Improved annotation of *C. elegans* microRNAs by deep sequencing reveals structures associated with processing by Drosha and Dicer

M. BRYAN WARF,[1] W. EVAN JOHNSON,[2] and BRENDA L. BASS[1]
[1]Department of Biochemistry, University of Utah, Salt Lake City, Utah 84112, USA
[2]Department of Statistics, Brigham Young University, Provo, Utah 84602, USA

ABSTRACT

MicroRNAs (miRNAs) are small regulatory RNAs that are essential in all studied metazoans. Research has focused on the prediction and identification of novel miRNAs, while little has been done to validate, annotate, and characterize identified miRNAs. Using Illumina sequencing, ~20 million small RNA sequences were obtained from *Caenorhabditis elegans*. Of the 175 miRNAs listed on the miRBase database, 106 were validated as deriving from a stem–loop precursor with hallmark characteristics of miRNAs. This result suggests that not all sequences identified as miRNAs belong in this category of small RNAs. Our large data set of validated miRNAs facilitated the determination of general sequence and structural characteristics of miRNAs and miRNA precursors. In contrast to previous observations, we did not observe a preference for the 5′ nucleotide of the miRNA to be unpaired compared to the 5′ nucleotide of the miRNA*, nor a preference for the miRNA to be on either the 5′ or 3′ arm of the miRNA precursor stem–loop. We observed that steady-state pools of miRNAs have fairly homogeneous termini, especially at their 5′ end. Nearly all mature miRNA–miRNA* duplexes had two nucleotide 3′ overhangs, and there was a preference for a uracil in the first and ninth position of the mature miRNA. Finally, we observed that specific nucleotides and structural distortions were overrepresented at certain positions adjacent to Drosha and Dicer cleavage sites. Our study offers a comprehensive data set of *C. elegans* miRNAs and their precursors that significantly decreases the uncertainty associated with the identity of these molecules in existing databases.

Keywords: miRNA biogenesis; miRNA processing; pri-miRNA processing; pre-miRNA processing

INTRODUCTION

MicroRNAs (miRNAs) are small regulatory RNAs that are essential in all studied metazoans (for review, see Ghildiyal and Zamore 2009; Kim et al. 2009). Originally discovered in *C. elegans* (Lee et al. 1993; Wightman et al. 1993), hundreds of miRNAs have been subsequently identified in dozens of species. miRNAs regulate gene expression post-transcriptionally, causing either mRNA degradation or inhibiting translation of target mRNAs.

Biogenesis of miRNAs has multiple steps, taking place in both the nucleus and cytoplasm. The primary transcript (pri-miRNA) from which the miRNA derives is usually many kilobases in size and generally transcribed by RNA polymerase II (Cai et al. 2004; Lee et al. 2004). A stem–loop miRNA precursor (pre-miRNA) that is ~60–70 nucleotides (nt) in length is excised from the pri-miRNA in the nucleus by the RNase III protein Drosha (Lee et al. 2003) and its partner protein Pasha (DGCR8 in humans) (Denli et al. 2004; Gregory et al. 2004; Han et al. 2004; Landthaler et al. 2004). The pre-miRNA is then exported from the nucleus to the cytoplasm by Exportin 5 (Yi et al. 2003; Bohnsack et al. 2004; Lund et al. 2004). In the cytoplasm, the pre-miRNA is processed by the RNase III protein Dicer to remove the loop and create a short double-stranded RNA (dsRNA) duplex between the mature miRNA (~20–24 nt) and its partner strand, called the miRNA* (Bernstein et al. 2001; Grishok et al. 2001; Hutvagner et al. 2001; Ketting et al. 2001; Knight and Bass 2001). The miRNA–miRNA* dsRNA duplex is loaded into an Argonaute protein that is part of the RISC (RNA induced silencing complex), where it is determined which of the two arms will be the miRNA that guides the RISC (Schwarz et al. 2003). Using the miRNA to guide its binding, the RISC then targets an mRNA, silencing

Reprint requests to: Brenda L. Bass, Department of Biochemistry, University of Utah, Salt Lake City, UT 84112, USA; email: bbass@biochem.utah.edu; fax: (801) 581-5379.
Article published online ahead of print. Article and publication date are at http://www.rnajournal.org/cgi/doi/10.1261/rna.2432311.

expression by promoting its degradation and/or inhibiting its translation (Ghildiyal and Zamore 2009). It has been reported that determination of which strand will be the miRNA is influenced by the thermodynamic stability at the termini of the miRNA–miRNA* duplex, with the miRNA having the less stable 5′ end (Krol et al. 2004; Tomari et al. 2004; Han et al. 2006; Kawamata et al. 2009; Ghildiyal et al. 2010). This is not a stringent rule for selection, as there are cases where both the miRNA and miRNA* are used for RISC-mediated gene silencing (Okamura et al. 2008; Kim et al. 2009; Ghildiyal et al. 2010).

Previous studies have thoroughly analyzed small RNA populations of *C. elegans* to identify miRNAs (Lau et al. 2001; Lee and Ambros 2001; Ambros et al. 2003b; Grad et al. 2003; Lim et al. 2003; Ohler et al. 2004; Ruby et al. 2006; Friedlander et al. 2008; Kato et al. 2009). A handful of miRNAs have been studied in depth, to determine if they derive from a stem–loop pre-miRNA and if their production requires proteins of the miRNA biogenesis pathway. However, many identified miRNAs have not been validated, since neither miRNA* nor pre-miRNA sequences have been annotated. At least one study made an effort to validate miRNAs and annotate miRNA* sequences by analyzing a high-throughput sequencing data set (Ruby et al. 2006). However, these efforts were hampered by the limited sequence coverage from available technologies, and consequently, there are still large gaps in our understanding of miRNA and miRNA* features. Analyses have been particularly restricted by the low abundance of miRNA* sequences in steady-state RNA populations, presumably because this species is rapidly degraded. Consequently, since a miRNA* sequence is required to infer the pre-miRNA, less is known about the sequences and structures of the pre-miRNA.

We capitalized on recent advances in sequencing technology that dramatically increased the number of reads obtained for small RNA populations. Our analysis of *C. elegans* small RNAs provides strong validation for many annotated miRNAs, while raising questions about the validity of others. We annotated numerous miRNA* sequences that had not previously been determined, allowing us to assemble sequences for many unannotated pre-miRNAs. Using our new data sets, we provide a comprehensive analysis of the features of Drosha and Dicer products. Our high confidence data sets of miRNA, miRNA*, and pre-miRNA sequences, as well as pre-miRNA structures, will greatly facilitate the study of miRNA biogenesis and function.

RESULTS

Sequencing and validation of miRNAs

A cDNA library of small RNAs (18–30 nt) was prepared from mixed-stage *C. elegans*, using a method that captures RNAs with 5′ monophosphates to enrich for products of Dicer and Drosha. Approximately 25,000,000 sequencing reads were obtained by an Illumina Genome Analyzer IIx, of which 19,300,000 aligned to the *C. elegans* genome. Reads were categorized as to whether they aligned to known miRNA loci, piRNA loci, or endogenous-siRNA (endo-siRNA) loci.

We obtained ~15 million sequencing reads that aligned to previously identified miRNAs. For many miRNAs, a sequence for the miRNA* or loop cleavage product (loop CP, the terminal portion of the pre-miRNA containing the loop removed by Dicer) has not been identified. We therefore also used our data set to validate these miRNAs and identify unannotated miRNA* and loop CP sequences. The most recent release of miRBase (release #15, April 2010) (Griffiths-Jones et al. 2008) was used as a reference for the sequences of identified miRNAs and miRNA*s.

miRNAs were considered validated if we observed sequencing reads for both the miRNA and miRNA*. Other characteristics were also required for validation: homogeneous termini, 3′ overhang cleavage products, and a pre-miRNA size of ~60-70 nt. These criteria differed slightly from previous studies (Ambros et al. 2003a), which did not require miRNA* sequencing reads. With the advance in sequencing technology, it is now feasible to obtain reads for miRNA*s to verify that the miRNA comes from a stem–loop precursor.

miRBase currently lists 175 *C. elegans* miRNAs (Griffiths-Jones et al. 2008). Using our criteria, 106 were validated, while 14 were partially validated, and 55 could not be validated (Tables 1, 2). Supplemental Table 1 lists the predominant sequencing read in our data set for each miRNA. The majority of validated miRNAs had a predominant sequencing read that was the same sequence as listed on miRBase (92 of 106). However, following the assumption that sequencing reads for miRNAs are more abundant than miRNA*s (Hutvagner et al. 2001; Lim et al. 2003; Schwarz et al. 2003), 14 miRNAs were misannotated as miRNA* on miRBase. Given the limited sequencing coverage in the past, it is understandable that miRNA*s were erroneously annotated as miRNAs. Five miRNAs had a similar number of reads as their miRNA*, leaving it unclear which was the miRNA or suggesting that both may be chosen by the RISC at comparable levels.

Of the 106 validated miRNAs, 65 did not have annotated miRNA* and loop CP sequences listed on miRBase. Therefore, miRBase only lists pre-miRNA sequences for 41 of the validated miRNAs. Our data set indicated that, of these 41 pre-miRNAs, only 18 had correct miRNA* sequences listed on miRBase, while 19 had correct loop CP sequences.

