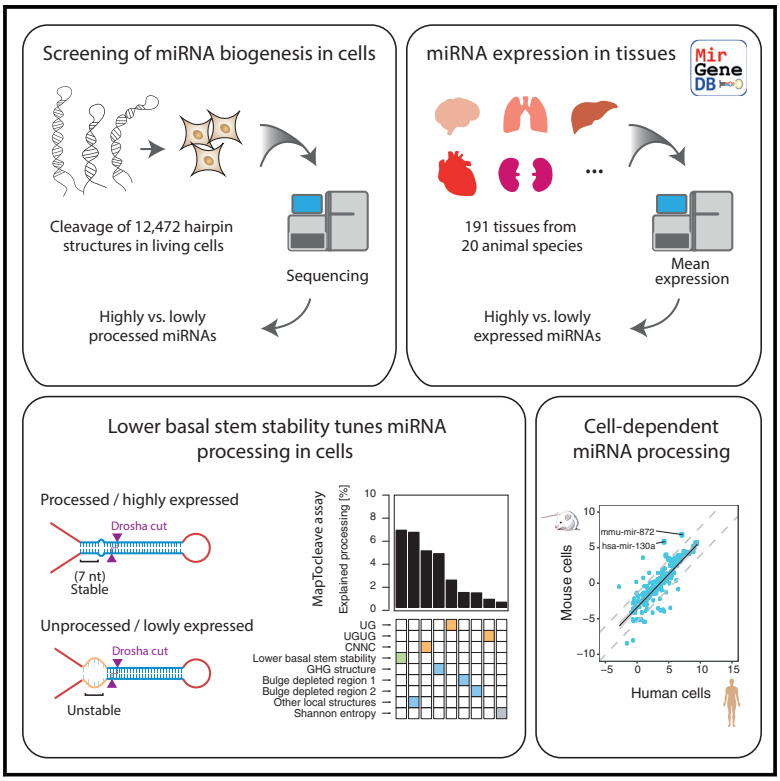
Cell Reports

Article

MapToCleave: High-throughput profiling of microRNA biogenesis in living cells

Graphical abstract



Highlights

- MapToCleave method allows simultaneous screening of 12,472 RNA structures in cells
- The biogenesis of ~15% of human miRNAs is influenced by cell-dependent factors
- We perform a systematic comparison of the importance of miRNA biogenesis features
- Stability of first seven base pairs of the stem tunes processing in cells and tissues

Authors

Wenjing Kang, Bastian Fromm, Anna J. Houben, ..., Rory Johnson, Inna Biryukova, Marc R. Friedländer

Correspondence

marc.friedlander@scilifelab.se

In brief

Numerous miRNA features that facilitate biogenesis are known, but most have been identified *in vitro*. Kang et al. re-evaluate miRNA biogenesis in living cells and in animal tissues, and they find that stability of the first seven base pairs of the stem is particularly important for processing in cells.

Kang et al., 2021, Cell Reports *37*, 110015
November 16, 2021
https://doi.org/10.1016/j.celrep.2021.110015

CellPress

Cell Reports

CellPress
OPEN ACCESS

Article

MapToCleave: High-throughput profiling of microRNA biogenesis in living cells

Wenjing Kang,[1] Bastian Fromm,[1,11] Anna J. Houben,[2] Eirik Høye,[3] Daniela Bezdan,[2,4,5] Carme Arnan,[2] Kim Thrane,[6] Michaela Asp,[6] Rory Johnson,[7,8,9,10] Inna Biryukova,[1] and Marc R. Friedländer[1,12,*]

