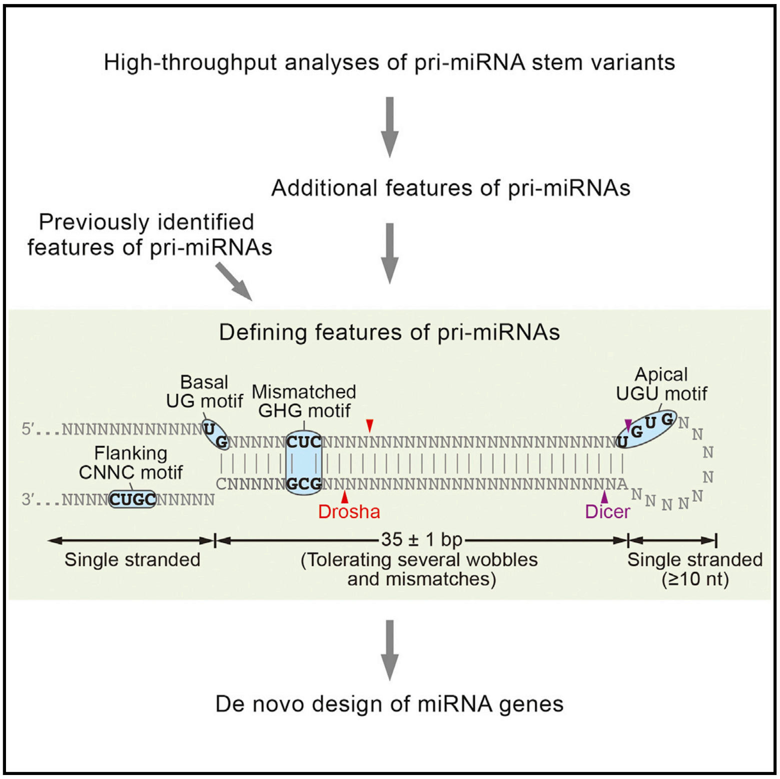
Molecular Cell

The Menu of Features that Define Primary MicroRNAs and Enable De Novo Design of MicroRNA Genes

Graphical Abstract


High-throughput analyses of pri-miRNA stem variants
Additional features of pri-miRNAs
Previously identified features of pri-miRNAs
Defining features of pri-miRNAs
Basal UG motif
Mismatched GHG motif
Apical UGU motif
Flanking CNNC motif
5′...NNNNNNNNNNNNUGNNNNNCUCNNNNNNNNNNNNNNNNNNNNNNNNNUGUGNNNN
3′...NNNNCUGCNNNNN CNNNNNGCGNNNNNNNNNNNNNNNNNNNNNNNNNNANNNN
Drosha
Dicer
Single stranded
35 ± 1 bp
(Tolerating several wobbles and mismatches)
Single stranded (≥10 nt)
De novo design of miRNA genes

Highlights

- High-throughput analyses of pri-miRNAs reveal determinants of efficient processing
- Stem pairing and a length of 35 base pairs are key structural features
- A mismatched motif and primary-sequence motifs compensate for structural defects
- These defining features of pri-miRNAs enable reliable de novo design of miRNA genes

Authors

Wenwen Fang, David P. Bartel

Correspondence

dbartel@wi.mit.edu

In Brief

MicroRNAs are processed from the stem-loop regions of primary microRNAs (pri-miRNAs). Using high-throughput analyses of pri-miRNA variants, Fang and Bartel identified features of pri-miRNA stems that promote processing. Analyses of these and previously identified features provided a unifying model defining pri-miRNAs, which enabled de novo design of functional miRNA genes.

Accession Numbers

GSE67937

Fang & Bartel, 2015, Molecular Cell *60*, 131–145
October 1, 2015 ©2015 Elsevier Inc.
http://dx.doi.org/10.1016/j.molcel.2015.08.015

CellPress

The Menu of Features that Define Primary MicroRNAs and Enable De Novo Design of MicroRNA Genes

Wenwen Fang[1,2,3] and David P. Bartel[1,2,3,*]
[1]Howard Hughes Medical Institute, Cambridge, MA 02142, USA
[2]Whitehead Institute for Biomedical Research, Cambridge, MA 02142, USA
[3]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
*Correspondence: dbartel@wi.mit.edu
http://dx.doi.org/10.1016/j.molcel.2015.08.015

SUMMARY

MicroRNAs (miRNAs) are small regulatory RNAs processed from stem-loop regions of primary transcripts (pri-miRNAs), with the choice of stem loops for initial processing largely determining what becomes a miRNA. To identify sequence and structural features influencing this choice, we determined cleavage efficiencies of >50,000 variants of three human pri-miRNAs, focusing on the regions intractable to previous high-throughput analyses. Our analyses revealed a mismatched motif in the basal stem region, a preference for maintaining or improving base pairing throughout the remainder of the stem, and a narrow stem-length preference of 35 ± 1 base pairs. Incorporating these features with previously identified features, including three primary-sequence motifs, yielded a unifying model defining mammalian pri-miRNAs in which motifs help orient processing and increase efficiency, with the presence of more motifs compensating for structural defects. This model enables generation of artificial pri-miRNAs, designed de novo, without reference to any natural sequence yet processed more efficiently than natural pri-miRNAs.

INTRODUCTION

MicroRNAs (miRNAs) are ~22-nucleotide (nt) RNAs that pair to sites within mRNAs to target these transcripts for post-transcriptional repression (Bartel, 2009). In the canonical biogenesis pathway, miRNA genes are transcribed as primary microRNAs (pri-miRNAs), which contain at least one region that folds back on itself to form a hairpin that is cleaved by the Microprocessor complex, a heterotrimeric complex consisting of one molecule of the Drosha endonuclease and two molecules of its co-factor, DGCR8 (Lee et al., 2003; Denli et al., 2004; Gregory et al., 2004; Han et al., 2004; Nguyen et al., 2015). Drosha-catalyzed cleavage releases the pre-miRNA hairpin, which is exported to the cytoplasm and cleaved by Dicer (Grishok et al., 2001; Hutvágner et al., 2001) to produce a ~20-base pair (bp) RNA duplex with 2-nt 3′ overhangs on each end (Lee et al., 2003; Lim et al., 2003b). One strand of the RNA duplex is ultimately loaded into the Argonaute protein, forming the core of the silencing complex (Hutvágner and Zamore, 2002; Mourelatos et al., 2002; Liu et al., 2004; Song et al., 2004).

When considering the broad diversity of miRNA hairpins, the question arises as to what features Microprocessor and its associated proteins recognize to discriminate the pri-miRNAs from the thousands of other hairpins encoded in the human genome (Lim et al., 2003a; Bentwich et al., 2005). Mutations of mammalian pri-miRNAs have shown the importance of an unstructured apical loop of ≥10 nt (Zeng et al., 2005), pairing at the base of the pri-miRNA hairpin (Lee et al., 2003; Zeng and Cullen, 2003), which optimally extends 11 bp beyond the pre-miRNA hairpin (Han et al., 2006), and unpaired segments flanking the basal stem (Zeng and Cullen, 2005; Han et al., 2006). However, these features do not in themselves impart much discrimination, because many non-pri-miRNA hairpins also have loops of ≥10 nt and basal pairing flanked by unpaired segments.

The distance from both ends of the stem influences the site of cleavage (Ma et al., 2013), which implies that Microprocessor could prefer stems with lengths falling within a specified window. The mean stem lengths of human pri-miRNAs are reported to be in the range of 33–35 bp (Han et al., 2006), although considerable heterogeneity in predicted stem lengths is observed. In addition, simultaneous analysis of millions of functional pri-miRNA variants shows that primary-sequence features can contribute to pri-miRNA recognition (Auyeung et al., 2013). These features include a 5′-UG-3′ motif (hereafter called the UG motif) at the base of the pri-miRNA hairpin, a 5′-UGU-3′/5′-GUG-3′ motif in the apical loop (the UGU motif), and a 5′-CNNC-3′ motif (the CNNC motif) downstream of the hairpin, positioned 16–18 nt from the Drosha cut (Auyeung et al., 2013). Because of technical challenges, however, the stem region of pri-miRNA hairpins has not been examined using high-throughput approaches, leaving open the possibility that sequence or structural features (such as optimally placed bulges, wobbles, or mismatches) within this region might provide the elusive determinants needed to explain the specificity of pri-miRNA recognition in vivo. Indeed, the three known motifs seem to exert their effects in some pri-miRNAs but not others, leading to a model in which the known determinants somehow interact with additional unknown features to yield an idiosyncratic outcome.

With such a complex and incomplete model and little indication of how many features remain undiscovered, it is perhaps not surprising that the successful de novo design of an artificial miRNA gene has not been reported. In practice, this lack of knowledge is circumvented when building Drosha-dependent short-hairpin RNAs (shRNAs) to be used for gene knockdown experiments by relying on the modification of natural pri-miRNAs, typically pri-miR-30 (Zeng et al., 2002; Silva et al., 2005; Bassik et al., 2013; Fellmann et al., 2013; Knott et al., 2014; Kampmann et al., 2015). The design of such reagents de novo, without reference to a known pri-miRNA sequence, would require a more complete understanding of what a miRNA gene is. Because the requirements for Dicer processing are well understood (Zhang et al., 2004; MacRae et al., 2006; Park et al., 2011), the remaining challenge is to understand what Microprocessor looks for as it decides which transcripts are pri-miRNAs and which are not.

Here, we develop a high-throughput strategy that uses molecular barcodes to query the region of the pri-miRNA stem that had been intractable to previous high-throughput analyses. This strategy revealed a mismatched motif near the base of the hairpin, which enhances pri-miRNA processing; a preference for pairing throughout the remainder of the stem; and a narrow preference for stem length. Integrating these features with those that had been previously identified, we had all that was needed for the reliable de novo design of functional pri-miRNAs. Indeed, the processing of these artificial pri-miRNAs was more efficient than that of any natural pri-miRNAs assayed, including the one used as a scaffold for building shRNAs. Additional experiments revealed the ability of the mismatched motif and previously described primary-sequence motifs to compensate for structural defects in the hairpin, as well as their partial redundancies with each other. These insights resolved many of the complexities and seeming discrepancies of earlier studies, leading to a simplified and unifying model of what it takes to be a miRNA gene.

RESULTS

Barcoding Strategy for the Analysis of pri-miRNA Variants

To interrogate the stem region of pri-miRNAs, we randomized blocks of nucleotides within human pri-miR-125a, pri-miR-16-1, and pri-miR-30a (hereafter called pri-miR-125, pri-miR-16, and pri-miR-30, respectively) whose apical and flanking regions had already been thoroughly studied (Auyeung et al., 2013). Nucleotide identities within each of the 3-bp sliding windows across the stem were randomized, resulting in 4,096 variants for each window (Figure 1A). All three pri-miRNAs contained a bulge in the stem, which was also varied with respect to its sequence and length to generate all possible bulge sequences of all lengths spanning the length of the bulge (1 or 2 nt) to no bulge (0 nt) (Table S1). For each window (and each bulge length), a pool of DNA templates was synthesized by combinatorial synthesis, and then all template pools for each pri-miRNA were combined and transcribed into RNA, with the goal of achieving near-equal representation of each variant. In this way, ~50,000 variants were generated for pri-miR-125 and ~80,000 variants were generated for pri-miR-16 and pri-miR-30, with overlap of the sliding windows generating greater diversity for the latter two pools (Figure 1A).

In previous strategies for identifying variants cleaved by Microprocessor, the mutagenized positions all resided in one cleavage product (either all in flanking regions joined through circular permutation or all in the distal loop region), which enabled the original variants to be identified by simply sequencing of the relevant cleavage products (Auyeung et al., 2013). However, a different strategy was required for our variants because the mutagenized region spanned the cleavage sites, and thus for many variants, processing separated mutagenized residues from each other. Therefore, we devised a strategy in which each variant was linked to ~100 unique barcodes residing in the 5′ flanking region, which enabled the identity of a cleaved molecule to be inferred from the barcode sequence of its cleaved product.

The barcodes had 29–31 nt of random sequence and were appended to the hairpin regions in a primer-extension reaction that also added the T7 promoter to the pool of templates (Figure 1B). For each pool, a small fraction of extended product, containing ~10 million template molecules, was amplified, and one portion of the amplified material was sequenced to create a “dictionary” of ~10 million different barcode–variant linkages while another portion was transcribed to generate the pool of RNA variants to be used in the experiment. The bottleneck of 10 million molecules was imposed to reduce the barcode complexity so that most of the transcribed barcode sequences would also be in the dictionary. At the same time, the bottleneck was designed to be sufficiently large such that each of the hairpin sequences would be appended to multiple barcodes. The dictionaries each had a median of ~120 barcodes per hairpin variant (Figures 1C and S1A), with >20 barcodes observed for ≥99.92% of the hairpin variants. This large diversity of barcodes for nearly all hairpin variants minimized concern that an influence of certain barcode sequences on pri-miRNA processing might be misattributed to the hairpin sequence.

After brief incubation with a cell lysate from HEK293T cells overexpressing Microprocessor (Lee and Kim, 2007), <10% of each pri-miRNA pool was processed (Figure S1B) and the 5′ cleavage products were isolated, reverse-transcribed, and sequenced (Figure 1B). The sequences of these products revealed the precise site of cleavage and, for those with barcodes present in the dictionary (70.4%–82.2% of the sequenced cleavage fragments), the identity of the pri-miRNA variant. Barcodes from each input pool were also reverse-transcribed and sequenced, which provided the quantification of each variant in the input and the ability to normalize for differences in the input (Figures 1B and 1D). Comparing for each variant the number of input reads with the number of cleavage-product reads, considering only those indicating cleavage at the proper site (which comprised a large majority of reads for each pool; Figure 1E), provided a measurement of its cleavage efficiency. These measurements were normalized to that of the wild-type pri-miRNA to generate a cleavage score, and these scores were reported on a $\log_2$ scale, such that variants with cleavage efficiencies better or worse than their wild-type counterparts had positive or negative scores, respectively (Figure 1F). Scores for pri-miR-30 were determined using two time points, in which 1.6% and 9% of the pool was cleaved (Figure S1B). As expected, these scores were highly correlated (Figure S1C, Pearson’s $r = 0.94$), showing the robustness of the approach over the range of cleavage

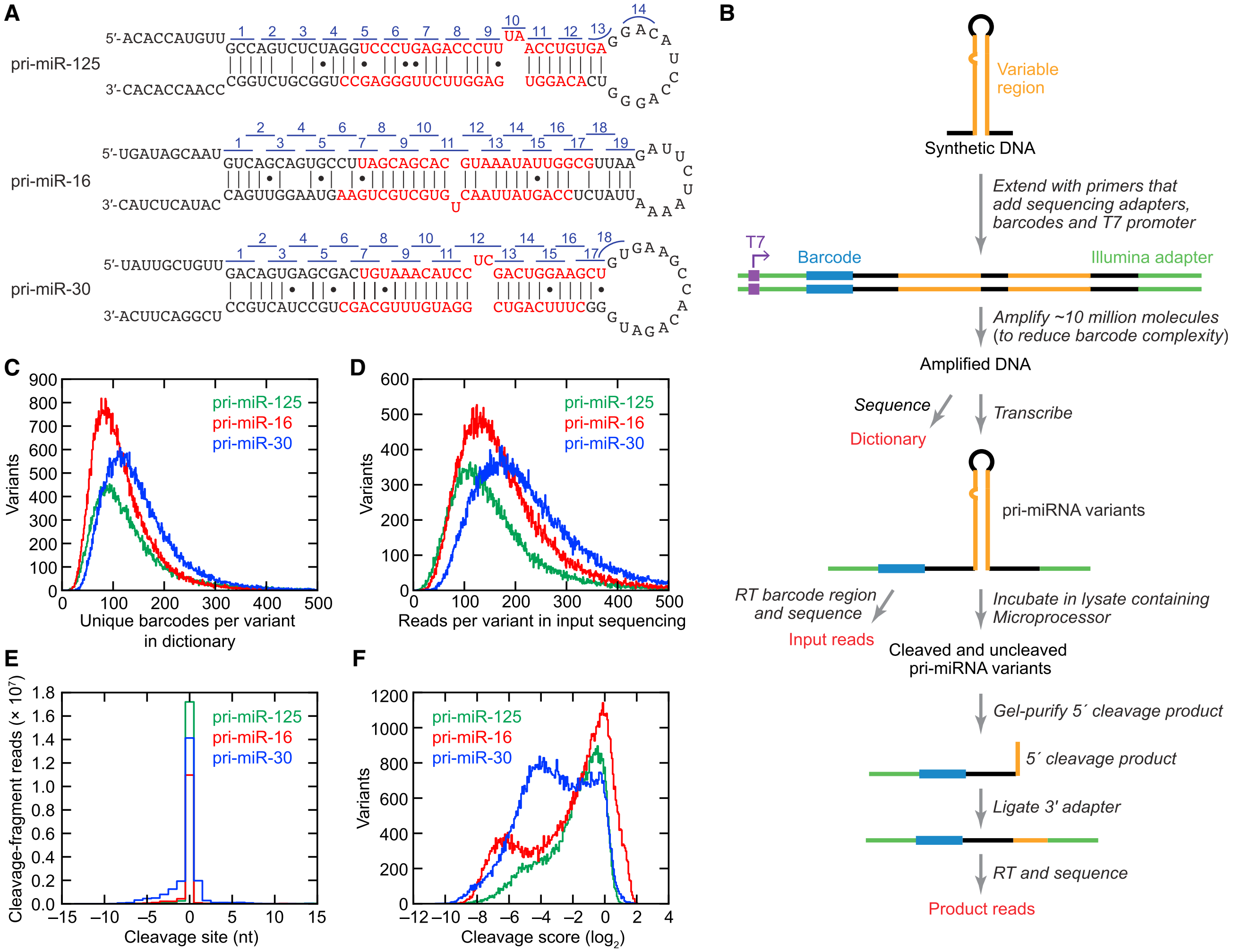


Figure 1. Design of pri-miRNA Pools and High-Throughput Analyses of Variants

(A) Secondary structure of the three parental pri-miRNAs. The miRNA–miRNA* duplexes are in red. Blue numbers above 5′ (5p) sequences indicate randomized windows, most of which contained three nucleotides on the 5p arm and the corresponding nucleotides on the 3p arm.