Since miRNA and miRNA* sequences were both identified in our data set, a pre-miRNA sequence was assembled for each miRNA (Supplemental Table 1). If a loop CP sequence was not in the data set, it was inferred from the termini of the predominant miRNA and miRNA*

TABLE 1. Overall statistics for the validation of miRNAs and annotation of miRNA*s and loop cleavage products (loop CPs)

Total miRNAs listed on miRBase	175
Validated	106
Partially validated	14
Not validated	55
Not validated	
No reads for miRNA or miRNA*	26
miRNA reads obtained, but no miRNA* reads	26
Likely endogenous siRNA cluster	3
Partially validated	
High variability of the miRNA's 5′ or 3′ nucleotide	5
miRNA–miRNA* duplex has atypical overhangs	4
Atypical size of pre-miRNA	1
miRNA reads obtained, but no miRNA* reads, but miRNA IPs with ALG-1[a]	9
Validated miRNAs	
miRNA with correct sequence	92
miRNA with incorrect sequence	14
miRNA* with correct sequence	18
miRNA* with incorrect sequence	23
miRNA* not annotated	65
Loop CP with correct sequence	19
Loop CP with incorrect sequence	22
Loop CP not annotated	65
miRNA and miRNA* switched	14
Unclear which is miRNA and miRNA*	5
Correct miRNA and miRNA*	87

[a]These sequences were observed to IP with ALG-1 in *C. elegans* (Zisoulis et al. 2010).

sequences. Of the 101 pre-miRNAs for which the miRNA and miRNA* could be distinguished, the miRNA was on the 5′ arm of the pre-miRNA 42% of the time and on the 3′ arm 58% of the time. Thus, in *C. elegans* there does not appear to be a strong bias for the miRNA to be on either arm of the pre-miRNA. This contrasts with a previous study in *C. elegans* that reported a twofold enrichment for the miRNA to be on the 3′ arm of the pre-miRNA (de Wit et al. 2009). However, that study had limited sequencing coverage (160,000 reads), and used all miRNA sequences listed on miRBase without additional validation, both of which could give rise to the discrepancy.

miRNAs that could not be validated

Of the 69 miRNAs that we could not validate, we considered 14 to be partially validated (Table 1). Five of these miRNAs had sequencing reads for both the miRNA and miRNA* but had inconsistencies when compared to other miRNAs. All five had 5′ and 3′ termini that were more heterogeneous compared to the validated miRNAs, and four of the five had miRNA–miRNA* duplexes with 5′ overhangs, which was atypical compared to all validated miRNAs. Additionally, *miR-229* had a pre-miRNA sequence that was longer than validated miRNAs. These inconsistencies do not discount these sequences as miRNAs, but we only considered them partially validated.

There were 35 miRNAs for which we obtained miRNA sequencing reads but did not observe a corresponding miRNA* sequence among aligned reads ("Partially Validated" and "Not Validated" categories in Tables 1, 2). Nine of these miRNA sequences immunoprecipitate (IP) with ALG-1 (Supplemental Table 1) (Zisoulis et al. 2010), indicating that they are real miRNAs, and thus, we considered these miRNAs to be partially validated (Tables 1, 2; Supplemental Table 1). One of the remaining 26 putative miRNAs for which we did not observe miRNA* reads was the *lsy-6* miRNA. The majority of previous sequencing studies have not obtained reads for the *lsy-6* miRNA (Ruby et al. 2006;

TABLE 2. List of individual miRNAs and to which validation category they belong

Validated miRNAs	*let-7, lin-4, 1, 2, 34-61, 63-67, 70-77, 79-87, 90, 124, 228, 230-238, 239a, 240-241, 243-247, 250, 253, 255, 259, 357-358, 360, 392, 784, 786-788, 790-792, 794-795, 797, 800, 1020, 1022, 1819-1820, 1822, 1824, 1828, 1829a-b, 1830, 2210, 2219, 2953*
miRNA and miRNA* are switched	*41, 44-45, 54, 75, 77, 83, 86, 124, 232, 240, 788, 1829a, 1830*
No clear miRNA or miRNA*	*46-47, 1824, 1828, 2210*
Partially validated	
High variability of the miRNA's 5′ or 3′ nucleotide	*229, 249, 252, 1834, 2214*
miRNA–miRNA* duplex has atypical overhangs	*249, 252, 1834, 2214*
Atypical size of pre-miRNA	*229*
miRNA reads obtained, but no miRNA* reads. miRNA IPs with ALG-1[a]	*62, 242, 248, 251, 785, 793, 799, 1821, 1829c*
Cannot validate	
No reads for miRNA	*256, 258-1, 258-2, 261, 264-273, 353-354, 356, 1019, 1021, 1832, 1832b, 2208a-b, 2209c, 2217, 2218a*
Only reads for miRNA	*lsy-6, 78, 239b, 254, 256-257, 355, 359, 789-1, 789-2, 796, 798, 1018, 1817-1818, 1823, 1833, 2207, 2209a-b, 2211, 2212, 2213, 2215-2216, 2218b, 2220, 2221*
Likely endogenous siRNA cluster	*257, 260, 262*

[a]These sequences were observed to IP with ALG-1 in *C. elegans* (Zisoulis et al. 2010).

Kato et al. 2009; Zisoulis et al. 2010), and *lsy-6* miRNA sequences do not IP with ALG-1 (Zisoulis et al. 2010). This is not surprising since *lsy-6* is only expressed in two neurons (Johnston and Hobert 2003). Nonetheless, we obtained ~260 reads for the *lsy-6* miRNA, albeit none for the miRNA*. While previous studies are consistent with *lsy-6* being a miRNA (Johnston and Hobert 2003), we cannot validate it using our criteria.

We did not obtain any reads for 26 other miRNAs currently listed on miRBase ("Not Validated" category in Tables 1, 2 and Supplemental Table 1). Thirteen of these miRNAs were based solely on computational predictions (Grad et al. 2003; Ohler et al. 2004), and the absence of sequencing reads indicates that these may be false positive predictions. Previous attempts to obtain sequencing reads for many of these miRNAs also failed (de Wit et al. 2009; Kato et al. 2009; Zisoulis et al. 2010), providing further evidence that these sequences are not miRNAs.

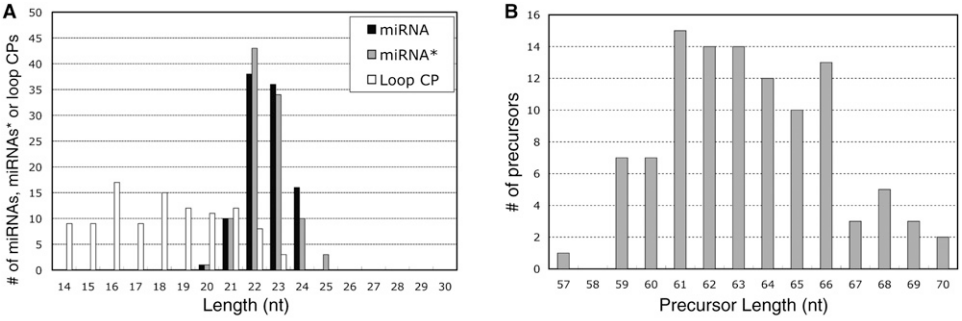
Finally, the characteristics of three miRNAs suggested that they were more likely to be endogenous siRNA clusters ("Not Validated" category in Tables 1, 2 and Supplemental Table 1), since there was no predominating sequence that could be identified as a miRNA. Instead, many overlapping reads had moderate coverage throughout the cluster, with one cluster (*miR-262*) spanning hundreds of nucleotides, which is greater than the normal ~60–70-nt length of a pre-miRNA. Additionally, these overlapping reads were very heterogeneous at their 3′ termini, with <50% of the reads ending at the same nucleotide. All of these characteristics are emblematic of an endogenous siRNA cluster (Ruby et al. 2006; Ghildiyal and Zamore 2009; Kim et al. 2009).

General structural characteristics of validated pre-miRNAs

For each miRNA locus, the predominant miRNA, miRNA*, and loop CP sequencing read was determined and used to assemble a pre-miRNA sequence (Supplemental Table 1). For validated miRNA loci, the majority of both miRNAs and miRNA*s were 22 or 23 nt in length, with an average length of 23 ± 1 nt for both (Fig. 1A,C). The loop CP showed a broad length distribution, ranging from 14 to 24 nt, with an average length of 18 ± 3 nt (Fig. 1A,C). Thus, miRNA processing can occur with a broad range of loop CP lengths. For the validated miRNAs, the majority of pre-miRNAs were 61–66 nt, with an average length of 63 ± 3 nt (Fig. 1B,C).

We determined how many base-pairs or structural distortions were in the pre-miRNA or miRNA-miRNA* duplex (Fig. 1D). First, structures for all the pre-miRNAs were created using the lowest energy state predicted by mfold (Supplemental Fig. 1) (Zuker 2003). No miRNA–miRNA* duplex was fully base-paired, and the number of base-pairs was similar between most duplexes, with all duplexes having an average of 17 ± 1 bps. We observed two types of distortions in the miRNA–miRNA* duplex. Internal loops were more frequent than bulged nucleotides; there were on average 2 ± 1 internal loops versus 1 ± 1 bulged nt per duplex.

The same analysis was performed for pre-miRNA structures, which had on average an additional 5 bps and 1 bulge compared to the miRNA–miRNA* duplex (Fig. 1D). The length of the terminal loop was on average 9 ± 4 nt. However, one should be careful when making conclusions



C	Length (nt) Mean	Median
miRNA	23 ± 1	23
miRNA*	23 ± 1	22
Loop CP	18 ± 3	18
Precursor	63 ± 3	63

D	Base-pairs	Int. Loops	Bulged nts	Loop length (nt)
miRNA: miRNA*	17 ± 1	2 ± 1	1 ± 1	n/a
Precursor	22 ± 2	2 ± 1	2 ± 1	9 ± 4

FIGURE 1. General characteristics of validated miRNAs. (*A*) For each locus, the predominant read length in nucleotides (nt) for the miRNAs, miRNA*s, and loop CPs was determined. The bar graph shows the number of loci with a predominant read of a given length. (*B*) The bar graph shows the number of loci with a pre-mRNA sequence of a given length. (*C*) The mean and median length for indicated RNA species are tabulated, with standard deviation. (*D*) The average numbers of various features are tabulated for the mature miRNA–miRNA* duplexes and pre-miRNAs, with standard deviation. The total number of base-pairs, internal loops (int. loop) and bulged nucleotides are tabulated. For the pre-miRNA, the length (in nt) of the terminal loop is also given.

about loops from computationally predicted structures; a study of 10 human pre-miRNAs found that mfold incorrectly predicted base-pairing near the loop 70% of the time (Krol et al. 2004).

miRNA–miRNA* duplexes predominately have 2-nt 3′ overhangs

We observed that the validated miRNA–miRNA* duplexes predominately had 2-nt 3′ overhangs. Of the 106 validated miRNAs (whose miRNA–miRNA* duplexes contain a total of 212 termini), 87.0% of the termini had a 2-nt 3′ overhang (Fig. 2A). Both 1-nt and 3-nt 3′ overhangs were each present ~5.5% of the time. A fraction of termini (1.9%) had other overhangs.

