Cell

Large-Scale Sequencing Reveals 21U-RNAs and Additional MicroRNAs and Endogenous siRNAs in *C. elegans*

J. Graham Ruby,[1,2] Calvin Jan,[1,2] Christopher Player,[1] Michael J. Axtell,[1,4] William Lee,[3] Chad Nusbaum,[3] Hui Ge,[1] and David P. Bartel[1,2,*]
[1] Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA 02142, USA
[2] Howard Hughes Medical Institute and Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[3] Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
[4] Present address: Department of Biology, The Huch Institutes of the Life Sciences, Pennsylvania State University, University Park, PA 16802, USA.
*Contact: dbartel@wi.mit.edu
DOI 10.1016/j.cell.2006.10.040

SUMMARY

We sequenced ~400,000 small RNAs from *Caenorhabditis elegans*. Another 18 microRNA (miRNA) genes were identified, thereby extending to 112 our tally of confidently identified miRNA genes in *C. elegans*. Also observed were thousands of endogenous siRNAs generated by RNA-directed RNA polymerases acting preferentially on transcripts associated with spermatogenesis and transposons. In addition, a third class of nematode small RNAs, called 21U-RNAs, was discovered. 21U-RNAs are precisely 21 nucleotides long, begin with a uridine 5′-monophosphate but are diverse in their remaining 20 nucleotides, and appear modified at their 3′-terminal ribose. 21U-RNAs originate from more than 5700 genomic loci dispersed in two broad regions of chromosome IV—primarily between protein-coding genes or within their introns. These loci share a large upstream motif that enables accurate prediction of additional 21U-RNAs. The motif is conserved in other nematodes, presumably because of its importance for producing these diverse, autonomously expressed, small RNAs (dasRNAs).

INTRODUCTION

RNAs ~22 nt in length play gene-regulatory roles in numerous eukaryotic lineages, including plants, animals, and fungi (Bartel, 2004; Nakayashiki, 2005). The first endogenous ~22 nt RNAs discovered in eukaryotes were the *lin-4* and *let-7* RNAs, both of which were found by mapping mutant *C. elegans* loci (Lee et al., 1993; Reinhart et al., 2000). The mature *lin-4* and *let-7* RNAs are each processed from a hairpin formed within their respective primary transcripts. Through molecular cloning and sequencing, many small RNAs with the potential to arise from foldback structures characteristic of the *lin-4* and *let-7* hairpins were identified, including more than 50 from *C. elegans*, thereby establishing a class of endogenous RNAs called miRNAs (Lagos-Quintana et al., 2001; Lau et al., 2001; Lee and Ambros, 2001). Additional miRNAs have been identified in *C. elegans* by cloning, genetics, or computational prediction supported by experimentation (Ambros et al., 2003; Grad et al., 2003; Johnston and Hobert, 2003; Lim et al., 2003; Ohler et al., 2004).

In addition to the miRNA, a less abundant species known as the miRNA star (miRNA*) derives from the miRNA hairpin precursor (Lau et al., 2001; Lim et al., 2003). The miRNA and miRNA* species pair to each other with ~2 nt 3′ overhangs. In animals, this miRNA:miRNA* duplex is generated by the sequential action of Drosha and Dicer RNase-III endonucleases (Grishok et al., 2001; Hutvagner et al., 2001; Lee et al., 2003). Drosha cleaves at sites near the base of the stem, thereby liberating a 60–70 nt fragment comprising the majority of the hairpin, which Dicer then cleaves at sites near the loop (Lee et al., 2003; Han et al., 2006). The miRNA strand of the resulting miRNA:miRNA* duplex is then loaded into a silencing complex, which contains at its core a member of the Argonaute family of proteins (Hutvagner and Zamore, 2002; Mourelatos et al., 2002).

Once incorporated into the silencing complex, the miRNA serves as a guide to direct the posttranscriptional repression of protein-coding messages. Most important for target recognition is pairing to the miRNA seed, defined as the 6 nt segment comprising nucleotides 2–7, counting from the 5′ terminus of the miRNA (Lewis et al., 2003, 2005; Doench and Sharp, 2004; Brennecke et al., 2005). When comparing related miRNAs, the seed is also the most conserved portion of the RNA, and *C. elegans* miRNAs can be grouped into families based largely on their shared seed sequences (Ambros et al., 2003; Lim et al., 2003).

Other types of endogenous small RNAs have been found within libraries made from *C. elegans*. Those that are antisense to *C. elegans* mRNAs have been classified as small

interfering RNAs (siRNAs), with the idea that they might be processed from long double-stranded RNA (dsRNA) and might direct the silencing of complementary mRNAs (Ambros et al., 2003; Lim et al., 2003). Other cloned and sequenced ~22 nt RNAs do not appear to correspond to protein-coding regions and do not have the potential to arise from hairpins characteristic of miRNA precursors and yet are expressed at sufficiently high levels to be detected on RNA blots. These have been annotated as tiny noncoding RNAs (tncRNAs; Ambros et al., 2003). In flies and mammals, other distinct classes of small RNAs have been reported, including repeat-associated siRNAs (rasiRNAs; Aravin et al., 2003; Vagin et al., 2006) and Piwi-interacting RNAs (piRNAs; Aravin et al., 2006; Girard et al., 2006; Lau et al., 2006).

Recent advances in high-throughput sequencing technology have allowed for a more complete assessment of the global small RNA population in plants (Lu et al., 2005). Here, we applied high-throughput pyrosequencing methods (Margulies et al., 2005) to the discovery of small RNAs expressed in mixed-staged *C. elegans*. Our results reshape the list of known miRNAs by reporting newly identified miRNA genes, defining the processing of most previously annotated miRNAs, refining the termini of some, and raising new questions as to the authenticity of others. In addition, we describe thousands of endogenous siRNAs that appear to be RNA-templated products of activities acting preferentially on messages associated with spermatogenesis and transposons. We also describe the 21U-RNAs, which originate from an estimated 12,000–16,000 genomic loci dispersed between and within protein-coding genes in two broad regions of chromosome IV. These loci each have a conserved upstream motif, which we propose specifies the production of 21U-RNAs from thousands of noncoding transcripts.

RESULTS

Our library of small RNAs isolated from mixed-staged *C. elegans* was previously constructed so as to represent only those RNAs with 5′ monophosphate and 3′ hydroxyl groups, the termini expected of miRNAs and siRNAs (Lau et al., 2001). Standard sequencing of this and similar libraries previously yielded sequences of 4078 small RNA clones that match the *C. elegans* genome (Lau et al., 2001; Lim et al., 2003). High-throughput pyrosequencing (Margulies et al., 2005) of the library yielded 394,926 sequence reads that perfectly matched the worm genome. Of those, 80% matched annotated miRNA hairpins. Another 6.4% matched other annotated noncoding RNA genes, such as rRNA and tRNA, and were present at similar frequencies for each length from 18 to 28 nt, which was the pattern expected for degradation fragments of these noncoding RNAs. Another 9.3% corresponded to 21U-RNAs, and at least 0.7% corresponded to endogenous siRNAs that were antisense to annotated exons. The remaining sequences included what appeared to be endogenous siRNAs that were antisense to annotated introns, mRNA/intron degradation fragments, and a small contingent of uncharacterized sequences.

Previously Annotated miRNAs

Our previous sequencing of small RNA libraries from *C. elegans* discovered, refined, or confirmed the identities of 80 miRNAs (Lau et al., 2001; Lim et al., 2003). All 80 were observed in the new set of high-throughput reads at relative frequencies similar to those observed previously (Table S1). As exemplified by *lin-4* (Figures 1A and 1B), these 80 miRNA genes were typically represented by one dominantly sequenced species, the miRNA, as well as a sequence from the opposing arm of the hairpin, the miRNA* (Table S1; Supplemental Data). In addition, sequences were sometimes observed that matched the portion of the transcript in between the miRNA and miRNA* (Figures 1A and 1B; Table S1).

On average, the miRNA* species was present at about 1.0% the frequency of the miRNA. When paired to the miRNA it generally exhibited the 3′ overhangs typical of miRNA hairpin processing (Table S1; Lee et al., 2003; Lim et al., 2003). Identifying the dominant miRNA* species for many of the miRNAs, together with information on end heterogeneity, provided useful data for considering the specificity and precision of Drosha and Dicer processing. For example, the observed miRNA 5′ ends were far more homogenous (99.5% identical) than the miRNA* 5′ ends (91% identical), which were more homogenous than the miRNA 3′ ends (85% identical) and miRNA* 3′ ends (77% identical). About half of the 5′ heterogeneity was from reads that were longer than the dominant species, implicating imprecise Drosha/Dicer processing as the major cause of heterogeneity at this end. Greater 3′ heterogeneity was attributed to three factors: (1) less precise Drosha/Dicer processing, as indicated by templated nucleotides extending beyond the dominant species, (2) preferential degradation at the 3′ end, and (3) addition of untemplated nucleotides to the 3′ ends of miRNA and miRNA* species. The more precise cut at the miRNA 5′ end, compared to the miRNA* 5′ end, presumably reflected selective pressure for accurately defining the miRNA seed. Cleavage by either Drosha or Dicer appeared equally consistent when that cut would set the seed. The observation that when Dicer set the seed it was more precise than Drosha disfavored models in which Dicer simply measures from the termini left by Drosha and suggested that additional determinants are employed when needed to more accurately define Dicer cleavage.

Examining the dominant mature miRNA sequences revealed that 1.33% were extended by a single untemplated nucleotide, with U being the preferred untemplated nucleotide (54%; Table S1). A second untemplated nucleotide appeared with greater efficiency (4% of those already extended by one untemplated nucleotide) and with greater preference for U (73%). Similar efficiency and U preference was observed for a third nucleotide. The untemplated uridylation of miRNAs was reminiscent of that reported for unmethylated small RNAs in *Arabidopsis* (Li et al., 2005).