(B) Schematic of the protocol for generating each dictionary of barcoded variants and quantifying the amount of each variant that was in the input and that was cleaved at each site.

(C) Distribution of unique barcodes per variant in each dictionary.

(D) Distributions of reads per variant in the input sequencing.

(E) Distributions of cleavage sites for the sequenced 5′ cleavage fragments.

(F) Distributions of cleavage scores.

See also Figure S1.

percentages used. Results of these two time points were treated as replicates and combined for subsequent analyses. Global distributions of cleavage scores showed that most variants were cleaved less efficiently than were their wild-type counterparts, and that among the three pri-miRNAs, pri-miRNA-125 was the least sensitive to mutations, whereas pri-miRNA-16 was most frequently improved by substitutions (Figure 1F).

Structural Preferences across the Stem

To visualize the results, we plotted the cleavage scores of all 4,096 variants within each 3-bp window on a 64 × 64 grid. Figure 2A shows a typical pattern in which the higher scores for combinations along the diagonals indicated a preference for base pairing. We also plotted cleavage scores of all 16 possible variants within each single-bp window across the stem (considering only variants that had the wild-type sequence at all other positions) on 4 × 4 grids, to see how all changes at each position across each stem affected cleavage in the wild-type background (Figure S2). Again, a preference for pairing (Watson–Crick and wobbles) was often observed, even at positions that were not paired in the wild-type sequence.

To summarize the preference for pairing at each stem position, we derived a simple base-pairing score, calculated as the difference between the average cleavage scores of the six paired

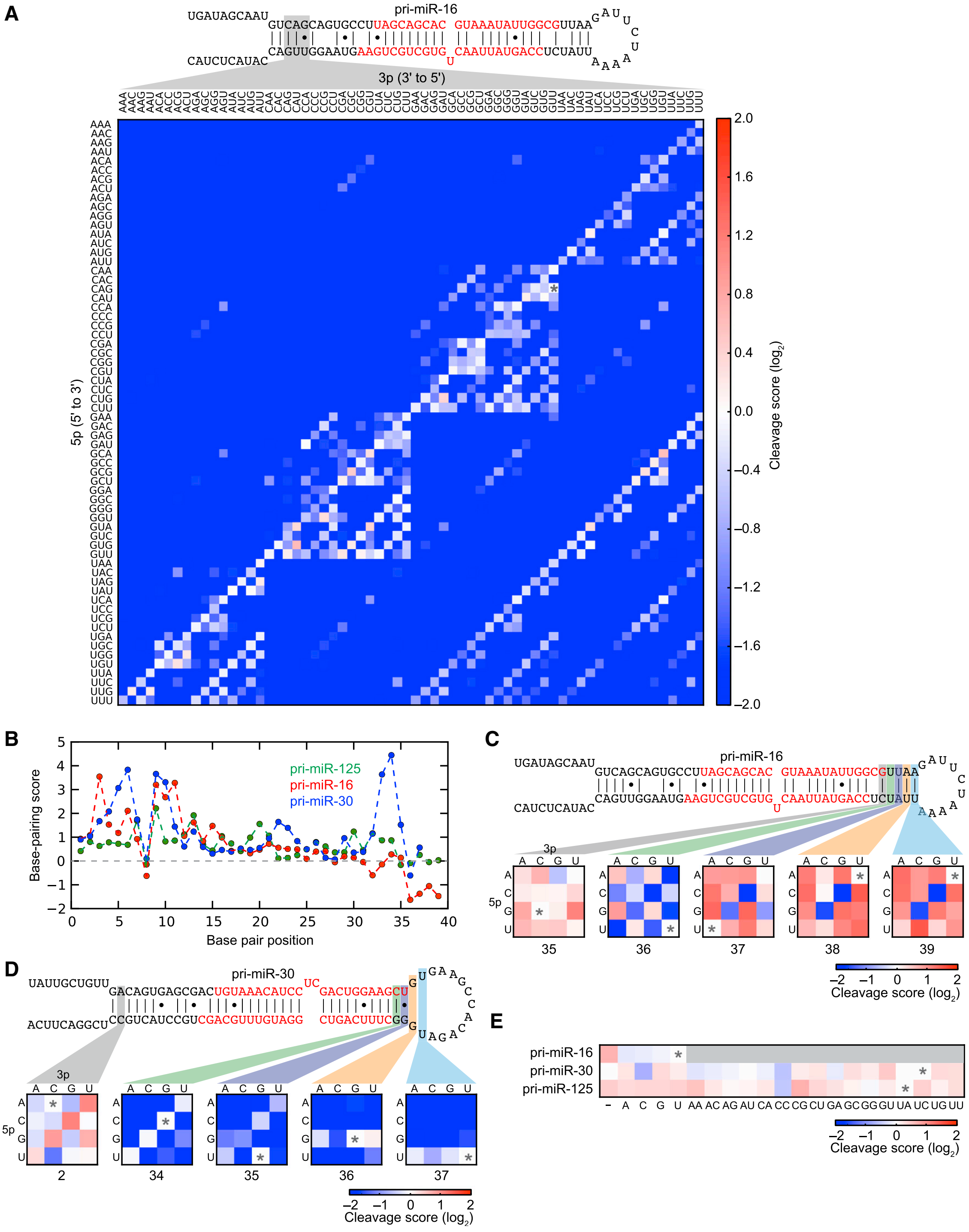


Figure 2. Structural Preferences across the Stem

(A) Cleavage scores for all 4,096 variants of pri-miR-16 within randomized window 2 (shaded gray). Each row shows the scores of the indicated 5p trinucleotide (written 5′ to 3′), and each column shows the scores of the indicated 3p trinucleotide (written 3′ to 5′), colored according to the key (right). The asterisk marks the wild-type sequence.

(B) Base-pairing scores at each position across each of the indicated stems.

(C) Detrimental effects of maintaining or strengthening the apical pairing of pri-miR-16. Each 4 × 4 heat map shows the scores of all 16 single-bp variants at the shaded position in the context of wild-type nucleotides at all other positions, colored according to the key (below). Each asterisk marks the wild-type sequence.

(D) Beneficial effects of pairing at position 2 and of maintaining the UGU motif in the apical region of pri-miR-30. Otherwise, this panel is as in (C).

(legend continued on next page)

variants (including G–U and U–G wobbles) and the average scores of the ten mismatch variants, all in the context of the wild-type background. At all but one of the first 35 positions of the stem, a preference for pairing was observed, as indicated by base-pairing scores > 0 (Figure 2B). The exception was at position 8, which is one of four positions reported to be frequently mismatched in human pri-miRNAs (Han et al., 2006). Base-pairing scores were particularly high in the basal stem (positions 1–13) of pri-miR-16 and pri-miR-30—the two pri-miRNAs with the most mismatches and wobbles in the basal stem—indicating that their cleavage scores were particularly sensitive to gain or loss of a pair (positive or negative change, respectively) in this region (Figure 2B). Maintaining pairing in the last three base positions of the pri-miR-30 stem (positions 33–35) was also particularly important (Figure 2B).

Overall, single-bp changes had relatively minor effects on pri-miR-125, suggesting redundancy in the features required for efficient processing of this hairpin (Figure S2). In contrast, many substitutions toward the apical end of the pri-miR-16 stem improved processing (Figure S2). In particular, disrupting any of the last three base pairs strongly enhanced processing, and replacing any of these U–A or A–U pairs with stronger pairs had the opposite effect (Figure 2C), suggesting that the length of the pri-miR-16 stem, which is 39 bp (counting Watson–Crick pairs, wobbles, and mismatches but not the 1-nt bulge) was too long. In agreement with this idea, the wild-type pri-miR-30 and pri-miR-125 had 35-bp stems, with either extension to 36 bp or shortening to <34 bp clearly disfavored (Figure S2). In contrast to the apical region of pri-miR-16, the apical region of pri-miR-30 was already near a local optimum on the fitness landscape of cleavage substrates, with preference for retention of both the pairing at the end of the stem and the previously described UGU motif (Auyeung et al., 2013) (Figure 2D). For this pri-miRNA, the most favorable single-bp changes repaired the mismatch in the basal stem at position 2 (Figure 2D).

Examining variants of the 1- or 2-nt bulges revealed a tendency for beneficial effects from changing their sizes (including eliminating the bulge) or nucleotide composition (Figure 2E). However, these effects were modest, indicating that in the wild-type contexts of these three pri-miRNAs, small bulges were neither necessary for nor detrimental to Drosha cleavage.

A Mismatched GHG Motif Enhances pri-miRNA Processing

The consistently lower and even negative pairing scores at position 8 (Figure 2B) prompted a closer look at pairs and mismatches favored at this position and its two flanking positions. Combining results from all three pri-miRNAs, we ranked the 4,096 variants at positions 7–9 based on their average cleavage scores and selected the top 1% (Table S2). When examining the frequencies of pairs and mismatches, these 41 variants all had pairs at positions 7 and 9, but mostly mismatches (particularly U–C, C–U, and G–A) at position 8 (Figure 3A). This analysis, combined with nucleotide composition analysis of these top variants, showed that in the 3′ (3p) arm of the hairpin, position 7 was enriched for a paired G, although the other Watson–Crick pairs and wobbles were also present; position 8 was never a G; and position 9 was enriched for a Watson–Crick paired G, although other Watson–Crick pairs were present (Figures 3A and 3B). We therefore named this the ‘‘mismatched GHG’’ motif (in which H is any nucleotide except G), based on this primary-sequence preference in the 3p arm and the frequent mismatch at position 8.

When examining the presence of this mismatched GHG motif within conserved human pri-miRNAs, we observed a significant position-specific enrichment (Figure 3C). Significant enrichment was also observed in other vertebrates, as well as in fruit fly (Figure 3C). Enrichment observed in other arthropods (mosquito and water flea) did not reach statistical significance, presumably because of the smaller sets of annotated pri-miRNAs in these species (Table S3).

The wild-type sequence at positions 7–9 of pri-miR-125, a 5′-CUC-3′ on the 5p arm imperfectly paired to 5′-GCG-3′ on the 3p arm, was chosen as a representative mismatched GHG motif for further study. It contained the pairs and mismatch most frequently observed in the top variants (Figure 3A) and was among the top three variants in the overall ranking (Table S2). To validate the high-throughput results and further characterize this motif, we tested engineered variants in a competitive-cleavage assay, in which cleavage was measured relative to that of an internal reference. The internal reference was wild-type pri-miR-125 with a long 5′ cleavage product, designed to be easily distinguished from that of the variants (Figure 3D). When the C–G pair at position 7, position 9, or both positions was flipped, in vitro cleavage efficiency decreased to 67%, 74%, or 26%, respectively, and when the U–C mismatch at position 8 was changed to either a U–A or a U–G pair, the in vitro processing decreased to 66% or 45%, respectively, results consistent with those determined from the sequencing data (Figure 3D). When motifs were added sequentially to a hairpin without motifs, the mismatched GHG motif, which was added first, had the greatest effect (12-fold), suggesting some redundancy among the motifs (Figure S3A).

To test whether this mismatched GHG motif enhanced pri-miRNA processing in vivo, we incorporated it into the corresponding region of derivatives of pri-miR-44, a *C. elegans* pri-miRNA known to be sub-optimally processed in mammalian cells (Auyeung et al., 2013), and asked whether the motif conferred more efficient processing in HEK293T cells. In this assay, the pri-miR-44 variants were each expressed on a transcript that also contained human pri-miR-1-1 (hereafter called pri-miR-1), and the relative accumulation of each mature miRNA was measured on RNA blots, using the accumulation of miR-1 as an internal normalization standard (Figure 3E) (Auyeung et al.,

(E) Modest effects of changing or eliminating each of the bulges normally found in each of the three pri-miRNAs. The heat map shows the cleavage scores of the indicated variants in the context of wild-type nucleotides at all other positions (—, no bulge), colored according to the key (below). Because pri-miR-16 had a 1-nt bulge, dinucleotide variants were not tested (gray). Each asterisk marks the wild-type sequence.
See also Figure S2.

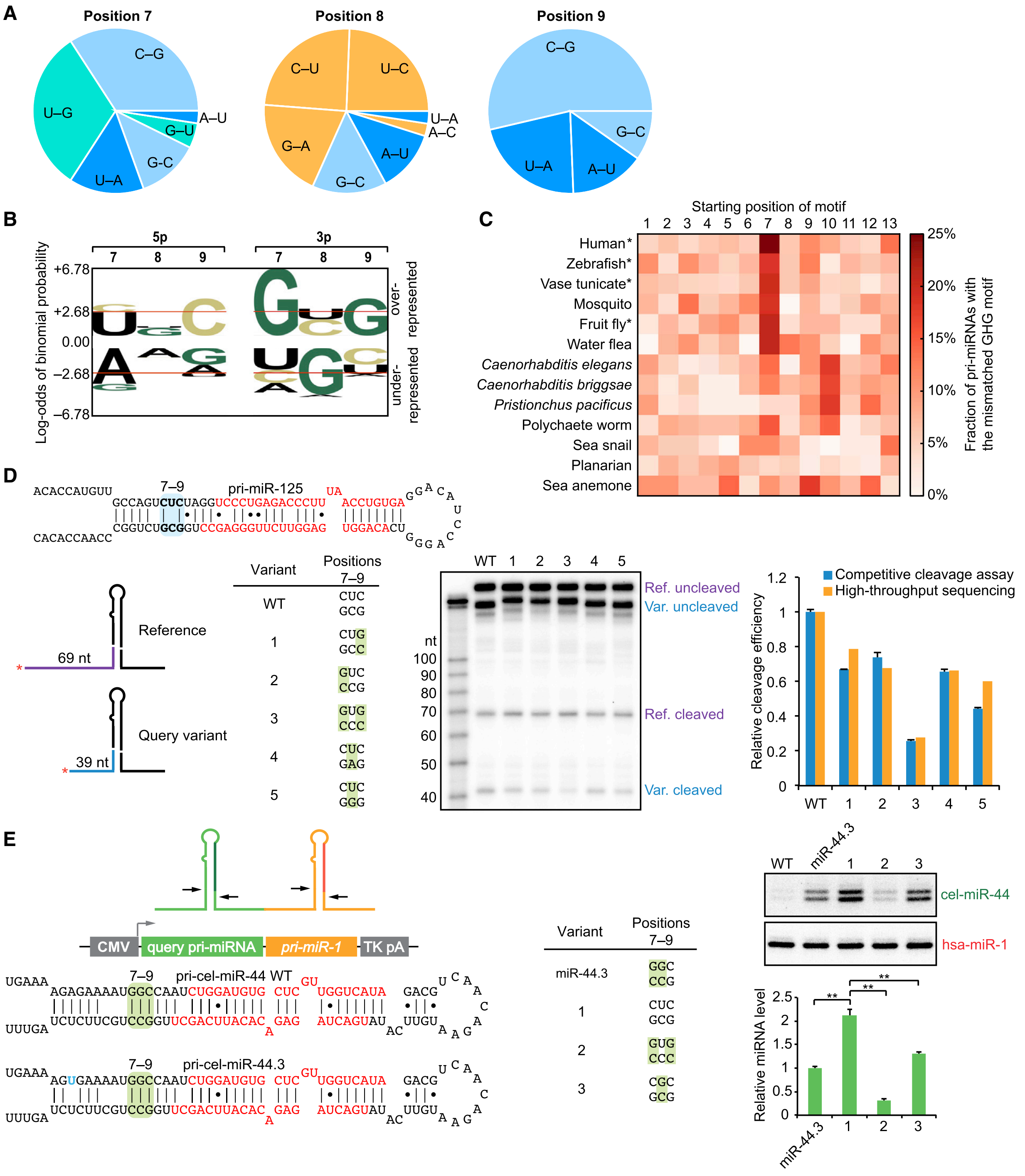


Figure 3. A Broadly Conserved Mismatched GHG Motif Enhances pri-miRNA Processing

(A) Nucleotide pairs preferred at the three positions of the mismatched GHG motif. Shown is the relative fraction of each nucleotide pair observed in the top 1% of the variants generated from randomizing positions 7–9 of the three pri-miRNAs (Table S2). For each pair, the first letter indicates the 5p nucleotide and the second letter indicates the 3p nucleotide.