In some cases, a Drosha or Dicer cleavage site was observed to encompass an internal loop, which could appear to give rise to an atypical overhang. For example, *pri-miR-67* has an internal loop at Drosha's cleavage site on the 5′ arm of the pri-miRNA (Fig. 2B). Cleavage at this site gives rise to a pre-miRNA with a 3-nt 3′ overhang and a 1-nt 5′ overhang. However, sequences in the pri-miRNA directly upstream of the cleavage site form base-pairs, leading us to believe that this pri-miRNA is actually positioned in Drosha's active site in a way that would typically lead to a 2-nt 3′ overhang product. We therefore scored *pri-miR-67*

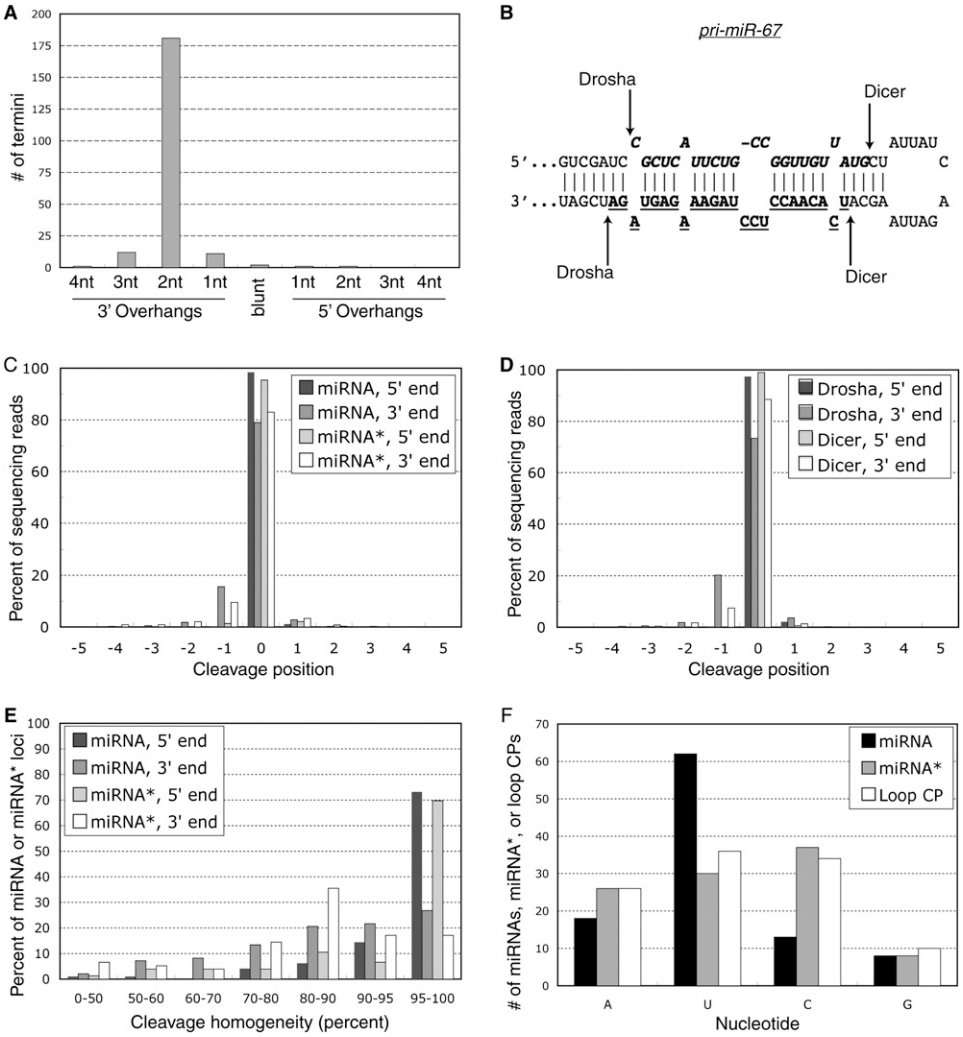


FIGURE 2. Characteristics of the termini for miRNAs, miRNA*s, and miRNA–miRNA* duplexes. Unless indicated, general features were compared using the pre-miRNA assembled from the predominant read for each RNA species. (*A*) Bar height indicates the number of termini with a given type of end among all mature miRNA–miRNA* duplexes. Each duplex is counted twice, once for each terminus. (*B*) The lowest free-energy secondary structure predicted by mfold is depicted for *pri-miR-67*, with mature miRNA sequence in underlined font and miRNA* sequence italicized. Arrows are Drosha and Dicer cleavage sites. (*C*) Reads for each validated locus were examined for the homogeneity of the 5′ and 3′ nucleotides (with respect to genomic position) for miRNAs and miRNA*s. The zero position denotes the predominant sequencing read, with negative and positive values indicating shorter and longer sequencing reads, respectively. (*D*) Same as in C, but for 5′ and 3′ termini produced specifically by either Drosha or Dicer. (*E*) Plot depicts the percent of all validated miRNAs or miRNA*s with a given homogeneity at the first and last nucleotide. (*F*) The bar graph depicts the number of miRNAs, miRNA*s, or loop CPs that have a specific nucleotide at their 5′ terminus.

and similar examples as a 2-nt 3′ overhang. Of relevance is a recent publication demonstrating that many 2-nt symmetrical internal loops form noncanonical base-pairs that are not recognized as distortions by RNases or metal ions, suggesting that some internal loops in miRNA precursors might not be recognized as distortions by Drosha or Dicer (Krol et al. 2004).

miRNAs and miRNA*s have homogeneous 5′ and 3′ nucleotides in steady-state populations, independent of the structure at the cleavage site

We next examined the 5′ and 3′ nucleotides of miRNA and miRNA* sequencing reads to evaluate the homogeneity of the first and last nucleotides, with respect to genomic position. It has been observed in mice and flies that miRNAs have 5′ nucleotides that are very homogeneous and usually start at the same nucleotide but have 3′ nucleotides that are heterogeneous (Wu et al. 2007; Seitz et al. 2008). While Dicer and Drosha produce these termini, it should be emphasized that other proteins may selectively stabilize or destabilize the miRNA or miRNA* based on other characteristics, such as the identity of the 5′ nucleotide (Mi et al. 2008; Ghildiyal et al. 2010). Therefore, in any analysis of the cleavage sites produced by Drosha or Dicer, the most stable and predominant miRNA or miRNA* sequence may not be the product that is preferentially made by Drosha or Dicer. For example, it has been observed that the heterogeneity at the 5′ termini for miRNAs and miRNA*s is decreased when the miRNA or miRNA* is loaded into fly Argonaute 2 (AGO-2), as AGO-2 is hypothesized to degrade any miRNA or miRNA* that begins with an alternate 5′ nucleotide (Seitz et al. 2008).

We observed that the 5′ nucleotide for the *C. elegans* miRNA and miRNA* were equally homogeneous; each sequence began at the same nucleotide 98.4% and 95.4% of the time, respectively (Fig. 2C). Both the miRNA and miRNA* had similarly decreased levels of homogeneity at the 3′ nucleotide, with each sequencing read ending at the same nucleotide 79.0% and 83.0% of the time, respectively. The alternative length for the 3′ terminus was mostly 1 nt shorter (leaving a 1-nt 3′ overhang) for both the miRNA and miRNA*. These results correlate well with a previous study in *C. elegans* (Ruby et al. 2006).

When looking at termini produced by either Drosha or Dicer, we observed that the 5′ nucleotides were similarly homogeneous in steady-state RNA populations. The 5′ nucleotides produced by Drosha began at the same nucleotide 97.4% of the time, and the 5′ nucleotides made by Dicer had the same nucleotide 99.0% of the time (Fig. 2D). The 3′ nucleotides produced by Drosha were less homogeneous than 3′ nucleotides produced by Dicer. The 3′ nucleotides made by Drosha were at the same nucleotide 73.4% of the time, while the 3′ nucleotides made by Dicer had the same nucleotide 88.5% of the time. As before, the predominate alternative product was 1 nt shorter for either protein, leaving a 1-nt 3′ overhang. The increased heterogeneity of Drosha's 3′ nucleotides (compared to Dicer's 3′ nucleotides) can be attributed almost entirely to *miR-58*, a highly abundant miRNA (4.6 million reads) with a variable 3′ end (60% homogeneous). If *miR-58* was excluded, the 3′ nucleotides made by Drosha occurred at the same nucleotide 88.0% of the time, a value nearly identical to Dicer's 88.5%.

Using the predominant sequencing read for each validated miRNA locus, we tabulated the homogeneity of 5′ and 3′ nucleotides among different loci (Fig. 2E). The homogeneity of 5′ nucleotides for miRNAs and miRNA*s was similar, with ~70% of loci showing miRNA or miRNA* sequencing reads beginning at the same 5′ nucleotide >95% of the time. As expected, the 3′ termini for miRNAs and miRNA*s were broadly distributed. Heterogeneous or homogeneous cleavage sites were just as likely to have mismatches or internal loops as they were to be base-paired, indicating that homogeneity at a cleavage site is not attributable to the structure of the miRNA precursor. We observed that heterogeneity at one terminus of an individual miRNA (or miRNA*) did not correlate with heterogeneity at the other terminus of the miRNA. Furthermore, we did not observe any correlation when comparing the heterogeneity of a miRNA terminus to the opposing miRNA* terminus in the miRNA–mRNA* duplex.

Alignments of miRNAs, miRNA*s, and loop CPs

It has been previously reported that there is a high frequency of uracil at the 5′ nucleotide of *C. elegans* miRNAs (Lagos-Quintana et al. 2001; Lau et al. 2001; Ambros et al. 2003b; Lim et al. 2003; Ruby et al. 2006; Zhang et al. 2009; Ghildiyal et al. 2010). We therefore generated sequence logos to determine if there were any positions within miRNA, miRNA*, or loop CP sequences with nucleotide frequencies that were significantly different from background frequencies. We used the WebLogo application, which requires sequences of the same length, to generate sequence logos (Crooks et al. 2004). We generated logos using the first and last 10 nt of the miRNA, miRNA*, or loop CP (Supplemental Fig. 2A–C). We also generated a logo for 1,000 randomly selected RNA sequences that are expressed in *C. elegans*, to determine background nucleotide frequencies (Supplemental Fig. 2F); the background frequencies of uracil and adenosine were 31%, while cytosine and guanosine had background frequencies of 19%–20%. Overall, we observed that the nucleotide frequencies within pre-miRNAs were very similar to background; adenosine was 31%, uracil was 28%, cytosine was 21%, and guanosine was 20%.