[1]Science for Life Laboratory, Department of Molecular Biosciences, The Wenner-Gren Institute, Stockholm University, Stockholm, Sweden
[2]Centre for Genomic Regulation (CRG), The Barcelona Institute for Science and Technology, Barcelona (BIST), Catalonia, Spain
[3]Department of Tumor Biology, Oslo Institute for Cancer Research, Oslo University Hospital, Oslo, Norway
[4]Institute of Medical Genetics and Applied Genomics, University of Tübingen, Tübingen, Germany
[5]NGS Competence Center Tübingen (NCCT), University of Tübingen, Tübingen, Germany
[6]Department of Gene Technology, School of Engineering Sciences in Chemistry, Biotechnology and Health, KTH Royal Institute of Technology, Science for Life Laboratory, Solna, Sweden
[7]Department of Medical Oncology, Inselspital, Bern University Hospital, University of Bern, Bern, Switzerland
[8]Department for BioMedical Research, University of Bern, Bern, Switzerland
[9]School of Biology and Environmental Science, University College Dublin, Dublin, Ireland
[10]Conway Institute for Biomolecular and Biomedical Research, University College Dublin, Dublin, Ireland
[11]The Arctic University Museum of Norway, UiT - The Arctic University of Norway, Tromsø, Norway
[12]Lead contact
*Correspondence: marc.friedlander@scilifelab.se
https://doi.org/10.1016/j.celrep.2021.110015

SUMMARY

Previous large-scale studies have uncovered many features that determine the processing of microRNA (miRNA) precursors; however, they have been conducted *in vitro*. Here, we introduce MapToCleave, a method to simultaneously profile processing of thousands of distinct RNA structures in living cells. We find that miRNA precursors with a stable lower basal stem are more efficiently processed and also have higher expression *in vivo* in tissues from 20 animal species. We systematically compare the importance of known and novel sequence and structural features and test biogenesis of miRNA precursors from 10 animal and plant species in human cells. Lastly, we provide evidence that the GHG motif better predicts processing when defined as a structure rather than sequence motif, consistent with recent cryogenic electron microscopy (cryo-EM) studies. In summary, we apply a screening assay in living cells to reveal the importance of lower basal stem stability for miRNA processing and *in vivo* expression.

INTRODUCTION

MicroRNAs (miRNAs) are short RNA molecules with important roles in animal gene regulation (Bartel, 2018). Since it has been estimated that mRNAs from more than 60% of all human genes are regulated by miRNAs in one or more cellular contexts (Friedman et al., 2009), it is not surprising that these molecules have been found to play roles in biological processes, ranging from development (Giraldez et al., 2005) and formation of cell identity (Lim et al., 2005) to various diseases, including neurological illnesses and cancer (Esteller, 2011). Mutant animals that are completely devoid of miRNAs either die at early developmental stages (mice) or develop severe developmental defects (zebrafish; Bernstein et al., 2003; Giraldez et al., 2005).

In the canonical biogenesis pathway, miRNA primary transcripts are transcribed by RNA polymerase II, often as molecules that are tens of thousands of nucleotides long (Cai et al., 2004; Lee et al., 2002, 2004). Each primary transcript harbors one or more hairpin fold-back structures, which are recognized by Drosha and its binding partner DGCR8 in the nucleus (Han et al., 2006). Drosha cleaves out the ~60-nt-long miRNA precursor, which is exported to the cytoplasm by Exportin-5 (Bohnsack et al., 2004; Lund et al., 2004; Okada et al., 2009; Yi et al., 2003). In the cytoplasm, the precursor is recognized and cleaved by Dicer, which is part of the canonical RNA interference pathway, thus releasing an ~22-nt-long RNA duplex (Bernstein et al., 2001; Hutvágner et al., 2001; Ketting et al., 2001; Knight and Bass, 2001). Subsequently, one of the strands of the duplex is selectively loaded into the Argonaute protein, which is a key component of the miRISC effector complex (Iwasaki et al., 2010). Once bound to Argonaute, the mature miRNA can guide the complex by partial base complementarity to target mRNAs, which are then degraded through de-adenylation and de-capping or are translationally inhibited through obstruction of translation initiation (Bartel, 2009). There are numerous non-canonical miRNA biogenesis pathways (Ha and Kim, 2014); these all share the presence of a precursor (pre-)miRNA hairpin structure and binding by an Argonaute effector protein promoting mRNA repression.