(B) Primary-sequence preferences within the mismatched GHG motif. Shown is a pLogo, which represents the nucleotide enrichment and depletion observed at the indicated positions within the top 41 variants generated from randomizing positions 7–9 (Table S2) and compared to the background of all 4,096 possible variants at these positions (O'Shea et al., 2013). Red lines indicate the p value threshold of 0.05.

(C) Enrichment of the mismatched GHG motif in natural miRNAs. The mismatched GHG motif was defined as a 3-bp structural element in which the first pair could be C–G or U–G, the second could be one of the seven mismatches or pairs shown in (A), and the third could be any Watson–Crick base pair. The heat map shows

(legend continued on next page)

2013). When 5′-CUC-3′–5′-GCG-3′ replaced 5′-GGC-3′–5′-GCC-3′ in *C. elegans* pri-miRNA-44 to create a GHG motif with a U–C mismatch at position 8, the mature miR-44 increased >2-fold (Figure 3E). Flipping the C–G pairs at positions 7 and 9 or changing the U–C mismatch to a G–C pair diminished the enhancement, consistent with our in vitro results. Similar results were observed in the other pri-miRNA contexts examined (Figures S3B and S3C).

De Novo Design of Artificial pri-miRNAs

Our high-throughput analyses of the stem regions complemented our previous analyses of the flanking and loop regions (Auyeung et al., 2013) to provide thorough analyses of three human pri-miRNAs. Based on these analyses, the preferred Microprocessor substrate is a 35-bp hairpin flanked by single-stranded sequences with a properly positioned GHG motif in the mid-basal stem and the previously described basal UG, apical UGU, and flanking CNNC motifs. Many additional, more nuanced preferences were also observed (Figure S2), raising the possibility that many additional weak features or complex combinations of features might be needed to define a miRNA. Alternatively, we might have already identified a subset of elements sufficient to define a pri-miRNA, and the additional preferences might have reflected idiosyncratic vulnerabilities of the starting pri-miRNAs, analogous to the heightened sensitivity to either mismatches in the basal stem of pri-miRNAs starting with more mismatches in this region (Figure 2B, pri-miR-16 and pri-miR-30) or strengthened pairing at the distal end of a pri-miRNA stem that is already too long (Figure 2C).

To test whether we knew enough to define a miRNA gene, we designed artificial pri-miRNAs using the features of the preferred Microprocessor substrate listed earlier—without reference to the sequence of any known miRNA—and asked whether these hairpins could be processed. To simplify the design, we used homopolymeric U segments at the single-stranded regions near the stem and perfect Watson–Crick pairs at all paired positions of the stem (Figure 4A). At most paired positions, the primary sequence was randomly generated—the exceptions were position 1, which included the G of the UG motif; positions 7–9, which comprised the mismatched GHG motif; position 35, which comprised the first U of the apical UGU motif; and positions 14–16. Although particular Watson–Crick pairs at positions 14–16 were not favored during Drosha processing (Figure S2), the possibilities at these positions were nonetheless constrained to facilitate loading of the mature miRNA into Argonaute, which is required for miRNA stability in vivo (Winter and Diederichs, 2011). Accordingly, the pairs at positions 14–16 were constrained to be A–U or U–A pairs so that Watson–Crick pairing of these positions in the miRNA duplex would be sufficiently weak to facilitate loading of the strand from the 5p arm into Argonaute (Khvorova et al., 2003; Schwarz et al., 2003). The pair at position 14, which included the first nucleotide of the mature miRNA, was further constrained to be a U, the most common first nucleotide of conserved mammalian miRNAs.

Surprisingly, all three artificial pri-miRNAs designed according to these guidelines (pri-miRNAs A1, A2, and A3 in Figure 4B) were well processed. In vitro competitive-cleavage assays indicated that they were each processed 1.5- to 4-fold more efficiently than pri-miR-125 (Figures 4C and S4A). Removing all motifs to yield pri-miRNAs that contained only structural features (derivatives A1.1, A2.1, and A3.1 in Figure 4B) reduced cleavage at the intended site ~6-fold, with appearance of miscleaved products observed for the A1 and A3 derivatives (Figure 4C), including products suggestive of unproductive cleavage (Han et al., 2006; Nguyen et al., 2015), in which Microprocessor recognizes the stem in the opposite orientation (Figure S4A). Nonetheless, processing efficiency at the intended site was within 25%–50% of that of pri-miR-125. Restoring the mismatched GHG motif (A1.2, A2.2, and A3.2) improved the processing of variants with no other motif by 2- to 3-fold, but removing this single motif from those with all motifs (A1.3, A2.3, and A3.3) had a statistically significant effect in only one of the three contexts (Figures 4C and S4A).

To assay processing in vivo, we expressed each of the artificial pri-miRNAs in HEK293T cells using our bicistronic system in which accumulation of miR-1 served as an internal normalization standard (Figure 3E). The in vivo results corroborated the in vitro ones, with differences between variants being somewhat muted, although still statistically significant, compared to those observed in vitro (Figure 4D). We then used quantitative RNA blots to measure the absolute accumulation of artificial miRNAs A1, A2, and A3 in cells. When the artificial pri-miRNAs were inserted between pri-miR-30 and pri-miR-1, their mature miRNA levels accumulated to about twice that of either miR-30 or miR-1 (Figures 4E and S4B). These results indicated that our simple design produced pri-miRNAs that are processed at

the frequency of the motif observed at the indicated position within the stems of representative pri-miRNA from the indicated species (Table S3). The asterisk indicates species with a significant enrichment at position 7 ($p < 0.05$, one-tailed binomial test with Bonferroni correction).

(D) Increased cleavage efficiency imparted by the mismatched GHG motif. The gel (center) shows results of competitive-cleavage assays that determined the relative cleavage of pri-miR-125 variants 1–5, which had the indicated substitutions within the mismatched GHG motif (center table). The wild-type (WT) hairpin with the mismatched GHG motif at positions 7–9 (blue shading) is shown for reference (upper left). As schematized (lower left), each assay included the query variant, which generated a 39-nt labeled product, and a longer pri-miR-125 wild-type reference substrate, which generated a 69-nt labeled product. The graph (right) shows the mean relative cleavage efficiency of each variant, normalized to that of the WT (blue bars; error bars, SEM, n = 3), compared to the value determined from the high-throughput sequencing experiment (orange bars).

(E) Increased miRNA accumulation imparted by the mismatched GHG motif in HEK293T cells. The mismatched GHG motif was tested in the context of pri-miR-44.3 (bottom left), a derivative of *C. elegans* pri-miR-44 with a U substitution (blue) that increases processing, presumably because it destabilizes pairing beyond the basal stem and introduces a basal UG motif (Auyeung et al., 2013). The variants (center table) introduced either the mismatched GHG motif (variant 1) or the control sequences. RNA blots (top right) examined miR-44 accumulation in cells for each variant when expressed as a query pri-miRNA on the same primary transcript as pri-miR-1, as schematized (top left). The graph (bottom right) plots relative levels of mature miR-44 after normalizing to the miR-1 internal reference (mean ± SEM, n = 3; **$p \leq 0.01$, one-tailed Student's t test).

See also Figure S3 and Tables S2 and S3.

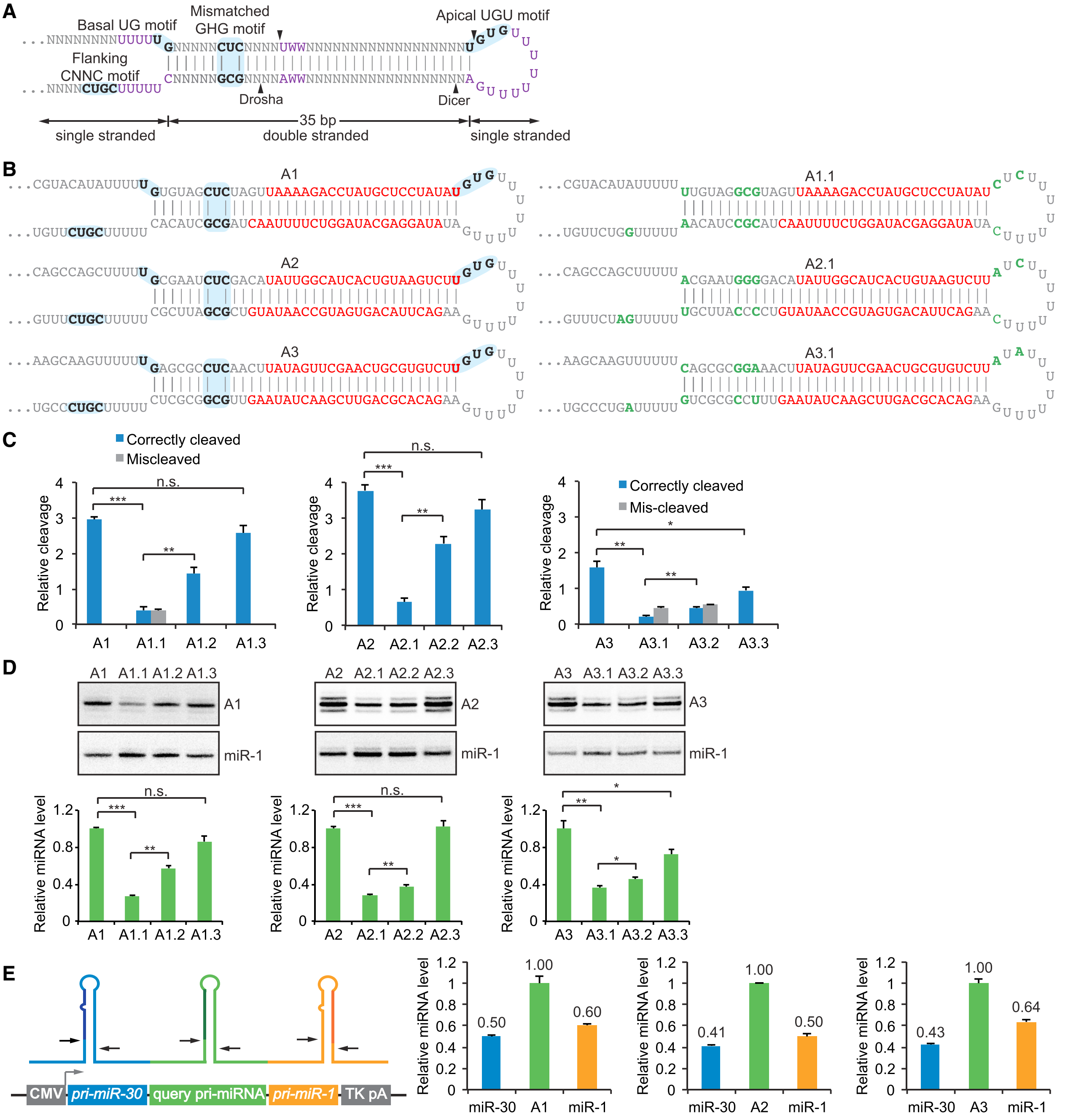


Figure 4. De Novo Designed pri-miRNAs Are Processed Efficiently and Accurately In Vitro and in Cells

(A) Guidelines for de novo design of pri-miRNAs. Motif residues are highlighted (blue). PolyU segments, which disfavor pairing, and other constrained sequences, some of which favor loading into Argonaute, are purple (W = A or U), and randomly assigned residues or pairs are gray (N = A, C, G, or U).

(B) Sequences of three artificial pri-miRNAs (A1, A2, and A3) and their variants in which the motifs were disrupted (A1.1, A2.1, and A3.1, green substitutions). Motif residues are highlighted (blue), and residues of the miRNA duplex are red.

(C) In vitro cleavage efficiencies of artificial pri-miRNAs, comparing variants with and without all motifs and those with and without the mismatched GHG motif. Variants with and without all motifs are shown in (B); A1.2, A2.2, and A3.2 each have the mismatched GHG motif as the only motif, and A1.3, A2.3, and A3.3 each have all motifs except the mismatched GHG motif. Plotted in blue are mean cleavage efficiencies at the correct site relative to the pri-miR-125 internal reference, determined as in Figure 3D (error bars, SEM, n = 3; ***p ≤ 0.001; **p ≤ 0.01; *p ≤ 0.05; not statistically significant, n.s., p > 0.05; one-tailed Student's t test). If miscleavage was detected, its efficiency was similarly plotted in gray. See Figure S4A for images of competitive-cleavage results.

(D) Accumulation of mature artificial miRNAs in HEK293T cells, comparing variants with and without all motifs and those with and without the mismatched GHG motif. Assays were as in Figure 3E; artificial pri-miRNA variants and evaluation of statistical significance were as in (C).

(legend continued on next page)

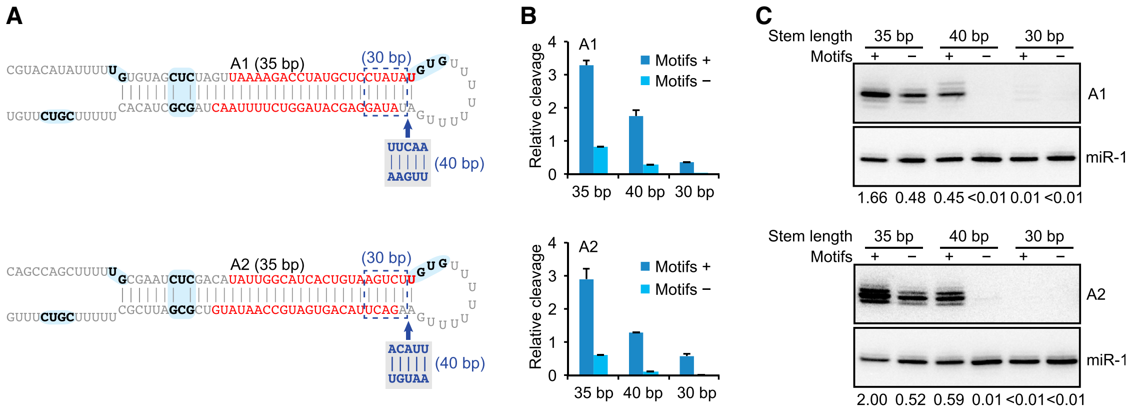


Figure 5. Sequence Motifs Rescue Suboptimal Stem Lengths

(A) Diagrams of extension and deletion variants of A1 and A2. Otherwise, this panel is as in Figure 4B.

(B) In vitro cleavage efficiencies of the extension (40 bp) and deletion (30 bp) variants, with or without motifs. Plotted are mean cleavage efficiencies relative to the pri-miR-125 internal reference, determined as in Figure 3D (error bars, SEM, n = 2). See Figure S5 for images of competitive-cleavage results.

