eventually leads to the "locking" of the miRNA and the target. From this point onward the miRNA will expand its target spectrum and its expression will be increased (Chen and Rajewsky 2007). Although this is an attractive model, miRNA evolution through duplication, arm switching, and hairpin shifting has the benefit of already having a complete framework for transcriptional regulation in place. Furthermore, successive rounds of arm switching combined with hairpin shifting and/or antisense transcription may produce miRNA genes that do not share seeds with any other miRNA and would therefore be indistinguishable from miRNA genes derived from de novo expression of genomic hairpins. Moreover, random genomic hairpins have lower genetic robustness as opposed to real miRNAs (Borenstein and Ruppin 2006), making them more vulnerable to mutations. As a consequence, evolving novel miRNAs by de novo expression of genomic hairpins is less likely compared with evolving a novel miRNA from a pre-existing duplicated miRNA gene.

With respect to the structure of miRNA hairpins, we have shown that there is a substantial difference in the thermodynamic profiles of miRNAs that originate from 3′ and 5′ arms of the hairpins. miRNAs that originate from the 5′ arm have a region of strong instability in the 5′ region of the miRNA, around the first nucleotide, consistent with previous results (Khvorova et al. 2003; Schwarz et al. 2003; Krol et al. 2004). In addition, the canonical hairpin has a lower stability in the middle of the miRNA. Intriguingly, this thermodynamic profile is much less pronounced for miRNAs that originate from the 3′ arm. We hypothesize that in nematodes the 3′ arm small RNA is the default miRNA that is loaded into the RISC complex, unless thermodynamic characteristics force the 5′ arm to be selected for incorporation. Modifications of the processing pathway throughout evolution may explain differences in arm bias in different species (Supplemental Fig. 4).

We have also experimentally confirmed the existence of a class of small RNAs known as 21U-RNAs in nematode species other than *C. elegans*. 21U-RNAs are believed to be the nematode equivalent of vertebrate piRNAs (Batista et al. 2008; Das et al. 2008; Wang and Reinke 2008), which are involved in transposon silencing and are essential for gametogenesis (Batista et al. 2008). The known 21U-RNA upstream motif (Ruby et al. 2006) was found independently in *C. briggsae*, *C. remanei*, and *P. pacificus*, which allowed us to identify 21U-RNAs in these species. The number of identified loci is limited compared with recent 21U-RNA identification efforts (Batista et al. 2008) due to a limited number of small RNA reads in our data set.

However, the distinctiveness of the upstream motif allows for characterization based on a limited number of reads. The 21U-RNA upstream motif is more diverged in *P. pacificus* and 21-RNAs form multiple genomic clusters scattered over the genome, as opposed to only several clusters in *C. elegans*. The current *P. pacificus* genome assembly consists of more than 18,000 contigs, which are not arranged into chromosomal order, and thus might influence the analysis of genomic distribution of 21U-RNA motifs in *P. pacificus*. However, the amount of assembled sequence covered by 21U-RNAs is significantly higher in *P. pacificus* compared with *C. remanei*, which has a genome assembly of similar quality, suggesting that the 21U-RNA distribution trend in *P. pacificus* would remain similar to an improved genome assembly.

In conclusion, we provide a systematic analysis of small RNAs in nematodes and describe various mechanisms of miRNA evolution, both over small and large evolutionary distances, and integrate them into a comprehensive model for miRNA evolution, which can lead to a great diversification of regulatory pathways. A question that remains open is whether, given the conservation of the seeds over such a large evolutionary distance, the miRNA target spectrum is also conserved or that it has significantly diverged over the course of evolution. The lack of sufficient sequence homology between the *Caenorhabditis* species and *P. pacificus* presently complicates genomic alignments of 3′ UTRs. Furthermore, the current genome annotations for *C. remanei* and *P. pacificus* are rather sparse with regard to 3′ UTRs, precluding target prediction. However, with improvements in genome annotation, future analyses are expected to shed light on the conservation of miRNA-based regulatory pathways over large evolutionary distances.

