Cell

# Beyond Secondary Structure: Primary-Sequence Determinants License Pri-miRNA Hairpins for Processing

Vincent C. Auyeung,[1,2,3,4] Igor Ulitsky,[1,2,3] Sean E. McGeary,[1,2,3] and David P. Bartel[1,2,3,*]
[1]Whitehead Institute for Biomedical Research, Cambridge, MA 02142, USA
[2]Howard Hughes Medical Institute, Chevy Chase, MD 20815, USA
[3]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[4]Harvard-MIT Division of Health Sciences and Technology, Cambridge, MA 02139, USA
*Correspondence: dbartel@wi.mit.edu
http://dx.doi.org/10.1016/j.cell.2013.01.031

## SUMMARY

To use microRNAs to downregulate mRNA targets, cells must first process these ~22 nt RNAs from primary transcripts (pri-miRNAs). These transcripts form RNA hairpins important for processing, but additional determinants must distinguish pri-miRNAs from the many other hairpin-containing transcripts expressed in each cell. Illustrating the complexity of this recognition, we show that most *Caenorhabditis elegans* pri-miRNAs lack determinants required for processing in human cells. To find these determinants, we generated many variants of four human pri-miRNAs, sequenced millions that retained function, and compared them with the starting variants. Our results confirmed the importance of pairing in the stem and revealed three primary-sequence determinants, including an SRp20-binding motif (CNNC) found downstream of most pri-miRNA hairpins in bilaterian animals, but not in nematodes. Adding this and other determinants to *C. elegans* pri-miRNAs imparted efficient processing in human cells, thereby confirming the importance of primary-sequence determinants for distinguishing pri-miRNAs from other hairpin-containing transcripts.

## INTRODUCTION

MicroRNAs (miRNAs) are ~22 nt RNAs that pair to messenger RNAs (mRNAs) to direct posttranscriptional repression (Bartel, 2004). MicroRNAs are processed from hairpin-containing primary transcripts (pri-miRNAs). In the canonical processing pathway of animals, pri-miRNAs are cleaved by the Microprocessor, a protein complex containing an RNase III enzyme, Drosha, and its cofactor, DGCR8/Pasha (Lee et al., 2003; Denli et al., 2004; Gregory et al., 2004; Han et al., 2004; Landthaler et al., 2004). The liberated portion of the hairpin (the pre-miRNA) is then cleaved by the RNase III enzyme Dicer (Grishok et al., 2001; Hutvágner et al., 2001), leaving two ~22 nt strands that pair to each other with ~2 nt 3′ overhangs (Lee et al., 2003; Lim et al., 2003b). One strand of each duplex is loaded into an Argonaute protein to form the core of the silencing complex, and the other strand is discarded (Khvorova et al., 2003; Schwarz et al., 2003; Liu et al., 2004). Noncanonical pathways also contribute to the miRNA repertoire through the processing of mirtrons (Okamura et al., 2007; Ruby et al., 2007) or other pri-miRNAs that bypass Drosha cleavage (Babiarz et al., 2008) and through one pre-miRNA that bypasses Dicer cleavage (Cheloufi et al., 2010; Cifuentes et al., 2010).

A long-standing mystery has been how pri-miRNAs are distinguished from the many other hairpin-containing transcripts for processing as Microprocessor substrates. Determinants of Dicer cleavage are better understood (Zhang et al., 2004; Macrae et al., 2006; Park et al., 2011), as illustrated by both the design (Brummelkamp et al., 2002; Paddison et al., 2002) and prediction (Chung et al., 2011) of Dicer substrates that bypass Drosha processing. For Microprocessor recognition, sequences within 40 nt upstream and 40 nt downstream of the pre-miRNA hairpins are required for ectopic miRNA expression (Chen et al., 2004), which is consistent with (1) the observation that these flanking sequences tend to pair to each other to extend the stem another turn of the helix beyond the cleavage site (Lim et al., 2003b) and (2) a requirement for both this extension and a lack of pairing immediately following it for processing (Han et al., 2006). However, many cellular transcripts have paired regions flanked by single-stranded RNA (ssRNA), and most of these are not Microprocessor substrates. Indeed, attempts to predict canonical miRNA hairpins from genomic sequence yield many thousands of false-positive predictions, which must be eliminated using additional criteria, such as analysis of conservation or experimental evaluation (Lim et al., 2003a, 2003b; Bentwich et al., 2005; Berezikov et al., 2006; Chiang et al., 2010). This illustrates a large gap in our understanding of how the Microprocessor distinguishes between authentic substrates and other transcribed hairpins.

Here, we report that transcripts that enter the miRNA pathway in *C. elegans* failed to do so in human cells. Thus, the definition of a pri-miRNA in one species differs from that in another.

To find features that define human pri-miRNAs, we generated more than $10^{11}$ variants of four pri-miRNAs and sequenced millions that were cleaved by the human Microprocessor. Comparison of cleaved and initial variants revealed important sequence and structural features. These features were evolutionarily conserved in non-nematode lineages and sufficient to increase the processing efficiency of *C. elegans* hairpins in human cells.

## RESULTS

### Unknown Features Specify Human Pri-miRNAs

To examine whether miRNA processing features are shared across animals, we ectopically expressed a panel of *C. elegans*, *D. melanogaster*, and human pri-miRNAs in human cells and compared the yields of mature miRNA. Despite variability in the degree of overexpression, presumably reflecting differences in efficiency at various steps of the pathway (Fellmann et al., 2011; Feng et al., 2011), most human miRNAs were efficiently expressed (Figure 1A), as expected (Chiang et al., 2010). Four of nine *Drosophila* miRNAs also fell within the range observed for human miRNAs. However, the tested *C. elegans* miRNAs were less efficiently expressed (Figure 1A, $p = 1.4 \times 10^{-5}$, Wilcoxon rank-sum test). Similar results were observed in *Drosophila* S2 cells ($p = 0.024$). Thus, most nematode pri-miRNAs lack determinants required for efficient processing in human or insect cells.

To isolate the processing defect, we probed for processing intermediates. Consistent with the sequencing results, *cel-lin-4* was processed, with detectable pre-miRNA and mature miRNA (Figure 1B). For other *C. elegans* miRNAs, neither pre-miRNA nor mature miRNA was detected, despite the presence of primary transcripts (Figure 1B; Figure S1B, available online), suggesting that these *C. elegans* pri-miRNAs were not productively recognized as Microprocessor substrates. To assay directly for Microprocessor binding, we examined binding to catalytically deficient Drosha and DGCR8. Whereas human *pri-mir-122* bound the Microprocessor somewhat better than did the reference pri-miRNA (human *pri-mir-125a*), all seven tested *C. elegans* pri-miRNAs bound worse (Figure 1C). Thus, most *C. elegans* pri-miRNAs are missing some of the determinants needed for efficient recognition and processing by the human Microprocessor.

Known features of *C. elegans* and human pri-miRNAs appear largely similar, as illustrated by the accuracy of an algorithm trained on *C. elegans* pri-miRNAs in predicting most miRNA genes conserved in mammals and fish (Lim et al., 2003a). Nonetheless, the poor specificity of this algorithm when predicting nonconserved miRNAs suggests that unknown features help define authentic pri-miRNAs. To look for clues regarding these unknown features, we analyzed the conservation of sequence immediately flanking human pre-miRNAs. Residues extending 13 nt upstream of the 5p Drosha cleavage site (i.e., the site corresponding to the 5′ end of the pre-miRNA) and 11 nt downstream of the 3p Drosha cleavage site were conserved above background, consistent with the importance of the ~11 bp basal stem for pri-miRNA processing (Figure 1D). However, the signal beyond the basal stem tailed off rapidly (particularly in the upstream flanking region), suggesting that any determinants in the flanking regions might be either at variable distances from the hairpin or present in only subsets of miRNAs, making them difficult to identify using alignments.

### Functional Substrates from Large Libraries of Pri-miRNA Variants

To identify features important for Microprocessor recognition and cleavage, we generated more than $10^{11}$ pri-miRNA variants, sequenced millions that retained function, and compared these sequences to those of the initial variants (Figure 2A). This approach resembled classical in vitro selection approaches (Wilson and Szostak, 1999), except we did not perform multiple rounds of selection. Because the starting and the selected pools underwent the same number of transcription, reverse-transcription, and amplification steps, any differences between the two pools were subject to neither the compounding effects of multiple rounds nor the confounding effects of amplification biases. Moreover, as with previous analyses of selection results using high-throughput sequencing (Zykovich et al., 2009; Pitt and Ferré-D'Amaré, 2010; Slattery et al., 2011), sequencing depth reduced the influence of stochastic sampling. Thus, compared to the results of classical approaches, enrichment or depletion of a residue was a more direct reflection of its contribution to biochemical specificity.

Four pools of variants were constructed, each based on a different human pri-miRNA (*mir-125a*, *mir-16-1*, *mir-30a*, and *mir-223*). Residues more than 8 nt upstream of the 5p Drosha cleavage site and more than 8 nt downstream of the 3p cleavage site were varied, whereas the remaining hairpin residues were not. At each variable position, 79% of the molecules had the wild-type residue, and the remainder had one of the other three alternatives. As done for self-cleaving ribozymes (Pan and Uhlenbeck, 1992), each variant was circularized so that all of its variable nucleotides resided in a single cleavage product (Figure 2A), thereby enabling a full analysis of sequence interdependencies.

In vitro cleavage reactions were conducted in Microprocessor lysate, i.e., whole-cell lysate from HEK293T cells overexpressing Drosha and DGCR8 to enhance cleavage activity (Figure 2B). At a time in which the lysate cleaved linear and circularized *pri-mir-125a* to near completion, many *pri-mir-125a* variants remained uncleaved (Figure 2C), which indicated that some substitutions in the basal stem and flanking regions attenuated Microprocessor cleavage in vitro.

Cleaved variants were purified and sequenced (Figure 2A). At each variant position, the odds of each nucleotide in the cleaved pool were compared to the odds of that nucleotide in the starting pool. These odds ratios were used to calculate the information content of each nucleotide possibility at each variant position—the greater the information content, the more favorable the influence on activity, with positive values indicating beneficial influences and negative values disruptive ones. An advantage of plotting information content is that it reports the relative influence of each nucleotide possibility irrespective of whether it was the wild-type possibility. Because molecular manipulations and computational filtering both selected for cleavage at the wild-type site, nucleotide changes

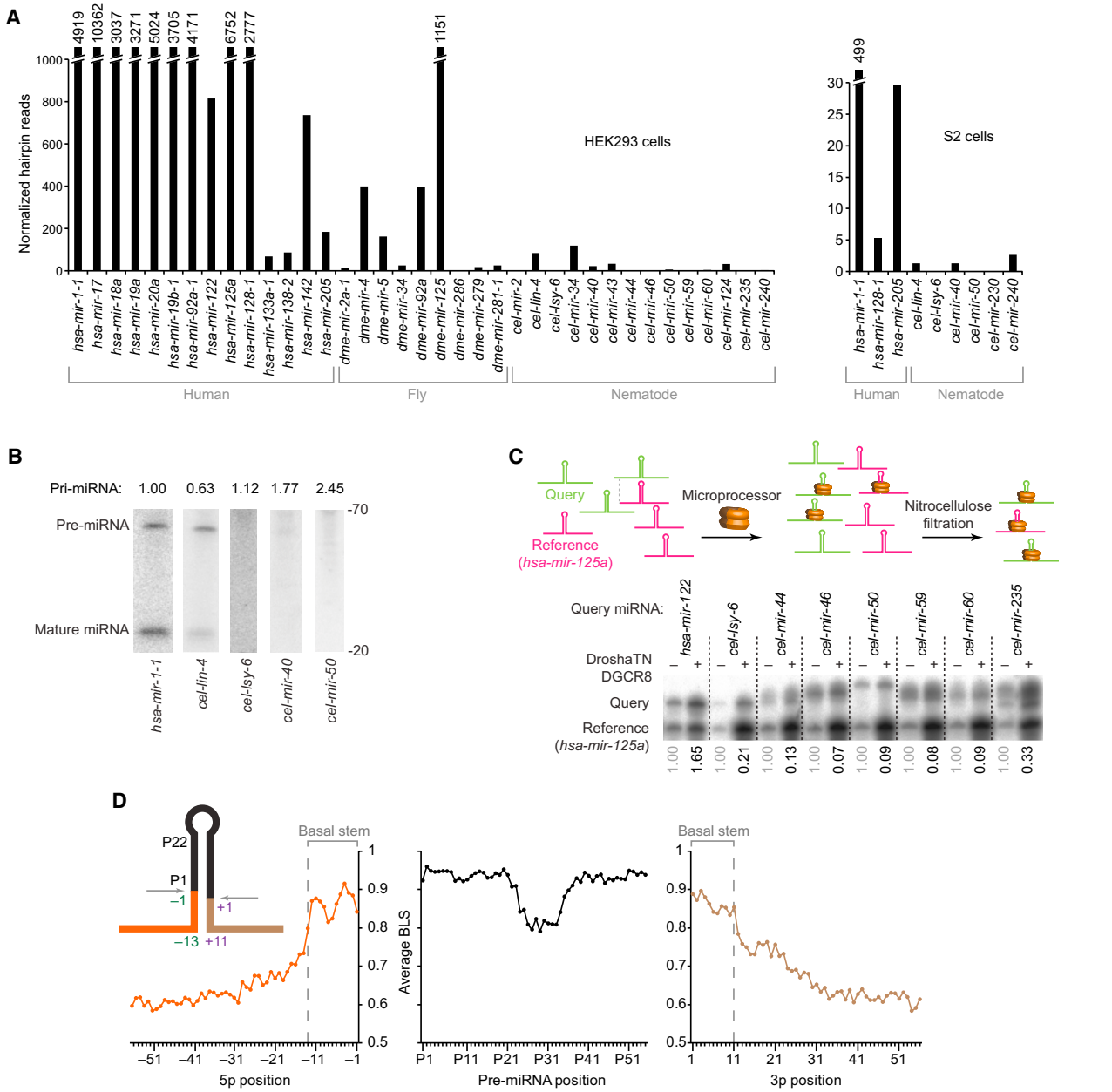


**Figure 1. Existence of Unknown Features Specifying Human Pri-miRNAs**

(A) Processing of human, fly, and nematode pri-miRNAs in human cells and *Drosophila* cells. Cells were transfected with plasmids expressing the indicated pri-miRNA hairpins with ~100 flanking genomic nucleotides on each side of each hairpin (Figure S1A), and total RNA was pooled for small-RNA sequencing. Plotted are small-RNA reads derived from the indicated pri-miRNAs.

(B) Accumulation of pri-miRNA, pre-miRNA, and miRNA after expressing the indicated pri-miRNAs in HEK293T cells. Pre-miRNA and mature species were measured by RNA blot of total RNA from cells transfected with plasmids expressing the indicated pri-miRNA (full gel images, including in vitro-transcribed cognate positive controls, in Figure S1B). Relative pri-miRNA levels (indicated above the lanes) are from ribonuclease protection assays, normalized to the signals for neomycin phosphotransferase mRNA also expressed from each expression plasmid.

(C) Relative binding of *C. elegans* and human pri-miRNAs to the Microprocessor. In the competitive binding assay (top, schematic), radiolabeled query pri-miRNA was mixed with the radiolabeled shorter reference pri-miRNA (human *mir-125a*) and incubated in excess over catalytically impaired Drosha (Drosha-TN) and DGCR8. Bound RNA was filtered on nitrocellulose and eluted for analysis on a denaturing gel. Phosphorimaging (bottom) indicated the relative amounts of input (–) and bound (+) RNAs. Numbers below each lane indicate the ratio of bound query to bound reference pri-miRNAs, normalized to their input ratio.

(D) Nucleotide conservation of human pri-miRNAs conserved to mouse, reported as the average branch-length score (BLS) at each position. Positions are numbered based on the inferred Drosha cleavage site (inset); negative indices are upstream of the 5p Drosha cleavage site, indices with "P" count from the 5′ end of the pre-miRNA, and positive indices are downstream of the 3p Drosha cleavage site.

See also Figure S1.