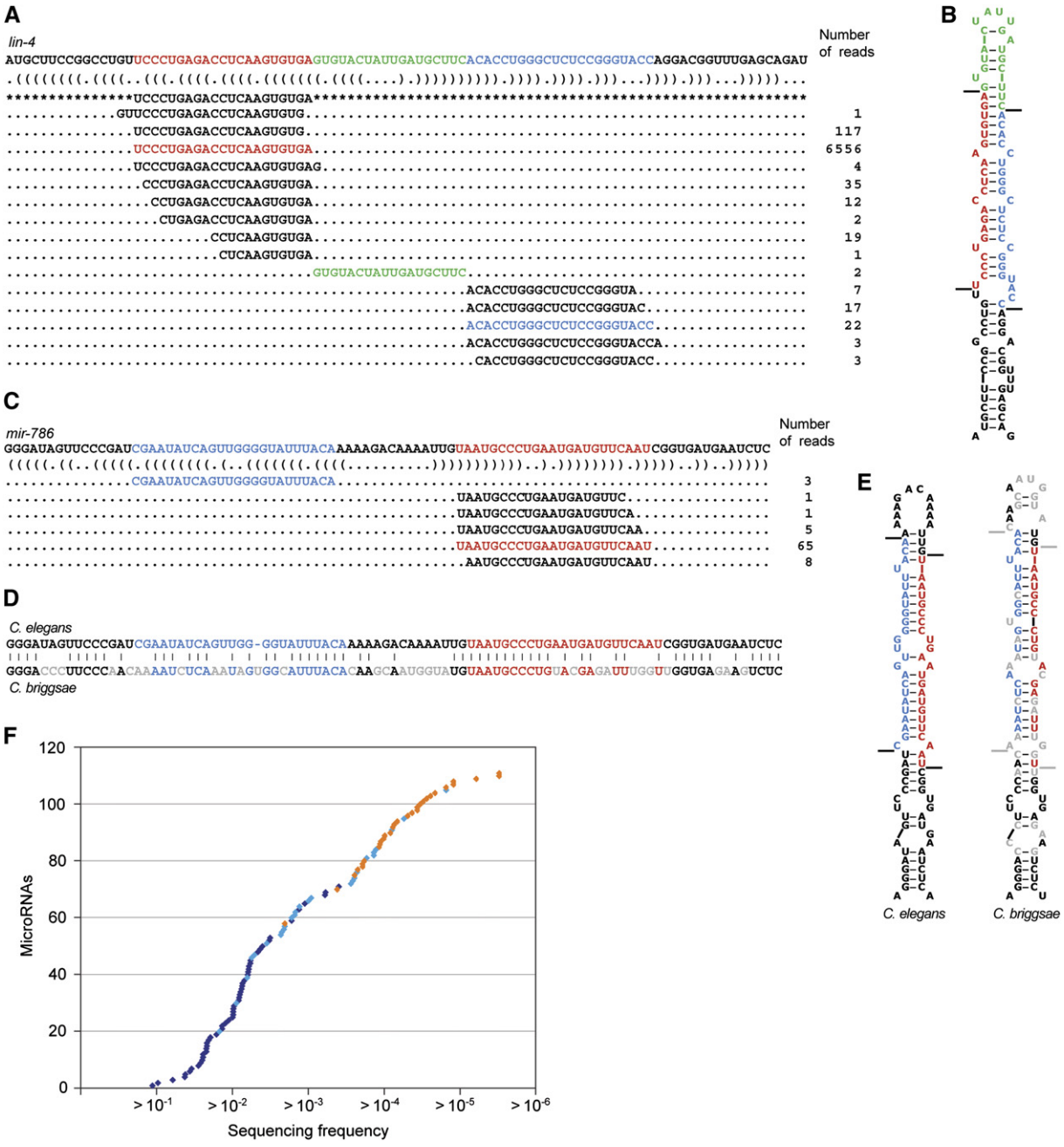

Figure 1. Distribution of Reads across the *lin-4* and *mir-786* Hairpins

(A) The sequence of the *lin-4* hairpin is depicted above its bracket-notation secondary structure as determined by RNAfold (Hofacker et al., 1994) and above the prior annotation of the mature *lin-4* miRNA (Lee et al., 1993), as refined by Lau et al. (2001). Below, each of the small RNA sequences that matched the *lin-4* hairpin is listed, with the number of reads representing each sequence shown. The dominant miRNA sequence is red; the dominant miRNA* species is blue; and the loop-containing sequence is green. Reads from the other previously annotated miRNA hairpins are provided (Table S1).

(B) The *lin-4* predicted hairpin, with the dominant species highlighted as in (A). Lines indicate inferred sites of Drosha and Dicer cleavage.

(C) The sequence of the *mir-786* hairpin depicted as in (A). Reads from the other newly identified miRNA hairpins are provided (Table S1).

(D) An alignment of the *mir-786* hairpin sequence with that of its inferred ortholog in *C. briggsae*. The dominant miRNA and miRNA* species are highlighted as in (A), and *C. briggsae* residues differing from those of *C. elegans* are in gray.

(E) The *C. elegans* and *C. briggsae mir-786* hairpins, depicted as in (B) with residues colored as in (D).

(F) Cumulative plot of *C. elegans* miRNAs with the indicated pyrosequencing frequency; blue, 53 miRNAs sequenced in Lau et al. (2001); cyan, 27 miRNAs first sequenced in Lim et al., (2003); orange, 31 miRNAs first sequenced in the current study (including 13 from previously annotated miRNA hairpins).

As expected, the high-throughput reads also included some annotated *C. elegans* miRNAs that were not among the 80 previously sequenced from our libraries. Thirteen such previously annotated miRNA hairpins gave rise to high-throughput reads (Table S1). All 13 were originally identified computationally and then experimentally supported by northern blotting and/or a PCR-based assay (Lim et al., 2003; Ohler et al., 2004). For five of these, the 5′ terminus did not match the one previously annotated, an observation with ramifications for the experimental validation of computational candidates (Supplemental Data). No reads matched 19 of the *C. elegans* miRNA hairpins annotated in miRBase (Supplemental Data). Of these 19, one was the *lys-6* miRNA, which had been identified genetically and appears to be expressed in only a few cells (Johnston and Hobert, 2003).

Newly Identified miRNAs

In a search for additional miRNAs, we evaluated reads that fell within potential miRNA-like hairpins, considering the following criteria: (1) the pairing characteristics of the hairpin; (2) the expression of the candidate, as measured by the abundance of sequence reads sharing the same 5′ terminus; (3) evolutionary conservation, as evaluated by the apparent conservation of the hairpin in *C. briggsae* and grouping of the miRNA candidate into a family based on its seed sequence; (4) the absence of annotation suggesting non-miRNA biogenesis; and (5) the presence of reads corresponding to the predicted miRNA* species. The observation of both a candidate miRNA and a candidate miRNA* in a set of reads provides particularly compelling evidence for Dicer-like processing from an RNA hairpin. As illustrated for miR-786 (Figures 1C–1E), seven newly identified genes satisfied all of our criteria (Table 1). Eleven others satisfied a subset of the criteria deemed sufficient for confident annotation as miRNAs. Three additional candidates that were sequenced more than once were, from our perspective, borderline cases and therefore not annotated here as miRNAs (Supplemental Data).

Sequencing frequencies of all newly and previously sequenced miRNAs are illustrated (Figure 1F). Seven newly identified genes were near another miRNA gene and on the same genomic strand (Table 1), an arrangement implying processing from a common polycystronic transcript (Lagos-Quintana et al., 2001; Lau et al., 2001). Seven newly identified genes added to previously known *C. elegans* miRNA families, in that they shared the same seed (Table 2). For example, miR-793, miR-794, and miR-795 all added to the *let-7*/48/84/241 family. Four other newly identified genes shared seeds with miRNAs annotated in distant species, thereby extending the scope of families previously identified in insects or vertebrates to the nematode lineage (Table 2).

21U-RNAs

After accounting for the miRNAs and other types of annotated noncoding RNAs, the remaining reads were dominated by 21-mers with 5′ uridines. We refer to the bulk of these as "21U-RNAs." The vast majority of RNAs with these properties mapped to two broad but distinct regions of chromosome IV, one spanning chromosomal coordinates 4.5–7.0 M, the other spanning 13.5–17.2M (Figure 2A). A few mapped to a third region, which spanned coordinates 9–9.7M of chromosome IV. The ~34,300 21U-RNA reads that derived from these three regions contained 5,454 unique sequences (Figure 2B), for which 5,302 loci were unambiguously mapped because their sequences were unique in the assembly. Many of these loci were represented by single reads in our set, suggesting the existence of more members of this small RNA class than were directly observed. Nonetheless, most of the 21U-RNA loci (67%) were represented by two or more identical reads, indicating that the 34,300 reads captured a nontrivial portion of the 21U-RNA diversity.

Four 21U-RNAs were sequenced more than 200 times, including 21UR-1 (pUGGUACGUACGUUAACCGUGC), which was represented by 521 reads and detectable on RNA blots. This 21U-RNA was sensitive to alkaline hydrolysis and phosphatase treatment and was a suitable substrate for RNA ligase—the expected properties of an RNA with a 5′ monophosphate (Figures 3C and S1). 21UR-1 was also resistant to periodate treatment (Figure 3C), indicating that its 3′ nucleotide was missing the *cis* diol and suggesting modification at either the 2′ or 3′ oxygen of this nucloetide, as reported for small RNAs in plants and rasiRNAs in flies (Li et al., 2005; Vagin et al., 2006).