Methods

Worm strains, construction, and sequencing of small RNA libraries

For sequencing, the wild-type strains Bristol N2 for *C. elegans*, AF16 for *C. briggsae*, EM464 for *C. remanei*, and PS312 for *P. pacificus* were used. Small RNA libraries were produced by Vertis Biotechnology AG (Freising-Weihenstephan, Germany) from mixed-stage worm samples that had been grown and collected according to standard procedures. Worms were cultured on Nematode Growth Medium (NGM) plates containing the *E. coli* strain OP50 at 20°C. Chunks of full-grown plates where transferred to NGM plates containing OP50 that had been cultivated in chicken egg yolk. Once these plates were full grown, the worms were rinsed off with M9 medium, cleared from OP50 by centrifuging (2 min, 814*g*) in a 30% glycerol solution at 4°C, washed in M9, rinsed in M9 for 30 min at RT to flush the guts, cooled down to collect the pellet, and stored and shipped at −80°C.

Library construction was performed as described previously (Berezikov et al. 2007). Briefly, we used poly(A) tailing to extend small RNAs. Subsequently, a sequencing adapter was ligated to the 5′ end of the small RNA. A locked oligo dT primer was used to perform the reverse transcriptase reaction. The resulting cDNA was used as a template for the library amplification step. Small RNA libraries were sequenced using the 454 Life Sciences (Roche) sequencing method. Raw sequences were processed by trimming of poly(A) stretches, which represent the 3′ adapter sequences.

Sources of genome assemblies and annotations

The *C. brenneri*, *C. remanei*, *C. japonica*, and *P. pacificus* draft genome assemblies used in the analysis were produced by The Genome Center at Washington University School of Medicine in St. Louis, Missouri (http://genome.wustl.edu; R. Wilson, pers. comm.) and obtained from the following URLs:

C. brenneri (assembly 6.0.1, Feb 27, 2008), http://genome.wustl.edu/pub/organism/Invertebrates/Caenorhabditis_PB2801/assembly/Caenorhabditis_PB2801-6.0.1; *C. remanei* (assembly 15.01, May 25, 2007), http://genome.wustl.edu/pub/organism/Invertebrates/Caenorhabditis_remanei/assembly/Caenorhabditis_remanei-15.0.1; *C. japonica* (assembly 3.0.2, Mar 12, 2008), http://genome.wustl.edu/pub/organism/Invertebrates/Caenorhabditis_japonica/assembly/Caenorhabditis_japonica-3.0.2; *P. pacificus* (assembly 5.0, Jan 11, 2007), http://genome.wustl.edu/pub/organism/Invertebrates/Pristionchus_pacificus/assembly/Pristionchus_pacificus-5.0. The annotations for respective genomes were obtained from the UCSC Genome Browser from the following URLs: *C. brenneri*, ftp://hgdownload.cse.ucsc.edu/goldenPath/caePb2; *C. remanei*, ftp://hgdownload.cse.ucsc.edu/goldenPath/caeRem3; *C. japonica*, ftp://hgdownload.cse.ucsc.edu/goldenPath/caeJap1; *P. pacificus*, ftp://hgdownload.cse.ucsc.edu/goldenPath/priPac1. The published *C. elegans* and *C. briggsae* genomes and annotations were obtained from Ensembl and UCSC Genome databases, respectively (ftp://ftp.ensembl.org/pub/release-50/mysql/caenorhabditis_elegans_core_50_190; ftp://hgdownload.cse.ucsc.edu/goldenPath/cb3).