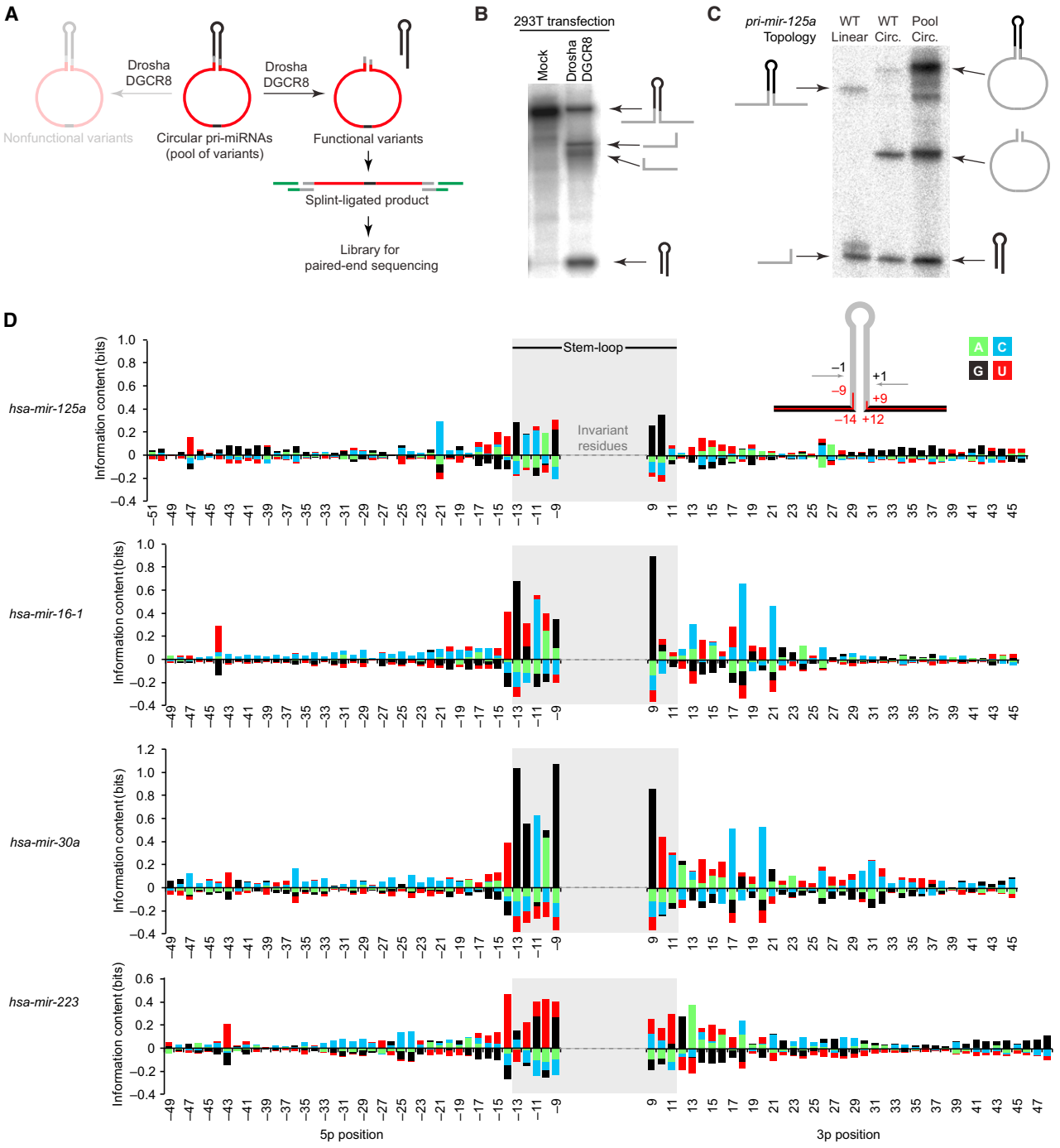


**Figure 2. Selection for Functional Pri-miRNA Variants**

(A) Schematic of the selection. Pri-miRNAs with variable residues (red) flanking the Drosha cleavage site were circularized by ligation and incubated in Microprocessor lysate. Cleaved variants were gel purified, ligated to adaptors, reverse transcribed, and amplified for high-throughput sequencing.

(B) Cleavage of *let-7a* in HEK293T whole-cell lysate (mock) and Microprocessor lysate (whole-cell lysate from HEK293T cells transfected with plasmids expressing Drosha and DGCR8). Incubations were 1.5 hr. Body-labeled reactants and products were resolved on a denaturing polyacrylamide gel and visualized by phosphorimaging.

(C) Cleavage of linear and circular *mir-125a* (WT linear and WT circ., respectively) and a pool of circular *mir-125a* variants (pool). RNAs were incubated for 5 min in Microprocessor lysate and analyzed as in (B). The linear RNA was 5′ end labeled; other RNAs were body labeled.

(D) Enrichment and depletion at variable residues in functional pri-miRNA variants. At each varied position (inset, red inner line), information content was calculated for each residue (green, cyan, black, and red for A, C, G, and U, respectively).

See also Figure S2.

that altered the cleavage site were not distinguished from those that abolished cleavage.

Some positions had substantial enrichment of one or more nucleotide possibilities, with corresponding depletion of others (Figure 2D). When tested in vitro, the results of changing specific residues closely matched those predicted from analysis of sequenced variants (Figures S2A and S2B). Moreover, the in vitro results predicted the direction and sometimes the magnitude of the effects observed in HEK293T cells (Figure S2C).

## Importance of an 11 bp Basal Stem Flanked by at Least Nine Unstructured Nucleotides

For all four miRNAs, some of the varied residues with the greatest influence fell within the basal stem (Figure 2D). Covariation matrices listing the odds ratio of each pair of nucleotide identities showed preference for Watson-Crick geometry at each basal pair, with the G:U wobble the most frequently preferred non-Watson-Crick alternative (Figures 3A and S3A). For example, the most favored alternatives to the wild-type C:G pair at positions −11 and +9 of *mir-125a* were the G:C and U:A pairs, and to a lesser extent, the A:U, G:U, and U:G pairs (Figure 3A). In fact, Watson-Crick pairing was strongly preferred even if it did not occur in the wild-type sequence. For example, the wild-type A:C pair at positions −12 and +10 of *mir-30a* was disfavored compared to the four Watson-Crick possibilities (Figure 3A), and the bulged A at position +10 of *mir-223* was preferentially incorporated into an alternative continuous helix (Figures S3A and S3B). Extending these methods to systematically evaluate all pairing possibilities involving all varied positions uncovered no evidence for Watson-Crick pairing outside the basal stem (Figure S3C).

Layered on the overall preference for Watson-Crick pairing were primary-sequence preferences specific to each basal pair. For example, at positions −11 and +9 the C:G pair was favored over the other Watson-Crick alternatives. The primary-sequence preference was most acute at the basal-most pair, where wobbles or mismatches involving G at −13 were favored over alternative Watson-Crick pairs (Figure 3A). We conclude that primary-sequence features supplement and sometimes supersede structural features important for basal-stem recognition.

The Microprocessor recognizes the junction between the miRNA hairpin and flanking ssRNA to position the active site approximately one helical turn (11 bp of A-form RNA) from the base of the duplex (Han et al., 2006; Yeom et al., 2006). To examine the preferred length of the basal stem, we calculated the relative cleavage efficiencies of different stem-length variants, normalizing to that of an 8 bp stem. Invariant mismatches within symmetric internal loops (e.g., the A:C mismatch at positions −6 and +4 of *mir-30a*) were assumed to be noncanonical pairs that stacked within the stem to contribute to its length, whereas mismatches at varied positions were assumed to disrupt further pairing and thereby terminate the inferred basal stem. For all four pri-miRNAs, an 11 bp basal stem was optimal (Figure 3B), consistent with the single-turn model. Indeed, an 11 bp basal stem was preferred for *mir-223* even though the wild-type sequence was predicted to form a 12 bp stem (Figures 3A and S3A). For most pri-miRNAs, however, the efficiency of the 12-pair stem approached that of the 11-pair stem (Figure 3B). This tolerance of a twelfth pair hinted that other features, such as the G at position −13, help specify the precise site of cleavage.

The single-turn model also posits that the nucleotides immediately flanking the basal stem are unstructured (Han et al., 2006; Yeom et al., 2006). To test this, we used RNAfold (Hofacker and Stadler, 2006) to predict the minimum free-energy structure of each sequenced pri-miRNA variant. For those with predicted wild-type stem pairing, we recorded the number of nucleotides between the base of the stem and the most proximal two consecutive structured residues. Although an imperfect estimate of the size of the unstructured segments flanking the base of the helix, this metric correlated well with cleavage (Figure 3C). Predicted pairing was tolerated in one flank, provided that the other flank contained at least 5–7 unpaired bases, consistent with reports of some cleavage when only one flanking segment is present (Zeng and Cullen, 2005; Han et al., 2006). When summing the flanking unpaired bases from both sides, the optimum plateaued at ~9–18 nt (Figure 3D).

## A Basal UG Motif Enhances Processing

Among the nucleotides upstream of the stemloop, the most striking enrichment was for a U at position −14 (Figure 2D). This U immediately preceded the position that, as mentioned above, displayed a strong primary-sequence preference for a G. The U and G at positions −14 and −13 contributed independently; variants with either a U or a G were enriched over variants with neither, and variants with both were even more enriched (Figure 4A). For *mir-223*, the UG at positions −14 and −13 was preferred (Figure 2D), even though wild-type *mir-223* has a UG at positions −15 and −14, respectively. This basal UG motif was also enriched among variants of *mir-125a* selected for Microprocessor binding rather than cleavage (Figure S4B).

The basal UG was conserved in vertebrate orthologs of *mir-16-1* and *mir-30a* (Figure 4B). Moreover, the motif was enriched in other mammalian pri-miRNAs, as illustrated by the sequence composition of human pri-miRNAs (Figure 4C). It was also enriched in pri-miRNAs of zebrafish (*D. rerio*) and tunicate (*C. intestinalis*) but only sporadically in more distantly related lineages, suggesting that its recognition emerged in a chordate ancestor (Figure 4D).

## The Broadly Conserved CNNC Motif Enhances Processing

In *mir-16-1*, *mir-30a*, and *mir-223* we observed a preference for two C residues, separated by two intervening nucleotides, beginning 17–18 nt downstream of the Drosha cleavage site (Figure 2D). The two C residues of this CNNC motif (N signifies any nucleotide) acted synergistically, in that variants that retained neither C residue were not disfavored much more than those that retained one (Figure 5A). The C residues enriched in the active variants were conserved in vertebrate orthologs of these three pri-miRNAs (Figure 5B).

The *mir-125a* pri-miRNA also had four C residues in this vicinity (positions 16–21), which gave rise to a CNNC at position

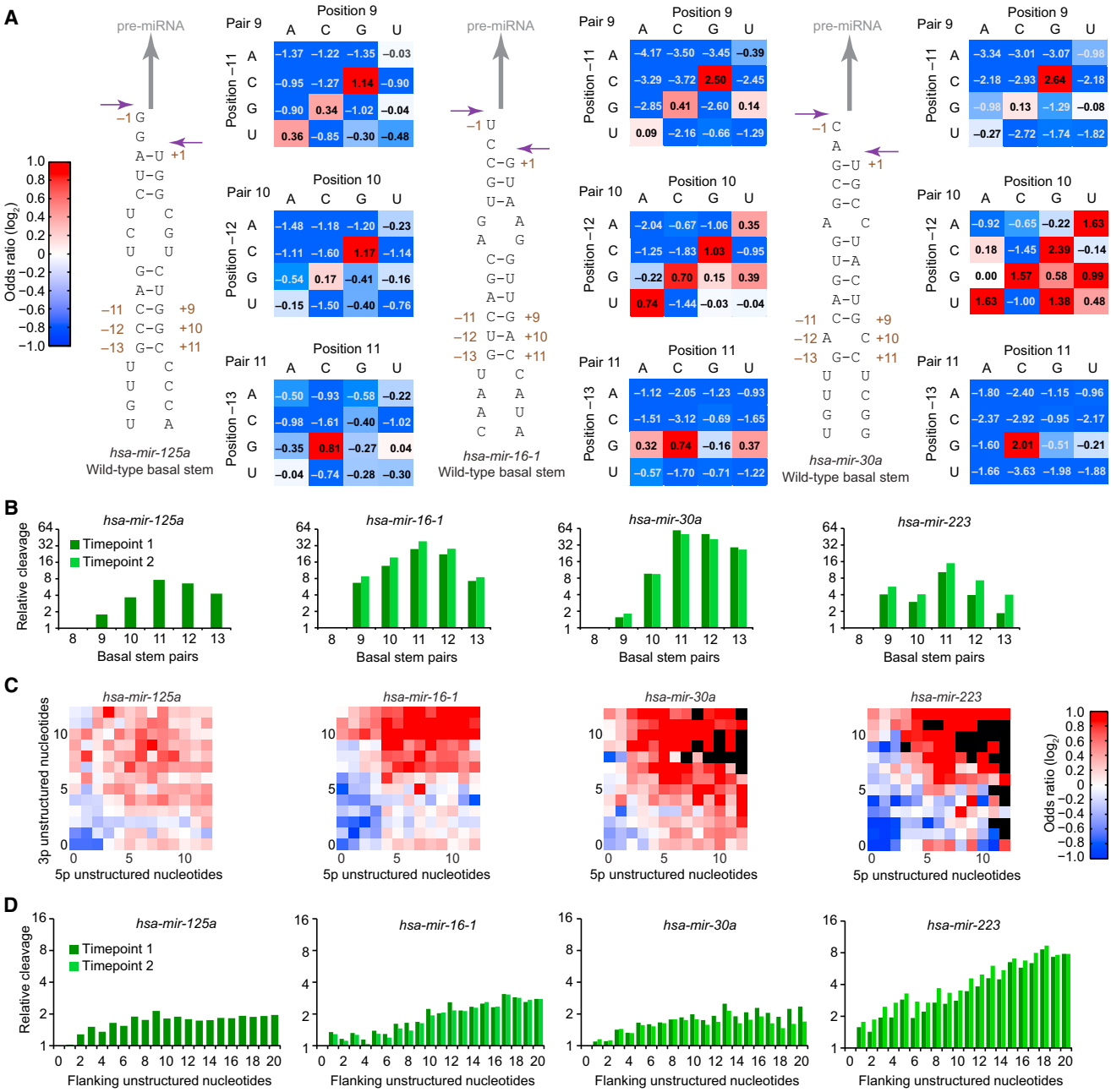


**Figure 3. Basal Stem Structure in Functional Pri-miRNA Variants**

(A) Predicted basal secondary structures and covariation matrices for *mir-125a*, *mir-16-1*, and *mir-30a*. For each pair of positions, joint nucleotide distributions were tabulated from sequences of the initial and selected pools, and the log odds ratio was calculated. Favored and disfavored pairs are colored red and blue, respectively, with color intensity (key) and values indicating magnitudes.

(B) Relative cleavage of variants with different stem lengths. The number of contiguous Watson-Crick pairs was counted, and the relative cleavage was calculated, normalized to the 8 bp stem. For selections with two time points, results are shown for both (key).

(C) Enrichment for unstructured nucleotides flanking the basal stem. Predicted folds of variant sequences were generated, and the subset of sequences with wild-type basal stem pairing were classified based on the distance to the nearest consecutive structured nucleotides upstream of position −13 and the nearest consecutive structured nucleotides downstream of position +11. Enrichment (red) and depletion (blue) of unstructured lengths among the selected variants are colored (key), with black indicating that sequencing data were insufficient to calculate enrichment.

(D) Relative cleavage of variants with differing numbers of total unstructured nucleotides flanking the basal stem. Upstream and downstream unstructured lengths predicted in (C) were summed, and the relative cleavage was calculated, normalized to zero unstructured nucleotides. For selections with two time points, results are shown for both (key).

See also Figure S3.