While most positions were not significantly different from background, we observed a few specific positions within miRNA, miRNA*, and loop CP sequences that differed in a statistically significant manner from the background frequencies. Specifically, the 5′ nucleotide of the miRNA, miRNA*, and loop CP all varied compared to background

frequencies. As expected, the frequency of a 5′ uracil for miRNAs (61%) was well above the background frequency of uracil (31%), with a χ^2 test indicating a significant difference (p-value = 7.4×10^{-9}) (Fig. 2F; Supplemental Fig. 2). The levels of a 5′ uracil were similar to the background (31%) for miRNA*s (31%) and loop CP sequences (27%). The 5′ terminus of the miRNA or miRNA* is equally likely to be made by Drosha or Dicer; thus the increased frequency of uracil is likely independent of either Drosha or Dicer. Studies in *Drosophila melanogastor* and *Arabidopsis thaliana* found that specific Argonautes have preferences for 5′ nucleotides (Mi et al. 2008; Ghildiyal et al. 2010). The miRNA-specific Argonaute ALG-1 in *C. elegans* may have a preference for a 5′ uracil in miRNAs, possibly to help to differentiate between the miRNA and miRNA* sequences.

The miRNA, miRNA*, and loop CP sequences also all had reduced frequencies of a 5′ guanosine (8% for each sequence) compared to the background frequency (20%). This depletion was statistically significant for miRNAs (p-value = 3.0×10^{-2}), miRNA*s (p-value = 3.0×10^{-2}), and loop CP sequences (p-value = 3.9×10^{-2}), suggesting that Drosha and Dicer discriminate against a guanosine in these positions, although other proteins that facilitate processing could also be responsible for this discrimination. Frequencies of the 5′ nucleotide were independent of the length of the sequence (Supplemental Fig. 3).

An alignment of all known miRNAs in metazoans shows an enrichment for uracil in the first position, as well as the ninth and last five positions of the miRNA (Zhang et al. 2009). As stated, we observed an increased frequency of uracil at the first position (61%), and we also observed a significantly increased frequency of uracil at the ninth position (51%) of miRNAs (p-value = 2.9×10^{-4}). However, we did not observe an increased frequency of uracil at the last five positions of the miRNA (Supplemental Fig. 2A,F). Interestingly, the previous study noted that the first and ninth positions are the boundaries of the seed sequence (Zhang et al. 2009), implying that it may be advantageous for an Argonaute to have uracil nucleotides directly flanking the seed sequence.

Finally, we observed that the first 10 nt in the loop CP were significantly depleted of guanosine compared to the background (p-value = 5.5×10^{-4}; Supplemental Fig. 2C,F), while frequencies of adenosine, cytosine, and uracil were not significantly different (p-values all > 0.1). Aside from their 5′ nucleotide, neither the miRNA nor the miRNA* had a similar depletion of guanosines.

Evaluation of nucleotide frequencies surrounding Drosha and Dicer cleavage sites

In addition to evaluating nucleotide frequencies along the length of the miRNA, miRNA*, and loop CP (Fig. 2F; Supplemental Fig. 2), we also generated sequence logos of the nucleotide frequencies surrounding Drosha and Dicer cleavage sites (Fig. 3; Supplemental Fig. 2D,E). We observed frequencies that were significantly different from background at some positions, but a strict requirement for certain nucleotides at specific cleavage sites was not apparent. Consistent with the decreased frequency of guanosine at the 5′ nucleotide of the miRNA, miRNA* and loop CP, we observed a statistically significant depletion of guanosine at N^2, 3′ of a Drosha cleavage site, and N^4 and N^6, 3′ of each Dicer cleavage site (Fig. 3).

There were only three additional positions with frequencies that were different from background in a statistically significant manner. There was an increase in uracil 3′ of the Dicer cleavage site on the 3′ arm of the pre-miRNA (position N^6), most likely because this position is the 5′ nucleotide of the miRNA 58% of the time. Additionally, there was also a depletion in the frequency of adenosine 5′ of the Drosha cleavage site on the 5′ arm of the pri-miRNA (position N^1) and an increase in the frequency of adenosine 5′ of the Dicer cleavage site on the 5′ arm of the pre-miRNA (position N^3). The nucleotides at Drosha and Dicer cleavage sites that occur at frequencies significantly different from background could relate to sequence preferences intrinsic to the enzyme active site or could relate somehow to preferred structures. For example, both Drosha and Dicer may have a preference against a guanosine 3′ of a cleavage site, as three of the four cleavage sites have significantly depleted frequencies of guanosine.

The miRNA 5′ nucleotide is just as likely to be base-paired as the miRNA* 5′ nucleotide

Previous studies led to the hypothesis that the strand of the miRNA–miRNA* duplex with an unpaired or less stable

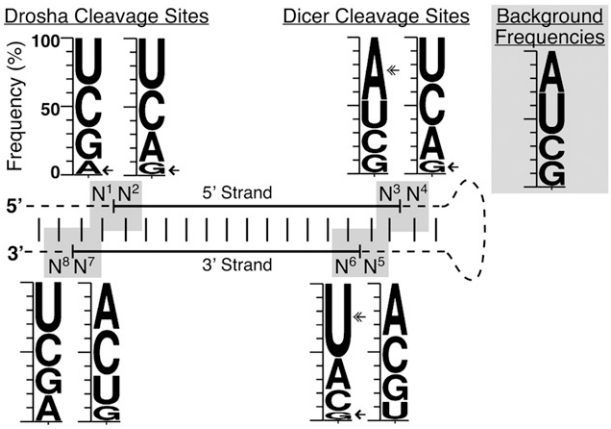


FIGURE 3. Sequence logos indicate nucleotide frequencies observed in steady-state small RNA populations for nucleotides adjacent to Drosha (N^1, N^2, N^7, N^8) and Dicer (N^3, N^4, N^5, N^6) cleavage sites. For each sequence logo, the Y-axis denotes the frequency of that position being a specific nucleotide, with the size of the nucleotide correlating to its frequency. Also depicted are the background frequencies of nucleotides, determined from 1,000 randomly selected sequences that are expressed in *C. elegans*. Nucleotides that are enriched compared to background frequencies (p-value < 0.01) are denoted with double arrowheads, while nucleotides that are depleted compared to background frequencies (p-value < 0.01) are denoted with single arrowheads.

5′ nucleotide is preferentially chosen as the miRNA (Khvorova et al. 2003; Schwarz et al. 2003; Tomari et al. 2004; Han et al. 2006; Ghildiyal et al. 2010). We examined if the miRNA 5′ nucleotide was less likely to be base-paired than the miRNA* 5′ nucleotide in the miRNA–miRNA* duplex. Of the 106 validated miRNAs, only 24 miRNAs and 24 miRNA*s had an unpaired 5′ nucleotide (Supplemental Fig. 1). Furthermore, we also calculated the thermodynamic stability of the 5′ nucleotide for all miRNA and miRNA* sequences (within the context of the miRNA–miRNA* duplex). We observed that the miRNA 5′ nucleotide had a stability of −0.7 kcal/mole, while the miRNA* 5′ nucleotide had a stability of −0.9 kcal/mole. Therefore, the 5′ nucleotides of both the miRNA and miRNA* sequences are equally likely to be base-paired and have very similar thermodynamic stabilities.

For example, *miR-70* has a miRNA* with a 5′ nucleotide that is unpaired, while the miRNA 5′ nucleotide is base-paired (Supplemental Fig. 1). Sequencing reads for this miRNA are 1,981 times more frequent than for its miRNA* (168,455 miRNA reads compared to 85 miRNA* reads) (Supplemental Table 1). Furthermore, *miR-36* has an unpaired nucleotide at the 5′ end of both the miRNA and miRNA* (Supplemental Fig. 1), yet the reads for the miRNA are 322 times more frequent than for the miRNA* (26,476 miRNA reads compared to 82 miRNA* reads) (Supplemental Table 1).

These data indicate that miRNAs are not differentiated from miRNA*s by the base-pairing status of their 5′ nucleotide in *C. elegans*, as might be the case in other species. Instead, other factors (such as having a 5′ uracil) may differentiate the miRNA from the miRNA* in worms.

Thermodynamic profile of pri-miRNAs and pre-miRNAs

To further investigate the characteristics of miRNA precursors in *C. elegans*, we calculated thermodynamic profiles of miRNA precursors for all validated miRNAs. We first calculated the average thermodynamic profile for pri-miRNAs for a 40-nt window around the Drosha cleavage site on the 5′ arm of the pri-miRNA. Additionally, we calculated the frequency of base-pairing among all pri-miRNAs for each position within the same window. Consistent with previous reports (Krol et al. 2004; Han et al. 2006; de Wit et al. 2009), we observed decreased base-pairing and an increase in the ΔG at the nucleotides adjacent to Drosha's cleavage site on the 5′ arm of the pri-miRNA (Fig. 4A; positions −1 and 1). However, the nucleotides adjacent to Drosha's cleavage site on the 3′ arm of the pri-miRNA (positions −2 and −3) were strongly base-paired, with stable ΔG values. This suggests that Drosha may determine its cleavage site on the 5′ arm of the pri-miRNA (but not on the 3′ arm) due to thermodynamic properties of the pri-miRNA. Other studies have observed that specific positions in pri-miRNAs have reduced stability (Krol et al. 2004; Han et al. 2006; de Wit et al. 2009). The only such study in *C. elegans* also reported that the two positions adjacent to Drosha's cleavage site on the 5′ arm of the pri-miRNA were less stable (higher ΔG values) compared to other positions in the pri-miRNA (de Wit et al. 2009).

We also observed a periodic decrease in stability within the pri-miRNA every ~10–12 nt (Fig. 4A). At four distinct positions across the pri-miRNA, there was a decrease in the frequency of base-pairing and a corresponding increase in the ΔG (position −23; positions −11 through −9; positions −1 through 1; positions 10 through 11). To determine if this periodicity was enriched in pri-miRNAs in a statistically significant manner, we analyzed 100 randomly selected stem-loops that are expressed in *C. elegans* for a similar pattern. We determined which structures had at least three patches of instability (1–3 consecutive nt that were unpaired), with each patch separated by 9–12 nt of stability (>70% base-pairing). While 77% of the pri-miRNAs had this type of periodicity (82 of 106), only 31 of the 100 random stem–loops had this characteristic. A χ^2 test indicated that this periodicity was therefore enriched in a highly statistically significant manner in the pri-miRNA data set (p-value $< 1.0 \times 10^{-10}$).

The decrease in stability at positions 10 and 11 is known to be important for efficient loading of the miRNA–miRNA* duplex into the miRNA-specific Argonaute in *C. elegans* (Steiner et al. 2007), but it is unclear if the other positions with decreased stability are also important for miRNA biogenesis or function. It is plausible that these positions are important for recognition of the pri-miRNA by Drosha. As a single turn of an A-form RNA helix is ~11 base-pairs, perhaps a single extended face of the pri-miRNA structure has decreased stability, with the rest of the pri-miRNA structure being more stable. Such a distinct structural feature may allow Drosha to discern between stem–loops that contain miRNA sequences and stem–loops that do not.

