Patterns of flanking sequence conservation and a characteristic upstream motif for microRNA gene identification

UWE OHLER,[1] SORAYA YEKTA,[1,2] LEE P. LIM,[1–3] DAVID P. BARTEL,[1,2] and CHRISTOPHER B. BURGE[1]
[1]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA
[2]Whitehead Institute for Biomedical Research, Cambridge, Massachusetts 02142, USA

ABSTRACT

MicroRNAs are ~22-nucleotide (nt) RNAs processed from foldback segments of endogenous transcripts. Some are known to play important gene regulatory roles during animal and plant development by pairing to the messages of protein-coding genes to direct the post-transcriptional repression of these messages. Previously, we developed a computational method called MiRscan, which scores features related to the foldbacks, and used this algorithm to identify new miRNA genes in the nematode *Caenorhabditis elegans*. In the present study, to identify sequences that might be involved in processing or transcriptional regulation of miRNAs, we aligned sequences upstream and downstream of orthologous nematode miRNA foldbacks. These alignments showed a pronounced peak in sequence conservation about 200 bp upstream of the miRNA foldback and revealed a highly significant sequence motif, with consensus CTCCGCCC, that is present upstream of almost all independently transcribed nematode miRNA genes. Scoring the pattern of upstream/downstream conservation, the occurrence of this sequence motif, and orthology of host genes for intronic miRNA candidates, yielded substantial improvements in the accuracy of MiRscan. Nine new *C. elegans* miRNA gene candidates were validated using a PCR-sequencing protocol. As previously seen for bacterial RNA genes, sequence features outside of the RNA secondary structure can therefore be very useful for the computational identification of eukaryotic noncoding RNA genes. The total number of confidently identified nematode miRNAs now approaches 100. The improved analysis supports our previous assertion that miRNA gene identification is nearing completion in *C. elegans* with apparently no more than 20 miRNA genes now remaining to be identified.

Keywords: microRNA; noncoding RNA; computational gene identification; regulatory motif; transcription

INTRODUCTION

MicroRNAs (miRNAs) are a class of small noncoding RNAs that are found in a variety of eukaryotic multicellular organisms (Lee et al. 1993; Reinhart et al. 2000; Lagos-Quintana et al. 2001; Lau et al. 2001; Lee and Ambros 2001; Llave et al. 2002; Park et al. 2002; Reinhart et al. 2002). They are known to be important gene-regulatory molecules in both animals and plants (Ambros 2003; Bartel 2004). In animals, miRNAs are processed in two steps (Lee et al. 2002, 2003), from primary transcripts to ~70-nucleotide (nt) precursors by the RNase III enzyme Drosha, and from precursors to the ~22-nt single-stranded miRNAs by the RNase III enzyme Dicer. The processed miRNAs can direct post-transcriptional regulation of specific target mRNAs (Lee et al. 1993; Wightman et al. 1993; Moss et al. 1997; Reinhart et al. 2000; Lai 2002; Abrahante et al. 2003; Brennecke et al. 2003; Lewis et al. 2003; Lin et al. 2003; Yekta et al. 2004).

Noncoding RNA genes (Eddy 2001) are typically independently transcribed by one of the three RNA polymerases, for example, rRNA genes by RNA polymerase I (pol-I), most snRNA genes by RNA pol-II, and tRNA genes by RNA pol-III (Brown 2002). Alternatively, they can be cotranscribed within host genes, as is the case with most vertebrate snoRNA genes (Bachellerie et al. 2002), which are located within introns of pol-II-transcribed host genes. Most miRNA genes are located far away from any annotated genes, implying independent transcription from their own

Reprint requests to: David P. Bartel, Whitehead Institute, 9 Cambridge Center, Cambridge, MA 02142, USA; e-mail: dbartel@wi.mit.edu; fax: (617) 258-6768; and Christopher B. Burge, Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02142, USA; e-mail: cburge@mit.edu; fax: (617) 452-2936.
[3]**Present address:** Rosetta Inpharmatics, Merck & Co., 401 Terry Avenue N., Seattle, WA 98109, USA.
Article and publication are at http://www.rnajournal.org/cgi/doi/10.1261/rna.5206304.

promoters, but some lie within predicted introns of protein-coding genes (Lau et al. 2001; Lagos-Quintana et al. 2003)—for example, 22 of the 88 nematode miRNAs known at the start of this study have intronic locations. In most of these cases (80%), the introns are in the same orientation as the miRNAs, implying that the protein-coding genes might serve as host genes for coexpressed miRNAs. Therefore, in this study, we provisionally group the miRNA genes into two categories as follows: Those located in the sense strand of annotated introns are classified as cotranscribed miRNAs (although some might be independently transcribed), and all other miRNA genes, including those that are clustered in the genome in a configuration suggestive of transcription as a single polycistronic RNA (Lagos-Quintana et al. 2001; Lau et al. 2001) are classified as independently transcribed because they are unlikely to share a primary transcript with a non-miRNA host gene.

Although functional miRNA genes can be expressed by pol-II or pol-III (Zeng et al. 2002; Zeng and Cullen 2003; Chen et al. 2004), the identity of the polymerase(s) that transcribes the endogenous genes is not known. Some miRNA foldbacks are located in close genomic proximity to each other and are transcribed as polycistronic units (Lee et al. 2002; Aravin et al. 2003). The largest of these miRNA clusters extend well over a kilobase on the genome, which makes transcription of these clusters by pol-III unlikely, in that annotated nematode pol-III transcripts are only up to 300–400 bases in size (Harris et al. 2003). Likewise, the primary transcripts of some singly transcribed miRNAs often appear to be longer than typical pol-III transcripts (Lee et al. 2002). Transcriptional regulation of miRNAs is only beginning to be studied in detail (Johnson et al. 2003; Sempere et al. 2003).

Computational identification of miRNAs is greatly aided by their occurrence in the context of conserved stem–loop foldbacks. Because of a more variable-sized foldback structure in plants (Reinhart et al. 2002), the prediction of plant miRNAs is more challenging, and has only recently been reported (Jones-Rhoades and Bartel 2004). Computational screens for conserved foldbacks, combined with large-scale cloning efforts, recently brought the number of identified *Caenorhabditis elegans* miRNA genes to 88 (Lim et al. 2003b). Since then, two groups have reported seven and 10 additional candidate miRNAs, respectively (Ambros et al. 2003b; Grad et al. 2003). These three independent studies give different upper-bound estimates, ranging from ~120 to 300 or more *C. elegans* miRNA genes. The number of *Drosophila* miRNA genes has been estimated at 110 (Lai et al. 2003), and about twice this number are thought to be present in vertebrates (Lim et al. 2003a). The computational approaches typically apply RNA folding methods to detect regions with potential to fold into stem–loop structures, use cross-species conservation to restrict the vast number of potential stem–loop structures found in each genome, and score conserved foldbacks for conservation and a variety of sequence and secondary structural features.

Our goal here was to identify specific sequence features in the vicinity of independent and cotranscribed miRNAs, which might be involved in their expression, and to integrate these features into an improved version of the miRNA gene finding algorithm MiRscan (Lim et al. 2003b). In particular, we focused on (1) the pattern of conservation upstream and downstream of miRNA foldbacks; (2) specific sequence motifs adjacent to foldbacks likely to be involved in transcription or processing of miRNAs; and (3) the location of cotranscribed miRNAs in orthologous host genes. For independently transcribed miRNAs, we also examined the benefits of requiring synteny of the flanking protein-coding genes, as well as the use of whole-genome alignments. We concentrated our efforts on miRNAs in *C. elegans*, as this organism had been subject to the most comprehensive miRNA cloning effort at the time this study was begun, and the closely related nematode *Caenorhabditis briggsae* had the advantage of an assembled and preannotated genome, which has now been published (Stein et al. 2003). The presence of transcription initiation and termination sequence elements has been successfully used in computational identification of prokaryotic noncoding RNA genes (Argaman et al. 2001). Here, we demonstrate the use of features outside of the actual RNA secondary structure, such as an upstream promoter/processing motif and upstream/downstream sequence conservation, for computational discovery of noncoding RNA genes in eukaryotes.

RESULTS

Analysis of microRNA genes

Conservation upstream and downstream of miRNA genes

We assembled sets of 43 orthologous *C. elegans/C. briggsae* miRNA upstream and downstream sequences likely to contain transcriptional regulatory sequences. The Upstream Sequence Set (USS) encompasses the regions 2000 bp upstream, and the Downstream Sequence Set (DSS) encompasses the regions 1000 bp downstream of the foldbacks. For each pair of sequences from the USS and DSS data sets, the orthologous sequences were aligned with the tools DBA (Jareborg et al. 1999) and BayesBlockAligner (Zhu et al. 1998), and the resulting sets of aligned regions were merged. Downstream sequences were generally less conserved than upstream sequences, and in both directions the degree of conservation decreased with increasing distance from the foldback (Fig. 1). There was also a pronounced peak of conservation at about 200 bp upstream of the foldbacks. On average, 248 bp of the first 1000 bp upstream were aligned within conserved blocks of at least 70% se-

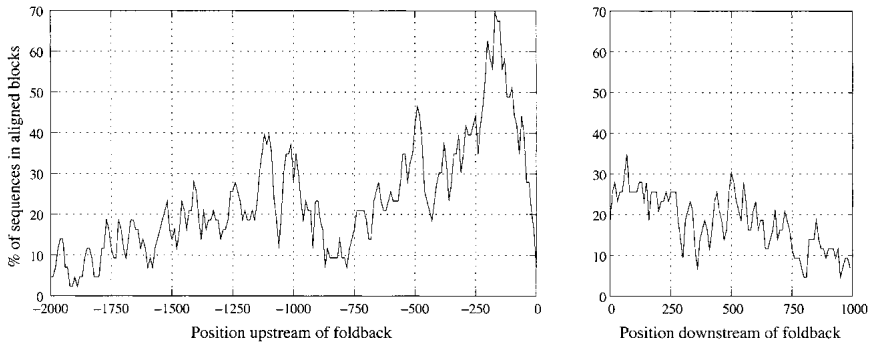

FIGURE 1. Conservation upstream and downstream of nematode microRNA foldbacks. The percentage of *C. elegans* sequences that are part of a conserved aligned block with *C. briggsae* at specific positions is plotted in bins of 10 bp. The positions are given relative to the beginning (*left*) or end (*right*) of the 110-nt segments containing the foldback. Genomic sequences were aligned using DBA and BayesBlockAligner as described in the text. Example alignments are part of the Supplementary Material (http://genes.mit.edu/burgelab/MiRscanII).

quence identity, compared with 146 bp of the first 1000 bp downstream.