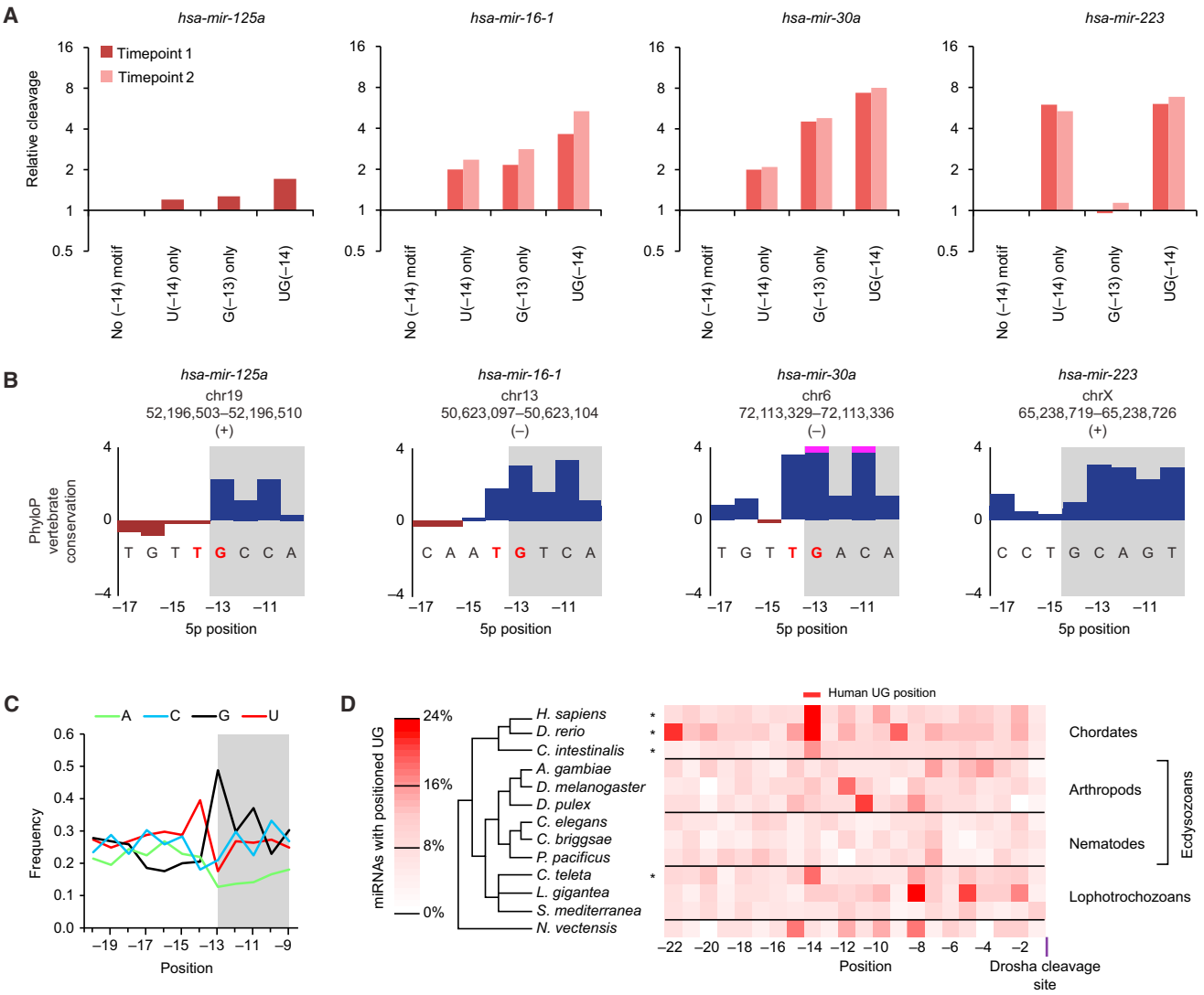


**Figure 4. The Basal UG Motif**

(A) Relative cleavage of variants with a full UG motif, a partial motif, and no motif. Values were normalized to those of variants with no motif, showing results from two time points, if available (key).

(B) PhyloP conservation across 30 vertebrate species in the region of the basal UG motif (red letters) for the four selected miRNAs. Bars extending beyond the scale of the graph are truncated (pink). Nucleotides predicted to be paired in the wild-type basal stem are shaded.

(C) Frequencies of A, C, G, and U (green, cyan, black, and red, respectively) at the indicated positions of human pri-miRNAs conserved to mouse. Analysis was of 204 pri-miRNAs, each representing a unique paralogous family (Table S2).

(D) Enrichment for the UG dinucleotide in the pri-miRNAs of representative animals with sequenced genomes. UG occurrences were tabulated for the upstream regions of pri-miRNAs aligned on the predicted Drosha cleavage site (Table S2). Species with statistically significant enrichment at position −14 are indicated (asterisks, empirical p value < $10^{-3}$).

See also Figure S4.

16 and the possibility of creating a CNNC at positions 17 or 18 (by changing either A20 or A18, respectively, to a C). However, the CNNC at position 16 was not preferred in the selection, nor were either of the single-nucleotide changes that could create a CNNC (Figures 2D and 5A). Moreover, the position 16 CNNC was not conserved in vertebrate orthologs (Figure 5B). These results indicate that unidentified features present in *mir-16-1*, *mir-30a*, and *mir-223*, but not *mir-125a*, are required for the CNNC to increase processing efficiency.

For the three pri-miRNAs in which the CNNC motif was effective, its position fell in a small window 17–18 nt downstream of the Drosha cleavage site. In variants in which neither wild-type C was present, alternative CNNC motifs were strongly enriched 1–2 nt downstream (Figure S5A), which further indicated that a CNNC motif within a small range of positions can contribute to pri-miRNA recognition.

Of the 64 possible dinucleotide motifs with zero to three intervening nucleotides, CNNC was the one most highly enriched

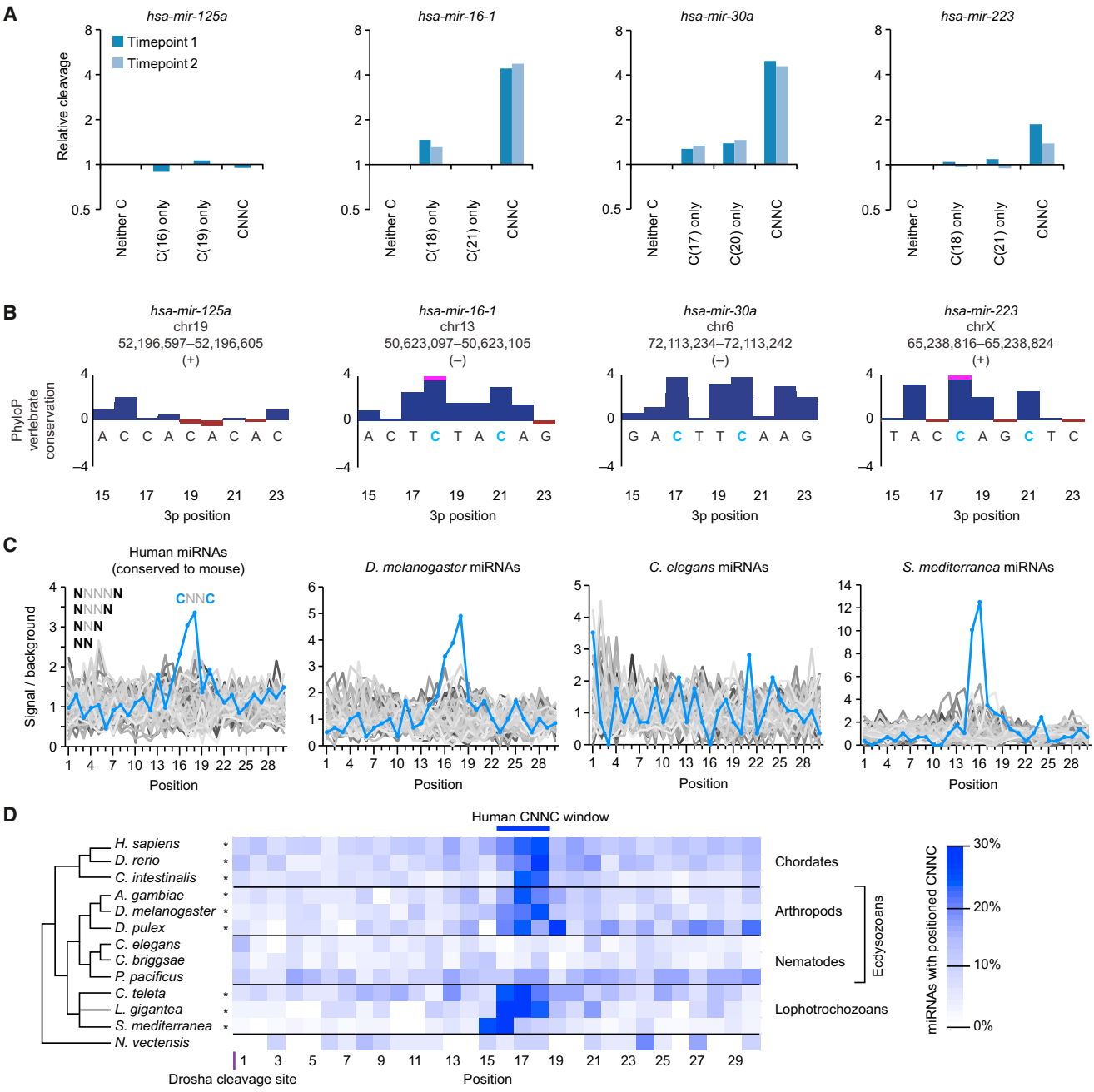


**Figure 5. The Downstream CNNC Motif**

(A) Relative cleavage of variants with a full CNNC motif, a partial motif, and no motif. Values were normalized to those of variants with no motif, showing results from two time points, if available (key).

(B) PhyloP conservation across 30 vertebrate species in the region of the downstream CNNC motif (blue letters) for the four selected pri-miRNAs. Bars extending beyond the scale of the graph are truncated (pink).

(C) CNNC enrichment compared to that of 63 other spaced dinucleotide motifs. Occurrences of each motif were tabulated for the downstream regions of pri-miRNAs aligned on the predicted Drosha cleavage site (Table S2). Background expectation was based on the nucleotide composition of pri-miRNA downstream regions in each species.

(D) Enrichment of the CNNC motif in the pri-miRNAs of representative bilaterian animals (Table S2). Species with statistically significant enrichment at positions 16, 17, or 18 are indicated (asterisk, empirical p value < $10^{-4}$).

See also Figure S5.

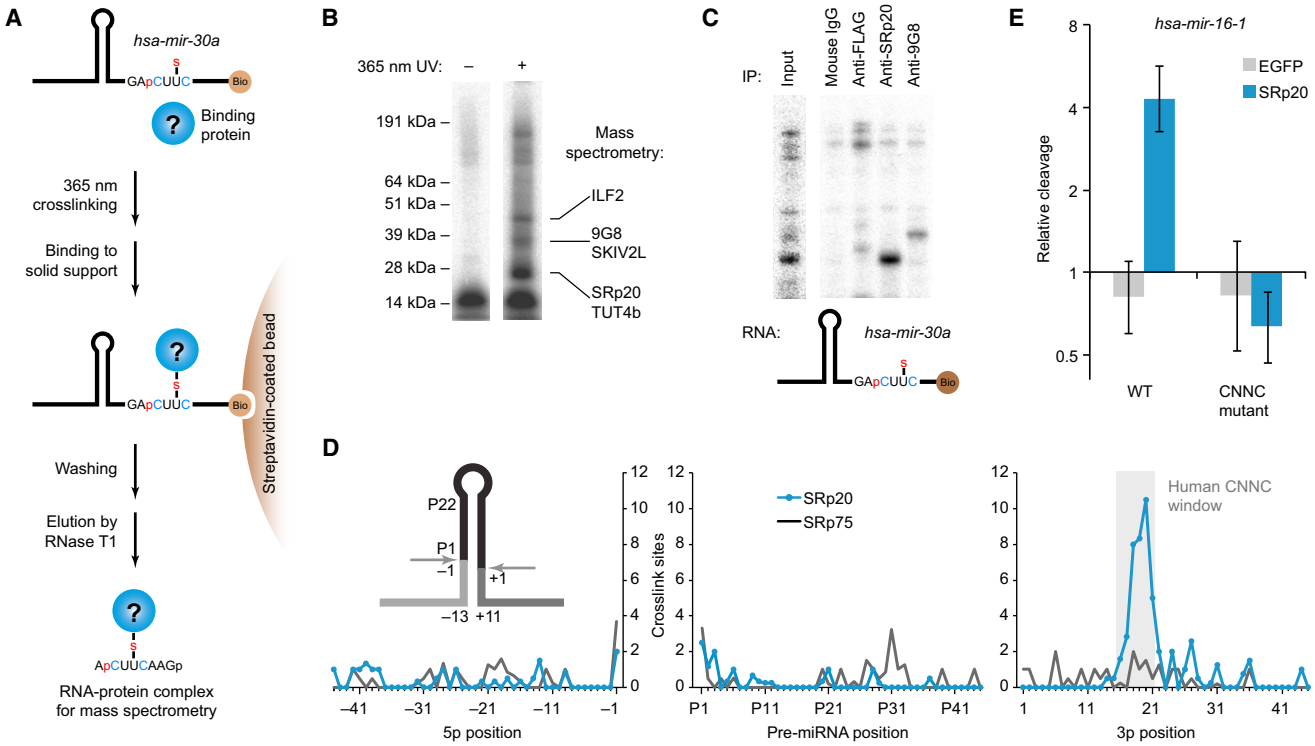


**Figure 6. Binding and Activity of SRp20 at the CNNC Motif**

(A) Site-specific crosslinking approach used to identify CNNC-binding proteins. The *mir-30a* crosslinking substrate contained a photoreactive base in the CNNC motif (4-thiouridine, U–S), a 3′ biotin (Bio), and for some applications, a $^{32}$P-labeled phosphate (red p). This substrate was incubated in Microprocessor lysate and irradiated with 365 nm UV light. Crosslinked complexes were captured on streptavidin-coated beads and eluted by RNase T1 digestion.

(B) Proteins within crosslinked RNA-protein complexes. Crosslinked complexes prepared as in (A) were separated on an SDS gel. For each CNNC-crosslinked band, proteins are listed that were identified by mass spectrometry and have known or inferred RNA-binding activity.

(C) Immunoprecipitation of proteins crosslinked to the CNNC motif. After crosslinking as in (A), complexes were enriched using monoclonal antibodies against either FLAG (the tag of the overexpressed Drosha and DGCR8), SRp20, or 9G8 and then resolved on an SDS gel. Input was run on a different region of the same gel for reference.

(D) SRp20 binding downstream of mouse pri-miRNA hairpins in vivo. Sites were obtained by reanalysis of crosslinking data for SRp20 and SRp75 in mouse cells (Änkö et al., 2012). Positions are numbered as in Figure 1D. Expected sites of crosslinks to any of the motif nucleotides in the region of motif enrichment (Figure 5D) are shaded (gray).

(E) Enhancement of in vitro pri-miRNA cleavage by SRp20. Wild-type *pri-mir-16-1* or *pri-mir-16-1* with mutated CNNC were incubated for 3 min with immunopurified Microprocessor, supplemented with either FLAG-EGFP or 3X-FLAG-SRp20 purified from HEK293T cells. Reactants and products were resolved on denaturing polyacrylamide gels and quantified by phosphorimaging relative to a buffer-only control (geometric mean ± standard error, n = 3).

See also Figure S6.

downstream of the cleavage sites of human pri-miRNAs (Figure 5C). Moreover, enrichment was limited to a small range of positions 16–18 nt downstream of the site, peaking at positions 17 and 18, which matched the positions of the motif within *mir-16-1*, *mir-30a*, and *mir-223*. These results suggest that the CNNC motif enhances processing of many human pri-miRNAs.

Similar analyses of nonmammalian pri-miRNAs indicated strong, position-specific enrichment of the CNNC motif in chordates, arthropods, and lophotrochozoans, but not in sea anemone (*Nematostella vectensis*) (Figures 5C and 5D), suggesting that its recognition emerged with the divergence of bilaterians. Interestingly, enrichment was also absent in nematodes (Figures 5C and 5D), suggesting an isolated loss in the nematode branch of the ecdysozoans.

Consistent with the results in extracts, mutation of the basal UG and downstream CNNC motifs each reduced accumulation of mature miR-16 and miR-30a in HEK293T cells, with mutation of both reducing accumulation ~4–8-fold relative to wild-type (Figures S5B and S5C). Furthermore, one or both motifs contributed to the accumulation of each of the additional pri-miRNAs tested in cell culture (*hsa-mir-28*, *hsa-mir-129-2*, and *hsa-mir-193b*; Figures S5D–S5F).

## SRp20 Binds the CNNC Motif and Enhances Processing

To learn how the CNNC motif is recognized, we used site-specific crosslinking (Wyatt et al., 1992). Proteins that crosslinked to *pri-mir-30a* RNA with a photoreactive nucleotide (4-thiouridine) placed within the CNNC motif were identified by mass spectrometry (Figure 6A). To guide gel-purification of

crosslinked proteins, we performed the procedure in parallel with a radiolabeled pri-miRNA designed to label only proteins that crosslinked in the vicinity of the CNNC (Figures 6A and 6B). The two strongest candidates were SRp20/SRSF3 and 9G8/SRSF7, closely related proteins implicated in splicing regulation (Zahler et al., 1993; Cavaloc et al., 1994), mRNA export (Huang and Steitz, 2001), and translation initiation (Bedard et al., 2007; Swartz et al., 2007). These proteins both have an RNA-recognition motif (RRM) conserved across bilaterian animals, which recognizes degenerate motifs closely related to the CNNC motif (Heinrichs and Baker, 1995; Cavaloc et al., 1999; Schaal and Maniatis, 1999). NMR studies of this RRM in complex with RNA indicate that the C residues, particularly the first C of the CNNC, are bound in a base-specific manner, with minimal preferences for the two intervening bases (Hargous et al., 2006). Immunopurification of SRp20 and 9G8 confirmed that these two proteins (particularly SRp20) were the ones that most efficiently crosslinked in our assay (Figure 6C).