Next, we calculated the average thermodynamic profile and the frequency of base-pairing for all pre-miRNAs, for a 32-nt window around the Dicer cleavage site on the 5′ arm of the pre-miRNA (Fig. 4B). We observed that there was a strong decrease in stability and in frequency of base-pairing directly 3′ of the Dicer cleavage site on the 5′ arm of the pre-miRNA, while the cleavage site on the 3′ arm was strongly base-paired with stable ΔG values. Thus, like Drosha, Dicer may determine its cleavage site on the 5′ arm of the pre-miRNA based on the structure of the RNA substrate.

Structural characteristics surrounding Drosha's cleavage sites

The thermodynamic profile showed a reduced stability at Drosha's cleavage site on the 5′, but not the 3′, arm of the

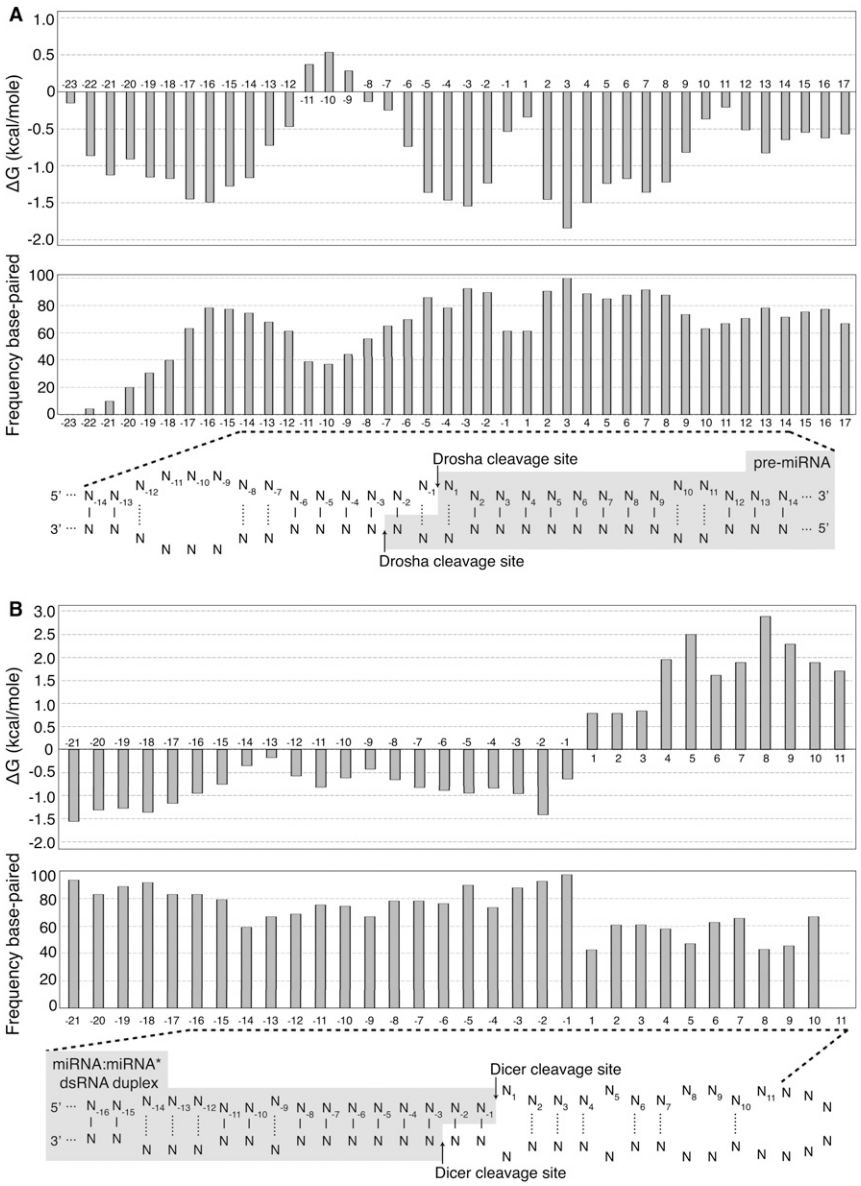


FIGURE 4. Thermodynamic profiles of pri-miRNAs and pre-miRNAs. (*A*) The bar graph shows the average thermodynamic profile of pri-miRNAs for all validated miRNAs, for a 40-nt window around the Drosha cleavage sites on the 5′ arm of the pri-miRNA. Numbering is relative to the Drosha cleavage site, with positive values going into the pre-miRNA sequence (5′ to 3′) and negative values going into the pri-miRNA sequence (3′ to 5′). The frequency of base-pairing within the pri-miRNA population, for each position within the same 40-nt window, was also determined and is shown in a bar graph. A model pri-miRNA depicting the Drosha cleavage site is shown. Nucleotides that were frequently base-paired (>70%) are connected with a solid line, moderately base-paired (50%–70%) with a dotted line, and those less frequently base-paired (<50%) are depicted without lines. The pre-miRNA sequence is within the gray box. (*B*) The bar graph shows the average thermodynamic profile of all validated pre-miRNA, for a 32-nt window surrounding the Dicer cleavage site on the 5′ arm of the pre-miRNA. Numbering is relative to the Dicer cleavage site, as in *A*. A model pre-miRNA is shown, as in *A*. Dicer cleavage sites are noted, and the miRNA–miRNA* duplex is shown within the gray box.

pri-miRNA. Consistent with this, among the 106 validated miRNAs, we found a significant enrichment for structural distortions surrounding the Drosha cleavage site on the 5′ arm of the pri-miRNA. Of the 106 miRNAs, 63 (~60%) had distortions on either side of the cleavage site on the 5′ arm (Fig. 5A–C). There were equal levels of distortions on either side of the cleavage site, with 31 distortions on the 3′ side, 28 on the 5′ side, and four surrounding the cleavage site (Fig. 5A–C). The majority of distortions were symmetrical internal loops (56 of 63), with 2-nt loops (single nucleotide mismatches) predominating (50 of 63). To determine statistical significance, we analyzed the random stem-loop data set for a 2–4-nt symmetrical internal loop 26–28 nt from the terminal loop, which is the most common structural distortion surrounding this Drosha cleavage site (Fig. 5A–C). We confirmed that this type of distortion was enriched in the pri-miRNA data set in a significant manner (p-value $< 1.0 \times 10^{-10}$). We also determined that 2–4-nt symmetrical internal loops were enriched in a highly significant manner specifically at the Drosha cleavage site compared to any other position within primary miRNA structures (p-value $< 1.0 \times 10^{-10}$). Furthermore, we found that asymmetrical internal loops >2 nucleotides were not statistically enriched at the Drosha cleavage site compared to any other location within the primary miRNA structure (p-value = 0.4). This indicates that only 2–4-nt symmetrical internal loops are specifically enriched surrounding the Drosha cleavage site on primary miRNAs.

Many well-studied miRNAs have distortions at the Drosha cleavage site on the 5′ arm of the pri-miRNA, such as *lin-4* (Supplemental Fig. 1). Recently, a study in *C. elegans* identified two miRNA families (*miRs-35-41* and *miRs-51-56*) where expression of at least one miRNA from each family is essential for viability (Alvarez-Saavedra and Horvitz 2010). Of the 13 essential miRNAs in these two families, 10 miRNAs had 2-nt or 4-nt symmetrical internal loops in their pri-miRNA structures at this Drosha cleavage site (Supplemental Fig. 1).

On the 3′ arm of the pri-miRNA, there were fewer distortions at the Drosha cleavage site, with only 25 pri-miRNAs (24%) having a distortion on either side of the cleavage site (Fig. 5D–F). On the 5′ side of the cleavage site (N^3), there were 15 distortions, nearly all in symmetrical internal loops (Fig.

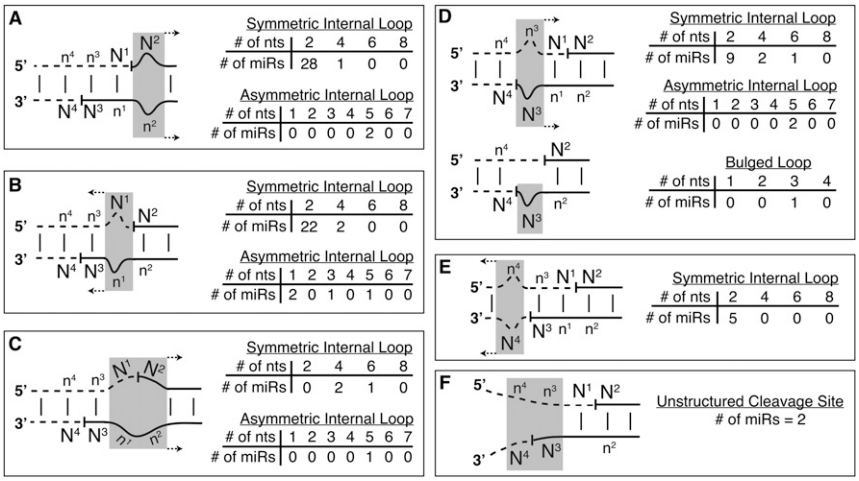


FIGURE 5. Structural distortions surrounding Drosha's cleavage sites. The capital N^x denotes a nucleotide directly adjacent to a cleavage site. The x value counts 5′ to 3′ along the pri-miRNA. The lower case n^x denotes the opposing nucleotide that base-pairs with N^x. For each position, only the types of structural distortions observed are listed, according to the number of nt within the structural distortion of the pri-miRNA. For internal loops that contain >2 nt, dotted arrows indicate the direction to which the distortion extends. (*A–C*) Structural distortions on the 5′ arm of the pri-miRNA, which are 3′, 5′, or encompass the Drosha cleavage site. These distortions were by far the most common. (*D–E*) Structural distortions on the 3′ arm of the pri-miRNA, 5′ and 3′ of the Drosha cleavage site, were not as common. Not depicted are three instances of distortions: a 1-nt bulged loop between n^3 and n^4, a 1-nt bulged loop between N^3 and N^4, and a 3-nt bulged loop encompassing the cleavage site (including N^3 and N^4). (*F*) Two pri-miRNAs have a 3′ arm with an unstructured Drosha cleavage site.