(C) Accumulation of mature miRNAs from the extension and deletion variants, with or without motifs, in HEK293T cells. Assays were as in Figure 3E. Mature miRNA levels relative to co-transcribed miR-1 are indicated below each lane, reporting the mean from two biological replicates. Results of Figure 4E were used to infer the ratio of A1 and A2 accumulation relative to that of miR-1, and the other values were calculated based on this ratio.

least as efficiently as natural pri-miRNAs, including pri-miR-30, which is commonly used as a platform for efficiently expressing shRNAs in vivo (Zeng et al., 2002; Silva et al., 2005; Bassik et al., 2013; Fellmann et al., 2013; Knott et al., 2014; Kampmann et al., 2015).

As observed for many natural pri-miRNAs, our artificial pri-miRNAs yielded mature miRNAs with some length heterogeneity (Figure 4D). Small-RNA sequencing showed that this heterogeneity was predominantly at the 3′ ends, as observed for natural miRNAs (Figure S4C). In addition, miRNAs from the 5p arm outnumbered those from the 3p arm by a factor of about five (Figure S4C). These results indicated that our de novo designed pri-miRNAs were each correctly processed into miRNAs, which were loaded into Argonaute with expected strand asymmetry.

Motifs Rescue Structural Defects

Although the motifs enhanced processing of our artificial miRNAs, the versions without these motifs (A1.1, A2.1, and A3.1) were processed with efficiency approaching that of natural pri-miRNAs (Figure 4). This showed that a hairpin with a 35-bp perfect stem was sufficient for recognition and cleavage, thereby illustrating the key role that structure can play in defining pri-miRNAs. However, sequences with the potential to form hairpins with perfectly paired stems of precisely 35 bp are rare in the genome, and natural pri-miRNA hairpins are mostly of other lengths and typically have mismatches and bulges in the stem. To understand better the features that define pri-miRNAs, we incorporated these structural “defects” into the A1 and A2 artificial pri-miRNAs and asked how they were processed, with and without sequence motifs.

When the pri-miRNA stems were extended by 5 bp in the apical region (Figure 5A), the pri-miRNAs were still processed in vitro, albeit at ~50% efficiency (Figures 5B and S5). Without the four motifs, however, cleavage was much less accurate (Figure S5), with efficiency reduced to <9% of that of the original A1 and A2 pri-miRNAs (Figure 5B). These differences observed in vitro translated to more striking differences in mature miRNA levels in vivo. With sequence motifs, miRNA levels dropped nearly 4-fold but were still within range of the miR-1 internal standard, and without sequence motifs, they dropped another 50-fold (Figure 5C). When the pri-miRNA stems were shortened by 5 bp (Figure 5A), some cleavage at the proper position occurred for the pri-miRNAs with the sequence motifs, but ~70% of the 5′ cleavage fragments were of a smaller size (Figure S5). Without the motifs, cleavage at the proper position was no longer detected, although some of the miscleaved fragment was observed for the A2 derivative (Figure S5). In vivo, little if any mature miRNA was observed from the shortened derivatives (Figure 5C), which presumably reflected reduced complementarity to the probe and the unsuitability of the properly cleaved product for Dicer cleavage, in addition to reduced cleavage by Microprocessor. The miscleavage observed with extended and shortened stems resembled that observed for analogous derivatives of natural miRNAs, which first revealed that measurement from the ends of the stem influences the site of cleavage (Zeng et al., 2005; Han et al., 2006; Ma et al., 2013). Our results added key insight with respect to the sequence motifs, showing that these motifs are important for positioning the cleavage site of pri-miRNA hairpins that are not the optimal length and that this positioning effect, combined with enhanced processing efficiency,

(E) miRNA yield from artificial pri-miRNAs relative to that from natural pri-miRNAs. As schematized (left), each artificial pri-miRNA was transcribed between the pri-miR-30 and the pri-miR-1 internal references as the query pri-miRNA. Plotted are the relative levels of mature miRNAs, determined using quantitative RNA blots (mean ± SEM, n = 3). See Figure S4B for images of quantitative RNA blots.
See also Figure S4.

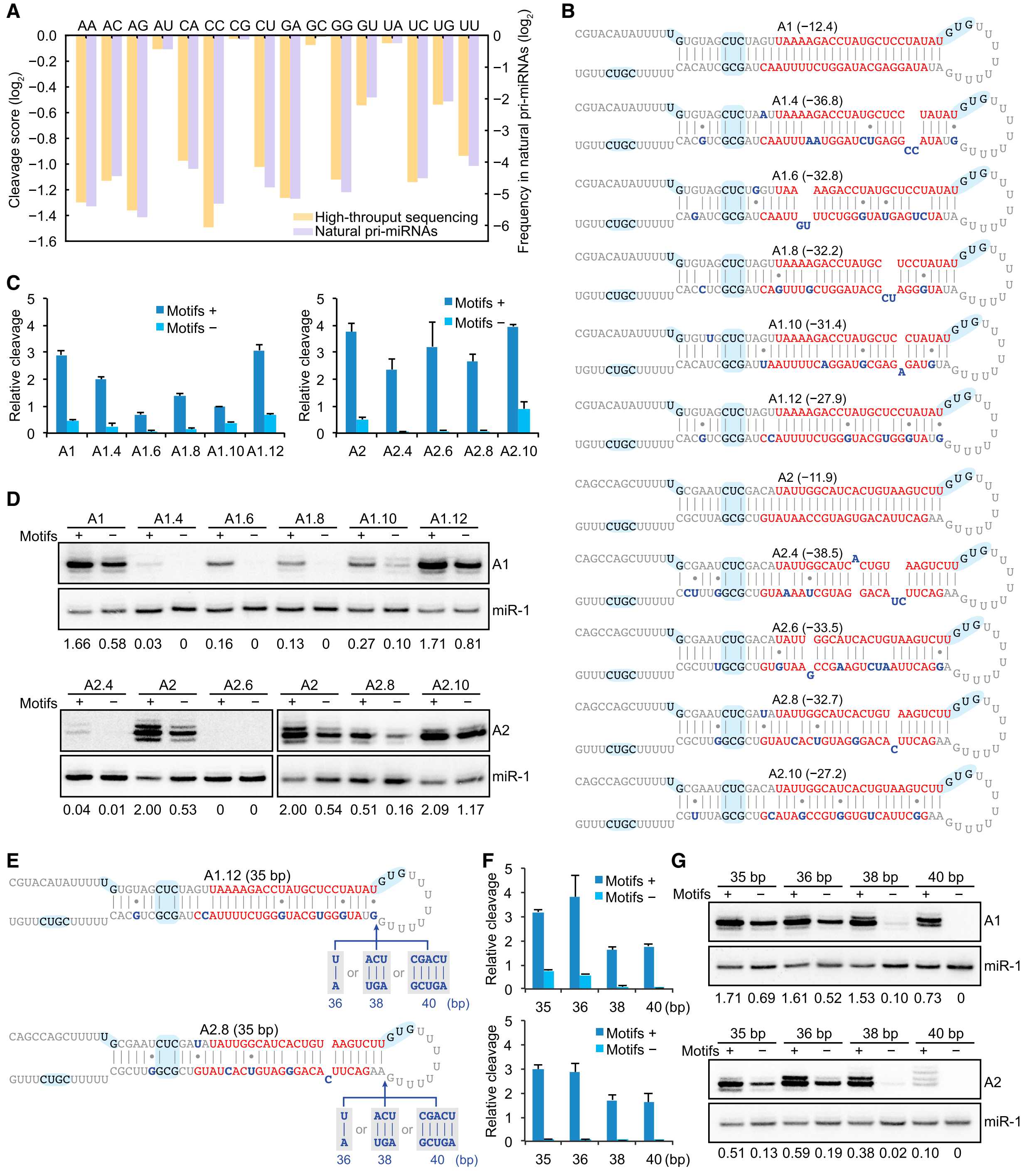


Figure 6. Motifs Rescue Structural Defects

(A) The average effects of each pair, wobble, or mismatch possibility on cleavage compared to the frequency of that possibility in natural pri-miRNAs. Cleavage effects were determined from the high-throughput results, averaging the cleavage scores calculated from single-bp variants of pri-miRNA-125, pri-miR-16, and pri-miR-30 (orange bars, left axis). The frequency of each possibility was tallied across the 35-bp stems of representative members of 186 conserved human pri-miRNA families (Table S4; purple bars, right axis).

(B) Diagrams of structural variants of A1 and A2. Motifs are highlighted (blue), miRNA duplexes are red, substituted residues are dark blue, and structure scores are in parentheses.

(C) In vitro cleavage efficiencies of structural variants of A1 and A2, with or without motifs. Plotted are mean cleavage efficiencies relative to the pri-miR-125 internal reference, determined as in Figure 3D (error bars, SEM, n = 2). See Figure S6B for images of competitive-cleavage results.

(legend continued on next page)

can increase mature miRNA levels from nearly undetectable to within range of the miR-1 internal standard.

With the goal of incorporating mismatches and bulges into the design of our artificial miRNAs, we developed a metric to score these structural defects among human pri-miRNAs. We first surveyed representative members of 186 conserved human pri-miRNA families and tallied the occurrences of all 16 possible single base pairs, wobbles, and mismatches; all four 1-nt bulges; and other less frequent bulges and internal loops within their 35-bp stem regions (Table S4). On average, human pri-miRNA stems had 3.7 G–U or U–G wobbles, 2.7 single-bp mismatches, 0.9 1-nt bulges, and 0.5 2-bp mismatches. The influence of each of these structural defects was then scored by taking the $\log_2$ of its frequency, relative to that of the most frequent base pair, G–C (G on the 5p arm, C on the 3p arm), which was assigned a score of 0. These scores agreed well with the analogous cleavage scores calculated from single-bp variants of pri-miR-125, pri-miR-16, and pri-miR-30 (Figure 6A), suggesting that the forces of natural selection acting on the pri-miRNA structure were accurately reflected in the cleavage preferences observed in vitro. Summing the frequency-based scores along the 35-bp stem region of each representative pri-miRNA yielded structure scores, which ranged between -56 and -18, with a median of -37 (Figure S6A).

We arbitrarily incorporated wobble pairs, mismatches, and bulges into A1 and A2, keeping the mature miRNA sequence unchanged to facilitate comparisons on RNA blots and allowing the structure score to range from -27 to -39 (Figure 6B). When retaining the sequence motifs, all of these A1 and A2 derivatives were cleaved in vitro, with efficiencies similar to or greater than that of pri-miR-125, and those with the most favorable structure scores (A1.12 and A2.10) had cleavage efficiencies similar to those of their respective parental pri-miRNAs (Figures 6C and S6B). Without motifs, these hairpins with defects were processed much less efficiently than pri-miRNA-125, with the exception of A1.12 and A2.10, which had the most favorable structure scores (Figures 6C and S6B). In vivo, mature miRNAs from A1.12 and A2.10 accumulated to levels similar to those from the parental A1 or A2 pri-miRNAs (Figures 6D and S6C). Mature miRNAs from the remaining variants accumulated to $\geq$4-fold lower levels, with those from hairpins with 2-nt bulges (A1.4, A1.6, A1.8, and A2.4) accumulating at levels that were >10-fold lower and showing enhanced dependence on the sequence motifs (Figures 6D and S6C). The variant with an internal loop spanning 3 nt on both arms (A2.6) failed to produce detectable miRNA in vivo, even when containing the sequence motifs (Figure 6B), which can be reconciled with its efficient cleavage by Microprocessor (Figure 6C) if its internal loop prevented subsequent cleavage by Dicer.

Finally, we extended the stems of A1.12 and A2.8, two derivatives that produced miRNAs with high intracellular accumulation (Figures 6C and 6D). The extensions by 1 bp had little effect in vitro and in cells, but the extensions by $\geq$3 bp were only tolerated in variants containing the sequence motifs (Figures 6E–6G and S6D). These results reinforced our conclusion that the sequence motifs can rescue structural defects in pri-miRNAs to confer efficient processing and can contribute most in a window in which the structural defects are sufficiently severe to substantially influence processing but not so severe that they eliminate processing in the presence of the motifs.

In Vivo Activity of Artificial miRNAs

To test further our understanding of the features required to define pri-miRNAs, we designed three new artificial pri-miRNAs and asked whether they generated mature miRNAs that function to mediate gene repression when expressed in cells. These hairpins each had a 35-bp stem, two to four wobble pairs, and two or three mismatches, and one had a 1-nt bulge (Figure 7A). Sequence identity within each stem was arbitrary at most positions, the exceptions being (1) the four sequence motifs, (2) the U as the first nucleotide of the mature miRNA, and (3) the use of sequences satisfying pairing-stability constraints that favor loading of the 5p species into Argonaute. To confirm that the homopolymeric U segments at the single-stranded regions near the stems of our previous artificial pri-miRNAs were not required for efficient processing, we included other nucleotides in these regions.

When expressing these pri-miRNAs in HEK293T cells, levels of the mature artificial miRNAs were 1.5- to 2.5-fold greater than that of the miR-1 internal standard (Figure S7A). Moreover, small-RNA sequencing indicated that processing occurred predominantly at the designed positions (Figure S7B). After sorting transfected cells based on GFP expressed from a co-transfected plasmid, we performed RNA sequencing (RNA-seq) analyses, comparing RNA from cells transfected with a bicistronic pri-miRNA plasmid to that of cells transfected with only the GFP-expression plasmid. These analyses revealed the expected miRNA-targeting effects for each of the three artificial miRNAs (Figure 7B) (Grimson et al., 2007). Indeed, the repression mediated by the artificial miRNAs appeared at least as strong as that mediated by the co-expressed miR-1 (Figure 7B).

DISCUSSION

Our results from tens of thousands of stem variants of three pri-miRNA hairpins revealed that pairing was favored over mismatches at all but one position of the stem (Figure 2), which implied that the three human pri-miRNAs each benefited from more pairing and were sensitive to less pairing. The benefit from more pairing had diminishing returns, however, as indicted from our analysis of the artificial pri-miRNAs. Artificial pri-miRNAs

(D) Accumulation of mature miRNAs from structural variants of A1 and A2, with or without motifs, in HEK293T cells. Assays were as in Figure 3E. Mature miRNA levels relative to co-transcribed miR-1 are indicated below each lane, as in Figure 5C.

(E) Diagram of extension variants of A1.12 and A2.8.

(F) In vitro cleavage efficiencies of extension variants of A1.12 and A2.8, with or without motifs. Otherwise, this panel is as in (C). See Figure S6D for images of competitive-cleavage results.

(G) Accumulation of mature miRNAs from extension variants of A1.12 and A2.8, with or without motifs, in HEK293T cells. Otherwise, this panel is as in (D). See also Figure S6 and Table S4.

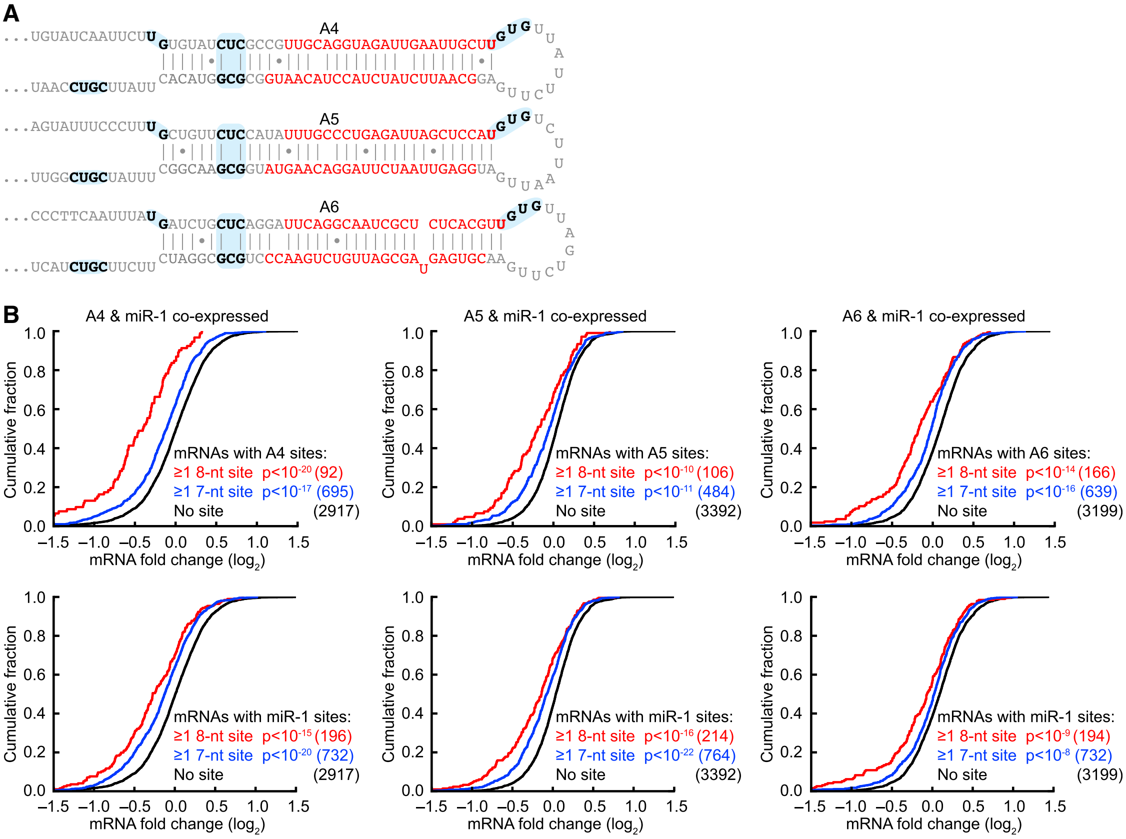


Figure 7. Artificial miRNAs Mediate Repression

(A) Sequences of artificial pri-miRNAs A4–A6. Otherwise, this panel is as in Figure 4B.