To evaluate SRp20 binding in vivo, we analyzed a large data set of SRp20 crosslinking sites in P19 cells (Änkö et al., 2012). Although the published analyses of this data set focused on sites within pre-mRNAs, we found that many SRp20 sites resided in pri-miRNAs, and, more importantly, that these sites overlapped the region of CNNC enrichment (Figure 6D). This analysis extended our results from in vitro binding to in vivo binding and from one pri-miRNA to many. Some of the crosslinking sites in the CNNC-enriched region were in pri-miRNAs that lacked a CNNC motif, suggesting that SRp20 (and presumably its paralog, 9G8) might play a role even more general than that implied by CNNC conservation and enrichment.

The requirement of SRp20 for cell viability (Jumaa et al., 1999; Jia et al., 2010) confounded attempts to test its function by depleting the protein in cell culture. Therefore, we tested its function in vitro, supplementing immunopurified Microprocessor complex with either immunopurified recombinant SRp20 (Figure S6) or an analogously purified control protein (EGFP). SRp20 enhanced *mir-16-1* processing in a CNNC-dependent manner (Figure 6E). Taken together, our results indicate that for many bilaterian miRNAs the CNNC motif is enriched and preferentially conserved because it helps recruit SRp20 (or its homologs), which enhances pri-miRNA recognition and processing.

### Loop and Apical Stem Elements Can Enhance Processing

To examine whether additional processing features reside in the loop and apical stem, we extended our approach to those regions (Figure S7A). Pairing at the apical portion of the stem contributed to pri-miRNA recognition and processing for *mir-125a* and *mir-30a*, but not for *mir-16-1* or *mir-223* (Figure S7B), consistent with differing conclusions drawn from studies of different miRNAs (Zeng et al., 2005; Han et al., 2006). Primary-sequence preferences were weaker than those observed for basal and flanking residues (Figure S7C). The best candidate for a loop-binding motif was observed only in *mir-30a*, in which the wild-type UGUG at positions P24–27 was both preferred in the selection (Figure S7D) and conserved in vertebrate orthologs (Figure S7E). Human and zebrafish miRNAs were enriched for UGU or GUG in this region of the loop (empirical $p < 10^{-5}$ for each species) (Figure S7F), thereby confirming it as the third primary-sequence motif identified in our study (Figure 7A).

### Rescue of *C. elegans* miRNA Expression in Human Cells

The primary-sequence motifs important for mammalian miRNAs were not enriched in the nematode clade, suggesting that their absence might account for the failure of *C. elegans* pri-miRNAs to be processed in human cells. To test this idea, we added the basal UG and the downstream CNNC motifs to *cel-mir-44* in the context of the *mir-1* bicistronic vector (Figure 7B). Before adding the motifs, we disrupted the predicted pairing between positions −14 and +12 and substituted the G:C pair at positions −13 and +11 (construct mir44.1). These changes, which were expected to simultaneously enhance processing by shortening the basal stem to its optimal length and inhibit processing by replacing the fortuitous G at position −13, had a marginal net effect on production of mature miR-44 in human cells (Figure 7B). Adding a basal UG enhanced production of mature miR-44 by 5-fold (8-fold over the wild-type), primarily from restoring the G at −13 (Figure 7B). Adding a CNNC 17 nt downstream of the cleavage site (mir44.4) enhanced production another 8-fold, yielding a 64-fold net increase over wild-type (Figure 7B). Similarly, converting the wild-type, asymmetrically bulged stem of *cel-mir-50* to a regular, 11-pair stem and adding the UG and CNNC motifs enhanced expression of mature miR-50 by 30-fold (Figure S7G), while adding the motifs to *cel-mir-40* enhanced expression of mature miR-40 by 5-fold (Figure S7H). We conclude that primary-sequence motifs discovered in this study help human cells to distinguish pri-miRNA hairpins from other hairpins and that the absence of these motifs in *C. elegans* pri-miRNAs helps to explain why human cells do not regard these transcripts as pri-miRNAs.

## DISCUSSION

Secondary structure is inadequate on its own to specify pri-miRNA hairpins: primary-sequence features, including the basal UG, the CNNC, and the apical GUG motifs, also contribute to efficient processing in human cells (Figure 7A). Complicating the story (and perhaps explaining why these primary-sequence features had not been observed earlier), different pri-miRNAs differentially benefit from the different motifs (Figure 7C). Among human pri-miRNAs, these motifs were nonetheless highly enriched, with 79% of the conserved human miRNAs containing at least one of the three motifs (Figure 7D).

The motifs were not enriched in *C. elegans* pri-miRNAs (Figures 7E) and, when added to the *C. elegans* pri-miRNAs, conferred more efficient processing in mammalian cells (Figure 7B, S7G, and S7H). These experiments also showed the benefit of disrupting pairing normally present at positions −14 and +12 of the *C. elegans* miRNAs. The presence of pairing that is inhibitory to mammalian processing suggests that measurement from the base of the helix might also differ

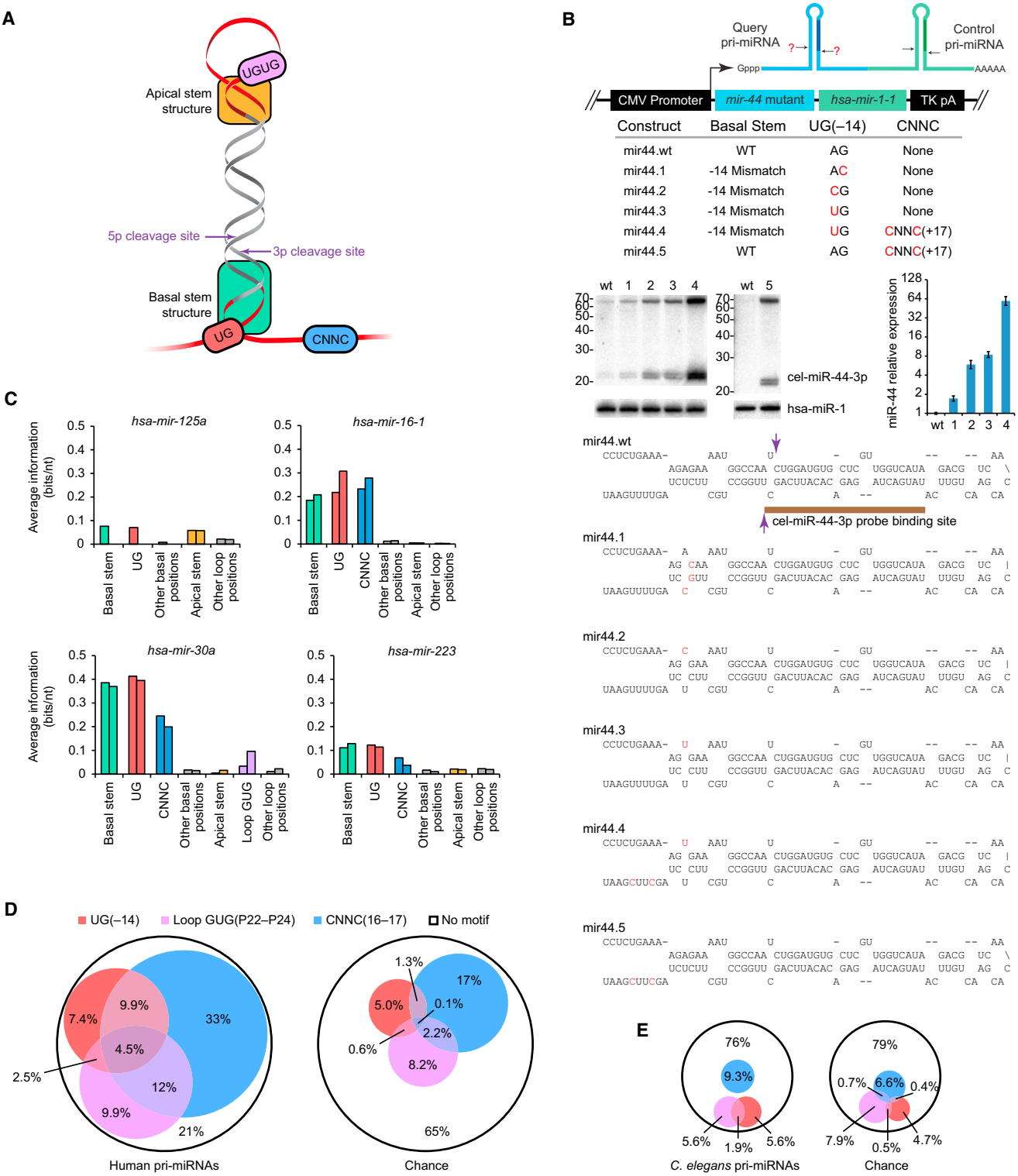


**Figure 7. Structural and Primary-Sequence Features Important for Human Pri-miRNA Processing**

(A) Summary of human pri-miRNA determinants identified or confirmed in this study.

(B) Processing enhancement from adding human pri-miRNA features to *C. elegans mir-44*. Changes that introduced the listed features were incorporated into *mir-44* within the bicistronic expression vector (top). Secondary structures are shown for mutations predicted to affect the wild-type basal stem (bottom; Drosha cleavage sites, purple arrowheads). After transfection into HEK293T cells, accumulation of miR-44-3p was assessed on RNA blots (middle), with the graph plotting increased miR-44-3p expression normalized to that of the hsa-miR-1 control (geometric mean ± standard error, n = 3). Adding a CNNC to the

*(legend continued on next page)*

in nematodes. Thus, despite the many broadly conserved features of miRNAs, some primary-sequence features and some secondary-structure features differ in mammals and nematodes.

About a fifth of human pri-miRNAs lack all three newly identified primary-sequence determinants (Figure 7D). These are attractive subjects for further study, in that the approach implemented here presumably would identify additional unique determinants used by these pri-miRNAs. Other determinants probably also exist at the Microprocessor cleavage site and nearby stem regions, which were inaccessible to our approach as implemented. Indeed, point mutations that disrupt pairing in the middle of the stem dramatically impair processing (Gottwein et al., 2006; Duan et al., 2007; Jazdzewski et al., 2008; Sun et al., 2009), and the SR-domain splicing factor SF2/ASF is reported to enhance the processing of *mir-7-1* by binding a motif in the stem near the cleavage site (Wu et al., 2010). Hinting at the possibility of additional primary-sequence preferences within the stem are results from both bacterial RNase III and fungal homologs (Rnt1 and Pac1), which prefer specific base-pair identities near the cleavage site (Lamontagne and Elela, 2004).

The emerging picture is that pri-miRNA recognition is a modular phenomenon in which each module contributes modestly, and each pri-miRNA depends on individual modules to varying degrees. Our results quantify the relative importance of each known module for each pri-miRNA (Figure 7C). Pairing within the basal stem was crucial, as expected (Lim et al., 2003b; Han et al., 2006). In addition, all four miRNAs made use of the basal UG motif, which provided information content per nucleotide resembling that provided by the basal-stem nucleotides. For the three miRNAs that used the CNNC SRp20-binding site, its importance was also comparable to that of the basal stem nucleotides. Compared to the nucleotides within these motifs, other flanking nucleotides contributed very little.

Apical and terminal loop elements were less important than the basal motifs (Figure 7C). We detected significant contributions only in *mir-125a*, in which the apical stem nucleotides were as important as the basal stem nucleotides, and in *mir-30a*, in which the loop UGUG motif contributed some information, albeit less than any of the three other features. Together, the features described here explained 61%–78% of the information content in the selected sequences. The remaining information content was diffusely distributed among the other partially randomized positions and might have mostly reflected avoidance of detrimental alternative structures.

Knowledge of biogenesis features will aid in interpreting human mutations. For example, reduced miR-16 expression associated with chronic lymphocytic leukemia (CLL) is typically due to deletions spanning the intron containing *mir-15a* and *mir-16-1* (Calin et al., 2002). However, 2 of 75 CLL patients studied had tumors that retain the pri-miRNA hairpins and instead carried a germline C > T single-nucleotide polymorphism (SNP) downstream of the *mir-16-1* hairpin (Calin et al., 2005). This SNP lowers overexpression of miR-16 in HEK293 cells, and in both patients heterozygosity for the SNP was lost in the leukemic cells (Calin et al., 2005). This SNP corresponds to the first C in the *mir-16-1* CNNC, which explains why it lowers miR-16 accumulation and leads to CLL: it affects pri-miRNA processing by disrupting SRp20 recruitment. Discovery of additional features for pri-miRNA recognition and processing might lead to improved diagnostic and therapeutic tools in cancer and other diseases in which miRNAs are dysregulated.

## EXPERIMENTAL PROCEDURES

### Ectopic Pri-miRNA Expression

Plasmids were derived from pcDNA3.2/V5-DEST and pMT-DEST (Invitrogen) for expression in HEK293 and S2 cells, respectively. Query pri-miRNA sequences and the human *pri-mir-1-1* sequence were cloned such that the query pri-miRNAs were transcriptionally fused upstream of *mir-1-1*. HEK293 and S2 cells were transfected using Lipofectamine 2000 and Cellfectin (Invitrogen), respectively. After 36–48 hr, total RNA was extracted, and miRNA expression was assayed by RNA blots, ribonuclease protection assays (Invitrogen), and high-throughput sequencing (Chiang et al., 2010). For additional details including the data analysis pipeline, see Extended Experimental Procedures.

### Binding and Cleavage Assays

To assay binding, we radiolabeled and mixed T7-transcribed competitor and reference pri-miRNA substrates in an equimolar ratio, then incubated them with limiting amounts of immunopurified catalytically impaired Microprocessor (Lee and Kim, 2007; Han et al., 2009). RNA-protein complexes were filtered on Immobilon-NC nitrocellulose discs (Whatman), and RNA extracted from the filter was resolved on 5% polyacrylamide gels. To assay cleavage, we incubated labeled substrates with Microprocessor lysate, which was prepared from cells overexpressing Drosha and DGCR8 (Lee and Kim, 2007). After extraction using Tri-Reagent (Ambion), substrates and products were resolved on denaturing 5% polyacrylamide gels. For additional details, see Extended Experimental Procedures.

### Synthesis and Selection of Pri-miRNA Variants

Templates for T7 transcription were assembled from oligonucleotides (IDT) synthesized using nucleoside phosphoramidite mixtures designed to introduce variability at specified positions (Table S1). Sequences encoding the HDV self-cleaving ribozyme were appended so that ribozyme cleavage would

wild-type sequence (construct mir44.5) enhanced processing ≥ 20-fold (geometric mean of triplicate experiment), a lower bound set by the wild-type background.

(C) Contributions of individual features to in vitro processing measured as average information content per nucleotide. If available, results from two time points are shown.

(D) Enrichment of primary-sequence motifs in human pri-miRNAs conserved to mouse (Table S2). Pri-miRNAs were classified based on whether they had the basal UG, the apical GUG or UGU, or the downstream CNNC motif (left). Expectations by chance (right) were estimated based on the nucleotide composition of upstream, pre-miRNA, and downstream regions of human pri-miRNAs for the basal UG, apical GUG or UGU, and CNNC motifs, respectively.

(E) A search for human motifs in *C. elegans* pri-miRNAs (Table S2). Pri-miRNAs were analyzed as in (D); the smaller diagrams reflect the smaller number of analyzed pri-miRNAs.

See also Figure S7.

generate transcripts with defined 3′ ends. Template pools were transcribed using T7 RNA polymerase, and after treatment with TurboDNase (Ambion) RNA was purified on denaturing polyacrylamide gels. After dephosphorylation of 5′ and 3′ ends using calf intestinal phosphatase (NEB) and T4 polynucleotide kinase (T4 PNK, NEB), followed by 5′ phosphorylation using T4 PNK, transcripts were circularized using T4 RNA ligase 1 (NEB) and gel purified. RNA pools were incubated with Microprocessor lysate, and after gel purification, cleavage products were ligated to oligonucleotide adaptors, reverse transcribed, amplified, and Illumina sequenced (75 nt paired-end reads). In parallel, the initial pool of RNA was also reverse transcribed, amplified, and sequenced. Selections for examining binding or apical stem-loops were similar, except transcripts were not circularized. For additional details including the data analysis pipeline, see Extended Experimental Procedures.

### Motif Enrichment

Enrichment of a motif within pri-miRNAs of a species was evaluated by comparing to 100,000 cohorts of miRNAs in which the upstream, downstream, and pre-miRNA sequences were independently shuffled, preserving dinucleotide frequencies. The numbers of miRNAs that contained a match to the motif in the actual and shuffled cohorts were used to compute an empirical p value. A list of the representative pri-miRNAs used for analyses is provided (Table S2). For additional details, see Extended Experimental Procedures.