Identification of a conserved upstream element

Next, we searched for conserved upstream sequence motifs, which might be involved in miRNA transcription or processing. Algorithms for the identification of conserved motifs can be grouped into enumerative and alignment approaches (Ohler and Niemann 2001). The ST algorithm, based on an approach described by Sinha and Tompa (2000), is an enumerative word-based algorithm that identifies statistically over-represented oligomers in a target set of sequences when compared with a background model. With this algorithm, we searched for over-represented words in the *C. elegans* sequence blocks conserved with *C. briggsae*, using a background model derived from the whole 2-kb upstream regions. The two significant distinct motifs that were found had the consensus sequences CTCCGCCC (motif A) and GCGTGGCS (motif B; S = C or G). Motif A was highly significant, frequently occurring, and well conserved. By comparison, motif B had a much lower score and was less frequent (Fig. 2A).

We repeated this search with the alignment-based motif-finding tool MEME (Bailey and Elkan 1995), choosing the "zero-or-one-occurrence" alignment mode, which identifies motifs present in some, but not necessarily all of the sequences. MEME reported a motif essentially identical to motif A as the strongest hit, either when searching only in the conserved sequence blocks or in the complete USS (Fig. 2B). A highly similar motif was identified in the *C. briggsae* sequences. In both *C. elegans* and *C. briggsae* sequences, the motif was preferentially located <500 bp upstream of the foldback (Fig. 2C). The motif was found on both strands, with a ~2:1 preference for the forward strand. In most cases, the location of the best match in *C. elegans* (on either the forward or reverse strand) was similar to that in *C. briggsae* (in 25/43 cases, the locations relative to the hairpin differed by <250 bp). Motif B (Fig. 2A) was not identified by MEME.

Finally, we asked whether motif A is also frequently found upstream of non-miRNA genes. The ST algorithm did not identify a similar motif in conserved sequence blocks upstream of 74 orthologous *C. elegans* and *C. briggsae* protein-coding genes (Webb et al. 2002). Also, no similar motif was found by MEME in sequences upstream of the 36 annotated *C. elegans* pol-II-transcribed snRNA genes (WormBase release 100), in the intronic sequences upstream of the 13 cotranscribed miRNA genes, or in the sequences upstream of the 13 protein-coding genes with cotranscribed intronic miRNA genes. These observations indicated that occurrence of motif A is a useful marker of independently transcribed miRNA genes.

Upstream elements in mammals and insects

An investigation of the regions upstream of 59 orthologous human/mouse orthologous miRNAs likewise identified an over-represented motif, CCCWCCC (ST algorithm Z-score 11.1; control background score 5.7; W = A or T), which was present 98 times in conserved blocks of 45 upstream regions. A second motif, ATGCAT, was present 18 times in 14 regions. Analysis of a set of 31 upstream sequences of independently transcribed *Drosophila melanogaster* miRNAs (Aravin et al. 2003) with the ST algorithm again yielded ATGCAT as an over-represented motif, with an exact match in 13 sequences. We also scanned the 1000-bp upstream regions of these *Drosophila* miRNA genes for motifs enriched in core promoters of protein-coding genes (Ohler et al. 2002), but did not detect a consistent preference for any of the known motifs.

Analysis of downstream sequences and foldbacks

Next, we investigated whether candidate termination signals could be identified by the approach described above. A search for over-represented oligonucleotides in the conserved blocks of the DSS using the ST algorithm did not identify a single statistically significant motif. Because the alignment algorithms require colinearity of sequences, conserved motifs might be missed if their positions were poorly conserved. Applying MEME to the complete DSS identified the motifs TTTT[TG]GAAA in *C. elegans* (E-value 1.7e-5) and TTTYYGAAA in *C. briggsae* (E-value 2.2e-6). Although instances of these motifs were found in all of the downstream sequences, there was no apparent positional conser-

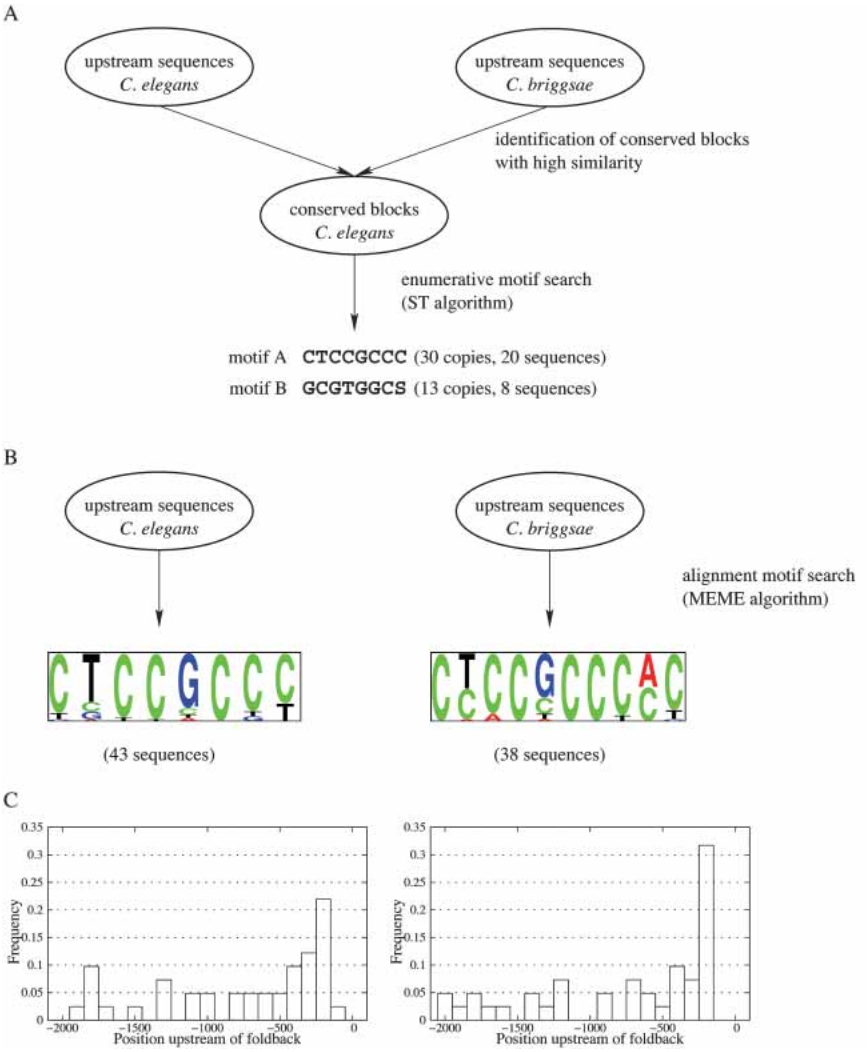

FIGURE 2. Identification of conserved upstream sequence elements. (*A*) Enumerative search for over-represented 8-mers within conserved upstream regions. Next to each consensus sequence is the number of instances of this sequence in conserved *C. elegans* blocks allowing for zero or one mismatch to the consensus or its reverse complement, and the number of distinct upstream sequences containing these instances. The Z-score of the consensus motif A was 29.0, the score of motif B was 14.7. As a control, a search in equally sized, randomly generated sequences delivered a Z-score of 11.2. (*B*) Application of the MEME local alignment algorithm to the complete 2000-bp upstream sequence sets. Shown are the pictograms (http://genes.mit.edu/pictogram.html) computed from the sequences that were used in the alignment by MEME for *C. elegans* (E-value of 3.0e-24) and *C. briggsae* (E-value of 1.5e-37). Both methods identify a highly similar motif as the most significant one. (*C*) Histograms of the locations of the best hit per sequence to the motifs given in *B*, in bins of 100 bp.

vation of the best hit in orthologous loci (data not shown). A similar motif was described previously in *C. elegans* introns (Fig. 2 in Lim and Burge 2001), and we also observed similar motifs downstream of protein-coding genes (data not shown). The common occurrence of this motif in introns argues against a role in transcriptional termination. Together, its relatively low statistical significance and ubiquitous distribution suggested that presence of this motif would not be a useful discriminatory feature for miRNA gene finding.