(B) Response of cellular mRNAs upon co-expression of the indicated artificial miRNA and miR-1. Plotted are cumulative distributions of fold changes for mRNAs with the indicated sites in their 3′ UTRs. The mRNAs with 3′-UTR sites to both the miR-1 and the artificial miRNA were not considered. For each set of mRNAs, the number of reliably quantified distinct mRNAs is shown in parentheses, and for sets containing sites, the p value reports the significance of the difference in the fold-change distribution compared to that of the corresponding set of mRNAs without sites (one-tailed Mann-Whitney test).

See also Figure S7.

with four wobble pairs and three single-bp mismatches in their stem regions (A1.12 and A2.10) were at least as efficiently processed as either their parental pri-miRNAs with one mismatch (A1 and A2) or their derivatives with perfectly paired stems (A1.3 and A2.3) (Figures 4 and 6). In practice, the pri-miRNAs with wobbles and mismatches were much easier to clone than were those with perfectly paired stems. They might also be less likely to trigger an interferon response, although examination of RNA-seq data obtained with and without expression of artificial pri-miRNAs with more or less extensive pairing showed no evidence of an interferon response (data not shown). The advantage of greater genomic stability without compromising cleavage efficiency helps explain why all conserved pri-miRNAs have some wobbles and mismatches.

The length of the stem region, with a narrow preference for 35 ± 1 bp (including wobbles and mismatches but not bulged nucleotides), was found to be a second important structural feature of pri-miRNAs. In addition to contributing specificity to the first step in the miRNA biogenesis pathway, the preference for a stem length of 35 bp ensures that most products of the first step have the two helical turns favored for subsequent Dicer cleavage (MacRae et al., 2006; Gu et al., 2012). The basal UG and apical UGU primary-sequence preferences at the junctions of single-stranded and double-stranded RNA regions imply that Microprocessor recognizes either or both of these junctions at the ends of the pri-miRNA stem (Han et al., 2006; Auyeung et al., 2013). Indeed, biochemical analyses show that Drosha recognizes the basal junction and the DGCR8 dimer recognizes the apical junction (Nguyen et al., 2015). Also supporting the recognition of both junctions, results from inserting or deleting pairs within pri-miRNA stems indicate that measurements from both ends of the stem influence the site of cleavage (Ma et al., 2013). Our results showing that the efficiency of cleavage depended on a specific length of stem support the conclusion that recognition of both junctions is simultaneous and indicate that this recognition is performed by a protein complex too rigid to efficiently accommodate stems of other lengths, suggesting that the Drosha–DGCR8 heterotrimer acts as molecular calipers to measure the length of the pri-mRNA stem.

For pri-miRNAs predicted to have stems with suboptimal lengths, a stem of 35 ± 1 bp might still be achieved through disruption of extra pairing or creation of additional pairing to accommodate to Microprocessor as it binds. We found that weakened pairing at positions 37–39 favored the cleavage of pri-miR-16, presumably because this pairing must be disrupted to accommodate Microprocessor binding. Also supporting this idea were our results at position 1 of pri-miR-30. Flanked by mismatches on both sides, this single base pair would form only transiently in the context of free RNA, yet a 3p C opposite the 5p G of the basal GU motif was preferred over the other possibilities (Figure S2), presumably because forming this lone pair at position 1 to extend the stem to 35 bp favored accommodation within Microprocessor. The benefit of pairing at position 1 of pri-miR-30 was further supported by the benefit of creating a pair at position 2, which extended contiguous pairing to position 1 (Figure 2D).

Understanding the preference for a 35-bp stem and how some pri-miRNA derivatives might accommodate this structural feature better than others helps to reconcile seemingly contradictory results in the literature. For example, a 4-bp shift in the cleavage site observed after deleting 4 bp from the basal stem of pri-miR-16 has been interpreted as evidence that the distance from the base of the stem is more important for determining the cleavage site than is the distance from the loop (Han et al., 2006), which seems at odds with the conclusion from a study of pri-miR-30 derivatives (Zeng et al., 2005). We now realize, however, that the 4-bp deletion within the pri-miR-16 basal stem would favor a stem that incorporates rather than excludes the pairing at wild-type positions 36–39 to achieve an optimal length of 35 bp and that the repositioned cleavage site would fall at an optimal distance from not only the base of the stem but also the loop.

The newly identified mismatched GHG motif is basal to the region that produces the miRNA duplex. No substantial nucleotide preferences were detected in the region that produced the miRNA duplex, which from an evolutionary perspective, would benefit the emergence of new miRNAs with any primary sequence.

The mismatched GHG motif and the three previously identified primary-sequence motifs augmented the structural features to increase both efficiency and accuracy of cleavage. As suggested by our results (Figure S4A) and demonstrated for the basal UG and apical UGU (Nguyen et al., 2015), one way that the motifs increase accuracy is to break the symmetry of the single-stranded–double-stranded–single-stranded Microprocessor substrate, preventing unproductive cleavage that occurs when the substrate binds in the opposite orientation (Han et al., 2006). These four motifs exerted their greatest influence in hairpins that were suboptimal with respect to either pairing or stem length and imparted less benefit to pri-miRNAs that already had more optimal structural features (Figure 6). Likewise, the benefit of adding a motif diminished if more motifs were already present (e.g., Figures 4C and S3A). These diminishing returns implied some functional redundancy among the sequence motifs and between the structural and the sequence features.

Knowing these features that define pri-miRNAs, with awareness of their potential redundancies, explains why most natural pri-miRNAs have only a subset of these features and why the primary-sequence motifs have more impact in the context of some natural pri-miRNAs than they do in others. Pri-miR-125, which is less reliant on the primary-sequence features than is either pri-miR-16 or pri-miR-30 (Auyeung et al., 2013), differs from the other two pri-miRNAs in having an unambiguously demarcated stem of 35 bp and in having the mismatched GHG motif—two beneficial features that appear to lower the functional impact of the other features (Figure S3A). Knowing these features that define pri-miRNAs also helps explain why pri-miR-16 and pri-miR-30 respond differently to perturbations in basal and apical regions (Figures 2 and S2) (Zeng et al., 2005; Han et al., 2006; Ma et al., 2013). Pri-miR-16 appears to have a good basal stem and a less optimal apical region, whereas pri-miR-30 appears to have an optimal apical region and a less optimal basal stem. Perhaps the more optimal regions are initially recognized and provide the primary guidance for determining the cleavage site while the other regions accommodate Microprocessor binding.

Once we identified structural and sequence features that define pri-miRNAs, designing artificial pri-miRNAs that were processed more efficiently than natural human pri-miRNAs was surprisingly straightforward and reliable. This accomplishment was not overstated by comparison to human miRNAs that were processed with unusually poor efficiency. Indeed, pri-miR-125, our internal standard for the competitive-cleavage assays, has been the most efficiently processed of all natural pri-miRNAs that we have assayed in vitro, and pri-miR-1, our internal standard for accumulation of processed miRNA in vivo, accumulates to a level matching or exceeding that of any ectopically expressed miRNA that we have assessed using quantitative RNA blots (including miR-125 and miR-30). The de novo design of functional pri-miRNAs, particularly pri-miRNAs that were so efficiently processed, achieved a key milestone in the understanding of miRNA biogenesis.

The ease by which we were able to surpass processing efficiencies of natural pri-miRNAs implies that over the course of evolution, natural pri-miRNAs have not acquired the most efficient possible processing. The processing efficiency of natural pri-miRNAs might not have been optimized for several reasons. First, some natural pri-miRNAs might be constitutively inefficiently processed to enable post-transcriptional regulation through the action of differentially expressed factors that enhance processing, presumably through recognition of features beyond those characterized here (Ha and Kim, 2014). Second, mutations favoring additional production of a mature miRNA can act at any step of miRNA production, and increasing transcriptional production might be more accessible than improving post-transcriptional processing, particularly when considering the diminishing returns of adding and maintaining each additional feature that favors pri-miRNA processing. Third, pri-miRNAs that share primary transcripts with either mRNAs or other pri-miRNAs might not be optimized for processing efficiency if rapid processing compromises expression of the co-transcribed RNA. For example, Drosha processing of a pri-miRNA from a pre-mRNA intron before splice-site definition would preclude production of the mature mRNA (Kim and Kim, 2007). Likewise, because 5′-to-3′ Xrn2-mediated exonucleolytic degradation of the cleavage product downstream of Drosha processing promotes RNA polymerase II release through a torpedo-like mechanism (Ballarino et al., 2009), rapid Drosha processing of an upstream pri-miRNA could compromise transcription or stability of a downstream pri-miRNA.

For the most part, shRNAs do not face the obstacles that prevent natural pri-miRNAs from achieving more efficient processing, the exception being unwanted Drosha cleavage of retroviral RNA during packaging of shRNA libraries (Liu et al., 2010), which can be controlled by inhibiting DGCR8 (Knott et al., 2014). Accordingly, we expect that applying our design principles to improve or replace the pri-miR-30 backbone will impart advantages to future generations of shRNA libraries. In addition, our high-throughput approach for identifying generic features that define human pri-miRNAs can be modified to reveal specialized features required for regulated processing of certain mammalian pri-miRNAs, as well as the enigmatic features defining pri-miRNAs of other lineages, such as nematodes and plants.

EXPERIMENTAL PROCEDURES

Pools of pri-miRNA Variants

For each pri-miRNA, subpools of DNA oligonucleotides with all possible sequences at each of the mutagenized windows were synthesized and mixed before extension with primers that added the T7 promoter, barcodes, and Illumina adaptor sequences. All synthetic oligonucleotide sequences are provided (Table S1). The extended DNA pool for each pri-miRNA was purified, and a small fraction (10 million to 16 million molecules) was amplified in a 1-ml PCR. Some of the amplified DNA was sequenced on a HiSeq2000 (Illumina) to generate the dictionary, and some was transcribed to generate the RNA pool. To quantify the amount of each variant in the input, the barcode region of a portion of each RNA pool was sequenced. For additional details, see Supplemental Experimental Procedures.

In Vitro Cleavage and Analyses

The 5′-end-labeled pools were incubated in Microprocessor lysate, which was prepared from cells overexpressing Drosha and DGCR8 as described (Lee and Kim, 2007; Auyeung et al., 2013). After a brief incubation at 37°C, each reaction was stopped, and 5′ cleavage products were gel-purified and ligated at their 3′ ends to a pre-adenylated adaptor using T4 RNA ligase 2, truncated KQ (NEB). Ligated cleavage products were then gel-purified, reverse-transcribed, and sequenced. At each cleavage site, the cleavage score for each variant was calculated as

$$\log_2 \frac{cleaved(var)/input(var)}{cleaved(wt)/input(wt)},$$

in which *input*(*var*) and *cleaved*(*var*) were the sum of the counts from all barcodes that were linked to the variant in the input or cleavage-product sequencing, respectively, with a pseudocount of 1 added to each, and *input*(*wt*) and *cleaved*(*wt*) were the analogous sums for the wild-type sequence, respectively, including the pseudocounts. For in vitro assays of designed variants, query and reference in vitro transcribed, gel-purified, and cap-labeled pri-miRNAs were mixed and added to Microprocessor lysate. After either 2 min (assays of pri-miRNA-125 variants) or 5 min (assays of other pri-miRNA variants) at 37°C, reactions were phenol-extracted, and RNA was precipitated and resolved on urea-acrylamide gels. For additional details, see Supplemental Experimental Procedures.

pri-miRNA Processing and Mature miRNA Activity in Cells

Constructs that co-expressed the query pri-miRNA, pri-miR-1, and sometimes pri-miR-30 were transfected into HEK293T cells using Lipofectamine 2000 (Life Technologies). After 36–48 hr, total RNA was extracted, and miRNA accumulation was analyzed using RNA blots and small-RNA sequencing, whereas miRNA activity was analyzed by RNA-seq using a NEXTflex Rapid Illumina Directional RNA-Seq Library Prep Kit (Bioo Scientific). For additional details, see Supplemental Experimental Procedures.

ACCESSION NUMBERS

The accession number for the data reported in this paper is GEO: GSE67937.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, seven figures, and four tables and can be found with this article online at http://dx.doi.org/10.1016/j.molcel.2015.08.015.

ACKNOWLEDGMENTS

We thank X. Wu for the analysis of Figure 3C; V. Auyeung, S. McGeary, B. Kleaveland, I. Ulitsky, S. Eichhorn, V. Agarwal, and L. Li for valuable discussions; S. McGeary for cell lysate; J. Stefano for technical assistance; V.N. Kim and T. Tuschl for plasmids; the Whitehead Institute Genome Technology Core for sequencing; and the Whitehead Institute FACS facility for cell sorting. This work is supported by NIH GM067031. W.F. is an HHMI fellow of the Damon Runyon Cancer Research Foundation (DRG-2174-13). D.B. is an investigator of the Howard Hughes Medical Institute.

Received: April 21, 2015
Revised: July 13, 2015
Accepted: August 19, 2015
Published: September 24, 2015

REFERENCES

Auyeung, V.C., Ulitsky, I., McGeary, S.E., and Bartel, D.P. (2013). Beyond secondary structure: primary-sequence determinants license pri-miRNA hairpins for processing. Cell *152*, 844–858.

Ballarino, M., Pagano, F., Girardi, E., Morlando, M., Cacchiarelli, D., Marchioni, M., Proudfoot, N.J., and Bozzoni, I. (2009). Coupled RNA processing and transcription of intergenic primary microRNAs. Mol. Cell. Biol. *29*, 5632–5638.

Bartel, D.P. (2009). MicroRNAs: target recognition and regulatory functions. Cell *136*, 215–233.

Bassik, M.C., Kampmann, M., Lebbink, R.J., Wang, S., Hein, M.Y., Poser, I., Weibezahn, J., Horlbeck, M.A., Chen, S., Mann, M., et al. (2013). A systematic mammalian genetic interaction map reveals pathways underlying ricin susceptibility. Cell *152*, 909–922.

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.

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.

Fellmann, C., Hoffmann, T., Sridhar, V., Hopfgartner, B., Muhar, M., Roth, M., Lai, D.Y., Barbosa, I.A., Kwon, J.S., Guan, Y., et al. (2013). An optimized microRNA backbone for effective single-copy RNAi. Cell Rep. *5*, 1704–1713.

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.

Grimson, A., Farh, K.K., Johnston, W.K., Garrett-Engele, P., Lim, L.P., and Bartel, D.P. (2007). MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Mol. Cell *27*, 91–105.

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.

Gu, S., Jin, L., Zhang, Y., Huang, Y., Zhang, F., Valdmanis, P.N., and Kay, M.A. (2012). The loop position of shRNAs and pre-miRNAs is critical for the accuracy of dicer processing in vivo. Cell *151*, 900–911.

Ha, M., and Kim, V.N. (2014). Regulation of microRNA biogenesis. Nat. Rev. Mol. Cell Biol. *15*, 509–524.

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.

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

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.

Kampmann, M., Horlbeck, M.A., Chen, Y., Tsai, J.C., Bassik, M.C., Gilbert, L.A., Villalta, J.E., Kwon, S.C., Chang, H., Kim, V.N., and Weissman, J.S. (2015). Next-generation libraries for robust RNA interference-based genome-wide screens. Proc. Natl. Acad. Sci. USA *112*, E3384–E3391.