The 21U-RNAs mapped to both strands of the DNA but overlapped with each other or with other sequenced small RNAs on the opposing DNA strand less frequently than would be expected by chance given a random distribution, thereby providing no evidence for a dsRNA precursor. WormBase-annotated genes were somewhat less abundant within the 21U-RNA-rich portions of chromosome IV (mean ± SD of 93 ± 28 genes per 500 kb) compared to the genome as a whole (116 ± 26 genes per 500 kb). The vast majority of the 21U-RNAs mapped either between genes or within introns, with no preference for the sense or antisense orientiation among intronic matches. Only 2.5% of the 21U-RNA loci overlapped annotated exons, a substantial depletion versus the total fraction of the regions overlapping exons (~21%), and the read abundance of sense versus antisense exonic matches was nearly even (~750 and ~810, respectively). Overall, the genomic data suggested that the 21U-RNA loci are maintained independently of other genetic elements, with informational constraints that can conflict with those of other genes.

The ~34,300 21U-RNA reads in our set of high-throughput reads came from a mixed-staged library, raising the question of which stage(s) in development the 21U-RNAs might accumulate. Our previous effort (Lim et al., 2003) included reads from this mixed-stage library as well as reads from a larval stage L1 library, a dauer (dormant L3) library and a mixed-staged library made from *him-8* mutant worms (which are enriched in males). Revisiting the 4078 reads from that earlier study revealed that 125 represented

Table 1. Eighteen Newly Identified miRNAs in *C. elegans*

miRNA	Sequence	miRNA Reads	miRNA* Reads	*C. briggsae* Ortholog	Fly or Vertebrate Family Members	Genomic Cluster Parter
miR-784	UGGCACAAUCUGCGUACGUAGA	11	1	Yes		
miR-785	UAAGUGAAUUGUUUUGUGUAGA	14	2	Yes	Yes	miR-359
miR-786	UAAUGCCCUGAAUGAUGUUCAAU	80	3	Yes	Yes	miR-240
miR-787	UAAGCUCGUUUUAGUAUCUUUCG	32		Yes	Yes	
miR-788	UCCGCUUCUAACUUCCAUUUGCAG	667	10	Yes		
miR-789-1	UCCCUGCCUGGGUCACCAAUUGU	63	1	Yes		
miR-789-2	UCCCUGCCUGGGUCACCAAUUGU	63		Yes		
miR-790	CUUGGCACUCGCGAACACCGCG	16	5	Yes	Yes	miR-228
miR-791	UUUGGCACUCCGCAGAUAAGGCA	1	1	Yes	Yes	miR-230
miR-792	UUGAAAUCUCUUCAACUUUCAGA	4		Yes	Yes	
miR-793	UGAGGUAUCUUAGUUAGACAGA	73			Yes	
miR-794	UGAGGUAAUCAUCGUUGUCACU	5			Yes	miR-795
miR-795	UGAGGUAGAUUGAUCAGCGAGCUU	4			Yes	miR-794
miR-796	UGGAAUGUAGUUGAGGUUAGUAA	9			Yes	
miR-797	UAUCACAGCAAUCACAAUGAGAAGA	12			Yes	miR-247
miR-798	UAAGCCUUACAUAUUGACUGA	33				
miR-799	UGAACCCUGAUAAAGCUAGUGG	36				
miR-800	CAAACUCGGAAAUUGUCUGCCG	12	3			

Reads for miR-789-1 and miR-789-2 cannot be distinguished.

21U-RNAs: 79 from mixed stage, 8 from dauer, 10 from L1, and 28 from *him-8*. Normalizing to the read counts of miRNAs with constant expression throughout larval development, the *him-8* library was ~2-fold enriched in 21U-RNAs compared to the wild-type mixed-stage library, whereas the L1 and dauer libraries were ~2- and ~3-fold depleted, respectively. The presence of 21U-RNAs in both L1 worms and dauer L3 worms implies their presence throughout much of worm development.

Two Sequence Motifs Associated with 21U-RNA Loci

Other than the U at their 5′ termini, the 21U-RNAs shared little sequence identity. Indeed, the composition of the four nucleotides was more equivalent for the 21U-RNAs than for their broader genomic contexts, which were A-T rich. However, the 21U-RNA genomic loci did share two upstream sequence motifs, one much larger than the other (Figure 3). The large motif was 34 bp and centered on an 8 nt core consensus sequence, CTGTTTCA. The small motif had a core sequence of YRNT, in which the T corresponded to the 5′ U of the 21U-RNA. The two subdomains of the motif were separated by a spacer typically 19–21 bp (Figure 3B).

A position-specific scoring matrix based on the combined properties of the two motifs was used to predict 21U-RNAs on *C. elegans* chromosome IV. With a score cut-off that correctly predicted 77% of the sequenced 21U-RNAs, 10,807 loci were identified on both strands of chromosome IV. The density of genomic matches to the motifs corresponded well to that of known 21U-RNA loci, demonstrating the specificity of our motif-scanning procedure (Figures 2B and 2C). As illustrated for a 100 kb region of chromosome VI, this correspondence held at high-resolution views (Figure 2D). As a test of sensitivity, we crosschecked the 10,807 predictions with an independent set of 245,420 *C. elegans* small RNA reads (Pak and Fire, in press) and found that nearly half (46%) of the 21U-RNAs uniquely identified in this independent data set had been predicted (see Experimental Procedures). We suggest that the correspondence of 21U-RNAs predicted through motif scanning with those detected by sequencing reflected the function of the motifs in specifying 21U-RNA production in the animal.

Discovery of the upstream motif allowed assessment of the other properties ascribed to 21U-RNAs (Figure S2). Nearly all of the motif-associated 21-mer reads (99.8%) began with a U, and 98.5% derived from the defined 21U-rich regions of chromosome IV. Over 99% of the motif-associated reads were 21 nt or less, with those that were shorter (5.4%) likely corresponding to 3′ degradation products.

To explore the potential conservation of 21U-RNAs, we scanned all the *C. briggsae* genomic contigs (Stein et al., 2003) for motif matches. Each *C. briggsae* contig with a high concentration of motifs (≥75 per 100 kb) was syntenous with one of the three 21U-rich regions of *C. elegans* chromosome IV (Figures 2A and 2B). We conclude that

Table 2. *C. elegans* miRNA Families, with the Corresponding Known miRNAs in other Animals

Seed	*C. elegans*	*C. briggsae*	*D. melanogaster*	*D. rerio*	Mammal
CCCUGA	*lin-4*/237	*lin-4*	miR-125	miR-125a/b/c	miR-125a/b,mmu-miR-351
UUUGUA	*lsy-6*	*lsy-6*			
GAGGUA	*let-7*/48/84/241/793/794/795	*let-7*/48/84/241	*let-7*	*let-7a/b/c/d/e/f/g/h/i*	*let-7a/b/c/d/e/f/g/i*/98/202
GGAAUG	miR-1/796	miR-1	miR-1	miR-1/206	miR-1/206
AUCACA	miR-2/43/250/797	miR-43	miR-2a/b/c/6/11/13a/b/308		
GGCAGU	miR-34	miR-34	miR-34	miR-34	miR-34a/c/449
CACCGG	miR-35/36/37/38/39/40/41/42	miR-35/36/38/39/40/41			
GACUAG	miR-44/45/61/247	miR-44/45/61	miR-279/286		
GUCAUG	miR-46/47	miR-46/47	miR-281		
AGCACC	miR-49/83	miR-49/83	miR-285	miR-29a/b	miR-29a/b/c
GAUAUG	miR-50/62/90	miR-50/62/90		miR-190	miR-190
ACCCGU	miR-51/52/53/54/55/56	miR-51/52/55	miR-100	miR-99/100	miR-99b/100, hsa-miR-99a
ACCCUG	miR-57	miR-57		miR-10a/b/c/d	miR-10a, hsa-miR-10b
GAGAUC	miR-58/80/81/82	miR-58/80/81/82	bantam		
CGAAUC	miR-59	miR-59			
AUUAUG	miR-60	miR-60			
AUGACA	miR-63/64/65/66/229	miR-64			
CACAAC	miR-67	miR-67	miR-307	miR-220	
AAUACG	miR-70	miR-70			
GAAAGA	miR-71	miR-71			
GGCAAG	miR-72/73/74	miR-73/74	miR-31a/b		mmu-miR-31
UAAAGC	miR-75/79	miR-75/79	miR-4		
UCGUUG	miR-76	miR-76			
UCAUCA	miR-77	miR-77			
GGAGGC	miR-78				
ACAAAG	miR-85	miR-85			
AAGUGA	miR-86/785	miR-86/785			
UGAGCA	miR-87/233	miR-87/233/356	miR-87		
AAGGCA	miR-124	miR-124	miR-124	miR-124	hsa-miR-506, mmu-miR-124a
AUGGCA	miR-228	miR-228		miR-183	miR-183
UAUUAG	miR-230	miR-230			
AAGCUC	miR-231/787	miR-231/787			
AAAUGC	miR-232/357	miR-232/357	miR-277		
UAUUGC	miR-234	miR-234		miR-137	mmu-miR-137
AUUGCA	miR-235	miR-235	miR-92a/b/310/311/312/313	miR-25/92a/b/363	miR-25/32/92, hsa-miR-367
AAUACU	miR-236	miR-236	miR-8	miR-200b/c/429	miR-200b/c/429