5D). On the 3′ side of the cleavage site, there were only eight distortions (Fig. 5E; note that three distortions are not depicted). Two pri-miRNAs had no discernable base-pairs for up to a dozen nucleotides past the cleavage site (Fig. 5F). None of these distortions on the 3′ arm were significantly enriched compared to the random stem–loop data set (*p*-values all > 0.1). Our data suggest that structural distortions surrounding Drosha's cleavage site on the 5′ arm—but not those on the 3′ arm—are important for pri-miRNA processing.

Structural characteristics surrounding Dicer's cleavage sites

Correlating with the thermodynamic profile, of the 106 validated miRNAs, 59 pre-miRNAs (56%) had structural distortions on the 5′ arm of the pre-miRNA, 3′ of the Dicer cleavage site (Fig. 6A; position N^2). This was a fourfold increase over any other position adjacent to a Dicer cleavage site. We observed a variety of distortions at this position, unlike the single type of distortion that surrounded the Drosha cleavage site on the 5′ arm of the pri-miRNA. The largest category was a single-nucleotide bulge on the opposing strand, directly opposite Dicer's cleavage site on the 5′ arm (23 of 59), as with *lin-4* (Supplemental Fig. 1). Internal loops were common, with 15 symmetric and 13 asymmetric internal loops. The symmetric internal loops were mostly 2 nt (as with *miR-58*, the most abundant miRNA in *C. elegans*), while the asymmetric internal loops varied from 3–9 nt (as with *miR-228*, the second most abundant miRNA which has a 5-nt asymmetric internal loop at this position). Finally, in five pre-miRNAs, the terminal loop began directly 3′ of the cleavage site.

To evaluate the significance of distortions at the Dicer cleavage site on the 5′ arm of the pre-miRNA (Fig. 6A), the random stem–loop data set was first evaluated for a 1–6-nt loop (either symmetric or asymmetric) within 5 bps of the terminal loop. This broad category of distortions was analyzed to account for the variety of distortions observed at this position. Surprisingly, such distortions were not statistically enriched in the pre-miRNA data set (*p*-value = 0.9). However, the most common structural distortion,

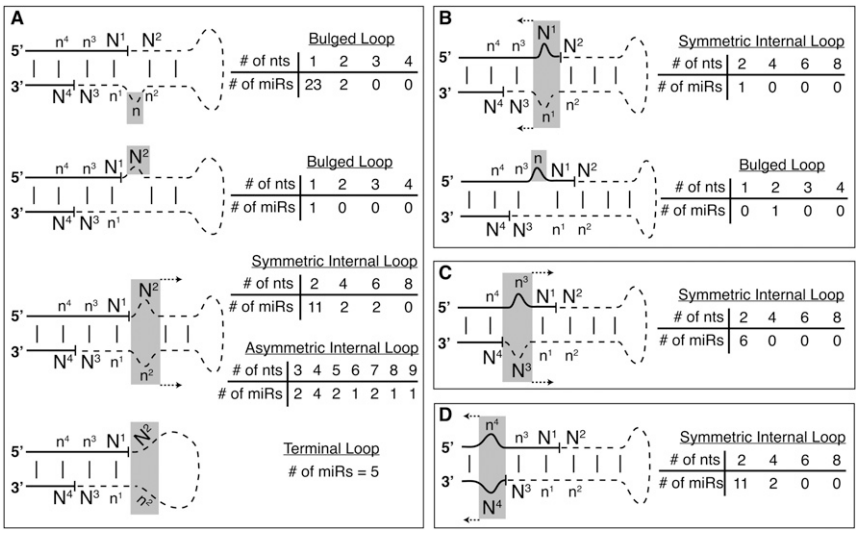


FIGURE 6. Structural distortions surrounding Dicer's cleavage sites. Naming of nucleotides (N^x or n^x) is the same as for Fig. 5. (*A,B*) Structural distortions on the 5′ arm of the pre-miRNA, 3′ or 5′ of the Dicer cleavage site. Distortions 3′ of the cleavage site were very common, while distortions 5′ of the cleavage site were not. In all analyses, the bulged nucleotide between n^1 and n^2 was considered to be 3′ of the cleavage site. (*C,D*) Structural distortions on the 3′ arm of the pre-miRNA, 5′ or 3′ of the Dicer cleavage site. Not depicted is a single instance of a 2-nt bulge between N^3 and N^4.

a single-nucleotide bulged loop within 5 bps of the terminal loop (Fig. 6A), was statistically enriched in the pre-miRNA data set (p-value = 1.3×10^{-5}), although only 23 pre-miRNAs had this specific distortion.

On the 5′ side of the Dicer cleavage site on the 5′ arm, there were only two structural distortions (Fig. 6B), suggesting that Dicer prefers a base-pair at this position. As this cleavage site becomes an unstructured overhang after Dicer cleavage, this preference cannot derive from a later step in miRNA biogenesis. However, this preference could arise from an earlier step in miRNA biogenesis.

On the 3′ arm of the pre-miRNA, there were few structural distortions adjacent to Dicer's cleavage site, with only 20 distortions (19%), including those on either side (Fig. 6C,D). On the 5′ side of the cleavage site, there were seven distortions, all but one of which were 2-nt internal symmetric loops (Fig. 6C; note that a 2-nt bulge between N^3 and n^1 is not depicted). On the 3′ side of the cleavage site, there were 13 pre-miRNAs with structural distortions. All were symmetric internal loops, 11 of which were 2-nt loops (Fig. 6D). We observed that 2-nt symmetrical internal loops, within 8 bps of the terminal loop, were not significantly enriched in the pre-miRNA data set (p-value = 0.9), indicating that Dicer prefers these nucleotides to be base-paired.

In summary, there are many types of structural distortions 3′ of the Dicer cleavage site on the 5′ arm of the pre-miRNA. However, except for the single-nucleotide bulge on the opposing strand, these structural distortions are just as likely to be found in a random-set of stem–loops.

DISCUSSION

Capitalizing on recent advances in sequencing technology that increase the number of reads that can be obtained for a small RNA population, we provide strong validation for 106 annotated miRNAs in *C. elegans*, while raising questions about the validity of 55 other annotated miRNAs. Our studies allowed annotation of 88 miRNA* sequences and 87 loop CP sequences not previously listed on miRBase, enabling the assembly of a pre-miRNA sequence and the prediction of a pre-miRNA structure for all validated pre-miRNAs. This is a key finding of this study, since most pre-miRNA sequences and structures were previously unknown.

Using our new data set, we also provide a comprehensive analysis of the features of *C. elegans* miRNAs and their precursors. Among other findings, we observed that the 5′ nucleotide of the miRNA is just as likely to be unpaired as the 5′ nucleotide of the miRNA*, suggesting that *C. elegans* miRNAs are not distinguished from miRNA*s by the stability of base-pairing at the 5′ nucleotide. Furthermore, we observed specific positions adjacent to both Drosha and Dicer cleavage sites had decreased stability and a corresponding enrichment in structural distortions. This suggests that the structure of miRNA precursors is important for efficient processing.

Drosha recognition and processing of pri-miRNAs

Drosha initiates miRNA processing and is thus responsible for identifying RNA stem–loops that contain miRNA sequences among the many that do not. This is a challenging task, and we propose that the periodic decrease in stability every 10–12 nt of a pri-miRNA creates a face of the helix that is distinct and easily recognized by Drosha (Fig. 4A). Once Drosha identifies a pri-miRNA, it must then determine its cleavage sites. We hypothesize that Drosha first determines its cleavage site on the 5′ arm of the pri-miRNA due to the significant enrichment of 2–4 nt symmetrical internal loops adjacent to or surrounding the cleavage site (Fig. 5A,B). Drosha would determine the cleavage site on the 3′ arm of the pri-miRNA by measuring 2 nt from the cleavage site on the 5′ arm.

Two other models have been proposed for how Drosha might determine its cleavage sites on pri-miRNAs. One study reported that Drosha measures ~11 bps from a ssRNA–dsRNA junction flanking the miRNA stem–loop (Han et al. 2006), while another study found that Drosha cleaves in reference to the terminal loop of the pri-miRNA (Zeng et al. 2005). These models hypothesize that Drosha measures from a structural feature within the pri-miRNA to determine its cleavage sites, while our studies raise the possibility that structural distortions at the cleavage site serve as a marker, without a need for measuring. Alternatively, Drosha might measure as well as use multiple features within the pri-miRNA to determine its cleavage sites.

Dicer processing of pre-miRNAs

After Drosha processing, pre-miRNAs are exported from the nucleus to the cytoplasm by Exportin-5, allowing for Dicer processing. It is unknown if Exportin-5 directly passes pre-miRNAs to Dicer, or if it releases them into the cytoplasm, at which point Dicer must identify them. The latter possibility seems less parsimonious, as both Drosha and Dicer would be required to identify miRNA precursors. Furthermore, our data show that most pri-miRNAs, but few pre-miRNAs, are significantly enriched with specific types of structural distortions (i.e., 77% of pri-miRNAs have periodic decreases in stability, while only 22% of pre-miRNAs have a 1-nt bulge directly 3′ of the Dicer cleavage site on the 5′ arm of the pre-miRNA). Therefore, structural distortions could enable Drosha to identify nearly all pri-miRNAs but would not help Dicer identify the majority of pre-miRNAs. This suggests that Dicer does not have to identify pre-miRNAs, but rather, Exportin-5 passes pre-miRNAs to Dicer, directly or through another factor.

Once Dicer binds a pre-miRNA, it must then determine its cleavage sites. The structure of Dicer's PAZ and RNase

III domains suggests a model whereby Dicer acts as a molecular ruler and measures a set distance from the termini produced by Drosha to determine its cleavage sites (MacRae et al. 2006). We observed 22- and 23-nt Dicer products with equal frequency (Fig. 1A,C), indicating that if measuring is involved, it is imprecise. We propose that cleavage site choice is also influenced by a preference for positioning pre-miRNAs into Dicer's RNase III domains such that a structural distortion will be directly 3′ of the cleavage site on the 5′ arm of the pre-miRNA. Dicer would then determine its cleavage site on the 3′ arm of the pre-miRNA by measuring two nucleotides from the cleavage site on the 5′ arm, as with the model we propose for Drosha processing of pri-miRNAs. Consistent with our model, at least for some pre-miRNAs, in vitro processing by Dicer is more efficient when there is a distortion directly 3′ of the cleavage site on the 5′ arm of the pre-miRNA (Zhang and Zeng 2010).