### Site-Specific Crosslinking

The *mir-30a* pri-miRNA crosslinking substrate was assembled using T4 RNA ligase 2 (NEB) and a DNA splint to join an in vitro-transcribed 5′ fragment to a synthetic 3′ fragment containing a 3′-terminal biotin and a 4-thiouridine within the CNNC motif (Dharmacon). This crosslinking substrate was incubated in Microprocessor lysate and exposed to 1000 mJ of 365 nm UV light in a Stratalinker (Stratagene). For purification of RNA-protein complexes for mass spectrometry, complexes were captured on streptavidin-coated magnetic beads (Invitrogen), washed, and eluted with RNase T1 (Ambion), which cleaves after G. Eluted complexes either were separated on SDS gels and analyzed by HPLC/tandem mass spectrometry or were immunoprecipitated and analyzed by SDS gel. For additional details, see Extended Experimental Procedures.

## ACCESSION NUMBERS

The Short Read Archive accession number for the sequencing data reported in this paper is SRA051323.

## SUPPLEMENTAL INFORMATION

Supplemental Information includes seven figures, two tables, and Extended Experimental Procedures and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2013.01.031.

## ACKNOWLEDGMENTS

We thank D. Shechner, C. Jan, O. Rissland, D. Weinberg, J. Ruby, J. Nam, and V.N. Kim for valuable discussions; O. Rissland for comments on this manuscript; L. Schoenfeld and J. Lassar for technical assistance; J. Stévenin for 9G8 antibody; V.N. Kim and T. Tuschl for plasmids; the Whitehead Institute Genome Technology Core for sequencing; and E. Spooner for mass spectrometry. This work was supported by NIH grants GM067031 and T32GM007753. D.B. is an Investigator of the Howard Hughes Medical Institute.

Received: April 10, 2012
Revised: October 28, 2012
Accepted: January 14, 2013
Published: February 14, 2013

## REFERENCES

Änkö, M.L., Müller-McNicoll, M., Brandl, H., Curk, T., Gorup, C., Henry, I., Ule, J., and Neugebauer, K.M. (2012). The RNA-binding landscapes of two SR proteins reveal unique functions and binding to diverse RNA classes. Genome Biol. *13*, R17.

Babiarz, J.E., Ruby, J.G., Wang, Y., Bartel, D.P., and Blelloch, R. (2008). Mouse ES cells express endogenous shRNAs, siRNAs, and other Microprocessor-independent, Dicer-dependent small RNAs. Genes Dev. *22*, 2773–2785.

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

Bedard, K.M., Daijogo, S., and Semler, B.L. (2007). A nucleo-cytoplasmic SR protein functions in viral IRES-mediated translation initiation. EMBO J. *26*, 459–467.

Bentwich, I., Avniel, A., Karov, Y., Aharonov, R., Gilad, S., Barad, O., Barzilai, A., Einat, P., Einav, U., Meiri, E., et al. (2005). Identification of hundreds of conserved and nonconserved human microRNAs. Nat. Genet. *37*, 766–770.

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

Brummelkamp, T.R., Bernards, R., and Agami, R. (2002). A system for stable expression of short interfering RNAs in mammalian cells. Science *296*, 550–553.

Calin, G.A., Dumitru, C.D., Shimizu, M., Bichi, R., Zupo, S., Noch, E., Aldler, H., Rattan, S., Keating, M., Rai, K., et al. (2002). Frequent deletions and down-regulation of micro- RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. Proc. Natl. Acad. Sci. USA *99*, 15524–15529.

Calin, G.A., Ferracin, M., Cimmino, A., Di Leva, G., Shimizu, M., Wojcik, S.E., Iorio, M.V., Visone, R., Sever, N.I., Fabbri, M., et al. (2005). A microRNA signature associated with prognosis and progression in chronic lymphocytic leukemia. N. Engl. J. Med. *353*, 1793–1801.

Cavaloc, Y., Popielarz, M., Fuchs, J.P., Gattoni, R., and Stévenin, J. (1994). Characterization and cloning of the human splicing factor 9G8: a novel 35 kDa factor of the serine/arginine protein family. EMBO J. *13*, 2639–2649.

Cavaloc, Y., Bourgeois, C.F., Kister, L., and Stévenin, J. (1999). The splicing factors 9G8 and SRp20 transactivate splicing through different and specific enhancers. RNA *5*, 468–483.

Cheloufi, S., Dos Santos, C.O., Chong, M.M., and Hannon, G.J. (2010). A dicer-independent miRNA biogenesis pathway that requires Ago catalysis. Nature *465*, 584–589.

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

Chiang, H.R., Schoenfeld, L.W., Ruby, J.G., Auyeung, V.C., Spies, N., Baek, D., Johnston, W.K., Russ, C., Luo, S., Babiarz, J.E., et al. (2010). Mammalian microRNAs: experimental evaluation of novel and previously annotated genes. Genes Dev. *24*, 992–1009.

Chung, W.J., Agius, P., Westholm, J.O., Chen, M., Okamura, K., Robine, N., Leslie, C.S., and Lai, E.C. (2011). Computational and experimental identification of mirtrons in *Drosophila melanogaster* and *Caenorhabditis elegans*. Genome Res. *21*, 286–300.

Cifuentes, D., Xue, H., Taylor, D.W., Patnode, H., Mishima, Y., Cheloufi, S., Ma, E., Mane, S., Hannon, G.J., Lawson, N.D., et al. (2010). A novel miRNA processing pathway independent of Dicer requires Argonaute2 catalytic activity. Science *328*, 1694–1698.

Denli, A.M., Tops, B.B., Plasterk, R.H., Ketting, R.F., and Hannon, G.J. (2004). Processing of primary microRNAs by the Microprocessor complex. Nature *432*, 231–235.

Duan, R., Pak, C., and Jin, P. (2007). Single nucleotide polymorphism associated with mature miR-125a alters the processing of pri-miRNA. Hum. Mol. Genet. *16*, 1124–1131.

Fellmann, C., Zuber, J., McJunkin, K., Chang, K., Malone, C.D., Dickins, R.A., Xu, Q., Hengartner, M.O., Elledge, S.J., Hannon, G.J., and Lowe, S.W. (2011). Functional identification of optimized RNAi triggers using a massively parallel sensor assay. Mol. Cell *41*, 733–746.

Feng, Y., Zhang, X., Song, Q., Li, T., and Zeng, Y. (2011). Drosha processing controls the specificity and efficiency of global microRNA expression. Biochim. Biophys. Acta *1809*, 700–707.

Gottwein, E., Cai, X., and Cullen, B.R. (2006). A novel assay for viral microRNA function identifies a single nucleotide polymorphism that affects Drosha processing. J. Virol. *80*, 5321–5326.

Gregory, R.I., Yan, K.P., Amuthan, G., Chendrimada, T., Doratotaj, B., Cooch, N., and Shiekhattar, R. (2004). The Microprocessor complex mediates the genesis of microRNAs. Nature *432*, 235–240.

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., Kim, Y.K., Jin, H., and Kim, V.N. (2004). The Drosha-DGCR8 complex in primary microRNA processing. Genes Dev. *18*, 3016–3027.

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.

Han, J., Pedersen, J.S., Kwon, S.C., Belair, C.D., Kim, Y.K., Yeom, K.H., Yang, W.Y., Haussler, D., Blelloch, R., and Kim, V.N. (2009). Posttranscriptional crossregulation between Drosha and DGCR8. Cell *136*, 75–84.

Hargous, Y., Hautbergue, G.M., Tintaru, A.M., Skrisovska, L., Golovanov, A.P., Stevenin, J., Lian, L.Y., Wilson, S.A., and Allain, F.H. (2006). Molecular basis of RNA recognition and TAP binding by the SR proteins SRp20 and 9G8. EMBO J. *25*, 5126–5137.

Heinrichs, V., and Baker, B.S. (1995). The *Drosophila* SR protein RBP1 contributes to the regulation of doublesex alternative splicing by recognizing RBP1 RNA target sequences. EMBO J. *14*, 3987–4000.

Hofacker, I.L., and Stadler, P.F. (2006). Memory efficient folding algorithms for circular RNA secondary structures. Bioinformatics *22*, 1172–1176.

Huang, Y., and Steitz, J.A. (2001). Splicing factors SRp20 and 9G8 promote the nucleocytoplasmic export of mRNA. Mol. Cell *7*, 899–905.

Hutvágner, G., McLachlan, J., Pasquinelli, A.E., Bálint, 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.

Jazdzewski, K., Murray, E.L., Franssila, K., Jarzab, B., Schoenberg, D.R., and de la Chapelle, A. (2008). Common SNP in pre-miR-146a decreases mature miR expression and predisposes to papillary thyroid carcinoma. Proc. Natl. Acad. Sci. USA *105*, 7269–7274.

Jia, R., Li, C., McCoy, J.P., Deng, C.X., and Zheng, Z.M. (2010). SRp20 is a proto-oncogene critical for cell proliferation and tumor induction and maintenance. Int. J. Biol. Sci. *6*, 806–826.

Jumaa, H., Wei, G., and Nielsen, P.J. (1999). Blastocyst formation is blocked in mouse embryos lacking the splicing factor SRp20. Curr. Biol. *9*, 899–902.

Khvorova, A., Reynolds, A., and Jayasena, S.D. (2003). Functional siRNAs and miRNAs exhibit strand bias. Cell *115*, 209–216.

Lamontagne, B., and Elela, S.A. (2004). Evaluation of the RNA determinants for bacterial and yeast RNase III binding and cleavage. J. Biol. Chem. *279*, 2231–2241.

Landthaler, M., Yalcin, A., and 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.

Lee, Y., and Kim, V.N. (2007). In vitro and in vivo assays for the activity of Drosha complex. Methods Enzymol. *427*, 89–106.

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

Lim, L.P., Glasner, M.E., Yekta, S., Burge, C.B., and Bartel, D.P. (2003a). Vertebrate microRNA genes. Science *299*, 1540.

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

Liu, J., Carmell, M.A., Rivas, F.V., Marsden, C.G., Thomson, J.M., Song, J.J., Hammond, S.M., Joshua-Tor, L., and Hannon, G.J. (2004). Argonaute2 is the catalytic engine of mammalian RNAi. Science *305*, 1437–1441.

Macrae, I.J., Zhou, K., Li, F., Repic, A., Brooks, A.N., Cande, W.Z., Adams, P.D., and Doudna, J.A. (2006). Structural basis for double-stranded RNA processing by Dicer. Science *311*, 195–198.

Okamura, K., Hagen, J.W., Duan, H., Tyler, D.M., and Lai, E.C. (2007). The mirtron pathway generates microRNA-class regulatory RNAs in Drosophila. Cell *130*, 89–100.

Paddison, P.J., Caudy, A.A., Bernstein, E., Hannon, G.J., and Conklin, D.S. (2002). Short hairpin RNAs (shRNAs) induce sequence-specific silencing in mammalian cells. Genes Dev. *16*, 948–958.

Pan, T., and Uhlenbeck, O.C. (1992). In vitro selection of RNAs that undergo autolytic cleavage with Pb2+. Biochemistry *31*, 3887–3895.

Park, J.E., Heo, I., Tian, Y., Simanshu, D.K., Chang, H., Jee, D., Patel, D.J., and Kim, V.N. (2011). Dicer recognizes the 5′ end of RNA for efficient and accurate processing. Nature *475*, 201–205.

Pitt, J.N., and Ferré-D'Amaré, A.R. (2010). Rapid construction of empirical RNA fitness landscapes. Science *330*, 376–379.

Ruby, J.G., Jan, C.H., and Bartel, D.P. (2007). Intronic microRNA precursors that bypass Drosha processing. Nature *448*, 83–86.

Schaal, T.D., and Maniatis, T. (1999). Selection and characterization of pre-mRNA splicing enhancers: identification of novel SR protein-specific enhancer sequences. Mol. Cell. Biol. *19*, 1705–1719.

Schwarz, D.S., Hutvágner, G., Du, T., Xu, Z., Aronin, N., and Zamore, P.D. (2003). Asymmetry in the assembly of the RNAi enzyme complex. Cell *115*, 199–208.

Slattery, M., Riley, T., Liu, P., Abe, N., Gomez-Alcala, P., Dror, I., Zhou, T., Rohs, R., Honig, B., Bussemaker, H.J., and Mann, R.S. (2011). Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. Cell *147*, 1270–1282.

Sun, G., Yan, J., Noltner, K., Feng, J., Li, H., Sarkis, D.A., Sommer, S.S., and Rossi, J.J. (2009). SNPs in human miRNA genes affect biogenesis and function. RNA *15*, 1640–1651.

Swartz, J.E., Bor, Y.C., Misawa, Y., Rekosh, D., and Hammarskjold, M.L. (2007). The shuttling SR protein 9G8 plays a role in translation of unspliced mRNA containing a constitutive transport element. J. Biol. Chem. *282*, 19844–19853.

Wilson, D.S., and Szostak, J.W. (1999). In vitro selection of functional nucleic acids. Annu. Rev. Biochem. *68*, 611–647.

Wu, H., Sun, S., Tu, K., Gao, Y., Xie, B., Krainer, A.R., and Zhu, J. (2010). A splicing-independent function of SF2/ASF in microRNA processing. Mol. Cell *38*, 67–77.

Wyatt, J.R., Sontheimer, E.J., and Steitz, J.A. (1992). Site-specific cross-linking of mammalian U5 snRNP to the 5′ splice site before the first step of pre-mRNA splicing. Genes Dev. *6*(12B), 2542–2553.

Yeom, K.H., Lee, Y., Han, J., Suh, M.R., and Kim, V.N. (2006). Characterization of DGCR8/Pasha, the essential cofactor for Drosha in primary miRNA processing. Nucleic Acids Res. *34*, 4622–4629.

Zahler, A.M., Neugebauer, K.M., Stolk, J.A., and Roth, M.B. (1993). Human SR proteins and isolation of a cDNA encoding SRp75. Mol. Cell. Biol. *13*, 4023–4028.

Zeng, Y., and Cullen, B.R. (2005). Efficient processing of primary microRNA hairpins by Drosha requires flanking nonstructured RNA sequences. J. Biol. Chem. *280*, 27595–27603.

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

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

Zykovich, A., Korf, I., and Segal, D.J. (2009). Bind-n-Seq: high-throughput analysis of in vitro protein-DNA interactions using massively parallel sequencing. Nucleic Acids Res. *37*, e151.

## EXTENDED EXPERIMENTAL PROCEDURES

### Ectopic Pri-miRNA Expression in HEK293 Cells and S2 Cells

A genomic fragment corresponding to the human *mir-1-1* hairpin and flanking sequences was amplified and cloned into both pcDNA3.2/V5-DEST (Invitrogen) and pMT-DEST (Invitrogen) expression plasmids downstream of the attR recombination sites. Query pri-miRNA sequences were recombined into these plasmids at the attR sites using the Gateway system (Invitrogen). Expression plasmids and pMAX-GFP were co-transfected into HEK293 cells using Lipofectamine 2000 (Invitrogen) and co-transfected into S2 cells using Cellfectin (Invitrogen) according to manufacturer's instructions. After 36–48 hr, total RNA was collected by addition of Tri-Reagent (Ambion) according to manufacturer's instructions. RNA blots for detecting mature and pre-miRNAs were as described. Ribonuclease protection assays were performed with the RPA III kit (Invitrogen) according to manufacturer's instructions.

For detection of expression by sequencing, total RNA from individual transfections was combined and libraries for small-RNA sequencing prepared as described (Chiang et al., 2010). Sequencing reads were mapped to a miRNA hairpin collection composed of the miRBase-annotated hairpins of miRNAs endogenously expressed in the cell line and the miRBase-annotated hairpins of the transfected miRNAs. Reads were included if they perfectly matched a hairpin in this library and excluded if they matched more than one hairpin corresponding to a transfected miRNA. Read counts were normalized to the total reads matching a set of endogenous hairpins that had no transfected counterparts. For each expressed pri-miRNA hairpin, number of reads reported is the number obtained after subtracting the number observed in a normalized, mock-transfected control library.

### Microprocessor Lysate

Microprocessor lysate was prepared as described (Lee and Kim, 2007), with minor modifications. HEK293T cells were transfected with a mixture of pCK-Drosha-FLAG (Lee and Kim, 2007) pFLAG-HA-DGCR8 (Landthaler et al., 2004), and a transfection-control plasmid pMAX-GFP (Amaxa) using Lipofectamine 2000 (Invitrogen) according to the manufacturer's instructions. After 72 hr, cells were harvested by rinsing the monolayer in phosphate buffered saline (PBS, 137 mM NaCl, 2.7 mM KCl, 1.5 mM $KH_2PO_4$, 8 mM $Na_2HP0_4$, [pH 7.4]). Cells were pelleted, resuspended in sonication buffer (100 mM KCl, 0.2 mM EDTA, 20 mM Tris-Cl pH 8.0, and 0.7 μl/ml 2-mercaptoethanol) supplemented with mini-EDTA Free Protease Inhibitor tablets (Roche), and sonicated. After clearing by centrifugation, cell lysis was confirmed by the liberation of GFP into the supernatant. The supernatant was distributed into single-use aliquots, and stored in liquid-nitrogen vapor phase. pri-miRNA cleavage assays were carried out as described (Lee and Kim, 2007) unless otherwise noted.