Table 2. *Continued*

Seed	*C. elegans*	*C. briggsae*	*D. melanogaster*	*D. rerio*	Mammal
UUGUAC	miR-238/239a/b	miR-239a	miR-305		
ACUGGC	miR-240	miR-240		miR-193a/b	miR-193
UGCGUA	miR-242	miR-242			
GGUACG	miR-243				
CUUUGG	miR-244	miR-244	miR-9a/b/c	miR-9	miR-9
UUGGUC	miR-245	miR-245		miR-133a/b/c	
UACAUG	miR-246	miR-246			
UACACG	miR-248.1				
ACACGU	miR-248.2	miR-248			
CACAGG	miR-249	miR-249			
UAAGUA	miR-251/252	miR-251			
UAGUAG	miR-253	miR-253			
GCAAAU	miR-254	miR-254			
AACUGA	miR-255	miR-255			
AAUCUC	miR-259	miR-259	miR-304	miR-216a/b	miR-216
UUGUUU	miR-355	miR-355			
UUGGUA	miR-358	miR-358			
CACUGG	miR-359	miR-359	miR-3/309/318		
AUCAUC	miR-392	miR-392			
GGCACA	miR-784	miR-784			
AAUGCC	miR-786	miR-786		miR-365	miR-365
CCGCUU	miR-788	miR-788			
CCCUGC	miR-789-1/-2	miR-789a/b			
UUGGCA	miR-790/791	miR-791	miR-263b	miR-96/182	miR-96/182
UGAAAU	miR-792	miR-792		miR-203a/b	
AAGCCU	miR-798				
GAACCC	miR-799				
AAACUC	miR-800				

Families sorted alphabetically by seed are listed in Table S2, and newly reported *C. briggsae* orthologs are listed in Table S5.

any roles that the motifs might play in the biogenesis of 21U-RNAs have been conserved in the ~100 million years since the divergence of these two nematode species (Coghlan and Wolfe, 2002). The 21U-RNAs themselves, in contrast, showed little evidence for conservation. Of the >10,000 21U-RNA sequences predicted on chromosome IV of *C. elegans* and the >11,000 sequences similarly predicted in *C. briggsae*, not a single sequence was shared between the two species.

Endogenous siRNAs

Of the remaining sequences with perfect matches to the *C. elegans* genome, some were antisense to known protein-coding transcripts. In fact, a larger number matched the antisense strand of spliced mRNAs (2934 reads, 2378 unique sequences; Figure 4A) than matched the sense strand (2150 reads, 1800 unique sequences; Figure 4B). As done previously (Lau et al., 2001; Ambros et al., 2003; Lim et al., 2003), we classified the RNAs matching the antisense strand as candidate endogenous siRNAs, which for simplicity we refer to herein as siRNAs. RNAs that matched the sense strand also might include endogenous siRNAs, but as they likely include other hydrolysis products, we refer to them as sense RNAs.

For different *C. elegans* libraries, the proportion of miRNAs to siRNAs varies greatly; our libraries contain 100 times more miRNAs than siRNAs, whereas the Ambros library contains roughly equal numbers of the two (Ambros et al., 2003; Lim et al., 2003). The large difference suggests that most *C. elegans* siRNAs lack the 5′ monophosphate

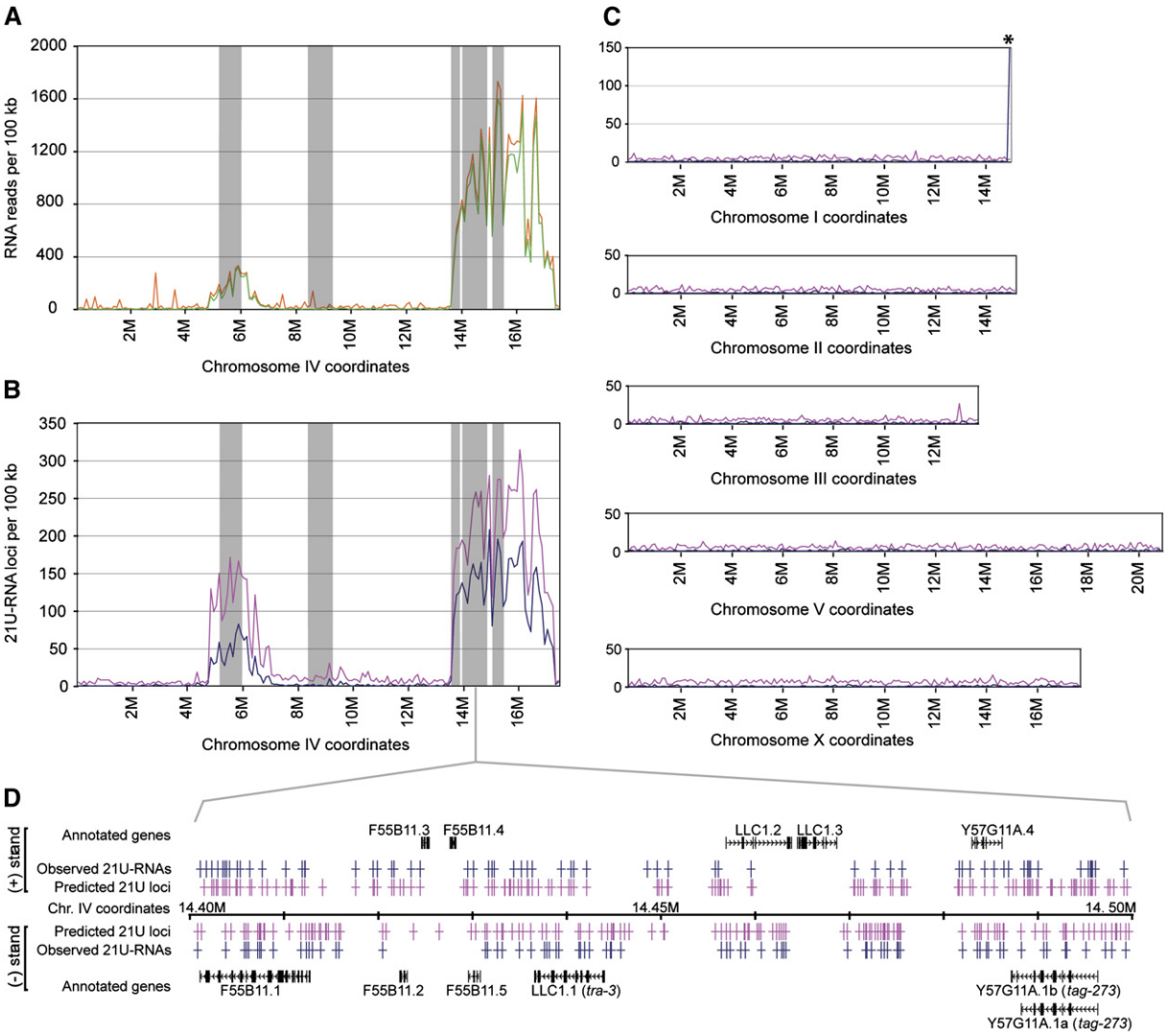

Figure 2. Observed and Predicted 21U-RNAs from Thousands of Loci across Two Broad Regions of *C. elegans* Chromosome IV

(A) Observed small RNA reads from chromosome IV. All normalized reads were counted in 100 kb bins (orange). The subset of normalized reads that were precisely 21 nt long and began with U were also counted (green). Gray shading is explained in (B).

(B) Observed and predicted 21U-RNA loci on chromosome IV. Loci that matched one or more 21U-RNA read were counted in 100 kb bins (blue). The same was done for 21U-RNA loci predicted by scanning for the associated motifs (pink). Sections of the chromosome shaded in gray are syntenic to *C. briggsae* contigs with a high density ($\geq$75 per 100 kb) of the 21U-RNA-associated motifs.

(C) Observed and predicted 21U-RNA loci on other chromosomes. Coloring is as in (B). The asterisk above chromosome I indicates the position of the ribosomal repeats, which are collapsed in the genome assembly; ribosomal RNA fragments mapped to this region, some of which were 21 nt with a 5′ U.

(D) Representative 100 kb fragment of a region that gives rise to 21U-RNAs. Shown are the 146 loci corresponding to observed 21U-RNA reads (blue) and the 257 predicted loci (pink) from coordinates 14.4–14.5 M (WormBase, build WS120). Shown also are WormBase-annotated genes.

required by our cloning protocol (Ambros et al., 2003). Perhaps many are short RNA-dependent RNA polymerase (RdRP) products that have retained their 5′ triphosphate. Consistent with this idea, we detected a population of endogenous ~22-mers that were suitable substrates for an in vitro 5′-capping reaction requiring a 5′ di- or triphosphate (Figure 4C). These sequences would be underrepresented in our library, although not totally absent if some molecules lost their γ and β phosphates or were transcribed with an initiating nucleoside monophosphate rather than nucleoside triphosphate, as has been observed for other RNA polymerases (Martin and Coleman, 1989; Ranjith-Kumar et al., 2002).