The polyadenylation-related motif AAAWTRAAA (Brown 2002) was the most significant motif computationally identified downstream of *C. elegans* protein-coding genes using MEME. No similar motif was identified in the sequences downstream of independently transcribed miRNAs. Therefore, although a subset of miRNA primary transcripts could be polyadenylated, polyadenylation does not appear to be a general feature of *C. elegans* miRNA transcripts. The absence of an identifiable polyadenylation signal does not rule out the possibility of pol-II-driven transcription, because other RNA genes, such as yeast snoRNAs, are derived from nonpolyadenylated pol-II transcripts (Steinmetz et al. 2001).

Finally, we examined the sequences around *C. elegans* miRNA foldbacks to search for candidate elements involved in the recognition and processing of the foldback from the primary transcript. As known foldbacks in polycistronic clusters are located immediately adjacent to one another, we restricted the search to ± 15 bases around the start and end of the foldback. No significant motif was identified, suggesting that the processing of the foldbacks is driven more by their secondary structure than by any conserved sequence. This conclusion is consistent with recent biochemical studies of pri-miRNA recognition and processing (Lee et al. 2003).

Improvement in microRNA gene finding

Previous approaches for the computational identification of miRNA genes have focused only on the stem–loop portion of the genes (Ambros et al. 2003b; Grad et al. 2003; Lai et al. 2003; Lim et al. 2003a,b). Our previous efforts started with conserved 110-nt genomic segments that were predicted to form stem–loops and did not fully overlap with protein-coding regions (Lim et al. 2003a,b). After passing an initial threshold on secondary structure similarity, the foldbacks were ranked using the program MiRscan. MiRscan evaluates miRNA candidates by sliding a 21-nt window along each arm of the foldback and assigning log-odds scores for seven features: base pairing of the candidate to the other arm of the stem, base pairing in the remainder of the stem–

loop structure, conservation of the 5′ and 3′ halves of the candidate miRNA, distance of the 21-nt window from the terminal loop, symmetry of the internal loops and bulges, and the sequence of the initial pentamer (Lim et al. 2003b). Overlapping 110-nt segments from both strands were then merged, and the higher scoring candidates were carried forward.

The observed upstream sequence motif and the patterns of sequence conservation flanking the stem–loop portion of the miRNA genes motivated us to develop an improved miRNA gene-finding algorithm, which we call MiRscanII. From here on, the previous version will be referred to as MiRscanI when needed for clarity. For the identification of independently transcribed *C. elegans* miRNA genes, we included three additional features as follows: (1) the score of the best hit to the *C. elegans* motif A within 1000 bp upstream of the predicted stem–loop; (2) the percentage of sequence contained in conserved blocks with >80% identity in the 1000 bp upstream of the stem–loop; and (3) the percentage of sequence contained in conserved blocks within 1000 bp downstream of the stem–loop. Log-odds scores for these features were derived from the MiRscanI training set of 50 conserved nematode miRNAs (Lim et al. 2003b), and these scores were simply added to the MiRscanI log-odds scores to give MiRscanII scores. The scores range from −3.3 to +2.0 bits for feature 1, −2.0 to +1.6 bits for feature 2, and −1.4 to +0.9 bits for feature 3.

MicroRNA candidates located on the sense strand of introns in protein-coding genes were not scored with these new features, but were instead filtered on the basis of their conserved genomic context. We observed that 11 of 13 known miRNAs in this group were located in introns of orthologous host genes as annotated in the Ensembl database (Clamp et al. 2003). For one of the two exceptions, the *C. briggsae* miRNA was located just downstream of the annotated orthologous gene, and in the remaining case, no ortholog was annotated. Thus, we kept only those foldbacks that were situated within, or at most 5000 bp from the *C. briggsae* ortholog, or for which no ortholog was annotated.

We included four additional filtering steps to eliminate the following types of unlikely candidates that had been scored in our previous effort: (1) candidate stem–loops that were located within extremely short intergenic regions between genes transcribed in opposite directions (<100 nt to each gene); (2) candidates on the antisense strand of an intron, where one end is too close to a splice site, leaving insufficient room for promoter or terminator sequences; (3) independent candidates with no upstream or downstream conservation whatsoever; and (4) candidates that overlapped an exon by >50 bp. Previously, all foldbacks were kept if they overlapped at all with noncoding sequence.

The third filter was the only one for which a known miRNA gene (*mir-238*) was lost. A possible explanation is that the *C. briggsae* ortholog assigned by BLAST in our procedure was not the true ortholog. The fourth criterion eliminated a surprisingly large number of candidates (~7000), implying that many exons overlap conserved secondary structures. The minimal overlap of 50 bp ensures that at least one arm of a miRNA stem–loop is located within an intron, and there is one case (*mir-62*) where one arm of a known miRNA stem–loop overlaps with an exon of a nearby gene (T07C5.1) on the sense strand in both species. Assuming that this portion of the pre-mRNA is not alternatively spliced, *mir-62* processing would be expected to compete with splicing, producing either the coding sequence or the miRNA foldback.

A flowchart of the filtering and rescoring of candidate foldbacks is shown in Figure 3. To allow a direct comparison, the same set of sequence windows was used as in our previous study. Of ~43,000 foldbacks obtained from align-

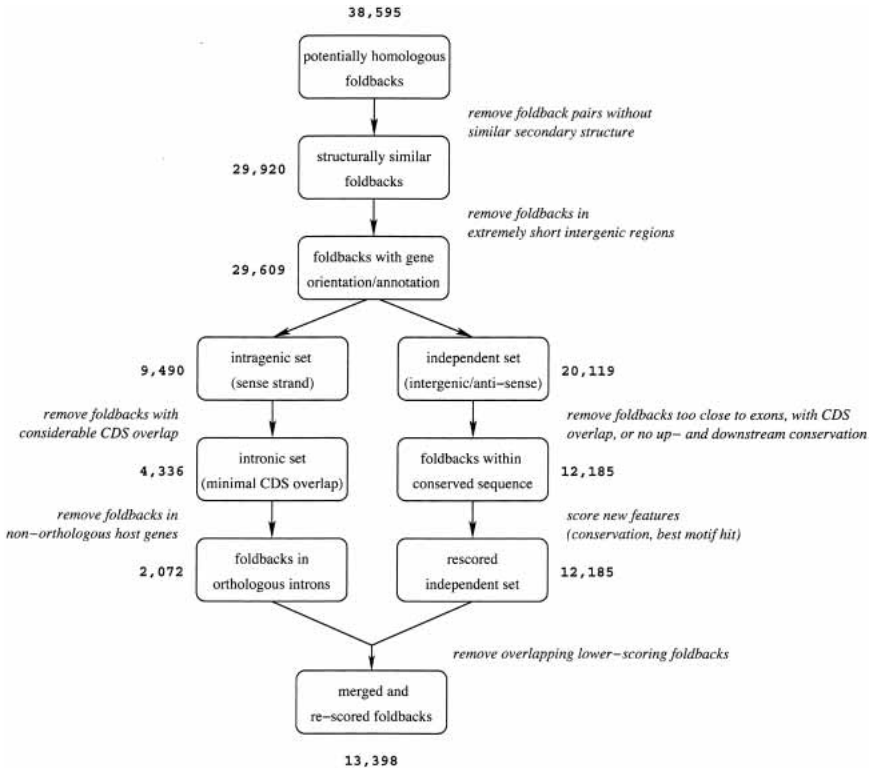

FIGURE 3. Flowchart of filtering and rescoring of candidate foldbacks with MiRscanII. Input was the set of conserved foldbacks that had received scores by MiRscanI. The numbers show how many candidates passed each step.

ments of the *C. elegans* genome with *C. briggsae* shotgun reads, ~35,700 passed the initial threshold on secondary structure similarity in MiRscanI, and were merged to ~28,000 nonoverlapping sequence windows. We realigned these ~43,000 foldbacks to the assembled *C. briggsae* genome sequence, recovering ~38,600 alignments. Of these, ~29,900 passed the secondary structure filter. All of the miRNAs previously scored by MiRscanI, as well as all previously tested candidates, were in this smaller set. After the additional filtering steps described above, the set of ~38,600 candidates was narrowed down to a mere ~13,400, as compared with ~28,000 previously (Fig. 3).

Compared with the previous analysis, the mode of the MiRscanII score distribution shifted from −4 to −9, and the score range expanded from [−28,18] to [−30,23] (Fig. 4; Lim et al. 2003b). Of the 86 miRNAs cloned and/or detected by Northerns in our previous study, 77 are scored by MiRscanII. The average score of these miRNAs increased by 0.9 bits when adding the new features, whereas the average score of all ~13,400 foldbacks decreased by 1.3 bits. In total, 73 miRNAs scored higher than nine bits, whereas four received low or negative scores. The remaining nine were not scored, either because a *C. briggsae* homolog was not identified by our automated methods or did not pass the folding free energy threshold (eight genes), or because flanking conservation was lacking (one gene).