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

Kim, Y.K., and Kim, V.N. (2007). Processing of intronic microRNAs. EMBO J. *26*, 775–783.

Knott, S.R., Maceli, A.R., Erard, N., Chang, K., Marran, K., Zhou, X., Gordon, A., El Demerdash, O., Wagenblast, E., Kim, S., et al. (2014). A computational algorithm to predict shRNA potency. Mol. Cell *56*, 796–807.

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.

Liu, Y.P., Vink, M.A., Westerink, J.T., Ramirez de Arellano, E., Konstantinova, P., Ter Brake, O., and Berkhout, B. (2010). Titers of lentiviral vectors encoding shRNAs and miRNAs are reduced by different mechanisms that require distinct repair strategies. RNA *16*, 1328–1339.

Ma, H., Wu, Y., Choi, J.G., and Wu, H. (2013). Lower and upper stem-single-stranded RNA junctions together determine the Drosha cleavage site. Proc. Natl. Acad. Sci. USA *110*, 20687–20692.

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.

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

Nguyen, T.A., Jo, M.H., Choi, Y.G., Park, J., Kwon, S.C., Hohng, S., Kim, V.N., and Woo, J.S. (2015). Functional anatomy of the human Microprocessor. Cell *161*, 1374–1387.

O'Shea, J.P., Chou, M.F., Quader, S.A., Ryan, J.K., Church, G.M., and Schwartz, D. (2013). pLogo: a probabilistic approach to visualizing sequence motifs. Nat. Methods *10*, 1211–1212.

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.

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.

Silva, J.M., Li, M.Z., Chang, K., Ge, W., Golding, M.C., Rickles, R.J., Siolas, D., Hu, G., Paddison, P.J., Schlabach, M.R., et al. (2005). Second-generation shRNA libraries covering the mouse and human genomes. Nat. Genet. *37*, 1281–1288.

Song, J.J., Smith, S.K., Hannon, G.J., and Joshua-Tor, L. (2004). Crystal structure of Argonaute and its implications for RISC slicer activity. Science *305*, 1434–1437.

Winter, J., and Diederichs, S. (2011). Argonaute proteins regulate microRNA stability: increased microRNA abundance by Argonaute proteins is due to microRNA stabilization. RNA Biol. *8*, 1149–1157.

Zeng, Y., and Cullen, B.R. (2003). Sequence requirements for micro RNA processing and function in human cells. RNA *9*, 112–123.

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., Wagner, E.J., and Cullen, B.R. (2002). Both natural and designed micro RNAs can inhibit the expression of cognate mRNAs when expressed in human cells. Mol. Cell *9*, 1327–1333.

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.

Molecular Cell, Volume *60*

Supplemental Information

The Menu of Features that Define Primary MicroRNAs and Enable De Novo Design of MicroRNA Genes

Wenwen Fang and David P. Bartel

SUPPLEMENTAL FIGURES AND LEGENDS

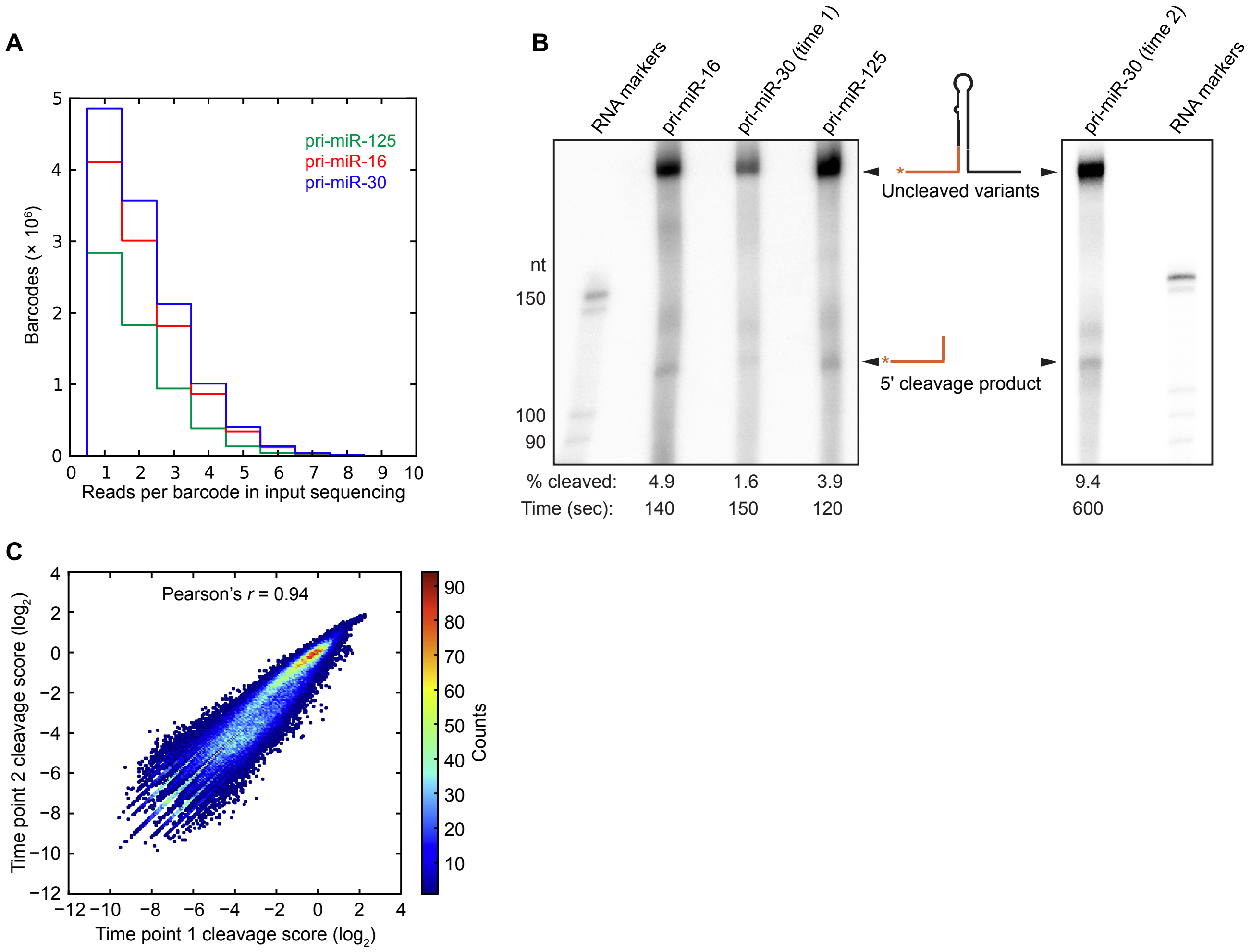


Figure S1. Additional analyses of pri-miRNA variants, related to Figure 1.
(A) Distributions of reads per barcode in input sequencing.
(B) Size selection of cleavage fragments after competitive cleavage of the pools of variants.
(C) Relationship between cleavage scores for pri-miR-30 variants obtained at two time points.

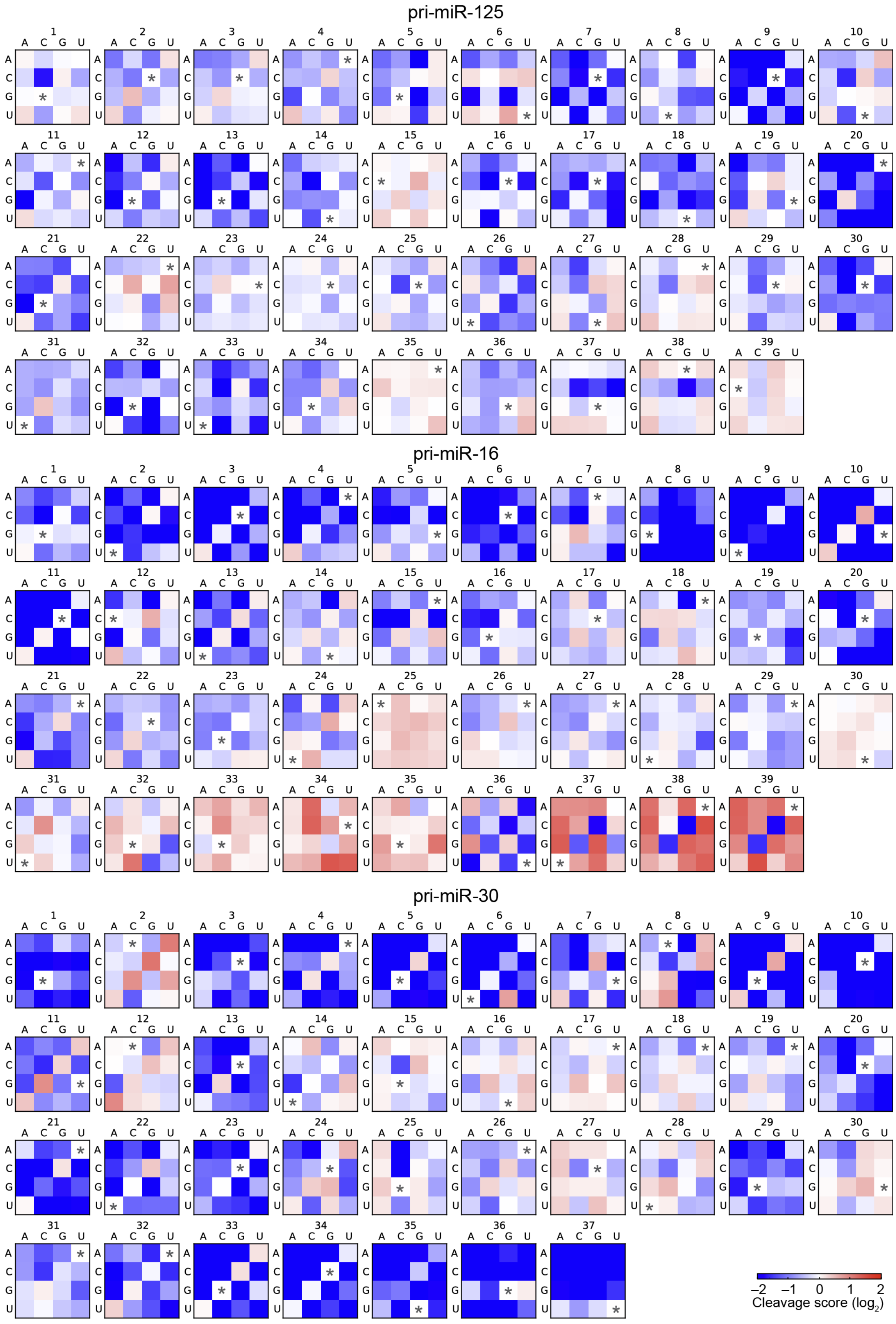


Figure S2. Cleavage scores of all single-bp variants across the stem of pri-miR-125 (top), pri-miR-16 (middle) and pri-miR-30 (bottom), related to Figure 2. Otherwise, this figure is as in Figure 2C.

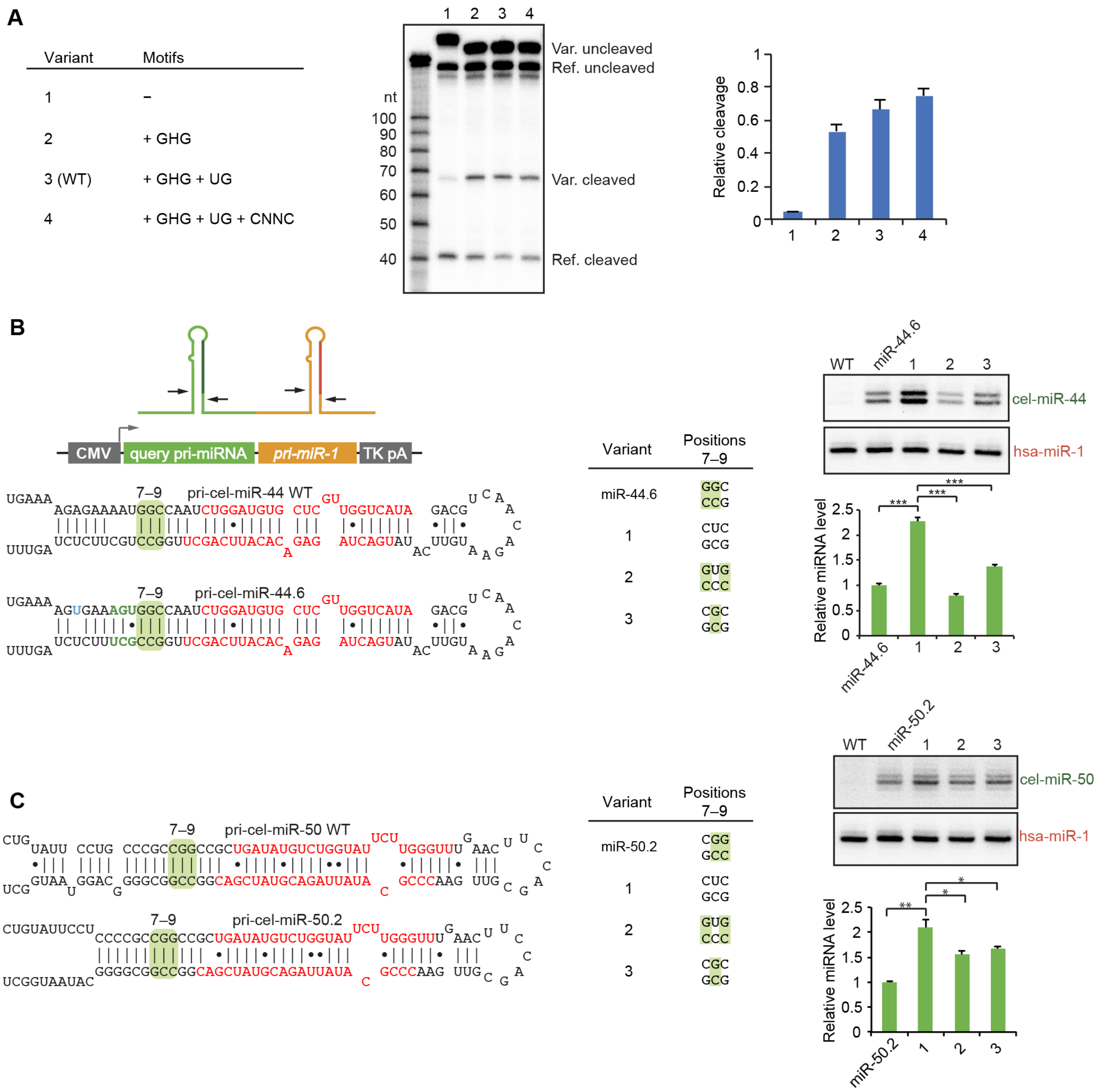
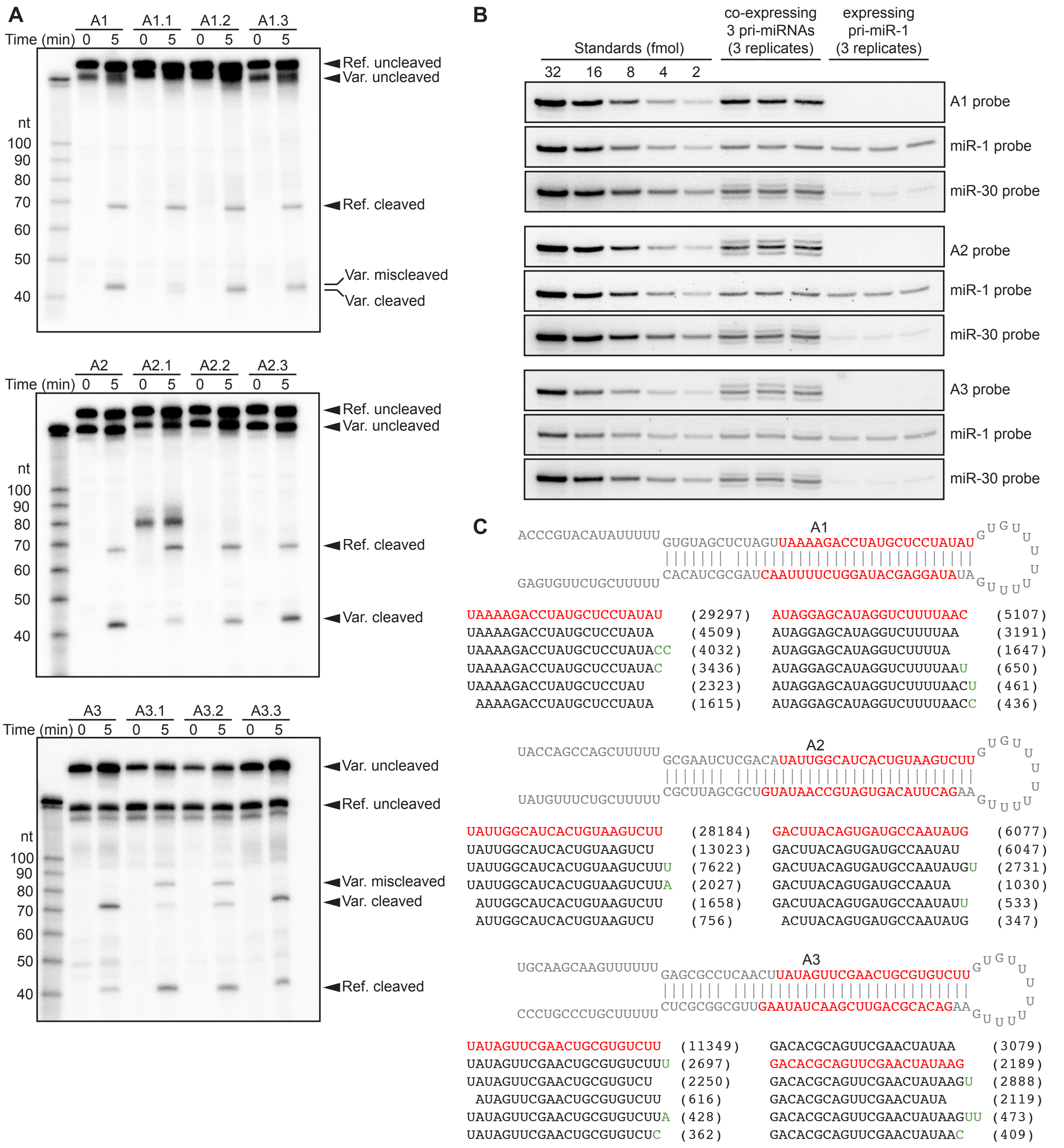


Figure S3. Additional examination of the effect of the mismatched GHG motif in vitro and in cells, related to Figure 3.