The RNase IIIb domains of Drosha and Dicer prefer to cleave adjacent to structural distortions

A statistically significant enrichment was observed for symmetrical internal loops adjacent to or surrounding Drosha's cleavage site on the 5′ arm of the pri-miRNA (Fig. 5A–C), as well as for a single-bulged nucleotide 3′ of Dicer's cleavage site on the 5′ arm of the pre-miRNA (Fig. 6A). According to current models (Zhang et al. 2004; MacRae and Doudna 2007), the RNase IIIb domain of Drosha and Dicer mediate cleavage at these two sites. Thus, structural disruptions may aid the RNase IIIb domain in determining a cleavage site or enacting cleavage. Conversely, the two sites that are cleaved by the RNase IIIa domains of Drosha and Dicer are most often base-paired. Therefore, the two nuclease domains may have different structural requirements for binding and/or processing miRNA precursors.

Argonaute loading of miRNAs

After processing by Dicer, the miRNA–miRNA* duplex is loaded into the miRNA-specific Argonaute ALG-1. In *C. elegans*, miRNA and siRNA pathways converge at Dicer, but the products, miRNA–miRNA* duplexes and siRNA duplexes, are subsequently passed to distinct pathways. Previous studies hypothesized that miRNA–miRNA* duplexes are selected by ALG-1 due to their imperfect structures (Forstemann et al. 2007; Steiner et al. 2007; Tomari et al. 2007; Kawamata et al. 2009; Ghildiyal et al. 2010), while perfectly paired duplexes are loaded into the siRNA-specific Argonaute RDE-1 (Steiner et al. 2007). Our data are consistent with this model, since every miRNA–miRNA* duplex had at least one structural distortion (Supplemental Fig. 1).

After the miRNA–miRNA* duplex is loaded into RISC by binding ALG-1, the protein must then differentiate between the miRNA and the miRNA*. Based on analyses of thermodynamic plots, studies of fly and human miRNAs concluded that the 5′ terminus of the miRNA is less frequently base-paired than the 5′ terminus of the miRNA* (Krol et al. 2004; Tomari et al. 2004; Han et al. 2006; Kawamata et al. 2009; Ghildiyal et al. 2010). In contrast, we observed that, in *C. elegans*, the 5′ nucleotide of the miRNA and the miRNA* were unpaired at equal frequencies and had similar thermodynamic stabilities. Instead, the only major differences we observed between the miRNA and miRNA* sequences were that only miRNAs were enriched for a uracil at the first and ninth positions (Fig. 2F; Supplemental Fig. 2). As Argonautes in other species have preferences for certain 5′ nucleotides (Mi et al. 2008; Ghildiyal et al. 2010), possibly ALG-1 differentiates between the miRNA and miRNA* by the identity of the 5′ nucleotide.

Confidence in the miRNA validation

Of the 106 miRNAs we validated with our sequencing data, 101 were observed to immunoprecipitate (IP) with the miRNA-specific Argonaute ALG-1 in a recent study in *C. elegans* (Supplemental Table 1; Zisoulis et al. 2010). Thus, 95% of our validated miRNAs directly interact with ALG-1, indicating that there were few false positives in our analysis. We observed few reads for the five miRNAs that we validated that did not IP with ALG-1 (Supplemental Table 1); these miRNAs may not have been observed by Zisoulis and colleagues due to lower coverage, since they obtained 3.8 million reads compared to our 19 million. Furthermore, the IP was conducted on L4 worms, and three of the five miRNAs are expressed at very low levels in L4 worms but are expressed at higher levels in other stages of development (Kato et al. 2009). As we sequenced mixed-stage worms, these miRNAs were more likely to be in our data set.

Nine of the 35 miRNAs for which we obtained reads for the miRNA but not for the miRNA* ("Not Validated" and "Partially Validated" categories in Tables 1, 2) were observed to IP with ALG-1 (Zisoulis et al. 2010). Thus, we may have a small number of false negative miRNAs in our data set, possibly because a miRNA* sequence exists for these miRNA loci, but we were unable to sequence it.

MATERIALS AND METHODS

Preparation of small RNA library and alignment of sequences to genome

Total RNA was extracted from mixed-stage wild type (N2) *C. elegans* using TRIzol reagent (Invitrogen). Ten μg of total RNA was run on a denaturing 15% polyacrylamide (19:1) 1× TBE gel, and small RNAs (~18–30 nt) were isolated. A cDNA library was made using the small RNA v1.5 Illumina kit, which captures RNAs containing only a 5′ monophosphate, and maintains directionality of the sequence. Sequencing was performed using an

Illumina Genome Analyzer IIx running SCS 2.6. The pipeline software was version 1.6 (OLB 1.6, CASAVA 1.6.0).

Novoalign (www.novocraft.com) was used to trim the 3′ adaptor and align sequences to the May 2008 *C. elegans* genome release, using an *r*-value of 0.2 and a *q*-value of 5. The USeq software package component NovalignParser (Nix et al. 2008) was used to parse out poorly aligned sequences, with a posterior probability of 0.1. Raw data and aligned data can be accessed at the Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo/), accession number GSE24704. Genomic coordinates for miRNAs were obtained from release 15 (April 2010) of miRBase (Griffiths-Jones et al. 2008). Genomic coordinates for piRNA loci (Ruby et al. 2006; Batista et al. 2008) and endo-siRNA loci (Asikainen et al. 2008) were compiled from previous studies. Of the ~19,300,000 sequencing reads that aligned to the genome, ~15,000,000 reads aligned to known miRNA loci, ~830,000 reads aligned to known piRNA loci, ~2,000,000 reads aligned to known endogenous siRNA loci, and ~1,500,000 reads did not align to known loci that produce small RNAs. This distribution is similar to previous studies (Ruby et al. 2006; Han et al. 2009; Kato et al. 2009; Welker et al. 2010).

Analysis of validated miRNAs

Pre-miRNA sequences were constructed by assembling the predominant miRNA, miRNA*, and loop CP sequences. The structures of the pre-miRNAs were determined using mfold version 3.2 (Mathews et al. 1999; Zuker 2003) with default folding conditions. The lowest energy structure was used, with alterations made to only a few predicted structures (noted by an * in Supplemental Fig. 1). In short, we observed that 59 of 106 pre-miRNAs had structural distortions directly 3′ of the Dicer cleavage site on the 5′ arm (Fig. 6A). However, five predicted pre-miRNA structures had a 1–2 nt bulge directly 5′ of this cleavage site. These five pre-miRNA structures were altered to maintain the same number of base-pairs, shifting the bulge to be directly 3′ of the cleavage site, as observed in other pre-miRNAs with a distortion at this position.

Sequence logos were created using WebLogo 3 (Crooks et al. 2004), with the output giving "probability" on the Y-axis (which is more accurately called nucleotide frequency). To generate background nucleotide frequencies, we compiled 1,000 randomly selected RNA sequences expressed in *C. elegans*, within coding and noncoding regions of genes listed on WormBase (www.wormbase.org). To determine the statistical significance of nucleotide frequencies, a χ^2 test was performed using a Bonferroni correction, comparing the nucleotide frequency at the queried position to the background frequency of that nucleotide. Using a less conservative method (a χ^2 test without a Bonferroni correction), additional positions within miRNA and miRNA* sequences (Supplemental Fig. 2) and adjacent to Drosha or Dicer cleavage sites (Fig. 3; Supplemental Fig. 2) would also be statistically different from background. While future studies may demonstrate additional positions have meaningful differences, we chose a more conservative method to minimize false positive predictions.

When determining homogeneity of the various 5′ and 3′ nucleotides, each sequencing read was weighted equally (Fig. 2C,D). For all other analyses performed for miRNA loci, the predominant read for a miRNA, miRNA*, or loop CP were weighted equally (independent of the number of sequencing reads obtained for that locus). When determining the homogeneity of cleavage for an individual miRNA or miRNA* (Fig. 2E), a miRNA or miRNA* was excluded if it had fewer than 10 sequencing reads, since each read would represent >10% of the population, and one aberrant sequencing read could easily skew the statistical analysis. A similar method was used by a previous study (Seitz et al. 2008).

When determining the significance of structural distortions within miRNA precursors (Figs. 5, 6), we used a method similar to a previous study (Ritchie et al. 2007). To generate a random stem–loop data set, 100 stem–loop structures that are expressed in *C. elegans* were randomly selected. The structures were generated by folding (using mfold with default conditions) ~400 nt sequences of RNA that are expressed in *C. elegans*, that were randomly chosen from coding and noncoding regions of genes listed on WormBase. For each sequence, a stem–loop (~70 nt or greater in length) was randomly selected from the most thermodynamically stable structure prediction. To determine if an enrichment for a specific distortion in the pre-miRNA or pri-miRNA data set was significant, we performed a χ^2 test comparing the number of distortions in the miRNA data set and the random data set.

Thermodynamic analysis

The thermodynamic analysis of pri-miRNAs and pre-miRNAs was performed as previously reported (Han et al. 2006). In brief, pri-miRNAs and pre-miRNAs were folded by mfold, and the most thermodynamically stable structure chosen. For pri-miRNAs, the adjacent 30 nt were added to the 5′ and 3′ ends of the pri-miRNA sequence. We calculated the average ΔG value for each position, relative to either the Drosha or Dicer cleavage sites. We also calculated the frequency that each position was base-paired. To calculate the thermodynamic stability of a specific nucleotide, we accessed the "Thermodynamic Details" generated for each structure prediction by mfold. Using the nearest neighbor method, mfold calculates the thermodynamic stability for the base-stacking interaction between two adjacent nucleotides, taking into account hydrogen bonding with the nucleotides on the opposing strand (www.bioinfo.rpi.edu/zukerm/cgi-bin/efiles-3.0.cgi). Therefore, to calculate the stability of a specific nucleotide, the thermodynamic values from the stacking energies on both the 5′ and 3′ side of the nucleotide were averaged. When determining the stability of loops and bulges, mfold treats structural distortions as a single position. This value cannot be subdivided into the individual nucleotides within the distortion; thus we assigned the thermodynamic value to the 3′-most position of the distortion and left the other positions within the structural distortion blank. When averaging thermodynamic values to find the stability of a single position, the structural distortion was considered as a single position and, as with any other position, averaged with the positions 5′ and 3′ of the distortion.

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGMENTS

We thank David Nix and Brian Dalley at the University of Utah's Core facilities for Illumina sequencing and data processing. This work was supported by funds from the National Institute of General Medical Sciences to B.L.B. (GM044073; GM067106).