### Competitive Binding and Cleavage Assays

The competitive binding assay was based on that of Bartel, et al. (Bartel et al., 1991). T7-transcribed ~200 nt pri-miRNA substrates were gel-purified, treated with calf intestinal phosphatase (NEB), extracted in Tri-Reagent (Invitrogen), and 5′ end-labeled using T4 Polynucleotide Kinase (NEB) and γ-[$^{32}$P]-ATP. The *mir-125a* reference substrate was prepared in the same way, except it was 10–25 nt shorter to enable separation on denaturing gels. Complexes containing Drosha-TN and DGCR8 were immunopurified from Microprocessor lysate in which DroshaTN-FLAG replaced the wild-type Drosha plasmid as described (Lee and Kim, 2007; Han et al., 2009). Competitor and reference RNAs were mixed and incubated with Drosha-TN–DGCR8 for 15-30 min [final concentrations, 250 nM each RNA, 100 mM KCl, 1 mM $MgCl_2$, 0.2 mM EDTA, 20 mM Tris-Cl (pH 8.0), 0.7 μl/ml 2-mercaptoethanol and 300 ng/μl yeast total RNA (Ambion)]. RNA-protein complexes were filtered on Immobilon-NC nitrocellulose discs (Millipore) and washed with at least 10 reaction volumes of sonication buffer. RNA was eluted from the membrane by incubating in elution buffer (300 mM NaCl, 8M urea, and 25 mM EDTA) for 10 min at 85°C, ethanol precipitated and resolved on denaturing 5% polyacrylamide gels.

For competitive cleavage, 5′ end-labeled query and reference pri-miRNA substrates were mixed and incubated with Microprocessor lysate [final concentrations, 50 nM each RNA, 100 mM KCl, 1 mM $MgCl_2$, 0.2 mM EDTA, 20 mM Tris-Cl (pH 8.0), 0.7 μl/ml 2-mercaptoethanol, 300 ng/μl yeast total RNA, 10 nM Microprocessor complex (concentration estimated exploiting the single-turnover behavior of the Microprocessor when cleaving linear *pri-miR-125a*)]. After incubation for 30 s at 37°C the reaction was stopped by addition of Tri-Reagent (Ambion) with mixing. Extracted RNA was precipitated with isopropanol, then resuspended and resolved on a denaturing 5% polyacrylamide gel.

### Synthesis and Selection of Pri-miRNA Variants

Templates for T7 RNA polymerase transcription were assembled from oligonucleotides (IDT) that were synthesized using nucleoside phosphoramidite mixtures to introduce variability at specified positions (Table S1). For body labeling, transcription reactions included α-[$^{32}$P]-UTP.

Transcripts for producing circular pri-miRNA variants ended with a minimal HDV ribozyme (Schürer et al., 2002) that co-transcriptionally self-cleaved at a defined position to produce homogenous 3′ ends. After treatment with TurboDNase (Ambion), these transcripts were gel-purified, treated with calf intestinal phosphatase (NEB) to remove the 5′ triphosphate, extracted with Tri-Reagent, precipitated with isopropanol, and treated with T4 polynucleotide kinase (NEB) to remove the 2′-3′ cyclic phosphate as described (Guo et al., 2010). After ethanol precipitation, they were 5′ phosphorylated with T4 polynucleotide kinase, diluted, and circularized using T4 RNA ligase 1 (NEB). Circular pri-miRNAs were purified from linear species on denaturing polyacrylamide gels.

Pools of variants were incubated in Microprocessor lysate, and at one or two time points (for circularized pri-miRNA variants, 1 min for *mir-125a*, 1 and 4 min for *mir-16-1*, 1 and 5 min for *mir-30a*, and 3 and 15 min for mir-223; for apical stem and loop variants, 5 s and 15 s for *mir-125a*, 15 s and 2 min for *mir-16-1*, 30 s and 2 min for *mir-30a*, and 30 s and 2 min for *mir-223*) reactions were stopped by addition of Tri-Reagent (Ambion) with mixing, and cleaved products were purified on denaturing gels. Cleavage products of circularized pri-miRNA variants were ligated to oligonucleotide adaptors containing barcode sequences using T4 DNA ligase (NEB) and DNA splints (Table S1), reverse transcribed, and amplified. To represent the initial pools, a sample of phosphorylated, uncircularized RNA was reverse transcribed and amplified. For the apical stem-loop variants, pre-miRNA cleavage products of linear pri-miRNA variants were reverse transcribed, amplified and sequenced. To represent the initial pools of these variants, a sample from each pool was reverse transcribed and amplified.

## High-Throughput Sequencing and Analysis

Amplicons from the initial pools and the cleaved products were pooled for Illumina paired-end sequencing (75 nt reads per end) for circularized substrate selections, and Illumina single-read sequencing (54 nt reads) for apical stem-loop selections. Sequencing reads were divided into experimental groups according to constant sequences specific to each pri-miRNA and barcodes indicating time points. After filtering for sequencing quality, discarding any sequences that had an error rate $\geq 0.1$ (phred score $\leq 10$) at any variant position, the sequencing error averaged $< 0.001$ per variant position (average phred score $> 30$). Sequences in which the length of a partially randomized region differed from that of the wild-type were also discarded, thereby eliminating many sequences with insertions or deletions. Libraries were collapsed so that sequences that appeared multiple times with the same bar code were considered just once in the analysis (although in retrospect this precaution was not required because there was no group of dominant, multi-copy sequences that would have biased the analyses). Analyses were also restricted to products cleaved at the wild-type processing sites, which were inferred from the dominant reads in small-RNA sequencing data (Landgraf et al., 2007; Bar et al., 2008; Chiang et al., 2010; Witten et al., 2010), except for miR-16-1* and miR-223, which appear to undergo post-cleavage 3′-end trimming (Han et al., 2011).

To calculate the information content at each position, we used the data from the initial sequences and the product sequences to calculate the relative cleavage of each base versus that of the other three bases. For example, for the A residue, the three relative cleavage values are given below, where $P(N)$ is estimated by the frequency of a base in the initial pool, and $P(N|\text{cleavage})$ is estimated by the frequency of that base in the product sequences.

$$\frac{P(\text{cleavage}|C)}{P(\text{cleavage}|A)} = \frac{P(C|\text{cleavage})/P(C)}{P(A|\text{cleavage})/P(A)}$$

$$\frac{P(\text{cleavage}|G)}{P(\text{cleavage}|A)} = \frac{P(G|\text{cleavage})/P(G)}{P(A|\text{cleavage})/P(A)}$$

$$\frac{P(\text{cleavage}|U)}{P(\text{cleavage}|A)} = \frac{P(U|\text{cleavage})/P(U)}{P(A|\text{cleavage})/P(A)}$$

We then used Bayes' Theorem (Pitman, 1993) to infer the nucleotide composition that would have resulted after selection from a pool of variants in which there was an equal probability of an A, C, G, or U at this position. For example, the formula to infer the frequency of A at a particular position after selection from such a pool was

$$P_{\text{inferred A}} = P(A|\text{cleavage}) = \left[1 + \frac{P(\text{cleavage}|C)}{P(\text{cleavage}|A)} + \frac{P(\text{cleavage}|G)}{P(\text{cleavage}|A)} + \frac{P(\text{cleavage}|U)}{P(\text{cleavage}|A)}\right]^{-1}$$

The inferred post-selection distribution was then used to calculate information content scores for each nucleotide at each position. For example, the information content for A at a particular position was calculated as

$$I_A = P_{\text{inferred A}} \times [log_2(P_{\text{inferred A}}) + 2]$$

If results from two time points were available, information content values were averaged.

For evaluating motifs, we calculated a relative cleavage value based on the frequencies of the motif in the reference and selected pools [$P(\text{motif}_i)$ and $P(\text{motif}_i)|\text{cleavage}$), respectively], and the frequencies of a reference motif in the reference and selected pools [$P(\text{motif}_{\text{ref}})$ and $P(\text{motif}_{\text{ref}})|\text{cleavage}$), respectively].

$$\text{Relative cleavage} = \frac{P(\text{motif}_i|\text{cleavage})/P(\text{motif}_i)}{P(\text{motif}_{\text{ref}}|\text{cleavage})/P(\text{motif}_{\text{ref}})}$$

We also used an odds ratio score to calculate the enrichment for particular motifs by using the frequency of the motif in the reference and selected pools [$P(\text{motif}_i)$ and $P(\text{motif}_i)|\text{cleavage}$), respectively].

$$\text{Odds ratio} = \frac{P(\text{motif}_i|\text{cleavage})/P(\text{motif}_i)}{[1 - P(\text{motif}_i|\text{cleavage})]/[1 - P(\text{motif}_i)]}$$

If two time points were available, the geometric mean of the ratios was reported, unless noted otherwise.

To screen specifically for Watson–Crick pairing between all possible combinations of randomized positions, we used a scoring metric to compare the geometric average of odds ratios for Watson–Crick pairing to that of odds ratios for non-Watson–Crick pairs.

$$\text{Pairing score} = \left(\prod_{\text{Watson-Crick}} \text{Odds ratio}\right)^{1/4} - \left(\prod_{\text{non-Watson-Crick}} \text{Odds ratio}\right)^{1/12}$$

## Pri-miRNA Collections and Positional Enrichments of Sequence Motifs

A list of representative pri-miRNAs used for analyses is provided (Table S2). Because of the large number of questionable annotations in miRBase (Chiang et al., 2010), analysis of human pri-miRNAs was restricted to those of miRNAs conserved in mouse. Coordinates of miRNA loci in miRBase version 17 (Kozomara and Griffiths-Jones, 2011) were used to extract the sequences of each annotated hairpin and 200 genomic bases flanking each side. miRBase hairpin sequences and flanking genomic sequences (20 nt on each side) were folded using RNAfold (Hofacker and Stadler, 2006). The Microprocessor cleavage site was inferred using the predicted structures and the mature sequences annotated in miRBase. In cases in which the 3′ overhang was shorter than 2 nt, the 3′ product was extended to generate a 2 nt overhang. Only hairpins for which the predicted folding and the annotated mature sequences could be reconciled or extended to form a 2 nt 3′ overhang were carried forward for analysis. For hairpins in miRBase-annotated miRNA families, a single representative was chosen to represent the family in each species. For human, *D. melanogaster*, and *C. elegans*, the family member with the most conserved pre-miRNA sequence (as determined by average branch-length score of pre-miRNA nucleotides) was chosen. For other species, the representative was chosen at random.

Whole-genome alignments and phylogenetic trees were obtained from the UCSC genome browser (Fujita et al., 2011). Conservation of a base was evaluated by its branch-length score, defined as the ratio between the total branch length of the species that contained the same base as the reference sequence and the total branch length of the species that had an aligned base at that position.

Enrichment of a motif at a set of positions relative to the cleavage site was computed by generating 100,000 cohorts of pri-miRNAs in which the upstream, downstream, and pre-miRNA sequences were independently shuffled, preserving dinucleotide frequencies. The numbers of miRNAs that contained a match to the motif in the actual and shuffled cohorts were used to compute an empirical p value.

## Analysis of Crosslinked Complexes

The *mir-30a* pri-miRNA crosslinking substrate was assembled using T4 RNA ligase 2 (NEB) and a DNA splint to join an in vitro-transcribed 5′ fragment to a synthetic 3′ fragment containing a 3′-terminal biotin and a 4-thiouridine within the CNNC motif (Dharmacon). This crosslinking substrate was incubated in Microprocessor lysate and exposed to 1000 mJ of 365 nm UV light in a Stratalinker (Stratagene). For purification of RNA–protein complexes for mass spectrometry, complexes were captured on streptavidin-coated magnetic beads (Invitrogen) and washed twice in Laemmli buffer (4% SDS, 20% glycerol, 125 mM Tris-Cl pH 6.8) and twice in urea buffer (8 M urea, 300 mM NaCl, 25 mM EDTA), then eluted with RNase T1 (Ambion). The eluted complexes were separated on SDS gels, and the corresponding gel slices excised. The complexes were reduced, alkylated and digested with trypsin. After extraction and concentration, peptides were analyzed by HPLC/tandem mass spectrometry using a Waters NanoAcquity UPLC system and a ThermoFisher LTQ linear ion trap mass spectrometer operated in a data-dependent manner. Peptides were identified using SEQUEST and data analyzed with Scaffold (Proteome Software).

For immunoprecipitation, eluted RNA–protein complexes were incubated for 1 hr with antibody [either anti-FLAG M2 (Sigma), polyclonal mouse IgG (Millipore), anti-SRp20 (Invitrogen), or anti-9G8 (gift of J. Stévenin)], followed by incubation with protein-G agarose beads (Sigma). After washing the beads three times in at least ten packed-bead volumes of sonication buffer, complexes were separated on SDS gels.

## Reanalysis of iCLIP Data

Sequencing reads from iCLIP of SRp20 (SRSF3) and SRp75 (SRSF4) were from ArrayExpress (accession ERP000815) (Änkö et al., 2012). Although that study did not find enrichment for SRp20-binding sites in miRNAs, the miRNA annotations examined did not extend beyond the miRNA hairpins (i.e., the pre-miRNAs and their basal stems) and thus did not include downstream regions containing the CNNC motif. After adaptor stripping, reads were mapped to the mouse genome using Bowtie (Langmead et al., 2009), allowing for two mismatches and considering only uniquely mapped reads. Each position immediately 5′ to an iCLIP read was considered a crosslink site, and the number of crosslink sites was tallied for each relative distance from the mouse pre-miRNAs confirmed or identified in Chiang et al. (2010). For pri-miRNAs with more than one site, the site supported by the most reads was the one plotted in Figure 6D (distributing fractions of a count to each site in cases in which multiple sites were tied for the most reads).

### Cleavage Assays with Purified SRp20

SRp20 cDNA with an N-terminal 3X-FLAG tag was cloned into the pcDNA3.2/V5-DEST vector (Invitrogen). HEK293T cells were transiently transfected with either the SRp20 construct or an analogous construct expressing N-terminally FLAG-tagged EGFP, using Lipofectamine 2000 (Invitrogen) according to manufacturer's instructions. After 48 hr cells were lysed in sonication buffer. The tagged proteins were immunoprecipitated using ANTI-FLAG M2 magnetic beads (Sigma), washed three times in sonication buffer, and eluted with 150 ng/ul 3X-FLAG peptide (Sigma), then dialyzed against 1000 volumes of sonication buffer using dialysis membrane with a 3 kDa cutoff (Pierce). For cleavage reactions, the Microprocessor complex was first immunoprecipatated from Microprocessor lysate. Biotinylated anti-FLAG-M2 antibody (1:500 dilution, Sigma) was incubated in lysate for 2.5 hr, then precipitated with streptavidin-coated magnetic beads (Invitrogen) for 30 min. Beads were washed twice in sonication buffer, then incubated at 37°C with 5′-labeled *pri-mir-16-1* substrates and either SRp20 or EGFP immunopurified from HEK293T cells, at a final volume of 40% beads and 40% either SRp20, EGFP or sonication buffer. Final concentrations were 2 nM pri-miRNA, 100 mM KCl, 1 mM MgCl2, 0.2 mM EDTA, 20 mM Tris-Cl (pH 8.0), 0.7 μl/ml 2-mercaptoethanol, 300 ng/μl yeast total RNA (Ambion) and 0.5% SUPERaseIn RNase Inhibitor (Ambion). After 3 min, reactions were stopped in Tri-reagent (Ambion), and products separated on 10% denaturing polyacrylamide gels.

### SUPPLEMENTAL REFERENCES


Änkö, M.L., Müller-McNicoll, M., Brandl, H., Curk, T., Gorup, C., Henry, I., Ule, J., and Neugebauer, K.M. (2012). The RNA-binding landscapes of two SR proteins reveal unique functions and binding to diverse RNA classes. Genome Biol. *13*, R17.

Bar, M., Wyman, S.K., Fritz, B.R., Qi, J., Garg, K.S., Parkin, R.K., Kroh, E.M., Bendoraite, A., Mitchell, P.S., Nelson, A.M., et al. (2008). MicroRNA discovery and profiling in human embryonic stem cells by deep sequencing of small RNA libraries. Stem Cells *26*, 2496–2505.

Bartel, D.P., Zapp, M.L., Green, M.R., and Szostak, J.W. (1991). HIV-1 Rev regulation involves recognition of non-Watson-Crick base pairs in viral RNA. Cell *67*, 529–536.