(A) Increased cleavage efficiency imparted by the mismatched GHG motif and diminishing returns of adding more motifs. In variant 1, the mismatched GHG motif of pri-miR-125 was replaced with GCG–CGC (which ranked 414 among the 4096 variants at this position in the high-throughput analysis, Table S2), and the basal UG was replaced with AC (also substituting the nucleotide originally pairing with the G of the basal UG motif to maintain Watson–Crick pairing). In variant 4, the sequence 15–18 nt downstream of the 3p Drosha cleavage site (CCACA in WT) was changed to UCUAC, which introduced a flanking CNNC motif. The assay was as described in Figure 3D. The graph (right) shows the mean relative cleavage efficiencies of each variant normalized to that of the wild-type (error bars, s.e.m., n = 2).

(B) The influence of the mismatched GHG motif on miRNA accumulation in the context of *C. elegans* pri-miR-44.6. This pri-miRNA differs from pri-miR-44.3 at stem positions 4–6 (green nucleotides), which are paired in pri-miR-44.6. Otherwise, this panel is as in Figure 3E.

(C) The influence of the mismatched GHG motif on miRNA accumulation in the context of pri-miR-50.2. This pri-miRNA differs from *C. elegans* pri-miR-50 at the basal stem, which is engineered to be of the optimal length for activity in human cells. Otherwise, this panel is as in Figure 3E.


A
A1
A1.1
A1.2
A1.3
Time (min) 0 5 0 5 0 5 0 5
nt
100
90
80
70
60
50
40
Ref. uncleaved
Var. uncleaved
Ref. cleaved
Var. miscleaved
Var. cleaved
A2
A2.1
A2.2
A2.3
Time (min) 0 5 0 5 0 5 0 5
nt
100
90
80
70
60
50
40
Ref. uncleaved
Var. uncleaved
Ref. cleaved
Var. cleaved
A3
A3.1
A3.2
A3.3
Time (min) 0 5 0 5 0 5 0 5
nt
100
90
80
70
60
50
40
Var. uncleaved
Ref. uncleaved
Var. miscleaved
Var. cleaved
Ref. cleaved
B
Standards (fmol)
32 16 8 4 2
co-expressing 3 pri-miRNAs (3 replicates)
expressing pri-miR-1 (3 replicates)
A1 probe
miR-1 probe
miR-30 probe
A2 probe
miR-1 probe
miR-30 probe
A3 probe
miR-1 probe
miR-30 probe
C
A1
ACCCGUACAUAUUUUU GUGUAGCUCUAGUUAAAAGACCUAUGCUCCUAUAU GUGUUUU
GAGUGUUCUGCUUUUU CACAUCGCGAUCAAUUUUCUGGAUACGAGGAUAUA GUUUU
A2
UACCAGCCAGCUUUUU GCGAAUCUCGACAUAUUGGCAUCACUGUAAGUCUU GUGUUUU
UAUGUUUCUGCUUUUU CGCUUAGCGCUGUAUAACCGUAGUGACAUUCAGAA GUUUU
A3
UGCAAGCAAGUUUUUU GAGCGCCUCAACUUAUAGUUCGAACUGCGUGUCUU GUGUUUU
CCCUGCCCUGCUUUUU CUCGCGGCGUUGAAUAUCAAGCUUGACGCACAGAA GUUUU

UAAAAGACCUAUGCUCCUAUAU	(29297)	AUAGGAGCAUAGGUCUUUUAAC	(5107)
UAAAAGACCUAUGCUCCUAUA	(4509)	AUAGGAGCAUAGGUCUUUUAA	(3191)
UAAAAGACCUAUGCUCCUAUACC	(4032)	AUAGGAGCAUAGGUCUUUUA	(1647)
UAAAAGACCUAUGCUCCUAUAC	(3436)	AUAGGAGCAUAGGUCUUUUAAU	(650)
UAAAAGACCUAUGCUCCUAU	(2323)	AUAGGAGCAUAGGUCUUUUAACU	(461)
AAAAGACCUAUGCUCCUAUA	(1615)	AUAGGAGCAUAGGUCUUUUAACC	(436)

UAUUGGCAUCACUGUAAGUCUU	(28184)	GACUUACAGUGAUGCCAAUAUG	(6077)
UAUUGGCAUCACUGUAAGUCU	(13023)	GACUUACAGUGAUGCCAAUAU	(6047)
UAUUGGCAUCACUGUAAGUCUUU	(7622)	GACUUACAGUGAUGCCAAUAUGU	(2731)
UAUUGGCAUCACUGUAAGUCUUA	(2027)	GACUUACAGUGAUGCCAAUA	(1030)
AUUGGCAUCACUGUAAGUCUU	(1658)	GACUUACAGUGAUGCCAAUAUU	(533)
AUUGGCAUCACUGUAAGUCU	(756)	ACUUACAGUGAUGCCAAUAUG	(347)

UAUAGUUCGAACUGCGUGUCUU	(11349)	GACACGCAGUUCGAACUAUAA	(3079)
UAUAGUUCGAACUGCGUGUCUUU	(2697)	GACACGCAGUUCGAACUAUAAG	(2189)
UAUAGUUCGAACUGCGUGUCU	(2250)	GACACGCAGUUCGAACUAUAAGU	(2888)
AUAGUUCGAACUGCGUGUCUU	(616)	GACACGCAGUUCGAACUAUA	(2119)
UAUAGUUCGAACUGCGUGUCUUA	(428)	GACACGCAGUUCGAACUAUAAGUU	(473)
UAUAGUUCGAACUGCGUGUCUC	(362)	GACACGCAGUUCGAACUAUAAC	(409)

Figure S4. Additional anayses of artificial pri-miRNAs, related to Figure 4.

(A) Representative competitive cleavage assays of A1, A2, A3, and their variants, using pri-miR-125 as an internal reference. Assays are as in Figure 3D, except for some variants products of miscleavage were also observed. See Figure 4C for quantification.

(B) Quantitative RNA blots probing for A1, A2, A3 and the co-expressed natural miRNAs. Blots included lanes with known amounts of synthetic standards, which enabled absolute quantification of each miRNA. The lanes from cells expressing only miR-1 served as a specificity control for the artificial miRNAs and enabled quantification of endogenous miR-30. The amount of endogenous miR-30 was subtracted from the amount observed when expressing pri-miR-30. See Figure 4E for quantification.

(C) Small-RNA sequencing results quantifying the major products of the artificial pri-miRNAs. For each product, the number of reads is in parenthesis. The expected miRNA and miRNA* sequences are red, untemplated residues are green.

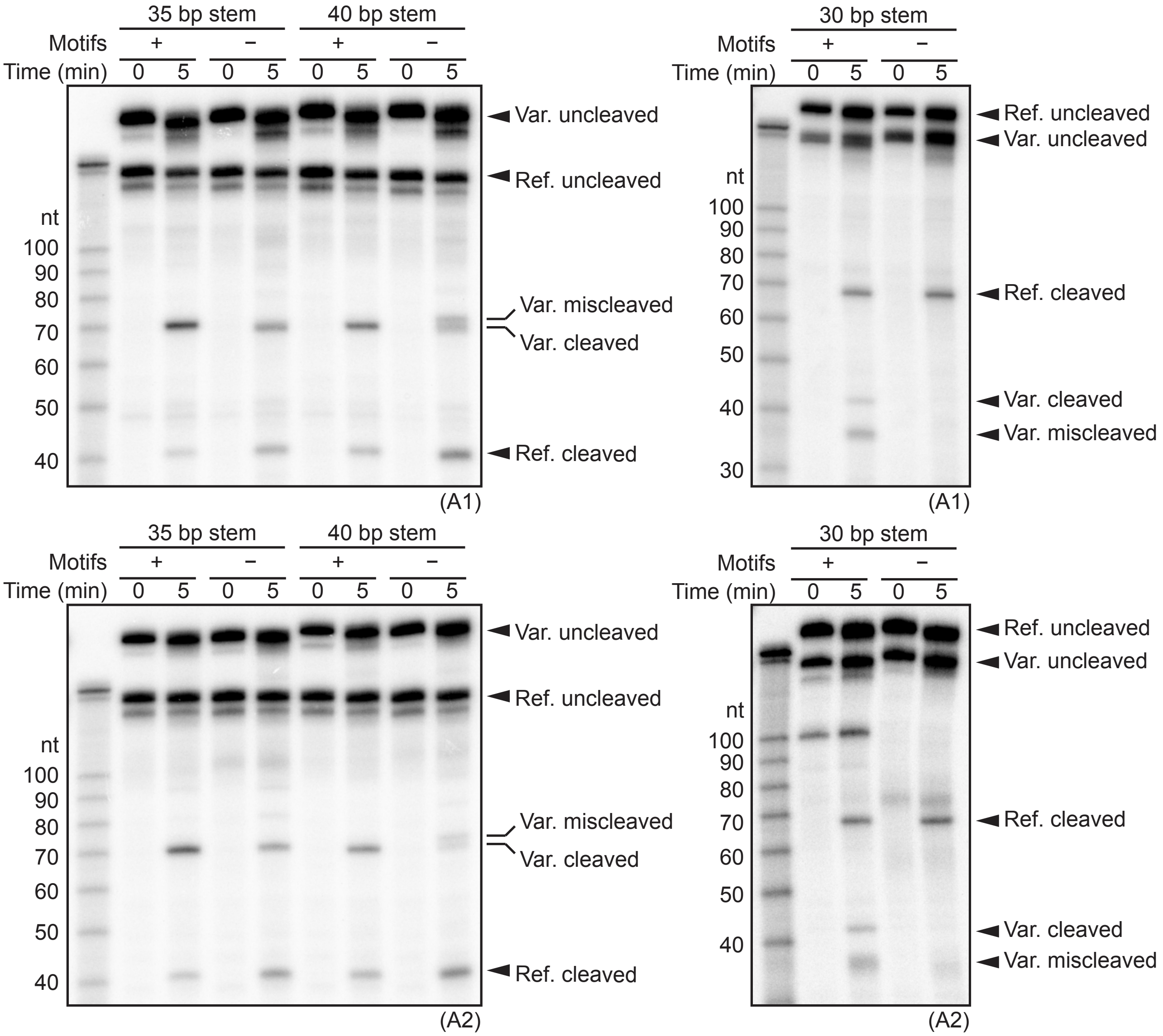


Figure S5. Sequence motifs rescue suboptimal stem lengths, related to Figure 5.
Representative competitive cleavage assays of extension (40 bp stem) and deletion (30 bp stem) variants of A1 and A2. Assays were as in Figure 3D, except some of the variants are longer than the internal reference, and for some variants products of miscleavage are observed. See Figure 5B for quantification.

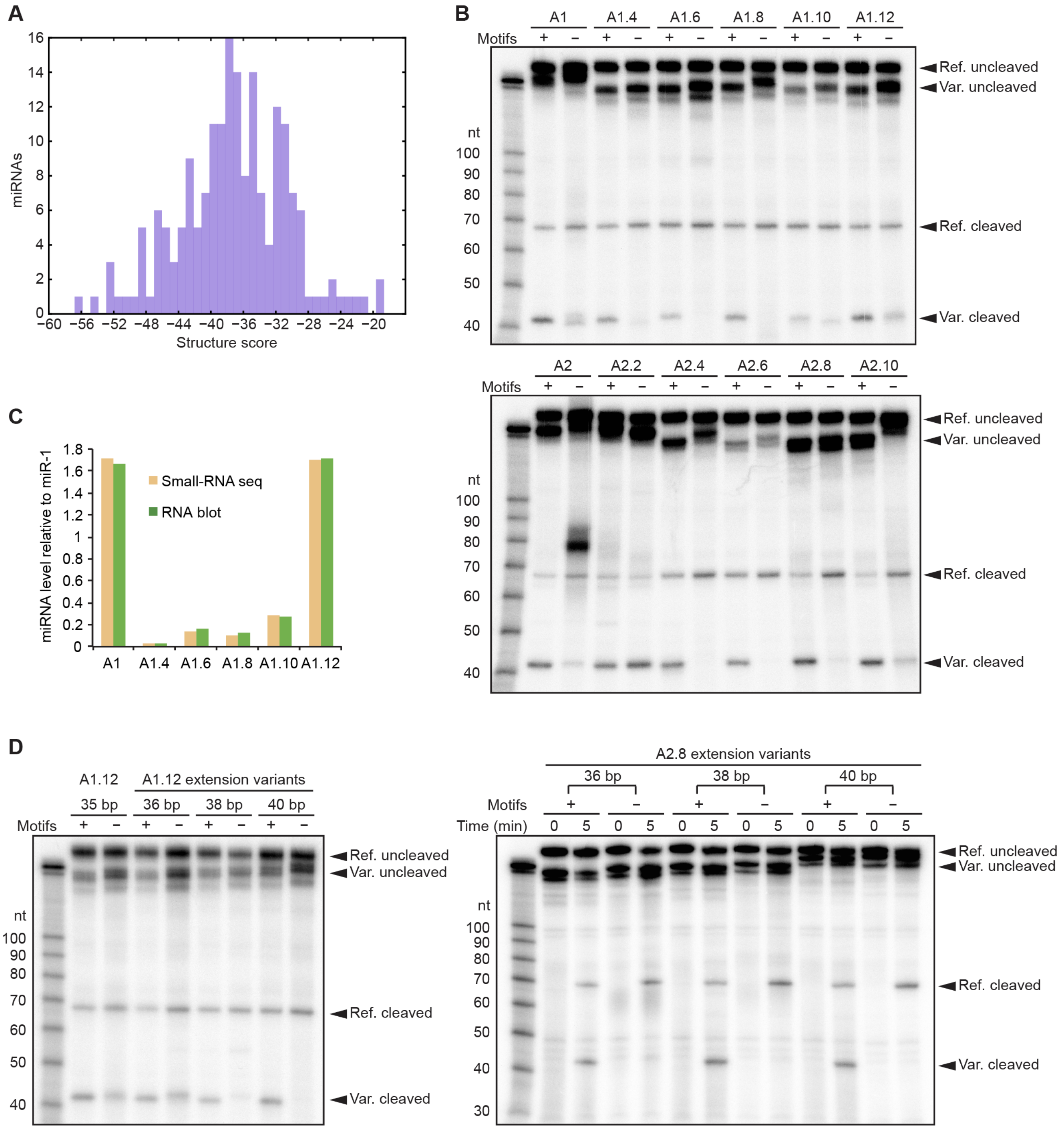


Figure S6. Sequence motifs rescue structural defects, related to Figure 6.