It has been estimated that the human genome harbors more than 400,000 regions that could give rise to hairpin structures if

transcribed (Bentwich et al., 2005). In contrast, the number of human precursors is estimated to be between 556 (Fromm et al., 2020) and 3,000 (Friedländer et al., 2014), suggesting that the hairpins that enter miRNA biogenesis pathways are stringently selected. Many studies have evaluated hairpin features that license miRNA biogenesis. These assays have measured hairpin cleavage *in vitro*, testing numerous variants of a limited number of distinct hairpins (Auyeung et al., 2013; Fang and Bartel, 2015; Kwon et al., 2019; Li et al., 2020a). Through comparison of the variants that were processed and unprocessed, a number of structural features and sequence motifs have been identified. The overall structure with two single-stranded flanking sequences, an ~35-nt double-stranded stem, and a single-stranded apical loop is the key entry point into miRNA biogenesis (Fang and Bartel, 2015; Han et al., 2006). The sequence motifs UG at the basal junction, UGUG at the apical junction, and CNNC at the 3′ flanking sequence have been reported to facilitate Drosha processing (Auyeung et al., 2013; Fang and Bartel, 2015). Recent studies have further found that miRNA precursors tend to have bulge-depleted regions in the upper and lower part of the miRNA duplex (Roden et al., 2017) and that bulges in the lower and middle part of the miRNA duplex influence Drosha processing efficiency and/or precision (Li et al., 2020b, 2020a). Other studies have shown that the GHG motif, defined as an unmatched nucleotide other than guanosine that is flanked by two base-paired guanosines at position −7 to −5 relative to the Drosha cleavage site, can facilitate miRNA precursor processing efficiency and precision (Fang and Bartel, 2015). However, there is some evidence that the GHG motif is better defined as a catalog of sequence/structure combinations (Kwon et al., 2019), and a recent cryoelectron microscopy (cryo-EM) study points to the importance of the structure itself (Jin et al., 2020).

Previous studies have been limited in that variants of only a few miRNA precursors have been tested, leaving open the possibility that some important biogenesis features may remain undiscovered. One recent study partly overcame this limitation by testing thousands of distinct RNA structures at the same time, providing evidence that structural uncertainty, measured as Shannon entropy, negatively influences processing (Rice et al., 2020). However, this experiment was conducted *in vitro*, so the contribution of the cellular context to miRNA biogenesis remains unstudied on a large scale.

Here, we present MapToCleave, a novel method that can measure the processing of thousands of distinct RNA structures in living cells in a single experiment, recapitulating the details of natural miRNA biogenesis. Our approach is comparable to the one used by Chiang et al. (2010) to distinguish bona fide miRNAs from likely false annotations. We are expanding on this previous pioneering work to profile >10,000 structures in one experiment while Chiang et al. profiled up to 10 structures per experiment. We find that miRNA precursors undergo differential processing in different cell types, underlining the importance of cell type-dependent processing. We also provide evidence that the precursors that are efficiently processed in our assay are significantly enriched in stable lower basal stem structures. We further extend this to *in vivo* conditions, showing that highly expressed miRNAs also tend to have stable lower basal stems in mammals, fruit flies, and Lophotrochozoans, animal groups that are separated by >600 million years of evolution. Comparing the importance of known and novel features in predicting miRNA processing efficiency and *in vivo* expression, the lower basal stem ranks higher than several of the known sequence and structural motifs. Surprisingly, the known and novel features together explain only ~20% of miRNA processing. Lastly, we provide evidence that the GHG motif defined as a structure motif is a better predictor of miRNA processing efficiency and precision than is the motif defined as a sequence motif, supporting a recent cryo-EM study (Jin et al., 2020). In summary, our study extends the current model of miRNA biogenesis by revealing the lower basal stem to be an important structure that can tune miRNA processing and expression.