While recognizing that the siRNAs of our library were likely depleted in the major subclass of endogenous siRNAs, we proceeded with their analysis. Their length distribution had prominent peaks at 21, 22, and 26 nt (Figures 4A and 4B). Comparison to the length distribution of reads matching tRNA and rRNA indicated that the 26-mer siRNA population was distinct, rather than the

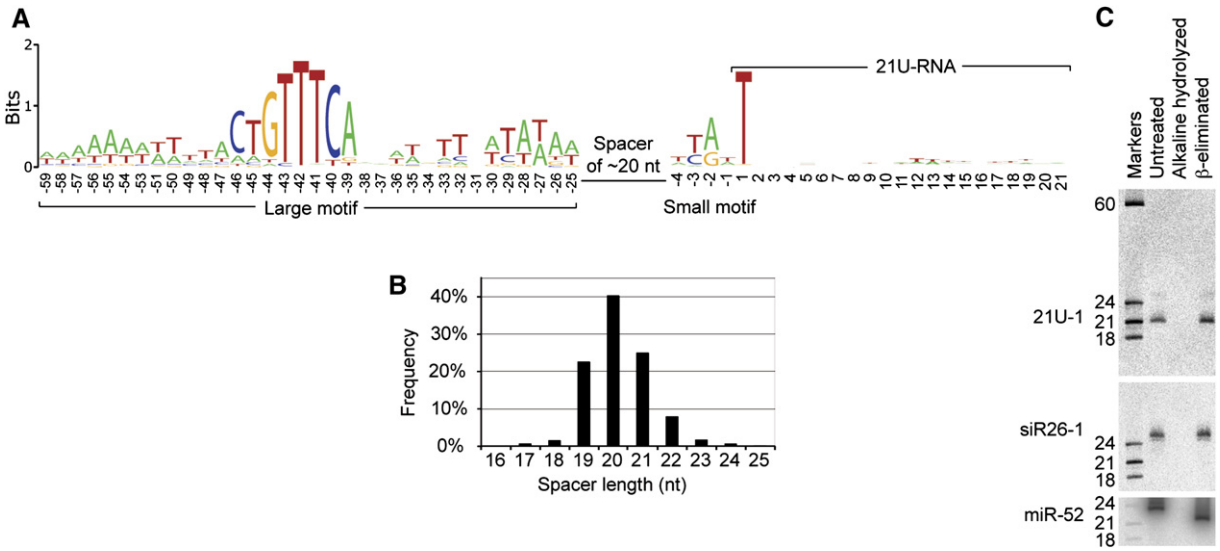

Figure 3. The 21U-RNA Sequence Motifs and Small RNA Chemical Reactivity

(A) The large and small motifs found upstream of 21U-RNA loci, depicted as a sequence motif (Crooks et al., 2004). The T at position 1 corresponds to the 5′ U of the 21U-RNA.

(B) The distribution of distances between the large and small motifs.

(C) Chemical reactivity of small RNAs. Total RNA (40 μg) was treated as indicated and analyzed by RNA blot, probing first for 21U-1, then stripping and reprobing for siR26-1, then miR-52.

shoulder of a larger, more broadly distributed population. A preference for a 5′ G, observed previously for siRNAs (Ambros et al., 2003), was persistent across all lengths of endogenous siRNAs but strongest among 26-mers. A 26-mer siRNA sequenced nine times had a 5′ monophosphate (siR26-1, pGCAAGAUGGAAAAGUUUGAGAUUCCG; Figure S1). As observed for the 21U-RNA, this siRNA was resistant to periodate treatment, again suggesting modification at either the 2′ or 3′ oxygen of the 3′ nucleotide (Figure 3C). With so many classes of plant and animal small RNAs now shown to be resistant to periodate oxidation, metazoan miRNAs appear increasingly unusual in not being modified at their 3′ residue.

Despite being spread out over a large number of genes, dense clusters of siRNAs were observed at some genomic loci (Figure 4D; Table S3). Examination of surrounding sequence revealed that siRNAs did not exclusively match annotated exons. For example, some also matched annotated introns. Nonetheless, more than 40 of the unique sequences represented by our reads did not match the genomic DNA but instead spanned splice junctions (exemplified in Figure 4E), implying that these RNAs were produced by an RdRP acting on a spliced transcript. Because these junction-spanning siRNAs had the length distribution and preference for a 5′ G characteristic of the siRNAs in general, it is reasonable to propose that the remainder of the siRNAs were also RdRP products and that at least some of the RdRP activity was nuclear and thus could act on both spliced and unspliced templates.

Correlations with siRNAs supported the idea that the biogenesis or function of some sense RNAs was linked to that of the siRNAs. The overlap of siRNA-complemented genes was greater with genes matching sense RNAs (24%) than with genes picked using SAGE data to control for expression (16%; $p < 0.01$, chi-square test). Among the sense-antisense pairs with at least 1 nt overlap at their genomic loci, 30% maximally overlapped (exemplified by all four sense reads in Figure 4D), which was 5-fold higher than expected by chance. For 47% of the sense-antisense pairs involving 26-mers, the most common configuration placed the 5′ nucleotide of the sense read across from nucleotide 23 of a 26-mer siRNA (exemplified by three sense reads in Figure 4D), which was 20-fold higher than chance expectation.

To gain insight into the biological consequences of siRNAs, we examined the functional categorization of genes they complemented. In addition to the enrichment for transposon genes, observed previously (Lee et al., 2006), genes matching siRNAs were frequently sperm enriched (Supplemental Data). This propensity was particularly striking for genes matching 26-mer siRNAs, 55% of which were sperm enriched.

DISCUSSION

There Are 112 Confidently Identified *C. elegans* miRNAs

The set of miRNA genes represented in our high-throughput reads included 93 previously annotated genes, plus 18 newly discovered genes (Table S1). The notable exception was the *lsy-6* miRNA, a genetically identified miRNA thought to be transcribed in only one to nine cells

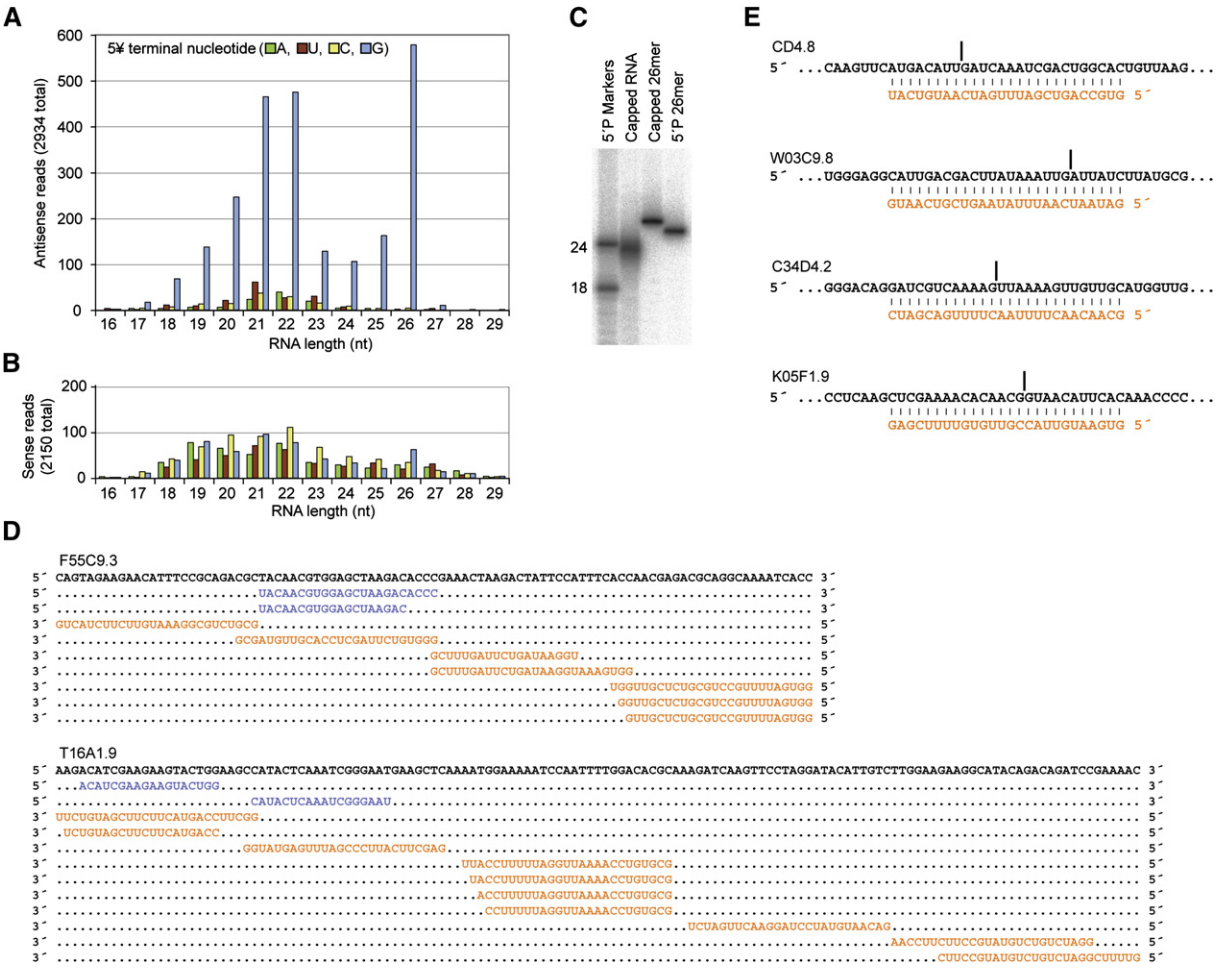

Figure 4. Many Reads Antisense to Known or Predicted mRNAs

(A) The length and initial nucleotide distribution of the antisense reads.

(B) The length and initial nucleotide distribution of the sense reads.

(C) A population of ~22-mer RNAs with terminal 5′ di- or triphosphates. Those RNAs with 5′ di- or triphosphates were selectively radiolabeled in a capping reaction that used α-^{32}P GTP (capped RNA) and compared to the indicated 5′ phosphorylated (5′ P) or capped size standards by 15% PAGE.

(D) Portions of two WormBase-annotated protein-coding genes aligned with small RNA reads that matched the sense (blue) and antisense (orange) strands. One hundred siRNA clusters, each comprising from 4 to 61 antisense reads, are shown in Table S3.

(E) Examples of siRNAs that did not match the genome but did match the splice junctions (vertical lines) of mature mRNAs.