The additional filters combined with the additional scoring features appear to have substantially increased the speci-

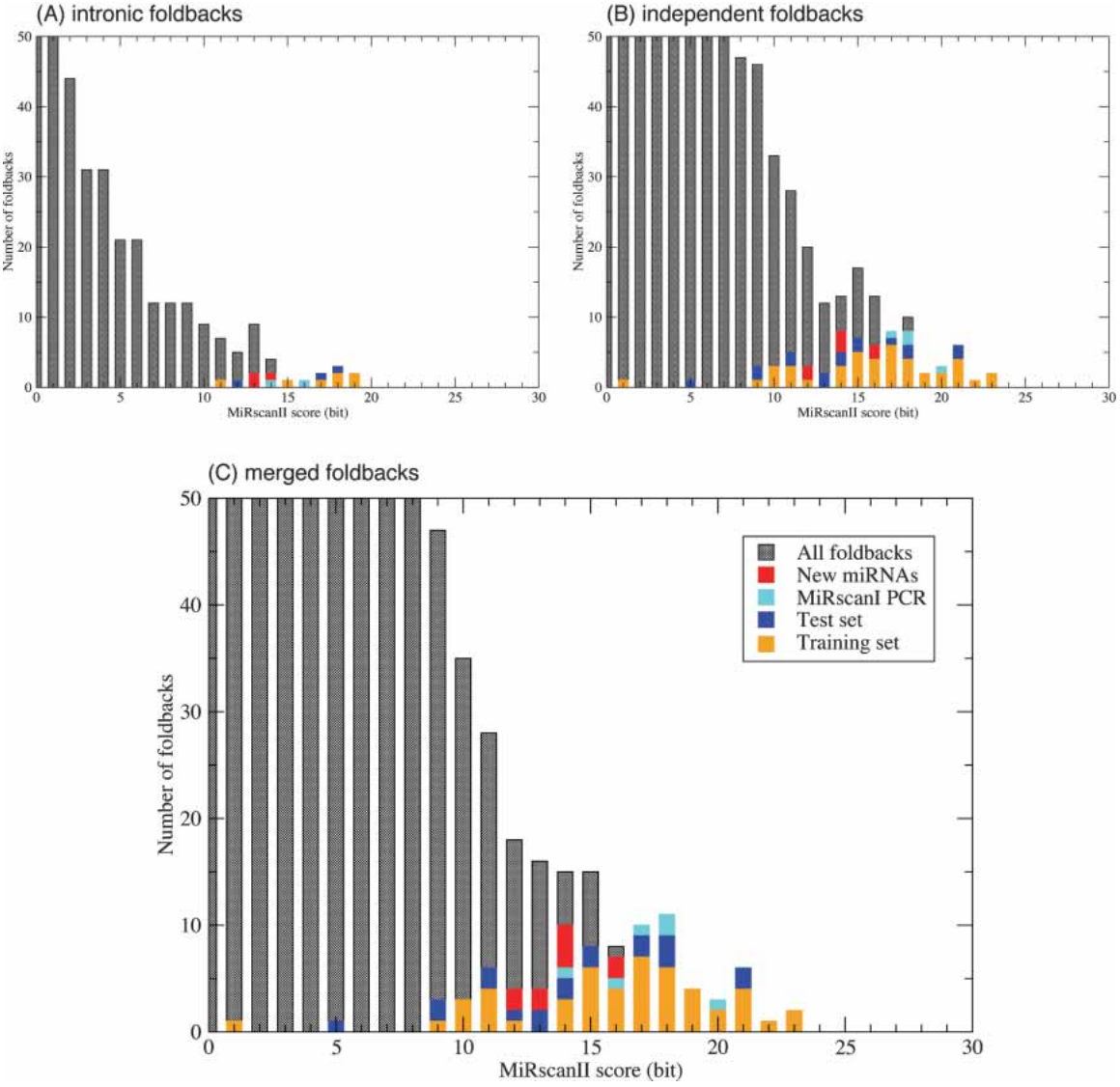

FIGURE 4. Histograms of MiRscanII scores greater than zero (nonsyntenic analysis). (*A*) Intronic foldbacks. (*B*) Independent foldbacks. (*C*) Merged set of 13,398 foldbacks. The training set (orange), test set (dark blue), previously verified MiRscanI predictions (light blue), and newly verified MiRscanII predictions (red) are marked in color. The score distributions were truncated at 50 foldbacks on the y axis. The scores of one miRNA gene in the training set (*mir-59*) was negative, and thus is not shown. Each bin covers a score range of one bit, e.g., the bin labeled 15 includes candidates with scores between 15 and 16 bits.

ficity of MiRscanII. In the MiRscanI analysis, 35 candidate miRNAs scored higher than the median score of cloned *C. elegans* miRNAs at the time, but 19 of these could not be confirmed by additional cloning or Northern blots (Lim et al. 2003b). Only seven of the 19 were left in the MiRscanII set of 13,400 rescored candidates, suggesting that the other 12 candidates are not, in fact, miRNAs. A total of 42 of 88 known miRNAs score higher than any unconfirmed sequence.

Because miRNAs are sometimes found within clusters, the relative position of miRNA candidates can provide a means of computationally identifying new genes, as was first shown in the prediction of *mir-39* and *mir-65*, two genes that were subsequently validated by expression analysis and/or cDNA cloning (Lau et al. 2001; Lim et al. 2003b). With this in mind, we scanned all candidates for their proximity to other candidates and retained those that were <1000 bp from each other, requiring a minimum score of five bits for all and eight bits for at least one candidate in such a potential cluster set. This simple algorithm recovered all known *C. elegans* miRNA clusters, and identified two additional potential clusters with two members each.

Experimental verification of new candidate miRNAs

Because MiRscanII more clearly distinguished previously identified miRNAs from other candidates, it was practical to examine new candidates with scores below the median of the test set. All unverified predictions that scored higher than the 43rd percentile of the test set miRNAs (12.7 bits) were subject to experimental screening. This set consisted of 35 new candidates plus six candidates that had not been detected in the previous attempt to validate computational candidates by Northern blotting (Lim et al. 2003b). One of the clusters, which resulted from the cluster analysis described above, was part of this set due to high scores. The other cluster was additionally included, giving a total set of 43 candidates that were subject to experimental testing by PCR and subsequent cloning and sequencing to confirm the identity of the amplified product.

With this approach, we verified 10 miRNA candidates (Table 1), two of which, in retrospect, had been previously identified. One corresponded to miR-259, recently reported in (Ambros et al. 2003b) based on the perfect conservation of the miRNA in *C. briggsae* combined with its detectable expression on Northern blots. Our PCR-sequencing validation defined the terminus of this miRNA, which turns out to be shifted by 3 nt from the previously proposed position. The second previously identified miRNA in the set of 10 new validations was miR-239b, which had been previously proposed to be a homolog of miR-239a, but not experimentally verified (Lim et al. 2003b). Interestingly, sequencing of the PCR product from the miR-239b amplification revealed that the miRNA had a different 5′ terminus than that seen for all four of the miR-239a clones. It was one nucleotide shorter on the 5′ end, providing evidence that a second *mir-239* gene was indeed expressed and that the primer was preferentially hybridizing to miR-239b, rather than to miR-239a. Had we not seen this difference in the 5′ termini of the miR-239 RNAs, it would have been difficult to argue against the possibility of primer cross-hybridization. Among the eight confirmed miRNAs that had not been previously proposed were two clustered candidates. The other clustered pair of candidates, which had scores lower than the 43rd percentile, was not validated by our PCR-sequencing protocol. Two of the eight newly identified

TABLE 1. Experimentally verified *C. elegans* miRNA candidates

miRNA	Sequence	Chr	Location	Arm	Sim
miR-353	caauugccauguguugguauu	I	intron of D1007.12 (s)	5′	+
miR-354	accuuguuuguugcugcuccu	I	intron of Y105E8A.16 (s)	3′	+++
miR-355	uuuguuuuagccugagcuaug	II	1 kb ds of T27D12.3	5′	+++
miR-356	uugagcaacgcgaacaaauca	III	intron of ZK652.2 (s)	5′	++
miR-357	uaaaugccagucguugcagga	V	0.6 kb us of C10B5.1	3′	+
miR-358	caauugguaucccugucaagg	V	0.9 kb us of C10B5.1	3′	+
miR-359	ucacuggucuuucucugacga	X	0.5 kb ds of Y41G9A.6	3′	+
miR-360	ugaccguaaucccguucacaa	X	0.5 kb us of Y23B4A.2	3′	+++
miR-392	uaucaucgaucacgugugaug	X	1.0 kb us of F54B11.5	3′	+
miR-239b	uuguacuacacaaaaguacug	X	7.0 kb us of C34E11.1	5′	++
miR-259	aaaucucauccuaaucuggua	V	1.2 kb us of F25D1.4	5′	+++
lsy-6 miRNA	uuuuguaugagacgcauuucg	V	0.5 kb us of C32C4.3	3′	++

The first nine rows show newly identified genes, the last three rows show the revised sequences for the successfully cloned, but previously described miRNAs miR-239b, miR-259, and *lsy-6*. The miRNAs are shown as 21-nt RNAs, but their actual length is generally not known because the PCR assay and sequencing validation determined the 5′ but not 3′ termini. The exception is the *lsy-6* miRNA for which the 21-nt length was deduced from the 5′ terminus of the miRNA* and assuming Drosha processing leaving a 2-nt 3′ overhang. For miR-358 and miR-360, some of the observed clones showed 5′ ends shifted by 2 nt toward the 3′ end. "Arm" denotes the side of the foldback on which the miRNA is located. The level of similarity (sim) with the miRNAs in the *C. briggsae* foldbacks are shown as +++ (100%), ++ (>90%), and + (>75%). For predicted stem–loops, see Supplementary Material at http://genes.mit.edu/burgelab/MiRscanII. (us) Upstream; (ds) downstream; (s) sense

miRNAs appear to be distant paralogs of previously identified *C. elegans* miRNAs; miR-357 and miR-356 have 5′ homology with miR-232 and miR-233, respectively.