Chiang, H.R., Schoenfeld, L.W., Ruby, J.G., Auyeung, V.C., Spies, N., Baek, D., Johnston, W.K., Russ, C., Luo, S., Babiarz, J.E., et al. (2010). Mammalian microRNAs: experimental evaluation of novel and previously annotated genes. Genes Dev. *24*, 992–1009.

Fujita, P.A., Rhead, B., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Cline, M.S., Goldman, M., Barber, G.P., Clawson, H., Coelho, A., et al. (2011). The UCSC Genome Browser database: update 2011. Nucleic Acids Res. *39*(Database issue), D876–D882.

Guo, H., Ingolia, N.T., Weissman, J.S., and Bartel, D.P. (2010). Mammalian microRNAs predominantly act to decrease target mRNA levels. Nature *466*, 835–840.

Han, J., Pedersen, J.S., Kwon, S.C., Belair, C.D., Kim, Y.K., Yeom, K.H., Yang, W.Y., Haussler, D., Blelloch, R., and Kim, V.N. (2009). Posttranscriptional crossregulation between Drosha and DGCR8. Cell *136*, 75–84.

Han, B.W., Hung, J.H., Weng, Z., Zamore, P.D., and Ameres, S.L. (2011). The 3′-to-5′ exoribonuclease Nibbler shapes the 3′ ends of microRNAs bound to *Drosophila* Argonaute1. Curr Biol *21*, 1878–1887.

Hofacker, I.L., and Stadler, P.F. (2006). Memory efficient folding algorithms for circular RNA secondary structures. Bioinformatics *22*, 1172–1176.

Kozomara, A., and Griffiths-Jones, S. (2011). miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res. *39*, D152–D157.

Landgraf, P., Rusu, M., Sheridan, R., Sewer, A., Iovino, N., Aravin, A., Pfeffer, S., Rice, A., Kamphorst, A.O., Landthaler, M., et al. (2007). A mammalian microRNA expression atlas based on small RNA library sequencing. Cell *129*, 1401–1414.

Landthaler, M., Yalcin, A., and 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.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

Lee, Y., and Kim, V.N. (2007). In vitro and in vivo assays for the activity of Drosha complex. Methods Enzymol. *427*, 89–106.

Moore, M.J. (1999). Joining RNA molecules with T4 DNA ligase. Methods Mol. Biol. *118*, 11–19.

Pitman, J. (1993). Probability (New York: Springer-Verlag).

Schürer, H., Lang, K., Schuster, J., and Mörl, M. (2002). A universal method to produce in vitro transcripts with homogeneous 3′ ends. Nucleic Acids Res. *30*, e56.

Sontheimer, E.J. (1994). Site-specific RNA crosslinking with 4-thiouridine. Mol. Biol. Rep. *20*, 35–44.

Witten, D., Tibshirani, R., Gu, S.G., Fire, A., and Lui, W.O. (2010). Ultra-high throughput sequencing-based small RNA discovery and discrete statistical biomarker analysis in a collection of cervical tumours and matched controls. BMC Biol. *8*, 58.

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

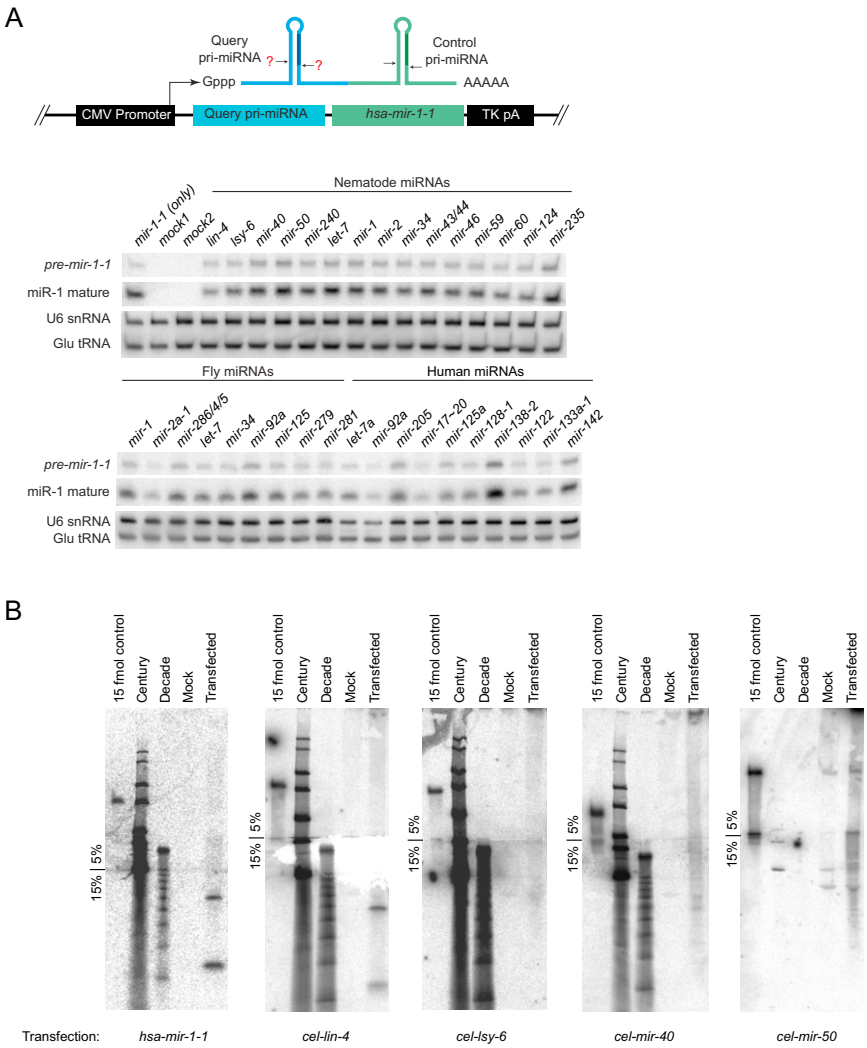


**Figure S1. Expression and Processing of Pri-miRNA Hairpins in HEK293 Cells, Related to Figure 1**

(A) Expression of miR-1 and *pre-mir-1-1* from bicistronic transcripts. HEK293 cells were individually transfected with plasmids bearing a human, *D. melanogaster*, or *C. elegans* pri-miRNAs transcriptionally fused to human *pri-mir-1-1*. Mature miR-1 and *pre-mir-1-1* derived from the transcriptional fusion were detected by RNA blot. (Results from vectors in which *let-7* and *mir-1* were the query pri-miRNAs are shown here but are not shown in Figure 1A because the corresponding mature miRNAs were indistinguishable from those of other transfected cells after total RNA was pooled for small-RNA sequencing.)

(B) Full membrane images for blots shown in Figure 1B. Total RNA was run on stacked polyacrylamide gels (5% top and 15% bottom) to resolve sizes from 20–1000 nt. Each blot included marker lanes (Century and Decade RNA markers, Ambion) and a positive-control lane with 15 fmol in vitro transcribed standard derived from the corresponding pri-miRNA (control).

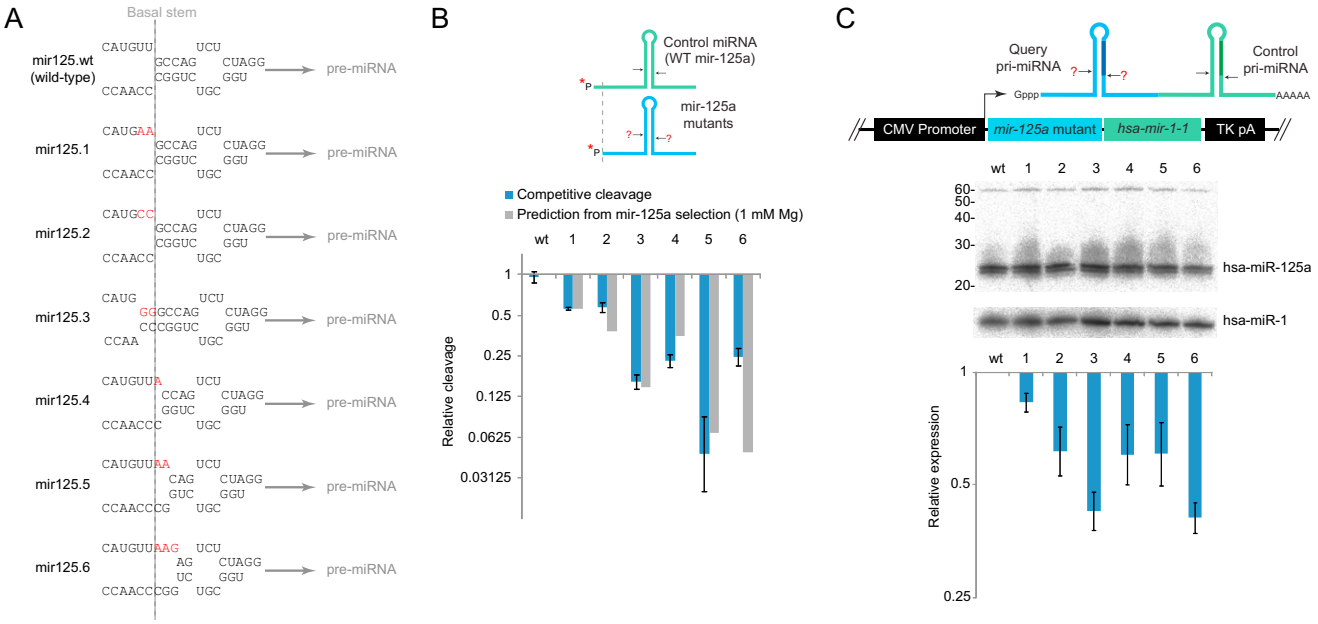


**Figure S2. Confirmation of *hsa-mir-125a* Selection Results In Vitro and in HEK293T Cells, Related to Figure 2**

(A) Predicted basal stem structure of *mir-125a* variants tested in the experiment.

(B) Competitive cleavage of individual *mir-125a* variants, relative to wild-type *mir-125a*. Variants were mixed with wild-type *mir-125a*, which was longer at its 5′ end, and incubated in Microprocessor lysate. Cleavage products were separated on denaturing gels, and the ratio of wild-type and variant products quantified (blue, geometric mean ± standard error, n = 3), together with the relative cleavage inferred from the selection experiment (gray).

(C) Evaluation of *mir-125a* variants in HEK293T cells. Variants were transcriptionally fused to *pri-mir-1-1* and expressed in HEK293T cells, as in Figure S1A. Accumulation of mature miR-125a was quantified by RNA blot and normalized to the level of mature miR-1 (geometric mean ± standard error, n = 3).

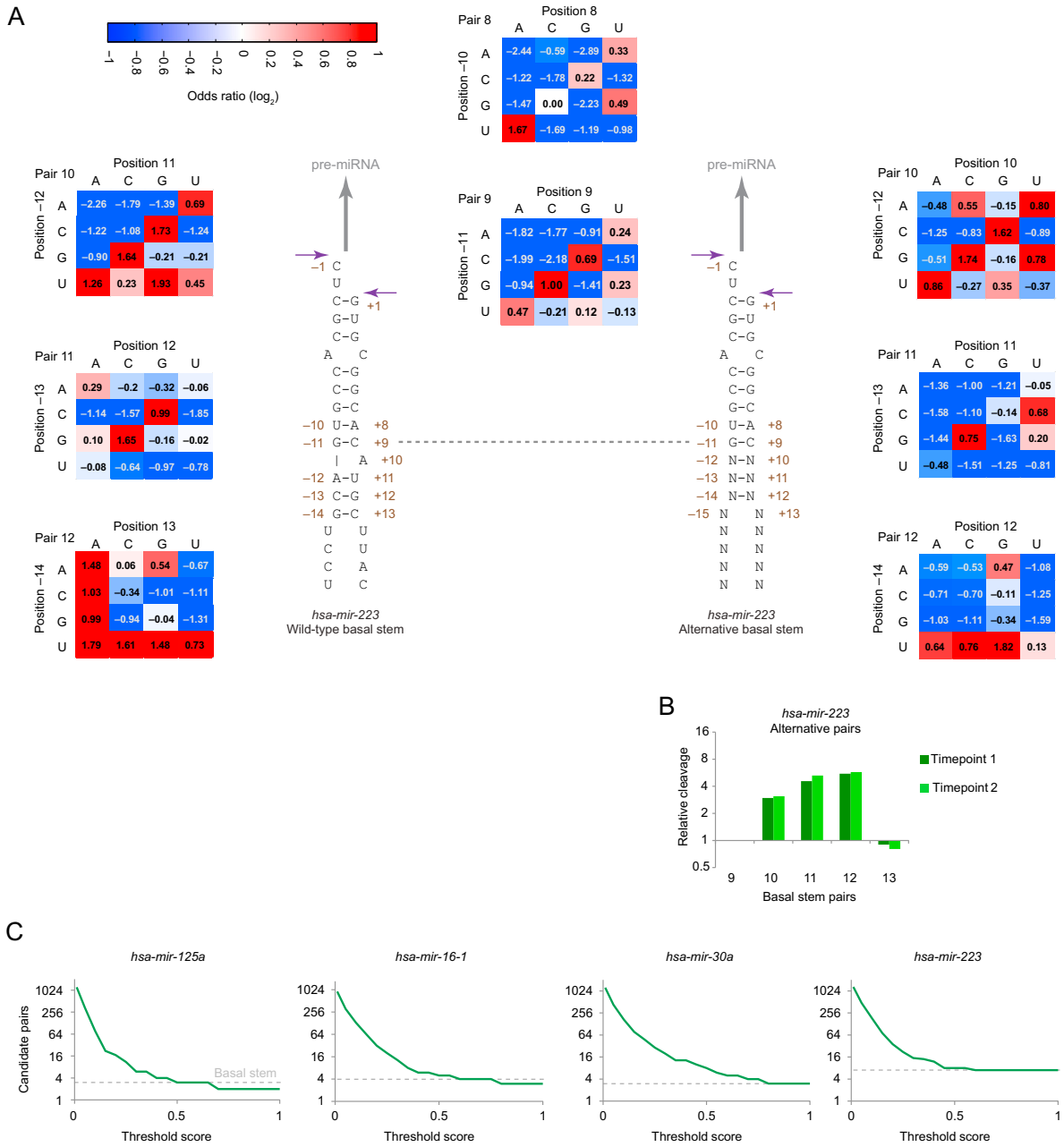


**Figure S3. Analysis of *mir-223* Basal Stem Structure, Related to Figure 3**

(A) Wild-type (left) and alternative (right) basal stem structures for *hsa-mir-223*. In the predicted structure of the wild-type the A at +10 is bulged, whereas in the predicted structure of some of the variants the pairing shifts to place nucleotide +10 within a contiguous helix. After sorting the selected variants based on whether or not their predicted secondary structures are consistent with shifted pairing, covariation matrices for both conformations were calculated as in Figure 3A.

(B) Relative cleavage of variants with different lengths of the alternative basal stem. Cleavage values were calculated as in Figure 3B and normalized to the 9 bp stem.

(C) Screen for Watson–Crick pairs involving any two varied positions. For each of the > 3000 possible pairs, the degree of Watson–Crick preference was evaluated using a scoring metric that compared the average odds of Watson–Crick pairs to that of non-Watson–Crick alternatives. The number of Watson–Crick candidates is plotted as a function of threshold score, in which a pair is considered a Watson–Crick candidate if its score exceeds the threshold. The number of pairs corresponding to the basal stem is shown (dashed line). In each case, the highest-scoring pairs were those of the basal stem. In the case of *mir-223*, the highest scoring pairs also included the alternative pairs that incorporated the bulged A at +10 into a contiguous helix. For each pri-miRNA, we inspected the next four highest-scoring pairs, and in each case, the covariation matrix did not appear consistent with Watson–Crick pairing (data not shown).