(A) The distribution of structure scores for representative members of 186 conserved human pri-miRNAs families (Table S4).
(B) Representative competitive cleavage assays of A1 and A2 variants, with or without motifs. Otherwise, this panel is as in Figure 3D. See Figure 6C for quantification.
(C) Abundance of mature miRNAs relative to the miR-1 internal standard, measured using either small-RNA sequencing (orange) or quantitative RNA-blot analyses (green). Small-RNA seq counted 20–25 nt reads mapping to the region of the mature miRNA; RNA-blot values are the same as those shown in Figure 6D.
(D) Competitive cleavage assays of extension variants of A1.12 and A2.8, with or without motifs. Otherwise, this panel is as in Figure 3D. See Figure 6F for quantification.

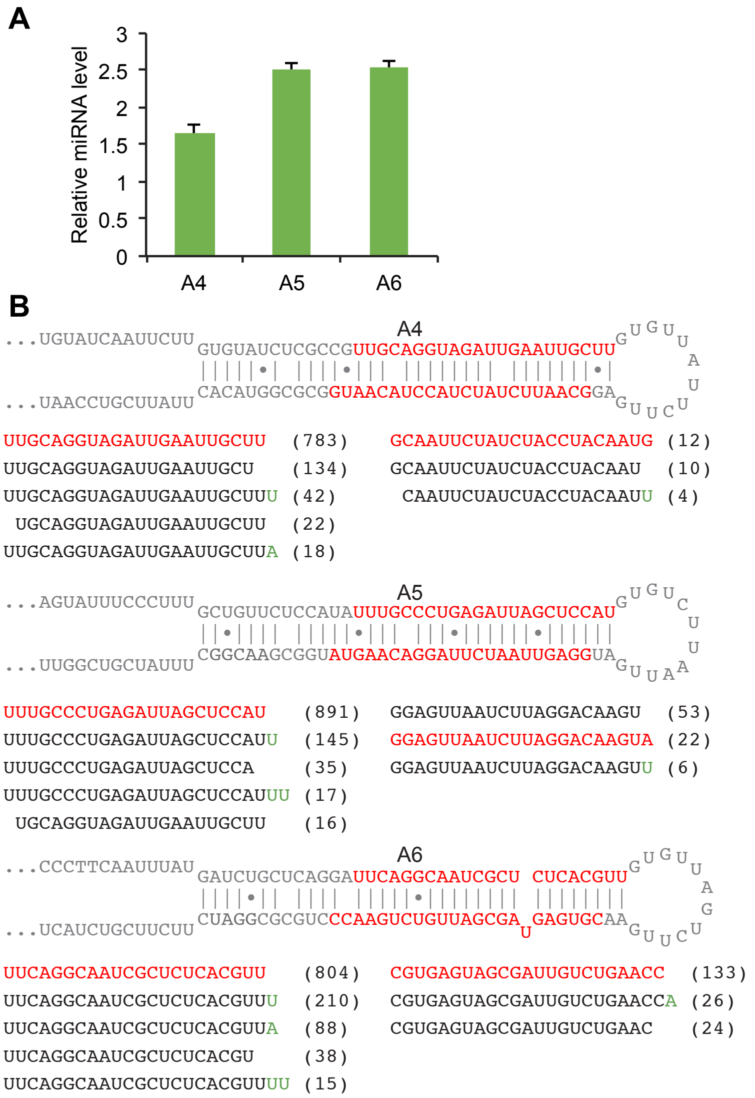


Figure S7. Additional analyses of artificial miRNAs, related to Figure 7.
(A) Accumulation of mature miRNAs from the indicated pri-miRNAs, relative to the miR-1 internal standard, in HEK293T cells. Assays were as in Figure 3E, but with synthetic miRNA standards for quantitative RNA-blot analyses.
(B) Small-RNA sequencing results quantifying the major products of the artificial pri-miRNAs. For each product, the number of reads is in parenthesis. The expected miRNA and miRNA* sequences are red, and untemplated residues are green.

SUPPLEMENTAL TABLES

Table S1. Oligonucleotides, related to Experimental Procedures.

Table S2. Cleavage scores of variants with mutations at positions 7 to 9, related to Figure 3.

Table S3. Representative animal pri-miRNAs, related to Figure 3C.

Table S4. Pairing and other structural elements across the 35-bp stem of representative members of conserved human pri-miRNAs families, related to Figure 6.

EXTENDED EXPERIMENTAL PROCEDURES

Preparation and analysis of pools of pri-miRNA variants

For each pri-miRNA, subpools of DNA oligonucleotides designed to have an equal mixture of all possible sequences at each of the mutagenized windows but otherwise corresponding to the pri-mRNA hairpin and its flanking sequences were synthesized and purified on urea-polyacrylamide gels. Subpools for each pri-miRNA were mixed and extended at each end with successive primer-extension reactions using KAPA HiFi PCR kit (KAPA Biosystems) and oligonucleotides that added the T7 promoter, barcodes and Illumina adapter sequences. The extended DNA products were purified on formamide-polyacrylamide gels and quantified using Library Quantification Kits (KAPA Biosystems). A bottleneck was then imposed, drawing 10 million, 16 million, and 16 million molecules from the pri-miR-125, pri-miR-16, and pri-miR-30 pools, and these samples were amplified for 19 cycles in 1-ml PCR reactions using KAPA HiFi PCR kit. After purification using a QIAquick PCR Purification Kit (Qiagen), a small fraction of each pool was sequenced on a HiSeq2000 (100 × 100 bp paired-end mode) to create the dictionary, and another fraction was transcribed in vitro using home-made T7 polymerase. The resulting RNA pools were purified on urea-polyacrylamide gels, dephosphorylated by calf intestine phosphatase (NEB), and trace labeled with ATP-γ-^{32}P (Perkin Elmer).

To sample the barcodes of each RNA pool and thereby quantify the amount of each variant in the input, the barcode region of a portion of each RNA pool was sequenced as follows. The region was reverse transcribed using AffinityScript reverse transcriptase (Agilent Technologies) and a primer that paired to the constant region between the barcode and the hairpin. This primer also added an adapter sequence for Illumina sequencing. After base hydrolysis of RNA and desalting using Micro Bio-Spin P-30 Gel Columns (BioRad), second-strand synthesis was performed using the KAPA HiFi PCR kit, which added the Illumina adapter sequence to the 5' end. Products were purified using the Agencourt AMPure XP system (Beckman Coulter), and barcode regions were sequenced on a HiSeq2000 (40 bp single-end mode).

In vitro cleavage and high-throughput sequencing

5'-end-labeled pools (25 pmol) were incubated at 37°C in a 250 µl cleavage reaction containing 200 µl reaction buffer (100 mM KCl, 1 mM $MgCl_2$, 20 mM Tris-Cl pH 8.0, 0.2 mM EDTA, 5 mM DTT), 0.3 µg/µl yeast RNA (Life Technologies) and 25 µl HEK293T whole-cell lysate overexpressing FLAG-tagged Drosha (Lee and Kim, 2007) and FLAG-HA-tagged DGCR8 (Landthaler et al., 2004), which was prepared as described (Auyeung et al., 2013). At one or two time points (120 seconds for miR-125, 140 seconds for miR-16, 2.5 minutes, and 10 minutes for miR-30) reactions were stopped by addition of ice-cold acid phenol (Life Technologies) with mixing, and the 5' cleavage products of each pool were purified on urea-polyacrylamide gels and ligated at their 3' ends to a pre-adenylated adapter using T4 RNA ligase 2, truncated KQ (NEB), in the presence of three oligonucleotides that paired to constant regions of the cleavage fragments to prevent their circularization (10 µM adapter, 0.67 µM of each of the protection oligonucleotides, 10% PEG 8000, 10U SUPERase• In (Life Technologies); otherwise, as recommended by the manufacturer). Ligated cleavage products were purified on urea-polyacrylamide gels and reverse transcribed. After second-strand synthesis with the KAPA HiFi PCR kit, which completed the 5' Illumina adapter, DNA was purified and sequenced, as done for the input barcodes, except that sequencing used the 100 × 100 bp paired-end mode (although only read 1 was used for analysis).

Sequence analysis

To create each dictionary of barcode–variant linkages, paired-end reads were first joined using fastq-join, allowing a maximum difference of 5% (-p 5) in the overlap. The joined reads were quality filtered using fastq_quality_filter, requiring a minimum quality score of 30 for 90% of the sequences (-q 30 -p 90). 3' adapter was trimmed using cutadapt (default parameters), discarding untrimmed sequences. A second cutadapt command split the barcode-variant pairs using the constant sequences linking them, generating an info-file from which the barcode-variant pairs could be extracted. From 47, 36, and 41 million raw paired-end reads for pri-miR-125, pri-miR-16, and pri-miR-30, respectively, 7.2, 9.5, and 11.7 million unique barcode–variant linkages were obtained. Of the sequenced barcodes, small fractions were excluded from the dictionary (4.4 %, 1.8%, 2.1% for pri-miR-125, pri-miR-16, and pri-miR-30, respectively) because they were associated with multiple pri-miRNA sequences, which were attributed primarily to sequencing errors, as indicated by the sequence-quality scores at positions that differed between sequences.

The read for the input barcodes were first adapter-trimmed using cutadapt, requiring a length between 25 and 35 nt after trimming (-m 25 -M 35). The trimmed reads were quality-filtered using fastq_quality_filter, requiring a minimum quality score of 30 for all bases in the read (-q 30 -p 100). The trimmed sequences were barcodes that could be linked to variants using dictionaries generated above. Read 1 from the cleavage-fragment sequencing was quality-filtered using fastq_quality_filter, requiring a minimum quality score of 30 for 95% of the sequence (-q 30 -p 95). Barcode–cleavage product pairs were split using cutadapt and constant sequences linking them to generate an info-file from which the barcode–cleavage product pairs could be extracted. For each variant, the cleavage score was calculated as

$$\log_2 \frac{cleaved(var)/input(var)}{cleaved(wt)/input(wt)}$$

where *input(var)* and *cleaved(var)* were the sum of the counts from all barcodes that were linked to the variant in the input or cleavage-product sequencing, respectively, with a pseudocount of 1 added to each; and *input(wt)* and *cleaved(wt)* were the analogous sums for the wild-type sequence, including the pseudocounts.

Analyses of natural pri-miRNAs

The lists of pri-miRNAs were as described (Auyeung et al., 2013), except that the set of representative human pri-miRNA conserved in mouse was further curated to remove Drosha-independent miRNAs and miRNAs not expressed in other mammals, and the cleavage sites of human pri-miRNAs were manually curated using published high-throughput sequencing data (Table S3). The occurrence of the mismatched GHG motif in each 3-bp window across the stem was counted after extending each annotated pre-miRNA 20 nt upstream and downstream (Table S3) and then folding it using RNAfold in ViennaRNA package (version 2.1.1) (Lorenz et al., 2011) with default settings. For each human pri-miRNA, a 35-bp stem was identified based on the site of Drosha cleavage (counting mismatches and wobbles, but not bulged nucleotides, as pairs), from which all base pairs, mismatches, and bulges were counted manually (Table S4).

Cloning of pri-miRNAs

Plasmids encoding human pri-miR-125 and *C. elegans* pri-miRNA variants used in Figures 3D, 3E, S3B and S3C were made using the QuikChange Site-Directed Mutagenesis Kit (Agilent Technologies) to modify previous constructs that fused the query pri-miRNA upstream of pri-miR-1 (Auyeung et al., 2013). Plasmids encoding human pri-miR-125 variants used in Figure S3A were synthesized as double-stranded DNA fragments (gBlocks from IDT) and recombined into pcDNA3.2 V5-DEST mir-1 reporter (Addgene Plasmid #46646) using Gateway pDonor221 vector and LR clonase II Enzyme mix (Life Technologies). DNA fragments encoding artificial pri-miRNAs were synthesized from primer-extension and amplification of synthetic oligonucleotides. The amplified DNA fragments were recombined into pcDNA3.2 V5-DEST mir-1 reporter as described above. The cloning reactions were transformed into MAX Efficiency DH5α Competent Cells (Life Technologies), except for constructs encoding perfect hairpins, for which either One Shot Stbl3 (Life Technologies) or NEB Stable Competent *E. coli* were used for transformation. Plasmids were prepared using the Qiagen Plasmid Midi Kit and sequences were confirmed by Sanger sequencing (Genewiz), using the alternative protocol and sequencing both strands when necessary.

Cleavage assays

Pri-miRNAs with desired lengths of 5'-cleavage fragments were prepared by T7 in vitro transcription of PCR products in which the appropriate regions of sequenced plasmids had been amplified. After transcription, RNA was purified on urea-acrylamide gels, cap-labeled with GTP-α-^{32}P (Perkin Elmer) using the Vaccinia Capping System (NEB), phenol extracted, ethanol precipitated, resuspended in water and quantified using Qubit Fluorometric Quantification (Life Technologies). Query and reference pri-miRNAs were heated in water (70°C, 15 minutes), slow-cooled to room temperature, and then mixed at 0.5 µM each before starting the reaction with 10-fold dilution into the other assay components, as described for the pri-miRNA pools but using 10 µl reactions. Incubation was for 5 minutes at 37°C, except for assays examining pri-miRNA-125 variants (Figure 3D), which were for 2 minutes. Reactions were stopped by addition of ice-cold acid phenol. After phenol extraction and ethanol precipitation, unprocessed and processed RNAs were resolved on urea-acrylamide gels.

Measuring miRNA accumulation in HEK293T cells

Constructs that co-expressed the query pri-miRNA and pri-miR-1 (and sometimes also pri-miR-30) under the CMV promoter were transfected into HEK293T cells using Lipofectamine 2000 (Life Technologies) together with a plasmid expressing GFP (pMAX-GFP) that was used for estimating transfection efficiencies. Cells were harvested after 36–48 hrs, and total RNA was extracted using TRIzol Reagent (Life Technologies). Detailed protocols for small-RNA blots and small-RNA sequencing are available at http://bartellab.wi.mit.edu/protocols. Modifications to the protocols were as follows. For absolute quantification of small RNAs on RNA blots, 2–32 fmol of synthetic standards were diluted in 0.1 µg/µl yeast RNA before loading. For small-RNA sequencing, T4 RNA Ligase 2, truncated KQ was used for the 3' ligation, and 10% PEG 8000 was included in both the 3' and 5' ligations. After reverse transcription by SuperScriptIII (Life Technologies), KAPA HiFi PCR kit was used to amplify the libraries. Adapter sequences were trimmed using cutadapt (default parameters) and quality-filtered by requiring all bases to have a minimum score of 10 (-q 10 -p 100). Additional random nucleotides that derived from the 5' and 3' adapters were removed, and the resulting reads were collapsed while recording read counts and mapped to the pri-miRNAs using Geneious 6.1.2 (Kearse et al., 2012). Figures show the mapped reads with the most counts.

Analyzing the function of artificial miRNAs

At 48 hr post-transfection, cells were trypsinized and sorted by GFP signal on a FACSAria IIU SORP. The top 30% of the fluorescent cells were isolated, and total RNA from these cells was extracted using TRIzol Reagent. RNA was polyA-selected using Dynabeads Oligo (dT)$_{25}$ (Life Technologies), and RNA-seq libraries were made using a NEXTflex Rapid Illumina Directional RNA-Seq Library Prep

Kit (Bioo Scientific) and sequenced on a HiSeq2000 (40 bp single-end mode). Reads were quality-filtered using fastq_quality_filter, requiring a minimum quality score of 30 for 90% of the sequences (-q 30 -p 90), and then mapped to the RefSeq mRNAs (considering for each gene the transcript with the longest open reading frame) using bowtie (Langmead et al., 2009), requiring unique mapping with ≤1 mismatch. Only reads mapping to the correct strand were considered, and only transcripts with more than 20 mapped reads per million total mapped reads were used to calculate expression fold changes.

SUPPLEMENTAL REFERENCES

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., et al. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics *28*, 1647-1649.

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.

Lorenz, R., Bernhart, S.H., Honer Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F., and Hofacker, I.L. (2011). ViennaRNA Package 2.0. Algorithms Mol. Biol. *6*, 26.