DISCUSSION

Conserved and nonconserved miRNAs—the limitations of current computational approaches

Like other computational miRNA gene finders, MiRscan misses genes that lack detectable homologs in related species. The observation that clear *C. briggsae* homologs were not readily found for 12 genes known at the start of this study (eight genes without MiRscanII scores, and four genes with low scores; Lau et al. 2001; Lim et al. 2003b) does not imply that these 12 miRNAs lack homologs in *C. briggsae*. Our previous analysis showed that 10 of these 12 miRNAs were related to other *C. elegans* miRNAs, which, in turn, had easily identifiable orthologs in *C. briggsae*, leaving only two miRNAs without an identifiable homolog (Lim et al. 2003b). Nonetheless, because of extensive divergence within certain of the families of paralogous genes, some of the *C. elegans* genes are not matched with their proper *C. briggsae* homologs in the BLAST searches of our automated analysis. A manual investigation of syntenic regions illustrated this limitation of our automated search for homologous miRNA stem–loops. Among the four related miRNA genes in a cluster on chromosome III, one (*mir-65*) had been matched to a homolog in *C. briggsae*, whereas three (*mir-64*, *mir-66*, and *mir-229*) had not been. A closer look at the syntenic locus in *C. briggsae* revealed two additional foldbacks flanking the previously identified *mir-65* ortholog. The putative miRNAs of these *C. briggsae* foldbacks matched residues 2–15 and 2–12 of the *C. elegans* miR-64 and miR-66 miRNAs. The other interesting case concerned *mir-72*, which was among the foldbacks with negative MiRscanII scores, and for which no orthologous foldback in *C. briggsae* had been previously reported. Inspection of the *C. elegans* locus showed that an alternative foldback structure, which placed miR-72 on the 5′ instead of the 3′ arm, was energetically more favorable than the structure proposed previously (Lau et al. 2001). An analysis of the syntenic *C. briggsae* region revealed a homologous foldback that resembled the revised *C. elegans mir-72* foldback, except that it had an extra stem protruding from near the terminal loop of the *C. briggsae* structure (see Supplementary Fig. 2 at the MiRscanII Web site, http://genes.mit.edu/burgelab/MiRscanII). This extra stem is reminiscent of that seen in the *C. elegans mir-229* foldback (see Supplementary Fig. 1 at http://genes.mit.edu/burgelab/MiRscanII; Ambros et al. 2003b; Lim et al. 2003b).

Computational miRNA prediction in other animals (Lai et al. 2003; Lim et al. 2003a) has utilized whole-genome alignments (WGAs) of related species to restrict the search space for conserved foldbacks. At the time MiRscan was developed, the *C. briggsae* genome was only available in the form of short sequence reads, so there was no choice but to use BLAST searches of the reads to identify homologous foldbacks. To enable a direct comparison between the old and new versions, we decided to start with these foldbacks and realign them to the *C. briggsae* contigs. We therefore used the annotation of orthologous protein-coding genes to restrict the number of initially determined candidates, instead of starting from WGAs. The conservation of intronic miRNAs in orthologous host genes turned out to be a useful step for filtering of potential candidates (Fig. 3). We also explored the filtering of independently transcribed candidates in a similar manner (data not shown). First, we determined the *C. briggsae* orthologs of the closest flanking protein-coding genes. If both of these were located in the same *C. briggsae* contig, but the *C. briggsae* best match to the foldback under consideration mapped to a different contig, the foldback was eliminated. Of the 12,185 independent foldbacks (cf. Fig. 3), 45% did not pass this test, thus greatly reducing the number of candidates. Among those not passing the filter were four previously tested candidates that could not be verified by PCR and sequencing. However, this filter also eliminated five known miRNAs, including one of the newly identified ones, for which the corresponding *C. briggsae* sequence was not part of the same contig as the closest orthologous protein-coding genes. We checked all BLAST hits of these five miRNAs to the *C. briggsae* genome above the E-value threshold of 1.8 in more detail. In one case, the sequence in the syntenic *C. briggsae* location between the protein-coding genes also showed weak similarity to the foldback and might have been part of a WGA. The other four genes had no detectable similarity in the syntenic locus and would have been missed by a WGA, illustrating potential pitfalls of this approach, which can be confounded either by misassemblies, unusual rearrangements, or the selective loss of paralogs in different species. Lai et al. (2003) also reported that one of the first 24 *Drosophila* miRNAs to have been cloned was not part of the fly WGA, even though it was detectable by BLAST.

The above analysis suggested that a strict requirement for synteny would lower the sensitivity of the analysis, but we expected that demanding synteny would still be useful for increasing its specificity and might even lead to identification of a few additional miRNAs in cases for which a fortuitous BLAST hit to a nonsyntenic locus obscured the identification of the orthologous foldback pairs. The recent publication of the *C. briggsae* genome (Stein et al. 2003) contained a collection of 4837 syntenic blocks, that is, regions of long-range colinearity between the genomes of *C. elegans* and *C. briggsae*, allowing us to reconsider the syntenic analysis in an alternative fashion. In total, these blocks covered 84.6% of the *C. elegans* and 80.8% of the *C. briggsae* genome. We repeated the complete MiRscanII analysis, this time restricting BLAST to match potentially homologous hairpins within the syntenic blocks only. We used the *C. elegans* sequence from release 77 of WormBase, because the

synteny coordinates were given with respect to this release. Of the 88 *C. elegans* miRNA foldbacks from Lim et al. (2003b) and the additional nine that we newly verified, 92 were contained in the syntenic blocks, and 81 were part of the final set of foldback pairs scored by MiRscanII, compared with 87 when we did not require synteny. This analysis yielded 17 foldback pairs with scores higher than 12.7 bits (our cutoff for experimental validation) that were not previously considered. The scores of five of these foldback pairs differed only slightly from the nonsyntenic analysis; their scores were pushed above the 12.7-bit threshold because of slight score fluctuations resulting from an independent analysis using a different genome assembly. When these 17 candidates were subject to experimental verification using our PCR-sequencing assay, only one, miR-392, was verified (Table 1). Even this one case did not result from the use of syntenic alignments; instead, sequence differences in the *C. elegans* genome versions used for the original MiRscan and the syntenic analysis led to an improvement in the *mir-392* foldback score.

Overall, demanding synteny for independently transcribed candidates provided essentially no improvement in MiRscan efficacy, and decreased the sensitivity of our approach without leading to the identification of any new genes missed by simply using the top BLAST hit in the genome, irrespective of its location. Nonetheless, considering synteny would likely provide substantial benefit to computational approaches with lower inherent specificity or to the application of MiRscan to more complex genomes.

A consideration of other recently reported miRNAs

Two other publications (Ambros et al. 2003b; Grad et al. 2003) have recently reported newly identified nematode miRNAs. Ideally, these could serve as additional independent test sets to assess the sensitivity of MiRscanII. Of the seven miRNAs uniquely reported in Ambros et al. (2003b), we found one (miR-259) by computational analysis and PCR sequencing. The other six (miR-256, miR-257, miR-258, miR-260, miR-261, and miR-262) are reportedly not conserved across species, and thus, were not in the initial set of foldbacks scored by MiRscanII. For so many of these newly reported genes to lack homologs in *C. briggsae* was unexpected, because *C. briggsae* homologs could be identified for all but two of the first 80 miRNAs cloned from *C. elegans* (Lau et al. 2001; Lee and Ambros 2001; Lim et al. 2003b). One possibility is that some have homologs, but these happen to fall in portions of the *C. briggsae* genome that have not yet been sequenced. To assess how MiRscanII would score these six miRNAs in the event that a homolog was eventually found, we applied the program to pairs of identical *C. elegans* sequences, assuming the best possible scenario of perfect conservation. Still, only two candidates scored above our experimental cutoff of 12.7 bits. This observation indicated that, conservation aside, most of these uniquely reported miRNAs have features that are atypical of classical miRNAs. These features include an unusually long distance between the miRNA and the loop and less base pairing flanking the miRNA. Although not considered when originally formulating the criteria for miRNA annotation (Ambros et al. 2003a), base pairing flanking the miRNA is now known to be important for the nuclear processing of human primary miRNA transcripts by the enzyme Drosha (Lee et al. 2003). Because Drosha is conserved in nematodes and other metazoa, similar pairing is likely to be required in *C. elegans*.

The cloning effort that identified miR-256, miR-257, miR-258, miR-260, miR-261, and miR-262 also identified 33 unique tiny noncoding RNAs (tncRNAs), which differ from miRNAs in that they are not evolutionarily conserved, do not have the potential to be derived from miRNA-like precursors, and often begin with a G (Ambros et al. 2003b). With their lack of *C. briggsae* conservation and their atypical hairpin structures, a case could be made that most of these six uniquely reported miRNAs are instead tncRNAs, that is, they comprise the few tncRNAs that happened to have fortuitous potential pairing to flanking genomic sequence that was sufficient to satisfy the guidelines at that time for classification as miRNAs. Most of these six RNAs are also similar to the tncRNAs in another important aspect; their expression requires particular proteins of the RNAi pathway not generally needed for miRNA expression, further indicating that most of these six would be more accurately classified as tncRNAs (V. Ambros, pers. comm.).