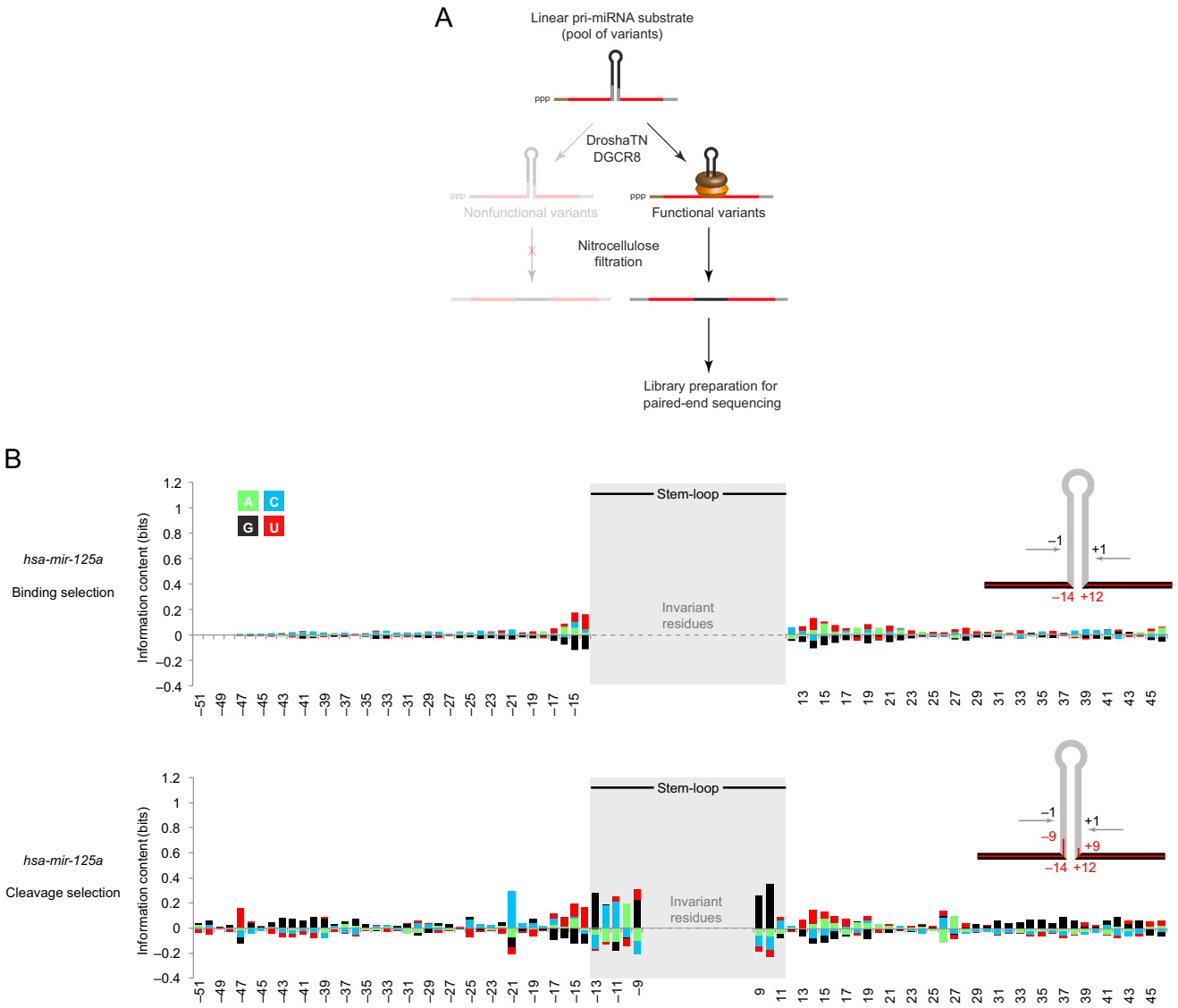


**Figure S4. Selection for Microprocessor-Binding Variants of *hsa-mir-125a*, Related to Figure 4**

(A) Schematic of the in vitro selection. Linear variants of *mir-125a* were incubated with immunopurified DGCR8 and catalytically inactive Drosha (DroshaTN). Bound variants were recovered after nitrocellulose filtration, reverse-transcribed, and amplified for high-throughput sequencing.

(B) Information content after selection for Microprocessor binding. Information content after selection for cleavage (Figure 2D) is reproduced here for comparison. The nucleotides varied in the initial pools are shown for each selection (insets, red inner lines).

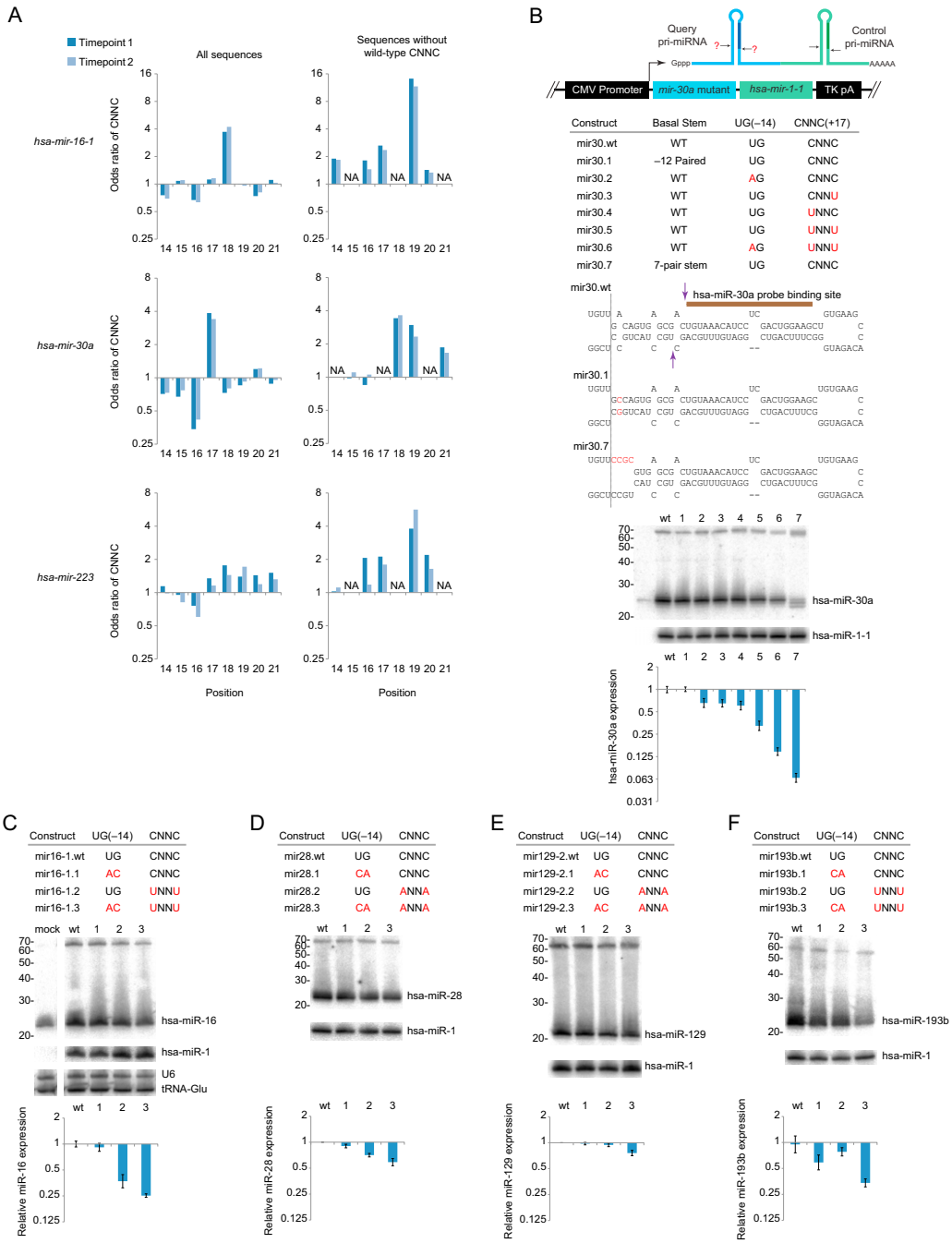


**Figure S5. Contribution of the CNNC Motif In Vitro and in HEK293T Cells, Related to Figure 5**

(A) CNNC odds ratios at alternative positions. Odds ratios were calculated for CNNC dinucleotides starting at the indicated positions downstream of the Drosha cleavage site. Plotted are odds ratios for all sequences (left panels) and for sequences that lack both wild-type C residues (right panels).

(B) Contributions of the basal UG and downstream CNNC motifs to the accumulation of hsa-miR-30a in HEK293T cells. The listed variants of *hsa-mir-30a* were transcriptionally fused to *hsa-mir-1-1* (top). Predicted secondary structures for variants with non-wild-type structure are shown (center), with the annotated Drosha cleavage sites (purple arrowheads). The accumulation of miR-30a was quantified by RNA blot, normalized to miR-1 (bottom, geometric mean ± standard error, n = 3).

(C) Contributions of the basal UG and downstream CNNC motifs to the accumulation of hsa-miR-16 in HEK293T cells, otherwise as in (B).

(D) Contributions of the basal UG and downstream CNNC motifs to the accumulation of hsa-miR-28 in HEK293T cells, otherwise as in (B).

(E) Contributions of the basal UG and downstream CNNC motifs to the accumulation of hsa-miR-129 in HEK293T cells, otherwise as in (B).

(F) Contributions of the basal UG and downstream CNNC motifs to the accumulation of hsa-miR-193b in HEK293T cells, otherwise as in (B).

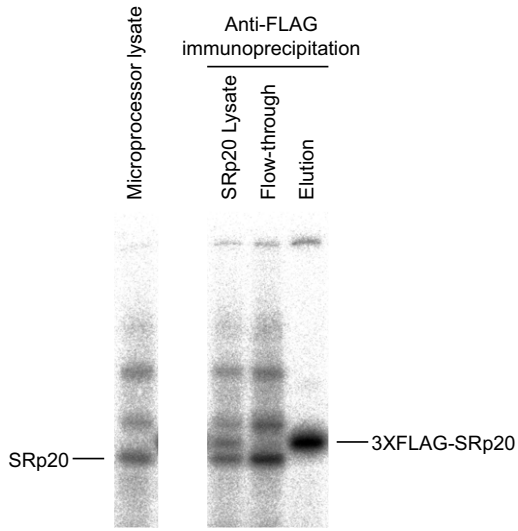


**Figure S6. Immunopurified SRp20, Related to Figure 6**

3X-FLAG-SRp20 was expressed in HEK293T cells, captured on anti-FLAG magnetic beads, and eluted with 3X-FLAG peptide. Binding activity of immunopurified SRp20 was measured by crosslinking to a *mir-30a* crosslinking substrate as in Figure 6A, except that the substrate contained only the *mir-30a* CNNC motif (pCU(4-S-U)CAAGGG). For comparison, crosslinked complexes were generated from Microprocessor lysate (left).

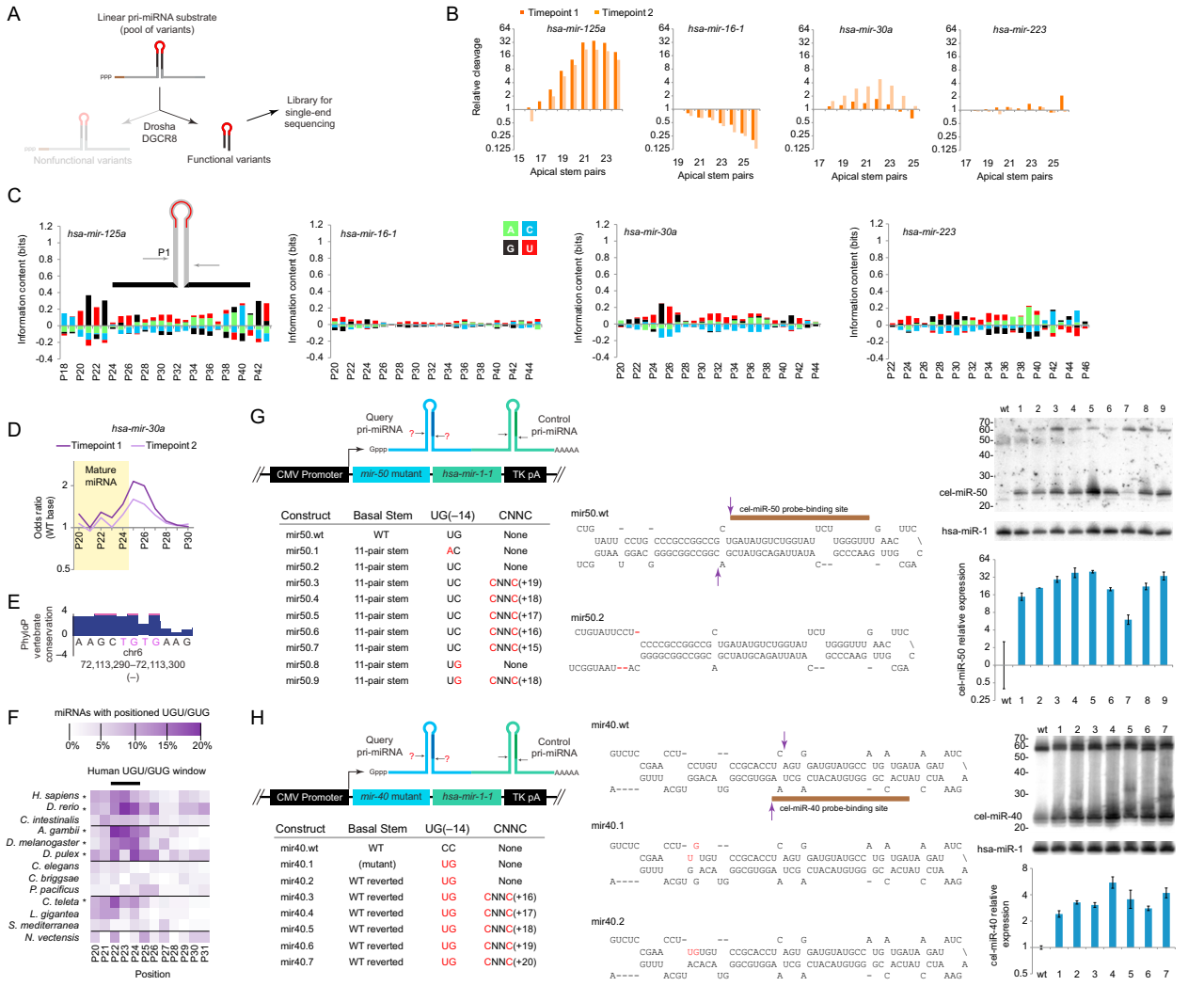


**Figure S7. Selection of Functional Variants with Changes in the Apical Stem Loop and Rescued Processing of *C. elegans* Pri-miRNAs in Human Cells, Related to Figure 7**

(A) Schematic of the selection for functional pri-miRNA variants with changes in the apical stem and terminal loop. Linear pri-miRNA variants were incubated in Microprocessor lysate, and cleaved pre-miRNA variants were gel-purified, reverse transcribed, and amplified for high-throughput sequencing.

(B) Relative cleavage of variants with different apical stem lengths. The number of contiguous Watson–Crick pairs was counted and the relative cleavage calculated, normalized to that of the 15 bp stem. For each pri-miRNA, results are shown for both time points (key). For *mir-125a*, 22 bp above the 5p Drosha cleavage site was strongly preferred; longer stems were tolerated, whereas shorter stems were disfavored. Watson–Crick pairing throughout the apical stem was supported by analysis of covariation (data not shown). A 22-pair apical stem was also preferred, albeit more weakly, in *mir-30a*. By contrast, no preference for apical pairing was observed in the stems of *mir-16-1* or *mir-223*. Indeed, lengthening of the *mir-16-1* apical stem at the expense of loop size was detrimental, which was consistent with a previous report (Zhang and Zeng, 2010).

(C) Enrichment and depletion at variable residues in the apical stems and loops. At each varied position (inset, red inner line), information content was calculated for each residue (green, cyan, black, and red for A, C, G, and U, respectively), as in Figure 2D.

(D) Relative cleavage of *mir-30a* variants with the apical UGUG motif beginning at the indicated positions, normalized to variants without the motif. Nucleotides of the mature miRNA are shaded in yellow.

(E) Conservation of the region centered on the apical UGUG of *mir-30a*, otherwise as in Figure 4B.

(F) Enrichment for UGU or GUG trinucleotides in the terminal loops of metazoan pri-miRNAs (Table S2). For each species, pri-miRNA sequences were aligned on the predicted Drosha cleavage site and occurrences of loop UGU or GUG trinucleotides tabulated. Species with a statistically significant enrichment within the window are indicated (asterisk, empirical p value < $10^{-4}$). Although enrichment was observed in fish and insects, the lack of enrichment in several other representative species raises the question of whether the usage of this motif arose independently in multiple lineages or was ancestral and lost multiple times.

(G) Effects of adding human pri-miRNA features to *C. elegans mir-50*. Changes that introduced the listed features were incorporated into *mir-50* within the bicistronic expression vector (left). Secondary structures are shown for changes that were predicted to affect the wild-type basal stem (middle; annotated Drosha cleavage sites, purple arrowheads). After transfection into HEK293T cells, accumulation of miR-50 was assessed on RNA blots, normalizing to the accumulation of the miR-1 control, and increased miR-50 expression is plotted (right; geometric mean ± standard error, n = 3).

(H) Effects of adding human pri-miRNA features to *C. elegans mir-40*, otherwise as in (G).