RESULTS

MapToCleave measures processing of thousands of distinct RNA structures in cells

To systematically study miRNA biogenesis, we developed a novel high-throughput screening method—massively parallel testing of hairpin cleavage (MapToCleave)—which we applied in a single experiment to simultaneously profile the processing of 12,472 distinct RNA structures in living cells. These structures include bona fide human miRNA precursors, non-human miRNA precursors, and control non-hairpin sequences (STAR Methods). First, the sequences were synthesized and cloned into an expression vector (Figure 1A; STAR Methods). The generated expression constructs were pooled in a single library and then transfected into human cells (i.e., human embryonic kidney 293T [HEK293T] cells). The tested library was transiently expressed, and the successfully transfected individual sequences were identified by DNA sequencing, while the structures that were successfully processed were detected by small RNA sequencing. By mapping the sequenced small RNAs back to the test structures, the biogenesis outcome of each RNA structure can be evaluated, as described below (Figure 1A; Figure S1).

Out of 150 bona fide human miRNA precursors successfully transfected into the HEK293T cells (as measured by DNA sequencing), a total of 74 were efficiently induced and processed (Figure 1B, red dots; STAR Methods). We found that the processing patterns of individual transfected precursors resembled known Drosha/Dicer processing signatures (Figure 1C), while the patterns for individual control sequences were staggered, suggesting random degradation (Figure 1D). This trend also holds when looking at compound distributions of read densities over the 74 processed human miRNA precursors (Figure 1E) and the 1,228 control sequences (Figure 1F). Overall, while we found 49% of human miRNA precursors to be robustly cleaved in our assay, only 3 of 1,228 (0.002%) control genomic non-hairpin sequences were processed, as were 0 of 1,369 (0%) random non-hairpin sequences, showing the specificity of MapToCleave (Figure 1G). It is well established that miRNA strands tend to have more precise start than end positions (Czech et al., 2009; Khvorova et al., 2003; Okamura et al., 2009; Schwarz et al., 2003). We find the same tendency for miRNA strands in the MapToCleave library (Figure 1H), indicating that our high-throughput screening recapitulates subtleties of natural miRNA biogenesis.