None of the validated MiRscanII candidates matched the 10 miRNAs uniquely reported by Grad et al. (2003), which were assigned names *cp-miR-264* to *cp-miR-273*, where cp stands for computationally predicted. With the exception of *cp-mir-268*, none of the cp-miRNA foldbacks have easily identified *C. briggsae* orthologs. Two (*cp-mir-264* and *cp-mir-272*) have atypical foldbacks, as revealed by their poor MiRscan scores when compared against themselves. The eight remaining cp-miRNAs were initially found as homology candidates, that is, *C. elegans* hairpins that had segments with loose sequence similarity to previously known mature animal miRNAs, usually miRNAs of *C. elegans*. One possibility is that these foldbacks are distant paralogs of *C. elegans* miRNAs, not all of which might be conserved between species. Another possibility is that some of these foldbacks are in fact not miRNA genes, even though their authenticity was supported by a PCR assay (Grad et al. 2003). The PCR verification protocol used was less stringent than ours because it used the complete miRNA 21mer as a primer and lacked an additional sequence-verification step. Without this additional step, we would have counted an additional 10 of our 43 candidates as new miRNAs because they resulted in clear bands of the right size (35–45 nt). However, they did not pass the subsequent sequence-verification test. cp-miR-268, which received a score above our cutoff for experimental validation, was one of our candi-

dates with a PCR band that did not pass the sequence-verification test. Further supporting the idea that some cp-miRNAs are not authentic paralogs is the observation that in three cases (cp-miR-267, cp-miR-268, and cp-miR-271), the presumed mature miRNA resides on the opposite arm of the foldback when compared with the presumed paralog (miR-55, miR-73, and miR-35, respectively). Of the five remaining foldbacks, *cp-mir-266* and *cp-mir-273* look the most promising, in that each has additional sequence similarity with its presumed paralog (*mir-72* and *mir-56*, respectively) that falls outside of the mature miRNA in a pattern that might be expected for authentic paralogs. In addition, cp-miR-269 can be regarded as a paralog of cp-miR-266, as they differ by only three nucleotides.

The recent discovery of the *lsy-6* miRNA gene (Johnston and Hobert 2003), which appears to be expressed in only eight cells of the adult nematode, raises the question as to whether our strategy of computational prediction and large-scale cloning might lack the sensitivity to detect this and similar cases. The reported *lsy-6* foldback pair scored 9.91 bits with MiRscanII, including a positive contribution of the upstream motif described in this study. Our computational pipeline also included the opposite strand of the *lsy-6* locus, which scored slightly better (10.27 bits), including a negative contribution of motif A, because the orientation was incorrect. This score was at the 29th percentile of our test set, and therefore not high enough to be included in the set targeted for experimental verification. To determine whether we would have been able to validate the *lys-6* gene if we had tested candidates down to the 29th percentile, we applied our PCR-sequencing assay and detected the *lsy-6* miRNA, showing that this assay is sufficiently sensitive to detect a miRNA expressed in only a few cells of the animal. The assay also detected the *lsy-6* miRNA* arising from the opposite arm of the hairpin and presumably present at even lower abundance in our library of small RNAs. These RNAs had not been detected previously, and the sequencing of their 5′ termini performed in the course of the assay enabled us to define the mature lsy-6 miRNA (Table 1). In summary, *lsy-6* is one of the anticipated miRNAs with a score somewhat below our current cutoff for experimental tests, but not otherwise unusual, and can be readily detected by the PCR-sequencing assay despite its restricted expression.

The estimated number of miRNA genes in *C. elegans*

Starting from MiRscanI predictions, we previously estimated that there were at least 93, but no more than ~120 miRNA genes in *C. elegans* (Lim et al. 2003b). The identification of additional miRNA genes, together with the increased specificity of MiRscanII, allows us to revisit these estimates. The 88 miRNA loci listed in our previous study and the 11 miRNA genes of Table 1 not present in the previous list add up to 99 unique loci. Nineteen of these were not among our 3423 sequenced miRNA clones (Lim et al. 2003b), and instead, were primarily identified by experimentally verifying MiRscan predictions. We attempted to validate only those MiRscan candidates with scores above the 43rd percentile of the miRNAs in our test set, and all of these 19, with the exception of lsy-6, scored higher than the threshold. It is therefore reasonable to assume that these 18 miRNA genes include no more than 57% of the miRNA genes not represented among our 3423 clones. This implies that at least another 12 genes resembling the lsy-6 miRNA have escaped our detection or validation efforts, because they either have no MiRscan scores or low scores. Thus, the current analysis enables the estimated lower limit on the number of miRNA genes in *C. elegans* to be revised upward to 99 + 12, or 111.

An upper limit of ~120 *C. elegans* miRNA genes was originally estimated by considering the number of MiRscanI candidates (validated genes together with nonvalidated candidates) that had scores exceeding the median score of the cloned miRNAs (Lim et al. 2003b). Because the cloned miRNAs included miRNAs without recognizable *C. briggsae* homologs, this calculation took into account poorly conserved miRNAs without MiRscan scores. Furthermore, the absence of a correlation between the number of times an miRNA was cloned and its MiRscan score argued against the idea that there might be a disproportionate number of *C. elegans* genes that have escaped detection because they are both difficult to clone and difficult to identify computationally (Lim et al. 2003b). Our confidence in this upper limit increases with the improved specificity of the current analysis. For instance, there is now reason to suspect that eight of the unvalidated candidates used to calculate this upper bound of ~120 are false positives, in that these eight had too much exon overlap to be considered in the current analysis. However, we do not attempt to revise the estimate on the upper bound of *C. elegans* miRNA genes because of the danger of some overfitting in the current analysis. For example, the more complicated and bifurcating set of filters and scoring schemes of the current analysis (Fig. 3) made it less amenable to jackknifing, a procedure implemented earlier so that the scores of genes from the training set could be considered when estimating the upper bound on the number of genes (Lim et al. 2003b). Because the status of many of the miRNAs uniquely reported by Ambros et al. (2003b) and Grad et al. (2003) is in doubt, we did not consider these candidates when estimating the lower and upper bounds of gene numbers in *C. elegans*. Thus, our estimate of ~110 to ~120 miRNA genes in *C. elegans* would have to be revised upward if future experiments overturn the idea that most of these candidates are not authentic miRNAs. Finally, the MiRscan pipeline to detect conserved foldback pairs excluded foldbacks with extreme GC- or AT-content, and filtered out sets of highly repetitive foldbacks, the members of which overlapped with RepeatMasked sequences (Lim et al. 2003b). We are currently investigating the extent to which such foldbacks potentially harbor noncoding RNA products.

Analysis of conserved upstream sequence elements

Our analysis of sequences upstream of independently transcribed nematode miRNAs identified a conserved sequence element, motif A with consensus CTCCGCCC, which is highly specific and useful for miRNA gene identification. The transcription factor database TRANSFAC (version 6.0 public; Matys et al. 2003) contains only a handful examples of nematode transcription factors, and none of them matched motif A. A literature search also failed to turn up any previously reported similar nematode sequence motifs. At this point, it is open as to whether motif A is a transcription-factor binding site, whether it is a signal that directs an miRNA processing enzyme to the miRNA genes, or whether its function is possibly related to both of these alternatives. Recent studies have shown that there is considerable coupling between transcription initiation and mRNA processing, in which transcription factors assist in the direction of splicing factors to the nascent transcript (Maniatis and Reed 2002). One can easily envision an analogous scenario for efficient recruitment of factors responsible for recognition and processing of miRNA stem–loops.

We also identified a common enriched sequence element in vertebrates, CCCWCCC, which was different from that found in nematodes. A second enriched sequence element, ATGCAT, occurred in only a subset of vertebrate upstream sequences and was also found in a subset of *Drosophila* upstream sequences. According to TRANSFAC, Sp-1 and POU1F1, respectively, are likely candidates for transcription factors that bind to these motifs. Sp-1 is a ubiquitous transcription factor, which has been shown to activate transcription. The occurrence of multiple instances of the first motif is consistent with binding by Sp-1, which often binds to several sites per regulatory region (Courey et al. 1989). POU1F1 is a growth hormone factor that contains one POU and one homeobox domain and also acts as a transcriptional activator (Lefevre et al. 1987). POU1F1 is not conserved in *Drosophila*, but other members of the same family of POU-homeobox-containing transcription factors with potentially similar binding preferences are present.

In none of the organisms under consideration—nematodes, arthropods, and vertebrates—were we able to identify strong motifs reminiscent of known eukaryotic core promoter sequence elements such as the TATA box. Even in the case of *Drosophila*, where a recent study has extended the set of motifs prevalent in core promoters, and reliable computational tools for pol-II transcription start site prediction are available (Ohler et al. 2002), no clear picture emerges at this point. Therefore, miRNA promoters do not share a common layout, but instead appear to be highly variable, as is characteristic of protein-coding gene promoters. In another parallel to protein-coding genes, a recent study showed that sequence elements as far as 1000 bp or more upstream are required for specific activation of the *let-7* gene (Johnson et al. 2003).

In summary, our efforts showed that features distinct from RNA primary and secondary structure, such as upstream and downstream conservation and an upstream sequence motif, lead to a considerable improvement in gene-prediction accuracy for an important family of noncoding RNAs. Our improved method enabled us to identify nine new miRNA genes that had gone undetected, despite previous computation and large-scale cloning efforts. The set of known conserved nematode miRNAs is now approaching completeness, which should aid efforts to identify their target genes and to understand their roles in the *C. elegans* regulatory circuitry.