(Johnston and Hobert, 2003). The absence of *lsy-6* in a set that included 37,225 reads of miR-52 illustrated the extreme diversity in metazoan miRNA expression. This difference can be attributed solely to the specific expression of *lys-6* in cells that are few in number and small in volume; we estimated that *lsy-6* RNA should have been ~100,000 times less abundant than a miRNA expressed in most cells of the worm (Supplemental Data). Clearly, more reads must be sequenced before all the miRNAs expressed during the course of nematode development will be catalogued.

Although the unsaturated status of our sequencing project prohibited any definitive judgments about miRNA annotations that were not represented by our reads, our observations were informative for evaluating the confidence in those annotations and the data originally used to justify them. These considerations increased the number of annotated genes whose authenticity is in doubt (Supplemental Data). Nonetheless, the 18 newly identified miRNA genes enabled the number of confidently identified *C. elegans* miRNAs to be revised upwards to 112, which included the 111 represented in our high-throughput reads plus *lsy-6*. Currently annotated loci with reasonable prospects of eventually joining the list include *mir-273*, for which reverse-genetic functional data has been reported (Chang et al., 2004). Our three borderline candidates also might eventually be added (Supplemental Data). These include one that was represented by only five reads and lacked conservation or miRNA* evidence and two that might be considered "young" miRNAs, potential Drosha/Dicer substrates that might have recently emerged from short inverted duplications and have not had sufficient

time to acquire the mismatches usually observed in miRNA hairpins (Table S1). Our results also prompted re-evaluation of miRNA gene-number estimates in worms (Supplemental Data).

The 112 confidently identified *C. elegans* miRNA genes arose from 83 genomic clusters, ranging from one to seven genes per cluster (Table S2). When grouped according to their seeds, they fell into 63 families, 58 (92%) of which have apparent orthologs in *C. briggsae* and 31 (49%) of which have counterparts in much more distantly related lineages, such as flies, fish, and mammals (Tables 2, S2, and S5). The 31 families with counterparts in flies or vertebrates encompassed most (64 of 112) of the *C. elegans* genes. The newly identified and revised miRNA sequences provided the opportunity to improve and expand the current set of predicted miRNA targets in *C. elegans* (Chan et al., 2005; Lall et al., 2006). Accordingly, the TargetScanS algorithm was used to predict conserved regulatory targets, which can be viewed at http://www.TargetScan.org.

Endogenous siRNA Biogenesis and Targeting

Our library-construction protocol appears to exclude the vast majority of the *C. elegans* siRNA molecules, which we suspect have 5′ triphosphates. Nonetheless, high-throughput sequencing generated more candidate siRNAs than observed previously, enabling insights into endogenous siRNA taxonomy, biogenesis, and function.

Many of the previously annotated tncRNAs fell into clusters of reads that resembled the siRNA clusters, and many of these tncRNA-containing clusters overlapped annotated mRNA exons (Table S4; compare to Table S3). Furthermore, the known factors required for tncRNA biogenesis and endogenous siRNA biogenesis are similar (Lee et al., 2006). Considering these similarities and reasoning that any minor differences reported between the biogenesis requirements of particular tncRNAs and siRNAs are likely to be no greater than those between different siRNAs, we propose that the tncRNAs do not represent a class of *C. elegans* RNAs separate from the endogenous siRNAs. Nonetheless, the endogenous siRNAs of *C. elegans* are not a monolithic class and appear to be combination of classes whose taxonomy includes an abundant shorter class underrepresented in our library, presumably because of 5′ triphosphates, and a newly identified ~26 nt class with 5′ monophosphates and modified 3′ termini.

Many of the small RNAs classified as *C. elegans* endogenous siRNAs have strong links with RNAi-mediated gene silencing. For example, they are enriched in matches to transposons, and their accumulation decreases in mutant worms that are defective in RNAi (Lee et al., 2006). Thus, their classification as siRNAs is appropriate. However, they differ from canonical siRNAs in that they lacked some of the classical features of Dicer products: most appear to lack a 5′ monophosphate; their length distribution (Figure 4A) largely differed from the 23 nt RNAs previously described for *C. elegans* exogenous siRNAs (Ketting et al., 2001), and their overlapping ends were uncharacteristic of Dicer processing (Figure 4D; Table S3), which should yield nonoverlapping ends when the RNAs are in phase with each other. We conclude that endogenous siRNAs biogenesis in nematodes involves little, if any, sequential Dicer processing of long dsRNA, which is perhaps unexpected given the facility by which *C. elegans* utilizes long dsRNA for exogenous RNAi (Fire et al., 1998), the Dicer-dependence of some siRNAs (Lee et al., 2006), and the models of transitive RNAi in worms, in which siRNAs serve as primers for the production of additional siRNAs (Sijen et al., 2001; Tijsterman et al., 2002). Instead, we propose that most endogenous *C. elegans* siRNAs are generated by unprimed RdRP activities insufficiently processive to generate long dsRNAs suitable for successive cleavage events and are thus reminiscent of short antisense RNAs generated by *Neurospora* QDE-1 (Makeyev and Bamford, 2002). Because longer dsRNA is mobile in worms (Feinberg and Hunter, 2003), shorter polymerization might ensure that the endogenous silencing is cell autonomous. If only a single siRNA was made from each RdRP product, then the 5′ terminus of each siRNA could be determined by the nucleotide used to initiate synthesis of the antisense strand, which we suspect is predominantly a GTP.

Recognizing that there could be multiple endogenous RNAi pathways in worms, we draw a speculative model focusing on the 26-mer siRNAs and the propensity of their 23rd residues to pair with sense RNA 5′ termini (Figure 5). A 26-mer siRNA is synthesized without priming by an RdRP, initiating with a G across from a C in the template transcript (step 1). The siRNA guides an endonuclease to cleave the template between residues that pair to nucleotides 23 and 24 of the siRNA (step 2). The cleaved template triggers a second round of unprimed siRNA synthesis, which starts across from the C residue closest to the cleavage site (step 3). Steps 2 and 3 repeat, generating the phased pattern of siRNAs that overlap in cases where C residues lie close to the cleavage site. Degradation of the ~26 nt sense fragments proceeds in the 3′ to 5′ direction but is slowed by pairing to the siRNA, thereby leading to accumulation of sense reads that fully pair to the siRNAs (step 4). Once liberated from the sense fragment, the siRNA might pair to a second transcript (step 5) and target its cleavage, thereby initiating another series of siRNA-synthesis and target-cleavage events. Although Dicer is not necessarily at the heart of this model, siRNA accumulation would still be Dicer-dependent if Dicer was required for either the initial mRNA cleavage or subsequent cleavages that trigger unprimed synthesis. A requirement of PIR-1 to remove the siRNA γ- and β-phosphates might explain both the importance of this presumed RNA phosphatase for siRNA production (Duchaine et al., 2006) and the monophosphate at the 5′ terminus of 26-mer siRNAs.

Endogenous siRNAs have previously been implicated in transposon silencing (Sijen and Plasterk, 2003; Lee et al., 2006). We found that endogenous siRNAs, particularly 26-mers, also had a propensity to match spermatogenesis-associated messages. Worms deficient in EGO-1, a nuclear RdRP, have delayed spermatogenesis-to-oogenesis transition (Smardon et al., 2000), tempting speculation that

Figure 5. Speculative Model for an Endogenous RNAi Pathway in Worms

Illustrated using the F55C9.3 transcript (blue) and sequenced siRNAs (orange) from Figure 4D. Small arrowheads indicate the transcript cleavage sites. See Discussion for explanation.

EGO-1 produces the endogenous siRNAs that silence sperm-enriched genes, thereby hastening the transition to oogenesis.

21U-RNAs: Diverse, Autonomously Expressed, Small RNAs

21U-RNAs are 21 nt RNAs that begin with a U and derive from thousands of loci in several broad regions of chromosome IV. The conservation in *C. briggsae* of the upstream motifs, presumably involved in 21U-RNA biogenesis, suggests that production of 21U-RNAs has an important biological function even if the RNA product itself might not. Such function might include opening of chromatin structure or changes to nucleosome phasing induced upon transcription of the 21U-RNA loci.

The more uniform nucleotide composition of 21U-RNA sequences versus their surrounding sequence, considered together with the diversity of 21U-RNAs, suggested that evolutionary pressure is maximizing their sequence complexities rather than maintaining their sequence identities. If 21U-RNAs act by base-pairing with a complementary nucleic acid strand, then this increased complexity would enable a higher degree of pairing specificity for the 21U-RNA sequences (important for both targeting and preventing off targeting) than would be possible using the less uniform nucleotide composition of neighboring sequence. Their 21 nt length and 5′ phosphate are both features of small RNAs that associate with Argonaute protein family members to target gene repression (Tomari and Zamore, 2005), suggesting that the 21U-RNAs might do the same and, perhaps, target the chromatin from which they derive. The regions defined by the 21U-RNA loci were vast and contained many protein-coding genes, with a wide variety of functions and expression patterns. Which of those functions that the 21U-RNAs might be influencing, if indeed they act locally, is unclear.

Equally mysterious as 21U-RNA function are aspects of their biogenesis. The large and small motifs might together serve as a promoter, driving expression of each 21U-RNA, with the AT-rich region at the 3′ end of the larger motif acting as a TATAA box. Or perhaps the motifs serve as a signal for targeting the cleavage of a larger transcript. The larger motif could serve as a promoter for a transcript that is processed at the site of the smaller motif. If the 21U-RNA primary transcript were to begin at the 5′ end of the mature 21U-RNA, the transcribing polymerase would either have to prefer incorporation of UMP to that of UTP at the 5′ end, or the 21U-RNA would have to be posttranscriptionally processed to remove the γ- and β-phosphates of the 5′-terminal UTP.