CellPress
OPEN ACCESS

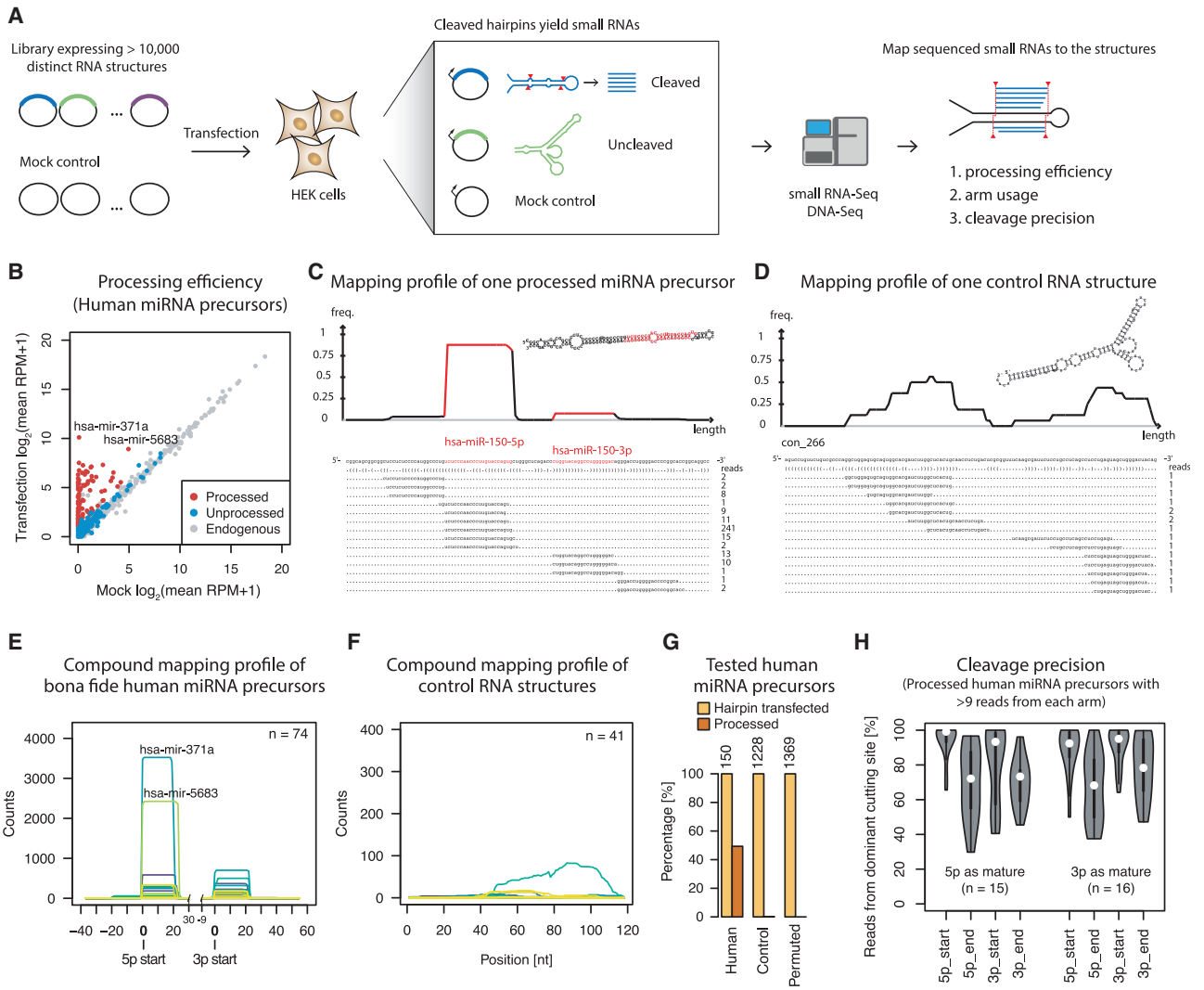


Figure 1. MapToCleave profiles miRNA processing of 12,472 distinct RNA structures

(A) Experimental design of MapToCleave.

(B) Small RNA abundance in HEK293T cells transfected with mock controls or MapToCleave library ("transfection"). miRNAs that are part of the library and increase significantly in expression are defined as processed (red), while miRNAs that are part of the library but do not increase in expression are defined as unprocessed (blue). miRNAs that are endogenous to the cells and not included in the library are in gray. Expression is in $\log_2$ RPM (reads per million).

(C) Example of MapToCleave processing of a bona fide human miRNA, showing clear patterns of Drosha and Dicer cleavage. A density plot of the read distribution of sequenced RNAs is shown above, and the exact read positions and read counts are shown below.

(D) Example of a control non-hairpin RNA. The read profile is staggered, suggesting random degradation.

(E) Compound read density plot of the 74 processed miRNA precursors. Each precursor is indicated with a distinct color.

(F) Compound read density plot of 41 control non-hairpin RNAs, showing staggered patterns suggestive of random degradation.

(G) Numbers of human miRNA precursors that are successfully transfected (yellow) and processed (orange). The same numbers are shown for control non-hairpin sequences from the human genome and control non-hairpins generated by randomizing (permuting) genome sequences.

(H) MapToCleave processing precision of miRNA precursors. The assay recapitulates details of natural miRNA processing, including the increased precision of miRNA start positions relative to end positions.

MapToCleave profiles cell type-dependent miRNA processing

A major advantage of MapToCleave is the ability to measure miRNA precursor processing in living cells, in the natural environment of protein cofactors, cellular compartments, and more, in contrast to previous large-scale efforts to profile miRNA biogenesis, which have all been *in vitro* (Auyeung et al., 2013; Fang and Bartel, 2015; Feng et al., 2011; Kwon et al., 2019; Li et al., 2020a; Nguyen et al., 2020; Rice et al., 2020). As a proof of principle, we tested human and murine MapToCleave precursors in HEK293T cells and mouse NIH 3T3 fibroblast cells (STAR Methods). In our replicate transfections in HEK293T cells, we find only 3% (5/195) of miRNA precursors to be differentially processed, showing the reproducibility of our method (Figure 2A, left). In contrast to these