Annotation of small RNA reads and identification of novel miRNA genes

Primary data analysis was performed using miR-Intess software (http://www.internagenomics.com), which is based on our previous works (Berezikov et al. 2006a,b). After quality and adapter trimming, raw sequence reads were aligned to the respective genomes using MEGABLAST software (Zhang et al. 1998) and custom post-processing scripts to allow trimming of 3′ nucleotides as described previously (Berezikov et al. 2006b). Specifically, we required perfect matching to the genome of the first 18 nt of the read, further extended the 3′ matching part of the read by removing extra "A" insertions as artifacts of 454 pyrosequencing of poly(A) tails, trimmed the remaining mismatched 3′ bases, and selected the longest possible hits as read locations. Aligned reads were classified according to genomic loci annotations retrieved from Ensembl or UCSC genome databases. Reads that aligned to intergenic or intronic regions but not to exons, repeats, or structural RNAs, were used in the miRNA discovery and annotation part of miR-Intess. First, overlapping reads were grouped into blocks, and every block and its surrounding genomic sequence were investigated for the potential to form a hairpin with the abstract shape "[ ]", using the RNAshapes program (Steffen et al. 2006) and sliding windows of 80, 100, and 120 nt, and requiring location of the reads within one of the arms of a hairpin. Next, several parameters were calculated for the identified haiprins, including number and length of supporting reads, their location in the hairpin, and variation of start/end positions, number and size of internal bulges, GC content, randfold value (Bonnet et al. 2004), Drosha/Dicer signatures and hairpin copy number in the genome, and annotation of homologous hairpins in other genomes. These parameters were used to classify hairpins into three confidence levels (confident, candidate, and unlikely), and parameter thresholds were selected such that the majority of miRBase miRNAs would fall into the confident category while keeping the number of unknown hairpins in this category to the minimum. It should be noted that conservation of the hairpins in other species, inferred as described below, is used in our hairpin classification only to "downgrade" hairpins that have obvious homologs in other species annotated as repeats or structural RNAs. Conservation of the hairpin itself is not required for assignment to confident or candidate hairpins.

Alignments of novel miRNAs identified by miR-Intess are available as Supplemental Figures 8–11 and the entire analysis is accessible at http://www.internagenomics.com/public/worm0903.

miRNA conservation analysis

Identification of miRNA homologs in other species was performed as described (Berezikov et al. 2006c). All hairpins were searched against other genomes using BLAST, and hits with conserved seed region were extended and checked for hairpin presence using RNAshapes. All potential hairpin homologs were further assessed using RNAforester software (Hochsmann et al. 2004), and the hairpin with the highest score above 0.3 was assigned as a most likely homolog.

21U RNAs

For the identification of the 21U-RNA motifs we selected 80 nt of the upstream region of all small RNA reads that were not annotated as hairpin or repeat by our computational pipeline. In these regions the 21U-RNA core motif GTTTC was strongly overrepresented and found in a limited region (Ruby et al. 2006). Only promoters that contain the GTTTC motif start between 48 and 42 and with a maximum of one mismatch in the motif were selected. For the calculation of the PWM, the promoters were aligned on the core motif and the nucleotide frequencies determined.

Using the PWM, we used the Perl TFBS module (Lenhard and Wasserman 2002) to search the genomes of *C. elegans, C. briggsae, C. remanei*, and *P. pacificus* for predicted 21U-RNA sites. Sites with a score higher than 24 were scored as predicted 21U-RNA motifs. All sequence reads that had a predicted 21U-RNA motif upstream were designated as 21U-RNAs. The propinquity of the predicted 21U-RNA motifs to each other allowed us to define regions in the genome with increased frequency of 21U-RNA motifs or clusters. We used a sliding window approach to count the number of 21U-RNA motifs within a 100-kb window, moving the window by 10 kb.

Thermodynamic profiles

Using mfold (v3.2) we predicted RNA secondary structures for the candidate miRNA genes (Zuker 2003). mfold calculates the stacking free energy, taking into account destabilizing elements such as bulges and internal loops. Using the loop free-energy decomposition, the ΔG for every base (for loops and bulges) or base pair (for stacks) is calculated. The canonical thermodynamic profile was calculated by aligning the profiles on the starting nucleotide and averaging over all thermodynamic profiles.