In our favored scenarios for 21U-RNA production, each locus represents an independent transcription unit; that is, each could be classified as an individual noncoding RNA gene. From this perspective, the discovery of the 21U-RNA loci dramatically increased the number of known nematode genes. A minimum of 5772 loci produced the observed reads (when also considering the 21U-RNA loci unique to reads from Pak and Fire, in press), and we estimate there to be 12,000–16,000 total loci (Supplemental Data). Nonetheless, the common upstream motif and broad clustering of 21U-RNA loci in the genome both suggest that these genes do not function alone but instead act concurrently to produce some aggregate effect. This scenario presents some fascinating evolutionary questions:

How do selective pressures act to maintain the motifs present at each of the thousands of individual 21U-RNA loci and, when they fail to do so, how do new loci emerge within the same broad regions of chromosome IV to replace those that are lost?

Another intriguing biogenesis question entails how the 3′ ends of the 21U-RNAs are defined. The absence of a discernable motif at or near the 3′ end suggests that it is defined in reference to the position of the 5′ end. This hypothesis requires a biochemical mechanism for precisely counting 21 ribonucleotides of any sequence. The known activity with closest precision in counting this number of ribonucloetides is Dicer-catalyzed cleavage. However, *C. elegans* Dicer is thought to produce 23-mer RNAs (Ketting et al., 2001), and Dicer products have a size diversity exceeding that of 21U-RNAs, even when processing dsRNA without mismatches (Zamore et al., 2000). Furthermore, we saw no evidence of 21 nt RNAs arising from the opposing RNA strand—no analog to the siRNA passenger strand. Even without conventional Dicer processing, counting 21 nt to determine the 3′ terminus in reference to the 5′ terminus is easiest to imagine if it occurs in the context of a double helix, presumably while the transcript is still paired to its DNA (or RNA) template.

21U-RNAs clearly represent a unique class of small RNAs. They are far more diverse than miRNAs, and unlike siRNAs and piRNAs, which are expressed in tight clusters, the 21U-RNAs appear to be autonomously expressed. We suggest that other types of diverse, autonomously expressed, small RNAs (dasRNAs) might be found in other species. The deep sequencing of small RNAs in species beyond *C. elegans* will provide important information for addressing this possibility.

EXPERIMENTAL PROCEDURES

Library Preparation

Five runs of high-throughput pyrophosphate sequencing (Margulies et al., 2005) were performed, the first at Broad Institute and the next four at 454 Life Sciences (Branford, CT). Primary RT-PCR DNA generated previously (Lau et al., 2001) was prepared for sequencing using three different methods. For runs 1 and 2, it was amplified as in (Lau et al., 2001) but substituting pATCGTAGGCACCTGAGA for the 5′ PCR primer and stopping the PCR during the linear phase of amplification. The amplified DNA was purified by phenol/chloroform extraction then native PAGE. Sequencing runs 1 and 2 began with the standard blunt-end ligation step and yielded 283,557 and 298,625 reads, respectively. For run 3, the PCR reaction was smaller (1 × 100 μl) and used primers GCCTCCCTCGCGCCATCAGTATCGTAGGCACCTG AGA and GCCTTGCCAGCCCGCTCAGTATTGATGGTGCCTACAG, which added sequences enabling the blunt-end ligation step of the protocol to be bypassed. This reaction was purified by phenol/chloroform extraction and denaturing (urea) PAGE and yielded 235,632 reads. For runs 4 and 5, PCR DNA was amplified as in run 3 but the second primer was replaced with A_{30}/iSp18/GCCTTGCCAGCCCGCTCAGTATTGA TGGTGCCTACAG (IDT, Inc., Coralville, IA). The 18-atom spacer prevented Taq polymerase from using the poly-A portion of the primer as a template (Williams and Bartel, 1995). PCR product (40 μl) was denatured (85°C, 10 min, formamide loading dye), and the differently sized strands were purified on a 90% formamide, 8% acrylamide gel, yielding single-stranded DNA suitable for the emulsion PCR reaction of the sequencing procedure. Sequencing of the longer strand yielded 196,083 reads (run 4), and the shorter yielded 110,299 reads (run 5). Although runs 4 and 5 yielded fewer reads than the other runs, the diversity of reads matching the genome was comparable.

Read Processing

The 1,124,196 individual sequence reads were processed in four steps. In step 1, 9 nt segments of each linker that immediately flanked the small RNA-derived sequence were found in 850,870 reads (181,668 unique small RNA sequences); the remaining reads were discarded. In step 2, each unique sequence was compared to annotated *C. elegans* miRNA hairpins (miRBase 7.0) (Griffiths-Jones, 2004), and those ≥10 nt and with perfect matches over their entire length were set aside (1002 sequences, 317,694 reads; Table S1). In step 3, sequences with perfect matches to the *E. coli* genome (Hayashi et al., 2001) as found by BLAST (Altschul et al., 1990) were discarded (20,845 sequences, 176,719 reads). In step 4, sequences were compared to the WormBase WS120 assembly of the *C. elegans* genome using BLAST, and those with perfect hits (no gaps or mismatches across their entire length) were retained (23,109 sequences, 77,232 reads). Up to 50 perfect hits to the *C. elegans* genome were recorded per query sequence. In downstream analyses, sequence and read counts were normalized to the number of genomic loci (Supplemental Data). Sequences spanning splice junctions were identified from those without matches in the *E. coli* or *C. elegans* genomes using BLAST to search annotated *C. elegans* cDNAs (Kent and Zahler, 2000).

21U-RNA Upstream Motifs

21U-RNA loci were defined as those whose sequences perfectly matched 21 nt reads beginning with a 5′ T and fell into regions of chromosome IV whose matching normalized reads were dominated by these two properties. Motifs were defined using alignments of genomic sequence surrounding the 21U-RNA loci, with each locus equally weighted. The motif-scoring matrix was constructed using $\log_2$-odds ratios of nucleotide frequencies at positions in the alignments (foreground) to genomic nucleotide frequencies (background). Predicted 21U-RNA loci were those scoring ≥15.5 (Supplemental Data).

An independent set of 245,420 *C. elegans* small RNA pyrosequencing reads was provided by J. Pak and A. Fire (personal communication). Processing as described above yielded 1475 21U-RNA sequences representing 7985 reads. Not present in our data set were 344 sequences. Of those, 157 (46%) matched predicted 21U-RNA loci of chromosome IV, which was a smaller portion than for sequences unique to any of our five data sets (64%, 65%, 66%, 69%, and 72%), indicating that some information represented in our motif model originated from peculiarities of our training set. Nonetheless, of the 4.7 million 21-mers beginning with a T from within those three regions, motif scanning predicted that only 0.1% were loci of unsequenced 21U-RNAs. Thus, correctly predicting almost half of the unique sequences from an independent set of reads (versus 0.1% if those sequences were picked randomly) indicated that most of the information in our model reflected the biological requirements of the motif.

siRNA Methods

Exon coordinates were from WormBase gene annotations (release WS120, 3/1/2004). Counts matching the sense and antisense strands of exons, excluding loci classified as 21U-RNAs, were normalized to the number of genomic loci. Splicing variants were collapsed, leaving 1720 siRNA-complemented genes and 1346 sense RNA-matched genes. To account for expression, SAGE data from the *C. elegans* Gene Expression Consortium (http://elegans.bcgsc.bc.ca) were used to select control cohorts (Supplemental Data).

Molecular Analyses

For alkaline hydrolysis, mixed-stage *C elegans* total RNA (40 μg) was incubated in 0.1 M KOH (90°C, 10 min), then neutralized with TrisHCl. Periodate oxidation and β elimination were as described (Kemper, 1976).

For enzymatic analyses, small RNAs from 800 μg of total RNA were gel purified, and one-fortieth was used to cap with the remainder divided equally for five treatments. Phosphatase (50U CIP, NEB) and rephosphorylation (20U T4 polynucleotide kinase, NEB) were performed according to manufacturer. RNA ligations were as in the second ligation step of the library construction (Lau et al., 2001). Capping was with vaccinia guanylyl transferase (Ambion) and α-^{32}P GTP per manufacturer's instructions. The 26-mer marker was an in vitro transcribed version of siR26-1. Northerns were as described (Lau et al., 2001), except 21U-1 and siR26-1 were hybridized to LNA probes (Exiqon) as described (Vagin et al., 2006).

Supplemental Data

Supplemental data include Supplemental Experimental Procedures, two figures, six tables, and three RNA sequence files and can be found with this article online at http://www.cell.com/cgi/content/full/127/6/1193/DC1/.

ACKNOWLEDGMENTS

We thank W. Johnston for assistance in preparing DNA for high-throughput sequencing, W. Brockman and P. Alvarez for base calling of sequencing run #1, and C. Perbost and others at 454 Life Sciences for sequencing runs #2–5, and J. Pak and A. Fire for sharing sequencing data before publication. We also thank S. Bagby, A. Grishok, A. Grimson, A. Mallory, and H. Vaucheret for useful comments on the manuscript. The SAGE data were produced at the Michael Smith Genome Sciences Centre with funding from Genome Canada. Our work was supported by the Prix Louis D from the Institut de France and a grant from the NIH (D.P.B). D.P.B is an HHMI Investigator.

Received: June 16, 2006
Revised: September 23, 2006
Accepted: October 27, 2006
Published: December 14, 2006

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

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

Aravin, A., Gaidatzis, D., Pfeffer, S., Lagos-Quintana, M., Landgraf, P., Iovino, N., Morris, P., Brownstein, M.J., Kuramochi-Miyagawa, S., Nakano, T., et al. (2006). A novel class of small RNAs bind to MILI protein in mouse testes. Nature *442*, 203–207.