Received August 23, 2010; accepted January 5, 2011.

REFERENCES

Alvarez-Saavedra E, Horvitz HR. 2010. Many families of C. elegans microRNAs are not essential for development or viability. *Curr Biol* **20:** 367–373.

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

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

Asikainen S, Heikkinen L, Wong G, Storvik M. 2008. Functional characterization of endogenous siRNA target genes in Caenorhabditis elegans. *BMC Genomics* **9:** 270.

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

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

Bohnsack MT, Czaplinski K, Gorlich D. 2004. Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs. *RNA* **10:** 185–191.

Cai X, Hagedorn CH, Cullen BR. 2004. Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *RNA* **10:** 1957–1966.

Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome Res* **14:** 1188–1190.

Denli AM, Tops BB, Plasterk RH, Ketting RF, Hannon GJ. 2004. Processing of primary microRNAs by the Microprocessor complex. *Nature* **432:** 231–235.

de Wit E, Linsen SE, Cuppen E, Berezikov E. 2009. Repertoire and evolution of miRNA genes in four divergent nematode species. *Genome Res* **19:** 2064–2074.

Forstemann K, Horwich MD, Wee L, Tomari Y, Zamore PD. 2007. Drosophila microRNAs are sorted into functionally distinct argonaute complexes after production by dicer-1. *Cell* **130:** 287–297.

Friedlander MR, Chen W, Adamidi C, Maaskola J, Einspanier R, Knespel S, Rajewsky N. 2008. Discovering microRNAs from deep sequencing data using miRDeep. *Nat Biotechnol* **26:** 407–415.

Ghildiyal M, Zamore PD. 2009. Small silencing RNAs: an expanding universe. *Nat Rev Genet* **10:** 94–108.

Ghildiyal M, Xu J, Seitz H, Weng Z, Zamore PD. 2010. Sorting of Drosophila small silencing RNAs partitions microRNA* strands into the RNA interference pathway. *RNA* **16:** 43–56.

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

Gregory RI, Yan KP, Amuthan G, Chendrimada T, Doratotaj B, Cooch N, Shiekhattar R. 2004. The Microprocessor complex mediates the genesis of microRNAs. *Nature* **432:** 235–240.

Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ. 2008. miRBase: tools for microRNA genomics. *Nucleic Acids Res* **36:** D154–D158.

Grishok A, Pasquinelli AE, Conte D, Li N, Parrish S, Ha I, Baillie DL, Fire A, Ruvkun G, Mello CC. 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 KH, Kim YK, Jin H, Kim VN. 2004. The Drosha-DGCR8 complex in primary microRNA processing. *Genes Dev* **18:** 3016–3027.

Han J, Lee Y, Yeom KH, Nam JW, Heo I, Rhee JK, Sohn SY, Cho Y, Zhang BT, Kim VN. 2006. Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. *Cell* **125:** 887–901.

Han T, Manoharan AP, Harkins TT, Bouffard P, Fitzpatrick C, Chu DS, Thierry-Mieg D, Thierry-Mieg J, Kim JK. 2009. 26G endo-siRNAs regulate spermatogenic and zygotic gene expression in Caenorhabditis elegans. *Proc Natl Acad Sci* **106:** 18674–18679.

Hutvagner G, McLachlan J, Pasquinelli AE, Balint E, Tuschl T, Zamore PD. 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 RJ, Hobert O. 2003. A microRNA controlling left/right neuronal asymmetry in Caenorhabditis elegans. *Nature* **426:** 845–849.

Kato M, de Lencastre A, Pincus Z, Slack FJ. 2009. Dynamic expression of small non-coding RNAs, including novel microRNAs and piRNAs/21U-RNAs, during Caenorhabditis elegans development. *Genome Biol* **10:** R54.

Kawamata T, Seitz H, Tomari Y. 2009. Structural determinants of miRNAs for RISC loading and slicer-independent unwinding. *Nat Struct Mol Biol* **16:** 953–960.

Ketting RF, Fischer SE, Bernstein E, Sijen T, Hannon GJ, Plasterk RH. 2001. Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in C. elegans. *Genes Dev* **15:** 2654–2659.

Khvorova A, Reynolds A, Jayasena SD. 2003. Functional siRNAs and miRNAs exhibit strand bias. *Cell* **115:** 209–216.

Kim VN, Han J, Siomi MC. 2009. Biogenesis of small RNAs in animals. *Nat Rev Mol Cell Biol* **10:** 126–139.

Knight SW, Bass BL. 2001. A role for the RNase III enzyme DCR-1 in RNA interference and germ line development in Caenorhabditis elegans. *Science* **293:** 2269–2271.

Krol J, Sobczak K, Wilczynska U, Drath M, Jasinska A, Kaczynska D, Krzyzosiak WJ. 2004. Structural features of microRNA (miRNA) precursors and their relevance to miRNA biogenesis and small interfering RNA/short hairpin RNA design. *J Biol Chem* **279:** 42230–42239.

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

Landthaler M, Yalcin A, Tuschl T. 2004. The human DiGeorge syndrome critical region gene 8 and its D. melanogaster homolog are required for miRNA biogenesis. *Curr Biol* **14:** 2162–2167.

Lau NC, Lim LP, Weinstein EG, Bartel DP. 2001. An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science* **294:** 858–862.

Lee RC, Ambros V. 2001. An extensive class of small RNAs in Caenorhabditis elegans. *Science* **294:** 862–864.

Lee RC, Feinbaum RL, Ambros V. 1993. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* **75:** 843–854.

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

Lee Y, Kim M, Han J, Yeom KH, Lee S, Baek SH, Kim VN. 2004. MicroRNA genes are transcribed by RNA polymerase II. *EMBO J* **23:** 4051–4060.

Lim LP, Lau NC, Weinstein EG, Abdelhakim A, Yekta S, Rhoades MW, Burge CB, Bartel DP. 2003. The microRNAs of Caenorhabditis elegans. *Genes Dev* **17:** 991–1008.

Lund E, Guttinger S, Calado A, Dahlberg JE, Kutay U. 2004. Nuclear export of microRNA precursors. *Science* **303:** 95–98.

MacRae IJ, Doudna JA. 2007. Ribonuclease revisited: Structural insights into ribonuclease III family enzymes. *Curr Opin Struct Biol* **17:** 138–145.

MacRae IJ, Zhou K, Li F, Repic A, Brooks AN, Cande WZ, Adams PD, Doudna JA. 2006. Structural basis for double-stranded RNA processing by Dicer. *Science* **311:** 195–198.

Mathews DH, Sabina J, Zuker M, Turner DH. 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol* **288:** 911–940.

Mi S, Cai T, Hu Y, Chen Y, Hodges E, Ni F, Wu L, Li S, Zhou H, Long C, et al. 2008. Sorting of small RNAs into Arabidopsis argonaute

complexes is directed by the 5′ terminal nucleotide. *Cell* **133:** 116–127.

Nix DA, Courdy SJ, Boucher KM. 2008. Empirical methods for controlling false positives and estimating confidence in ChIP-Seq peaks. *BMC Bioinformatics* **9:** 523.

Ohler U, Yekta S, Lim LP, Bartel DP, Burge CB. 2004. Patterns of flanking sequence conservation and a characteristic upstream motif for microRNA gene identification. *RNA* **10:** 1309–1322.

Okamura K, Phillips MD, Tyler DM, Duan H, Chou YT, Lai EC. 2008. The regulatory activity of microRNA* species has substantial influence on microRNA and 3′ UTR evolution. *Nat Struct Mol Biol* **15:** 354–363.

Ritchie W, Legendre M, Gautheret D. 2007. RNA stem-loops: To be or not to be cleaved by RNAse III. *RNA* **13:** 457–462.

Ruby JG, Jan C, Player C, Axtell MJ, Lee W, Nusbaum C, Ge H, Bartel DP. 2006. Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C. elegans. *Cell* **127:** 1193–1207.

Schwarz DS, Hutvagner G, Du T, Xu Z, Aronin N, Zamore PD. 2003. Asymmetry in the assembly of the RNAi enzyme complex. *Cell* **115:** 199–208.

Seitz H, Ghildiyal M, Zamore PD. 2008. Argonaute loading improves the 5′ precision of both MicroRNAs and their miRNA* strands in flies. *Curr Biol* **18:** 147–151.

Steiner FA, Hoogstrate SW, Okihara KL, Thijssen KL, Ketting RF, Plasterk RH, Sijen T. 2007. Structural features of small RNA precursors determine Argonaute loading in Caenorhabditis elegans. *Nat Struct Mol Biol* **14:** 927–933.

Tomari Y, Matranga C, Haley B, Martinez N, Zamore PD. 2004. A protein sensor for siRNA asymmetry. *Science* **306:** 1377–1380.

Tomari Y, Du T, Zamore PD. 2007. Sorting of Drosophila small silencing RNAs. *Cell* **130:** 299–308.

Welker NC, Pavelec DM, Nix DA, Duchaine TF, Kennedy S, Bass BL. 2010. Dicer's helicase domain is required for accumulation of some, but not all, C. elegans endogenous siRNAs. *RNA* **16:** 893–903.

Wightman B, Ha I, 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.

Wu H, Neilson JR, Kumar P, Manocha M, Shankar P, Sharp PA, Manjunath N. 2007. miRNA profiling of naive, effector and memory CD8 T cells. *PLoS ONE* **2:** e1020.

Yi R, Qin Y, Macara IG, Cullen BR. 2003. Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. *Genes Dev* **17:** 3011–3016.

Zeng Y, Yi R, Cullen BR. 2005. Recognition and cleavage of primary microRNA precursors by the nuclear processing enzyme Drosha. *EMBO J* **24:** 138–148.

Zhang X, Zeng Y. 2010. The terminal loop region controls microRNA processing by Drosha and Dicer. *Nucleic Acids Res* **38:** 7689–7697.

Zhang H, Kolb FA, Jaskiewicz L, Westhof E, Filipowicz W. 2004. Single processing center models for human Dicer and bacterial RNase III. *Cell* **118:** 57–68.

Zhang B, Stellwag EJ, Pan X. 2009. Large-scale genome analysis reveals unique features of microRNAs. *Gene* **443:** 100–109.

Zisoulis DG, Lovci MT, Wilbert ML, Hutt KR, Liang TY, Pasquinelli AE, Yeo GW. 2010. Comprehensive discovery of endogenous Argonaute binding sites in Caenorhabditis elegans. *Nat Struct Mol Biol* **17:** 173–179.

Zuker M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **31:** 3406–3415.