Nothern blot analysis

Worms were grown and collected as described in the library preparation section and spun down at 5000*g* for 30 sec to remove excess fluid. The pellets were stored at −80°C. Small RNA was collected from these according to the mirVana miRNA Isolation kit manufacturer's procedures with slight modifications, i.e., worms were lysed in 2 vol of lysis buffer, 20 μL of proteinase K (20 mg/mL) and 1/10 vol of 5 M NaCl, after which they were lysed at 55°C for 20 min and the standard protocol was continued. After quality assessment on 1% agarose gel and quantification by nanodrop, 3–5 μg of small RNAs (<200 nt) were heated at 65°C for 5 min in formamide loading buffer, run over a 12.5% denaturing polyacrylamide gel, and blotted onto negatively charged nylon membrane, after which the blot was dried, the RNA was cross-linked by UV, and incubated at 37°C in hybridization buffer (0.36 M Na_2HPO_4, 0.14 M NaH_2PO_4, 1 mM EDTA, 7%SDS) for 10 min. Hereafter, the radiolabeled probe (1 μL of 10 μM of DNA oligonucleotide probe, 1 μL of 10× buffer, 1 μL of PNK [Promega], 1 μL of 10 mCi/mL [γ-^{32}P]ATP [Perkin Elmer], 6 μL of DEPC-treated H_2O, incubated at 37°C for 1 h, and boiled for 2 min) was added and the blots were incubated overnight at 37°C. Prior to exposure, they were shortly washed three times in prewarmed 0.2% SDS, 2× SSC, at 37°C. Phosphor screens were read on a Typhoon scanner and images were enhanced by ImageQuant software (GE-healthcare).

Seed-shifting analysis

Orthologs of *C. elegans* miRBase miRNAs were collected from the miR-Intess output based on their annotation. Overlapping miRBase seeds (nt 2–8) with nt 1–7 sequences of cloned miRNAs, as well as nt 1–7 of miRBase miRNAs with seeds of cloned miRNAs were counted. Orthologs that were only identified as miRNAs that underwent seed shifting, i.e., could be incorrectly annotated, were excluded. Only the most abundantly cloned reads from each hairpin arm were included. This analysis was repeated for all cloned miRNAs to determine seed shifting in all miRNAs with overlapping 7-mers in the seed region.

Acknowledgments

We thank Graham Ruby for providing the 21U PWM and Rene Ketting for critically reading the manuscript. The nematode strains used in this work were kindly provided by Ralf Sommer (the *P. pacificus*) and the *Caenorhabditis* Genetics Center, funded by the NIH National Center for Research Resources (NCRR). We thank Georges Janssens for maintaining worm stocks for Northern blot preparations. We thank Ronald Plasterk for initiating the experiments. This work is supported by a Veni grant (NWO) to E.B., funding from the Cancer Genomics Center (NGI) to E.C. and a NWO TOP grant (91206104) to Ronald Plasterk. The *C. brenneri*, *C. remanei*, *C. japonica*, and *P. pacificus* draft genome assemblies were produced by The Genome Center at Washington University School of Medicine in St. Louis and can be obtained from the URLs indicated in the Methods section.

References

Bartel DP. 2009. MicroRNAs: Target recognition and regulatory functions. *Cell* **136:** 215–233.

Batista PJ, Ruby JG, Claycomb JM, Chiang R, Fahlgren N, Kasschau KD, Chaves DA, Gu W, Vasale JJ, Duan S, et al. 2008. PRG-1 and 21U-RNAs interact to form the piRNA complex required for fertility in *C. elegans*. *Mol Cell* **31:** 67–78.

Bentwich I. 2005. Prediction and validation of microRNAs and their targets. *FEBS Lett* **579:** 5904–5910.

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

Berezikov E, Cuppen E, Plasterk RH. 2006a. Approaches to microRNA discovery. *Nat Genet* **38:** S2–S7.

Berezikov E, Thuemmler F, van Laake LW, Kondova I, Bontrop R, Cuppen E, Plasterk RH. 2006b. Diversity of microRNAs in human and chimpanzee brain. *Nat Genet* **38:** 1375–1377.