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.

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

Brennecke, J., Stark, A., Russell, R.B., and Cohen, S.M. (2005). Principles of microRNA-target recognition. PLoS Biol. *3*, e85.

Chan, C.S., Elemento, O., and Tavazoie, S. (2005). Revealing posttranscriptional regulatory elements through network-level conservation. PLoS Comput Biol *1*, e69.

Chang, S., Johnston, R.J., Jr., Frokjaer-Jensen, C., Lockery, S., and Hobert, O. (2004). MicroRNAs act sequentially and asymmetrically to control chemosensory laterality in the nematode. Nature *430*, 785–789.

Coghlan, A., and Wolfe, K.H. (2002). Fourfold faster rate of genome rearrangement in nematodes than in *Drosophila*. Genome Res. *12*, 857–867.

Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. Genome Res. *14*, 1188–1190.

Doench, J.G., and Sharp, P.A. (2004). Specificity of microRNA target selection in translational repression. Genes Dev. *18*, 504–511.

Duchaine, T.F., Wohlschlegel, J.A., Kennedy, S., Bei, Y., Conte, D., Jr., Pang, K., Brownell, D.R., Harding, S., Mitani, S., Ruvkun, G., et al. (2006). Functional proteomics reveals the biochemical niche of *C. elegans* DCR-1 in multiple small-RNA-mediated pathways. Cell *124*, 343–354.

Feinberg, E.H., and Hunter, C.P. (2003). Transport of dsRNA into cells by the transmembrane protein SID-1. Science *301*, 1545–1547.

Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S.E., and Mello, C.C. (1998). Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. Nature *391*, 806–811.

Girard, A., Sachidanandam, R., Hannon, G.J., and Carmell, M.A. (2006). A germline-specific class of small RNAs binds mammalian Piwi proteins. Nature *442*, 199–202.

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.

Griffiths-Jones, S. (2004). The microRNA Registry. Nucleic Acids Res. *32*, D109–D111.

Grishok, A., Pasquinelli, A.E., Conte, D., Li, N., Parrish, S., Ha, I., Baillie, D.L., Fire, A., Ruvkun, G., and Mello, C.C. (2001). Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control *C. elegans* developmental timing. Cell *106*, 23–34.

Han, J., Lee, Y., Yeom, K.H., Nam, J.W., Heo, I., Rhee, J.K., Sohn, S.Y., Cho, Y., Zhang, B.T., and Kim, V.N. (2006). Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. Cell *125*, 887–901.

Hayashi, T., Makino, K., Ohnishi, M., Kurokawa, K., Ishii, K., Yokoyama, K., Han, C.G., Ohtsubo, E., Nakayama, K., Murata, T., et al. (2001). Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. DNA Res. *8*, 11–22.

Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, L.S., Tacker, M., and Schuster, P. (1994). Fast folding and comparison of RNA secondary structures. Monatsh. Chem. *125*, 167–188.

Hutvagner, G., and Zamore, P.D. (2002). A microRNA in a multiple-turnover RNAi enzyme complex. Science *297*, 2056–2060.

Hutvagner, G., McLachlan, J., Pasquinelli, A.E., Balint, E., Tuschl, T., and Zamore, P.D. (2001). A cellular function for the RNA-interference enzyme Dicer in the maturation of the *let-7* small temporal RNA. Science *293*, 834–838.

Johnston, R.J., and Hobert, O. (2003). A microRNA controlling left/right neuronal asymmetry in *Caenorhabditis elegans*. Nature *426*, 845–849.

Kemper, B. (1976). Inactivation of parathyroid hormone mRNA by treatment with periodate and aniline. Nature *262*, 321–323.

Kent, W.J., and Zahler, A.M. (2000). The intronerator: exploring introns and alternative splicing in *Caenorhabditis elegans*. Nucleic Acids Res. *28*, 91–93.

Ketting, R.F., Fischer, S.E., Bernstein, E., Sijen, T., Hannon, G.J., and Plasterk, R.H. (2001). Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in *C. elegans*. Genes Dev. *15*, 2654–2659.

Lagos-Quintana, M., Rauhut, R., Lendeckel, W., and Tuschl, T. (2001). Identification of novel genes coding for small expressed RNAs. Science *294*, 853–858.

Lall, S., Grun, D., Krek, A., Chen, K., Wang, Y.L., Dewey, C.N., Sood, P., Colombo, T., Bray, N., Macmenamin, P., et al. (2006). A genome-wide map of conserved microRNA targets in *C. elegans*. Curr. Biol. *16*, 460–471.

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.

Lau, N.C., Seto, A.G., Kim, J., Kuramochi-Miyagawa, S., Nakano, T., Bartel, D.P., and Kingston, R.E. (2006). Characterization of the piRNA complex from rat testes. Science *313*, 363–367.

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, R.C., Hammell, C.M., and Ambros, V. (2006). Interacting endogenous and exogenous RNAi pathways in *Caenorhabditis elegans*. RNA *12*, 589–597.

Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., Lee, J., Provost, P., Radmark, O., Kim, S., and Kim, V.N. (2003). The nuclear RNase III Drosha initiates microRNA processing. Nature *425*, 415–419.

Lewis, B.P., Burge, C.B., and Bartel, D.P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell *120*, 15–20.

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.

Li, J., Yang, Z., Yu, B., Liu, J., and Chen, X. (2005). Methylation protects miRNAs and siRNAs from a 3′-end uridylation activity in *Arabidopsis*. Curr. Biol. *15*, 1501–1507.

Lim, L.P., Lau, N.C., Weinstein, E.G., Abdelhakim, A., Yekta, S., Rhoades, M.W., Burge, C.B., and Bartel, D.P. (2003). The microRNAs of *Caenorhabditis elegans*. Genes Dev. *17*, 991–1008.

Lu, C., Tej, S.S., Luo, S., Haudenschild, C.D., Meyers, B.C., and Green, P.J. (2005). Elucidation of the small RNA component of the transcriptome. Science *309*, 1567–1569.

Makeyev, E.V., and Bamford, D.H. (2002). Cellular RNA-dependent RNA polymerase involved in posttranscriptional gene silencing has two distinct activity modes. Mol. Cell *10*, 1417–1427.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature *437*, 376–380.

Martin, C.T., and Coleman, J.E. (1989). T7 RNA polymerase does not interact with the 5′-phosphate of the initiating nucleotide. Biochemistry *28*, 2760–2762.

Mourelatos, Z., Dostie, J., Paushkin, S., Sharma, A., Charroux, B., Abel, L., Rappsilber, J., Mann, M., and Dreyfuss, G. (2002). miRNPs: a novel class of ribonucleoproteins containing numerous microRNAs. Genes Dev. *16*, 720–728.

Nakayashiki, H. (2005). RNA silencing in fungi: mechanisms and applications. FEBS Lett. *579*, 5950–5957.

Ohler, U., Yekta, S., Lim, L.P., Bartel, D.P., and Burge, C.B. (2004). Patterns of flanking sequence conservation and a characteristic upstream motif for microRNA gene identification. RNA *10*, 1309–1322.

Pak, J., and Fire, A. (2006). Distinct populations of primary and secondary effectors during RNAi in *C. elegans*. Science, in press.

Ranjith-Kumar, C.T., Gutshall, L., Kim, M.J., Sarisky, R.T., and Kao, C.C. (2002). Requirements for de novo initiation of RNA synthesis by recombinant flaviviral RNA-dependent RNA polymerases. J. Virol. *76*, 12526–12536.

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.

Sijen, T., and Plasterk, R.H. (2003). Transposon silencing in the *Caenorhabditis elegans* germ line by natural RNAi. Nature *426*, 310–314.

Sijen, T., Fleenor, J., Simmer, F., Thijssen, K.L., Parrish, S., Timmons, L., Plasterk, R.H., and Fire, A. (2001). On the role of RNA amplification in dsRNA-triggered gene silencing. Cell *107*, 465–476.

Smardon, A., Spoerke, J.M., Stacey, S.C., Klein, M.E., Mackin, N., and Maine, E.M. (2000). EGO-1 is related to RNA-directed RNA polymerase and functions in germ-line development and RNA interference in *C. elegans*. Curr. Biol. *10*, 169–178.

Stein, L.D., Bao, Z., Blasiar, D., Blumenthal, T., Brent, M.R., 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.

Tijsterman, M., Ketting, R.F., Okihara, K.L., Sijen, T., and Plasterk, R.H. (2002). RNA helicase MUT-14-dependent gene silencing triggered in *C. elegans* by short antisense RNAs. Science *295*, 694–697.

Tomari, Y., and Zamore, P.D. (2005). Perspective: machines for RNAi. Genes Dev. *19*, 517–529.

Vagin, V.V., Sigova, A., Li, C., Seitz, H., Gvozdev, V., and Zamore, P.D. (2006). A distinct small RNA pathway silences selfish genetic elements in the germline. Science *313*, 320–324.

Williams, K.P., and Bartel, D.P. (1995). PCR product with strands of unequal length. Nucleic Acids Res. *23*, 4220–4221.

Zamore, P.D., Tuschl, T., Sharp, P.A., and Bartel, D.P. (2000). RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. Cell *101*, 25–33.

Accession Numbers

All sequences with linker matches were deposited in the Gene Expression Omnibus (GSE5990). The 21U-RNA sequences were deposited in GenBank (EF044580–EF050033). MicroRNA sequences were submitted to miRBase (Griffiths-Jones, 2004). 21U-RNA, siRNA, and sense RNA sequences are also provided in FASTA format in Supplemental Data.