MATERIALS AND METHODS

Data sets

We constructed sets of orthologous upstream and downstream regions of independently transcribed miRNAs from a total set of 88 nematode miRNA genes (Lim et al. 2003b). First, we identified *C. elegans* miRNAs located in intergenic regions or on the antisense strand of introns, that are therefore likely to be transcribed independently of nearby protein-coding genes (WormBase annotation release 83). Next, we aligned the ~22-nt miRNA sequences to the assembled *C. briggsae* genome (July 2002) with BLAST (Altschul et al. 1997), retaining only those with >90% identity, that is, with no more than two mismatches. This stringent requirement should exclude the possibility of aligning upstream regions of related but nonorthologous miRNA genes. We then extracted up to 2000 bp upstream of both *C. elegans* and *C. briggsae* foldbacks for the Upstream Sequence Set (USS), and up to 1000 bp downstream for the Downstrean Sequence Set (DSS), excluding overlaps with annotated *C. elegans* genes. For miRNAs in clusters, only the regions upstream of the first miRNA were included in the USS, and only the regions downstream of the last miRNA were included in the DSS, leaving 43 miRNA pairs. For three *C. elegans* genes (*mir-45*, *mir-77*, and *mir-90*), two sequences in *C. briggsae* met all of the above requirements, and both were included in the analysis.

For training and evaluation of the revised model, we started from the same set of 88 miRNA genes. We used a training set of 50 sequences as described in our previous study (Lim et al. 2003b), excluding *mir-88* with an unknown processed miRNA sequence. The 24 miRNA genes newly cloned in the same study were kept as an independent test set. miRNAs that had not been cloned, but had been identified only by experimental validation of computational predictions, were excluded from both the training and test sets. Three miRNAs in the test set were not scored, because our automated procedure did not find an orthologous candidate foldback.

A set of 59 sequence pairs upstream of orthologous human–mouse miRNA genes (Lim et al. 2003a) were chosen in the same fashion as described for nematode miRNAs. Finally, 31 sequences upstream of independently transcribed *D. melanogaster* miRNAs according to the above criteria were taken from Aravin et al. (2003).

Alignment of upstream and downstream regions

We aligned the orthologous sequence pairs with the probabilistic sequence alignment tools BayesBlockAligner (BBA; Zhu et al. 1998) and Dynamic Block Aligner (DBA; Jareborg et al. 1999). Both programs have been specifically designed to identify short, highly conserved blocks in an alignment of two sequences, a pattern that can be expected in promoter sequences where transcription-factor binding sites are surrounded by stretches of nonconserved sequence. They perform a global alignment of two sequences, effectively ignoring stretches of unalignable sequences.

DBA uses a pair-hidden Markov model and computes the optimal alignment under a model of several match states corresponding to four different levels of conservation (with an average identity of 65%, 75%, 85%, and 95%). It requires colinearity of the two sequences, but allows for gaps within the conserved blocks. We retained blocks with at least 70% identity for the identification of motifs, and at least 80% for the feature computation in miRNA gene finding. The following parameter settings were used: block open probability 0.03, block close probability 0.98, gap probability 0.01.

BBA samples from the set of all possible alignments, covering a range of different substitution matrices and numbers of blocks. The output is the posterior probability that a specific position in one sequence is contained in an ungapped conserved sequence block with any position in the other sequence. In principle, these blocks are not required to be colinear. We considered all positions with posterior probability of at least 0.4 to be in an aligned conserved sequence block. We used PAM matrices from PAM5 to PAM30 in steps of 5 and base blocksize of 20.

In the case of multiple orthologs in *C. briggsae*, we merged the aligned blocks in *C. elegans* from all pairwise alignments. To avoid missing modestly conserved segments, we merged the output of both programs for the motif identification task. Because DBA and BBA deliver largely similar results, and the time complexity of the BBA algorithm is much higher, we restricted ourselves to DBA for the alignments scored by MiRscanII.

Two approaches for motif finding

We used an efficient implementation of the algorithm described by Sinha and Tompa (2000), here called the ST algorithm, which identifies statistically over-represented oligomers in a target set of sequences when compared with a background Markov chain model (H. Köstler, G. Stemmer, and U. Ohler, unpubl.). The algorithm uses a third-order Markov chain as a model for the background sequences and corrects for self-overlapping and complementary motifs. The motifs are composed of the standard A,C,G,T characters, but may also contain up to two ambiguous characters (N, S, W, R, Y). We retained all motifs with Z scores higher than a threshold obtained by a search in sequence sets of identical size, generated randomly with the same background distribution. We post-processed the resulting list of often highly similar significant oligonucleotides to determine how many distinct motifs were present. Details of this strategy to obtain motifs from lists of over-represented words have been given for a similar application elsewhere (Fairbrother et al. 2002).

We also used the probabilistic local alignment tool MEME (Bailey and Elkan 1995), with standard single-nucleotide frequencies as background, motif length 5–10 bases, and "zero or one occurrence" mode. MEME motif E-values refer to the expected number of motifs of the same width with equal or higher likelihood in the same number of random sequences with the same nucleotide composition as the considered set of sequences.

Parameter estimation for additional features scored in MiRscanII

We derived log-odds scores for the upstream and downstream features in the following way: (1) 1 kB upstream and downstream of the foldback window of 110 bp—or less, if an annotated exon was closer—were aligned with DBA. (2) From these blocks, we obtained the percentage of nucleotides contained in blocks of 80% or more sequence identity, and used these values as features representing upstream and downstream sequence conservation. (3) For the foldbacks that had passed the initial filter of containing at least some conservation (see Fig. 3), a discrete distribution was obtained by binning the feature values in intervals of five percentage points. (4) As the foreground distribution for true miRNAs was restricted to a small set of values, we took two measures to prevent overfitting to the scarce data and to allow for reasonable scores for foldbacks that might have features just outside the range of observed values. First, the discrete distributions of both foreground and background were smoothed with two iterations of a mean filter of width 3 bins, with 0.75 weight for the central value and 0.125 for the values to the left and right. By doing so, we spread a small amount of probability to unseen values adjacent to the range of observed values. As an example, if we saw 20%–40% conservation in the foreground sequences, this filter would extend the range of positive foreground values to 10%–50%. Next, we truncated the foreground and background distributions at the last foreground value with positive probability on both ends of the range. The background values at the low and high cutoffs were set to the sum of all bins below or above the cutoff, respectively. In our example, we would set the background value at the 10% bin to the sum of all values below and up to 10%, and the 50% bin to the sum of all values equal or higher than 50%. Thus, we do not rely on arbitrary scores for feature values in the range where we do not see any positive probabilities even after smoothing. (5) From these modified distributions, log-odds scores were computed as the base 2 logarithm of the ratio of foreground to background probability.

To judge the presence of the promoter motif, we used the tool patser-v3d (Hertz and Stormo 1999) to compute the score of the best hit within the 1-kb upstream sequence on either strand. From these values, discrete distributions for foreground and background were obtained using bins of 5 bits, and these distributions were smoothed and converted to log-odds scores as above. We also reapplied the above smoothing procedure for the set of seven features used by MiRscan, and used these slightly different parameter sets instead of the original ones.

PCR-sequencing assay

A PCR-sequence assay identical to the one described in Lim et al. (2003b) was performed to detect the sequences of predicted miRNAs within a cDNA library constructed from 18 to 26 nt RNAs. This library was the same as the one used for cloning (Lau et al.

2001). As specific primers, 17-nt-long sequences complementary to the 3′ ends of the predicted miRNAs were used, sometimes shifted by one or two nucleotides to prevent overlap with the primer to the generic 5′ adapter sequences in the library. In some cases, the algorithm might correctly identify a miRNA foldback, but predicts the wrong strand or the wrong side of the foldback as the location of the mature miRNA. To account for this possibility, a second primer was also tested, corresponding to the second highest score from either the other side of the foldback or the other strand of the sequence.

Following PCR amplification, the products were cloned and sequenced to ensure that no primer-dimers were obtained, and to verify that the nucleotides between the two primers indeed matched the corresponding genomic sequence. This step also identified the 5′ end of the miRNA; along with the greater sensitivity, this is a second advantage of this validation method compared with Northern blotting.

Primers for the successful reactions were as follows:

GCAATAATACCAACACA (miR-353),
AGGAGCAGCAACAAACA (miR-354),
ATTTGTTCGCGTTGCTC (miR-355),
CGAACTCCTGCAACGAC (miR-356),
TGAGACCTTGACAGGGA (miR-357),
CGTCAGAGAAAGACCAG (miR-358),
TTGTGAACGGGATTACG (miR-359),
AGCTCAGGCTAAAACAA (miR-360),
TCATCACACGTGATCGA (miR-392),
CCAGTACTTTTGTGTAG (miR-239b),
ACCAGATTAGGATGAGA (miR-259),
ATGATTTTGATACTAGA (*lsy-6* miRNA),
CATCGAAATGCGTCTCA (*lsy-6* miRNA*).

In all cases but miR-360 and *lsy-6*, the algorithm correctly identified the strand of the mature miRNA. In these two cases, the difference to the second highest score from the reverse strand was <0.4 bits.

Additional data files

Additional data files containing the following supplementary information are available through the Burge Lab Web site, http://genes.mit.edu/burgelab/MiRscanII: Supplementary Figure 1, foldback structures of newly identified miRNA genes; Supplementary Figure 2, foldback structure of the revised mir-72 locus; Supplementary Figure 3, examples of upstream alignments.