Berezikov E, van Tetering G, Verheul M, van de Belt J, van Laake L, Vos J, Verloop R, van de Wetering M, Guryev V, Takada S, et al. 2006c. Many novel mammalian microRNA candidates identified by extensive cloning and RAKE analysis. *Genome Res* **16:** 1289–1298.

Berezikov E, Chung WJ, Willis J, Cuppen E, Lai EC. 2007. Mammalian mirtron genes. *Mol Cell* **28:** 328–336.

Bernstein E, Caudy AA, Hammond SM, Hannon GJ. 2001. Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature* **409:** 363–366.

Bilen J, Liu N, Burnett BG, Pittman RN, Bonini NM. 2006. MicroRNA pathways modulate polyglutamine-induced neurodegeneration. *Mol Cell* **24:** 157–163.

Bonnet E, Wuyts J, Rouze P, Van de Peer Y. 2004. Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics* **20:** 2911–2917.
Borenstein E, Ruppin E. 2006. Direct evolution of genetic robustness in microRNA. *Proc Natl Acad Sci* **103:** 6593–6598.
Carrington JC, Ambros V. 2003. Role of microRNAs in plant and animal development. *Science* **301:** 336–338.
Chen K, Rajewsky N. 2007. The evolution of gene regulation by transcription factors and microRNAs. *Nat Rev Genet* **8:** 93–103.
Croce CM, Calin GA. 2005. miRNAs, cancer, and stem cell division. *Cell* **122:** 6–7.
Das PP, Bagijn MP, Goldstein LD, Woolford JR, Lehrbach NJ, Sapetschnig A, Buhecha HR, Gilchrist MJ, Howe KL, Stark R, et al. 2008. Piwi and piRNAs act upstream of an endogenous siRNA pathway to suppress Tc3 transposon mobility in the *Caenorhabditis elegans* germline. *Mol Cell* **31:** 79–90.
Dieterich C, Clifton SW, Schuster LN, Chinwalla A, Delehaunty K, Dinkelacker I, Fulton L, Fulton R, Godfrey J, Minx P, et al. 2008. The *Pristionchus pacificus* genome provides a unique perspective on nematode lifestyle and parasitism. *Nat Genet* **40:** 1193–1198.
Filipowicz W. 2005. RNAi: The nuts and bolts of the RISC machine. *Cell* **122:** 17–20.
Filipowicz W, Bhattacharyya SN, Sonenberg N. 2008. Mechanisms of post-transcriptional regulation by microRNAs: Are the answers in sight? *Nat Rev Genet* **9:** 102–114.
Gregory RI, Yan KP, Amuthan G, Chendrimada T, Doratotaj B, Cooch N, Shiekhattar R. 2004. The Microprocessor complex mediates the genesis of microRNAs. *Nature* **432:** 235–240.
Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ. 2008. miRBase: Tools for microRNA genomics. *Nucleic Acids Res* **36:** D154–D158.
Grimson A, Srivastava M, Fahey B, Woodcroft BJ, Chiang HR, King N, Degnan BM, Rokhsar DS, Bartel DP. 2008. Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. *Nature* **455:** 1193–1197.
Gu S, Jin L, Zhang F, Sarnow P, Kay MA. 2009. Biological basis for restriction of microRNA targets to the 3′ untranslated region in mammalian mRNAs. *Nat Struct Mol Biol* **16:** 144–150.
Hammell M, Long D, Zhang L, Lee A, Carmack CS, Han M, Ding Y, Ambros V. 2008. mirWIP: MicroRNA target prediction based on microRNA-containing ribonucleoprotein-enriched transcripts. *Nat Methods* **5:** 813–819.
Han J, Lee Y, Yeom KH, Kim YK, Jin H, Kim VN. 2004. The Drosha–DGCR8 complex in primary microRNA processing. *Genes & Dev* **18:** 3016–3027.
Hillier LW, Miller RD, Baird SE, Chinwalla A, Fulton LA, Koboldt DC, Waterston RH. 2007. Comparison of *C. elegans* and *C. briggsae* genome sequences reveals extensive conservation of chromosome organization and synteny. *PLoS Biol* **5:** e167. doi: 10.1371/journal.pbio.0050167.
Hochsmann M, Voss B, Giegerich R. 2004. Pure multiple RNA secondary structure alignments: A progressive profile approach. *IEEE/ACM Trans Comput Biol Bioinform* **1:** 53–62.
Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E. 2007. The role of site accessibility in microRNA target recognition. *Nat Genet* **39:** 1278–1284.
Khvorova A, Reynolds A, Jayasena SD. 2003. Functional siRNAs and miRNAs exhibit strand bias. *Cell* **115:** 209–216.
Kiontke K, Fitch DH. 2005. The phylogenetic relationships of *Caenorhabditis* and other rhabditids. *WormBook* http://www.wormbook.org/chapters/www_phylogrhabditids/phylorhab.html.
Kiontke K, Sudhaus W. 2006. Ecology of *Caenorhabditis* species. *WormBook* 1–14.
Kiontke K, Gavin NP, Raynes Y, Roehrig C, Piano F, Fitch DH. 2004. *Caenorhabditis* phylogeny predicts convergence of hermaphroditism and extensive intron loss. *Proc Natl Acad Sci* **101:** 9003–9008.
Krol J, Sobczak K, Wilczynska U, Drath M, Jasinska A, Kaczynska D, Krzyzosiak WJ. 2004. Structural features of microRNA (miRNA) precursors and their relevance to miRNA biogenesis and small interfering RNA/short hairpin RNA design. *J Biol Chem* **279:** 42230–42239.
Lai EC. 2002. Micro RNAs are complementary to 3′ UTR sequence motifs that mediate negative post-transcriptional regulation. *Nat Genet* **30:** 363–364.
Lau NC, Lim LP, Weinstein EG, Bartel DP. 2001. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science* **294:** 858–862.
Lee RC, Ambros V. 2001. An extensive class of small RNAs in *Caenorhabditis elegans*. *Science* **294:** 862–864.
Lee RC, Feinbaum RL, Ambros V. 1993. The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* **75:** 843–854.
Lee Y, Jeon K, Lee JT, Kim S, Kim VN. 2002. MicroRNA maturation: Stepwise processing and subcellular localization. *EMBO J* **21:** 4663–4670.
Lee Y, Ahn C, Han J, Choi H, Kim J, Yim J, Lee J, Provost P, Radmark O, Kim S, et al. 2003. The nuclear RNase III Drosha initiates microRNA processing. *Nature* **425:** 415–419.
Lee Y, Kim M, Han J, Yeom KH, Lee S, Baek SH, Kim VN. 2004. MicroRNA genes are transcribed by RNA polymerase II. *EMBO J* **23:** 4051–4060.
Lenhard B, Wasserman WW. 2002. TFBS: Computational framework for transcription factor binding site analysis. *Bioinformatics* **18:** 1135–1136.
Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP, Burge CB. 2003. Prediction of mammalian microRNA targets. *Cell* **115:** 787–798.
Liu N, Okamura K, Tyler DM, Phillips MD, Chung WJ, Lai EC. 2008. The evolution and functional diversification of animal microRNA genes. *Cell Res* **18:** 985–996.
Lu J, Shen Y, Wu Q, Kumar S, He B, Shi S, Carthew RW, Wang SM, Wu CI. 2008. The birth and death of microRNA genes in *Drosophila*. *Nat Genet* **40:** 351–355.
Martinez NJ, Ow MC, Reece-Hoyes JS, Barrasa MI, Ambros VR, Walhout AJ. 2008. Genome-scale spatiotemporal analysis of *Caenorhabditis elegans* microRNA promoter activity. *Genome Res* **18:** 2005–2015.
Okamura K, Phillips MD, Tyler DM, Duan H, Chou YT, Lai EC. 2008. The regulatory activity of microRNA* species has substantial influence on microRNA and 3′ UTR evolution. *Nat Struct Mol Biol* **15:** 354–363.