ACKNOWLEDGMENTS

We thank Harald Köstler and Georg Stemmer for work on the ST algorithm, Colleen T. Webb for sharing the data set of orthologous protein coding genes, Matthew W. Rhoades for help with the set of conserved mammalian miRNAs, the Genome Sequencing Center at the Washington University School of Medicine, St. Louis, and the Sanger Institute, Cambridge, UK, for sharing the assembled *C. briggsae* genome sequence before publication, N.C. Lau for providing the cDNA library of small *C. elegans* RNAs, and Victor Ambros for sharing unpublished results. This work was supported by grants from the NIH (C.B.B. and D.P.B.) and the Searle Scholars Program (C.B.B.).

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

Received October 14, 2003; accepted January 13, 2004; additional material accepted June 17, 2004.

REFERENCES

Abrahante, J.E., Daul, A.L., Li, M., Volk, L.M., Tennessen, J.M., Miller, E.A., and Rougvie, A.E. 2003. The *Caenorhabditis elegans* hunchback-like gene *lin-57/hbl-1* controls developmental time and is regulated by microRNAs. *Dev. Cell* **4:** 625–637.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Ambros, V. 2003. MicroRNA pathways in flies and worms: Growth, death, fat, stress, and timing. *Cell* **113:** 673–676.

Ambros, V., Bartel, B., Bartel, D.P., Burge, C.B., Carrington, J.C., Chen, X., Dreyfuss, G., Eddy, S.R., Griffiths-Jones, S., Marshall, M., et al. 2003a. A uniform system for microRNA annotation. *RNA* **9:** 277–279.

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

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

Argaman, L., Hershberg, R., Vogel, J., Bejerano, G., Wagner, E.G., Margalit, H., and Altuvia, S. 2001. Novel small RNA-encoding genes in the intergenic regions of *Escherichia coli*. *Curr. Biol.* **11:** 941–950.

Bachellerie, J.-P., Cavaillea, J., and Hüttenhofer, A. 2002. The expanding snoRNA world. *Biochimie* **84:** 775–790.

Bailey, T.L. and Elkan, C. 1995. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learn.* **21:** 51–83.

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

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

Brown, T.A. 2002. *Genomes II*. Wiley, New York.

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

Clamp, M., Andrews, D., Barker, D., Bevan, P., Cameron, G., Chen, Y., clark, L., Cox, T., Cuff, J., Curwen, V., et al. 2003. Ensembl 2002: Accommodating comparative genomics. *Nucleic Acids Res.* **31:** 38–42.

Courey, A.J., Holtzman, D.A., Jackson, S.P., and Tjian, R. 1989. Synergistic activation by the glutamine-rich domains of human transcription factor Sp1. *Cell* **59:** 827–836.

Eddy, S.R. 2001. Non-coding RNA genes and the modern RNA world. *Nat. Rev. Genet.* **2:** 919–929.

Fairbrother, W.G., Yeh, R.-F., Sharp, P.A., and Burge, C.B. 2002. Predictive identification of exonic splicing enhancers in human genes. *Science* **297:** 1007–1013.

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

Harris, T.W., Lee, R., Schwartz, E., Bradnam, K., Lawson, D., Chen, W., Blasier, D., Kenny, E., Cunningham, F., Kishore, R., et al. 2003.

WormBase: A cross-species database for comparative genomics. *Nucleic Acids Res.* **31:** 133–137.
Hertz, G.Z. and Stormo, G.D. 1999. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* **15:** 563–577.
Jareborg, N., Birney, E., and Durbin, R. 1999. Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. *Genome Res.* **9:** 815–824.
Johnson, S.M., Lin, S.Y., and Slack, F.J. 2003. The time of appearance of the *C. elegans let-7* microRNA is transcriptionally controlled utilizing a temporal regulatory element in its promoter. *Dev. Biol.* **259:** 364–379.
Johnston. R.J. and Hobert, O. 2003. A microRNA controlling left/right neuronal asymmetry in *Caenorhabditis elegans*. *Nature* **426:** 845–849.
Jones-Rhoades, M.W. and Bartel, D.P. 2004. Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. *Mol. Cell.* **14:** 787–799.
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., 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.
Lai, E.C., Tomancak, P., Williams, R.W., and Rubin, G.M. 2003. Computational identification of *Drosophila* microRNA genes. *Genome Biol.* **4:** R42.
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.
Lee, Y., Jeon, K., Lee, J.T., Kim, S., and Kim, V.N. 2002. MicroRNA maturation: Stepwise processing and subcellular localization. *EMBO J.* **21:** 277–284.
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.
Lefevre, C., Imagawa, M., Dana, S., Grindlay, J., Bodner, M., and Karin, M. 1987. Tissue-specific expression of the human growth hormone gene is conferred in part by the binding of a specific *trans*-acting factor. *EMBO J.* **6:** 971–981.
Lewis, B.P., Shih, I-H., Jones-Rhoades, M.W., Bartel, D.P., and Burge, C.B. 2003. Prediction of mammalian microRNA targets. *Cell* **115:** 787–798.
Lim, L.P. and Burge, C.B. 2001. A computational analysis of sequence features involved in recognition of short introns. *ProC. Natl. Acad. Sci.* **98:** 11193–11198.
Lim, L.P., Glasner, M.E., Yekta, S., Burge, C.B., and Bartel, D.P. 2003a. Vertebrate microRNA genes. *Science* **299:** 1540.
Lim, L.P., Lau, N.C., Weinstein, E.G., Abdelhakim, A., Yekta, S., Rhoades, M.W., Burge, C.B., and Bartel, D.P. 2003b. The microRNAs of *Caenorhabditis elegans*. *Genes & Dev.* **17:** 991–1008.
Lin, S.Y., Johnson, S.M., Abraham, M., Vella, M.C., Pasquinelli, A., Gamberi, C., Gottlieb, E., and Slack, F.J. 2003. The *C. elegans* hunchback homolog, *hbl-1*, controls temporal patterning and is a probable microRNA target. *Dev. Cell* **4:** 639–650.
Llave, C., Kasschau, K.D., Rector, M.A., and Carrington, J.C. 2002. Endogenous and silencing-associated small RNAs in plants. *Plant Cell* **14:** 1605–1619.
Maniatis, T. and Reed, R. 2002. An extensive network of coupling among gene expression machines. *Nature* **416:** 499–506.
Matys, V., Fricke, E., Geffers, R., Gossling, E., Hanbuck, M., Hehl, R., Hornischer, K., Karas, D., Kel, A.E., Kel-Margoulis, O.V., et al. 2003. TRANSFAC: Transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.* **31:** 374–378.
Moss, E.G., Lee, R.C., and Ambros, V. 1997. The cold shock domain protein LIN-28 controls developmental timing in *C. elegans* and is regulated by the *lin-4* RNA. *Cell* **88:** 637–646.
Ohler, U. and Niemann, H. 2001. Identification and analysis of eukaryotic promoters: Recent computational approaches. *Trends Genet.* **17:** 56–60.
Ohler, U., Liao, G.-C., Niemann, H., and Rubin, G.M. 2002. Computational analysis of core promoters in the *Drosophila* genome. *Genome Biol.* **3:** RESEARCH0087.1–12.
Park, W., Li, J., Song, R., Messing, J., and Chen, X. 2002. CARPEL FACTORY, a Dicer homolog, and HEN1, a novel protein, act in microRNA metabolism in *Arabidopsis thaliana*. *Curr. Biol.* **12:** 1484–1495.
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.
Sempere, L.F., Sokol, N.S., Dubrovsky, E.B., Berger, E.M., and Ambros, V. 2003. Temporal regulation of microRNA expression in *Drosophila melanogaster* mediated by hormonal signals and broad-Complex gene activity. *Dev. Biol.* **259:** 9–18.
Sinha, S. and Tompa, M. 2000. A statistical method for finding transcription factor binding sites. *ProC. Int. Conf. Intell. Syst. Mol. Biol.* **Vol. 8** 344–354.
Stein, L.D., Bao, Z., Blasiar, D., Blumenthal, T., Brent, M.R., Chen, N., Chenwalla, 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.
Steinmetz, E.J., Conrad, N.K., Brow, D.A., and Corden, J.L. 2001. RNA-binding protein Nrd1 directs poly(A)-independent 3′-end formation of RNA polymerase II transcripts. *Nature* **413:** 327–331.
Webb, C.T., Shabalina, S.A., Ogurtsov, A.Y., and Kondrashov, A.S. 2002. Analysis of similarity within 142 pairs of orthologous intergenic regions of *Caenorhabditis elegans* and *Caenorhabditis briggsae*. *Nucleic Acids Res.* **30:** 1233–1239.
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.
Yekta, S., Shih, I-H., and Bartel, D.P. 2004. MicroRNA-directed cleavage of HOXB8 mRNA. *Science* **304:** 594–596.
Zeng, Y. and Cullen, B.R. 2003. Sequence requirements for microRNA processing and function in human cells. *RNA* **9:** 112–123.
Zeng, Y., Wagner, E.J., and Cullen, B.R. 2002. Both natural and designed micro RNAs can inhibit the expression of cognate mRNAs when expressed in human cells. *Mol. Cell* **9:** 1327–1333.
Zhu, J., Liu, J.S., and Lawrence, C.E. 1998. Bayesian adaptive sequence alignment algorithms. *Bioinformatics* **14:** 25–39.