Pervouchine DD, Graber JH, Kasif S. 2003. On the normalization of RNA equilibrium free energy to the length of the sequence. *Nucleic Acids Res* **31:** e49. doi: 10.1093/nar/gng049.
Plasterk RH. 2006. Micro RNAs in animal development. *Cell* **124:** 877–881.
Ruby JG, Jan C, Player C, Axtell MJ, Lee W, Nusbaum C, Ge H, Bartel DP. 2006. Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. *Cell* **127:** 1193–1207.
Ruby JG, Jan CH, Bartel DP. 2007a. Intronic microRNA precursors that bypass Drosha processing. *Nature* **448:** 83–86.
Ruby JG, Stark A, Johnston WK, Kellis M, Bartel DP, Lai EC. 2007b. Evolution, biogenesis, expression, and target predictions of a substantially expanded set of *Drosophila* microRNAs. *Genome Res* **17:** 1850–1864.
Schneider TD, Stephens RM. 1990. Sequence logos: A new way to display consensus sequences. *Nucleic Acids Res* **18:** 6097–6100.
Schwarz DS, Hutvagner G, Du T, Xu Z, Aronin N, Zamore PD. 2003. Asymmetry in the assembly of the RNAi enzyme complex. *Cell* **115:** 199–208.
Stark A, Brennecke J, Russell RB, Cohen SM. 2003. Identification of *Drosophila* MicroRNA targets. *PLoS Biol* **1:** e60. doi: 10.1371/journal.pbio.0000060.
Stark A, Kheradpour P, Parts L, Brennecke J, Hodges E, Hannon GJ, Kellis M. 2007. Systematic discovery and characterization of fly microRNAs using 12 *Drosophila* genomes. *Genome Res* **17:** 1865–1879.
Stefani G, Slack FJ. 2008. Small non-coding RNAs in animal development. *Nat Rev Mol Cell Biol* **9:** 219–230.
Steffen P, Voss B, Rehmsmeier M, Reeder J, Giegerich R. 2006. RNAshapes: An integrated RNA analysis package based on abstract shapes. *Bioinformatics* **22:** 500–503.
Stein LD, Bao Z, Blasiar D, Blumenthal T, Brent MR, Chen N, Chinwalla A, Clarke L, Clee C, Coghlan A, et al. 2003. The genome sequence of *Caenorhabditis briggsae*: A platform for comparative genomics. *PLoS Biol* **1:** e45. doi: 10.1371/journal.pbio.0000045.
Thum T, Galuppo P, Wolf C, Fiedler J, Kneitz S, van Laake LW, Doevendans PA, Mummery CL, Borlak J, Haverich A, et al. 2007. MicroRNAs in the human heart: A clue to fetal gene reprogramming in heart failure. *Circulation* **116:** 258–267.
Tyler DM, Okamura K, Chung WJ, Hagen JW, Berezikov E, Hannon GJ, Lai EC. 2008. Functionally distinct regulatory RNAs generated by bidirectional transcription and processing of microRNA loci. *Genes & Dev* **22:** 26–36.
Wang G, Reinke V. 2008. A *C. elegans* Piwi, PRG-1, regulates 21U-RNAs during spermatogenesis. *Curr Biol* **18:** 861–867.
Wheeler BM, Heimberg AM, Moy VN, Sperling EA, Holstein TW, Heber S, Peterson KJ. 2009. The deep evolution of metazoan microRNAs. *Evol Dev* **11:** 50–68.
Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M. 2005. Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature* **434:** 338–345.
Zhang Z, Berman P, Miller W. 1998. Alignments without low-scoring regions. *J Comput Biol* **5:** 197–210.
Zuker M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **31:** 3406–3415.

Received March 13, 2009; accepted in revised form July 7, 2009.

Repertoire and evolution of miRNA genes in four divergent nematode species

Elzo de Wit, Sam E.V. Linsen, Edwin Cuppen, et al.

Genome Res. 2009 19: 2064-2074 originally published online September 15, 2009
Access the most recent version at doi:10.1101/gr.093781.109