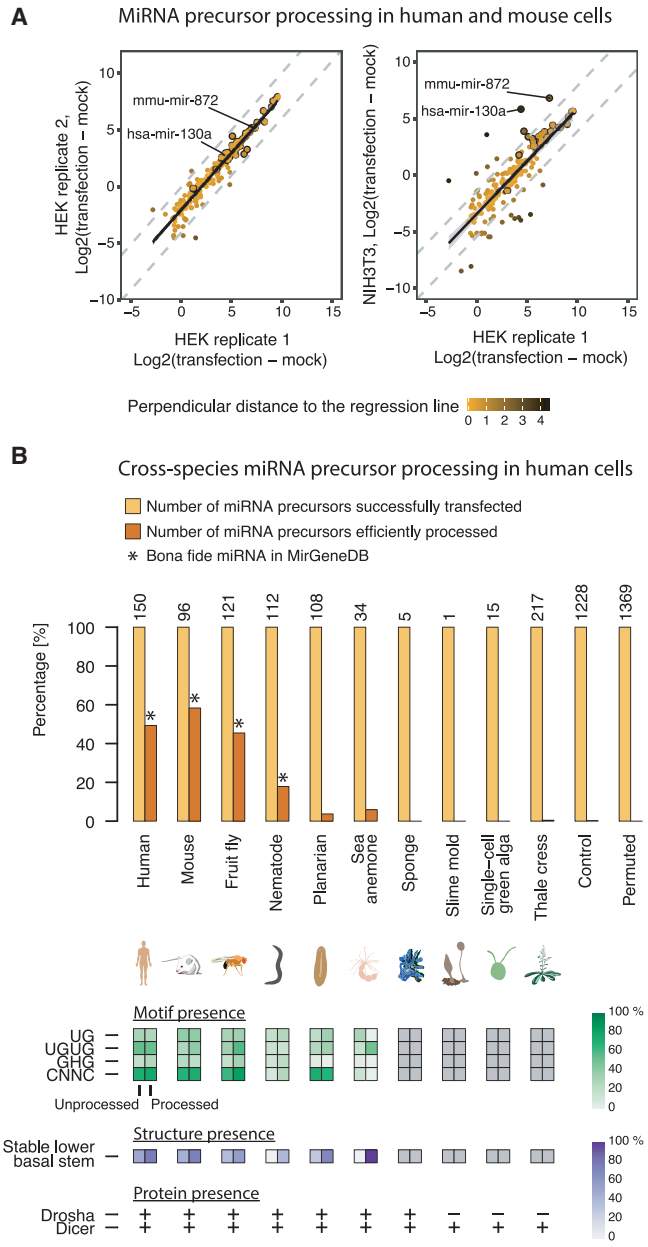


Figure 2. Cell type-dependent and cross-species miRNA precursor processing

(A) MapToCleave profiles cell type-dependent miRNA processing. The scatterplots show processing efficiency of the MapToCleave precursors in HEK293T or NIH 3T3 cells with different transfection conditions. The processing efficiency is measured by the difference between mean precursor expression (RPM) in the transfection cells and in the mock cells. The color gradient from orange to black indicates the perpendicular distance from the dots to the fitted linear regression line. The high-confidence precursors with expression level higher than 5 RPM in the transfected cells are highlighted by a black circle.

(B) Cross-species miRNA precursor processing in human cells. (Above) Number of transfected (yellow) and number of processed (orange) precursors for 10 animal and plant species, and for control non-hairpin sequences. (Below) Percentages of unprocessed and processed precursors that have sequence motifs known to facilitate processing (green). Also, percentages of unprocessed and processed precursors that have a stable lower basal stem structure (novel feature, in purple). (Bottom) Presence or absence of miRNA biogenesis factors in the 10 animal and plant species.

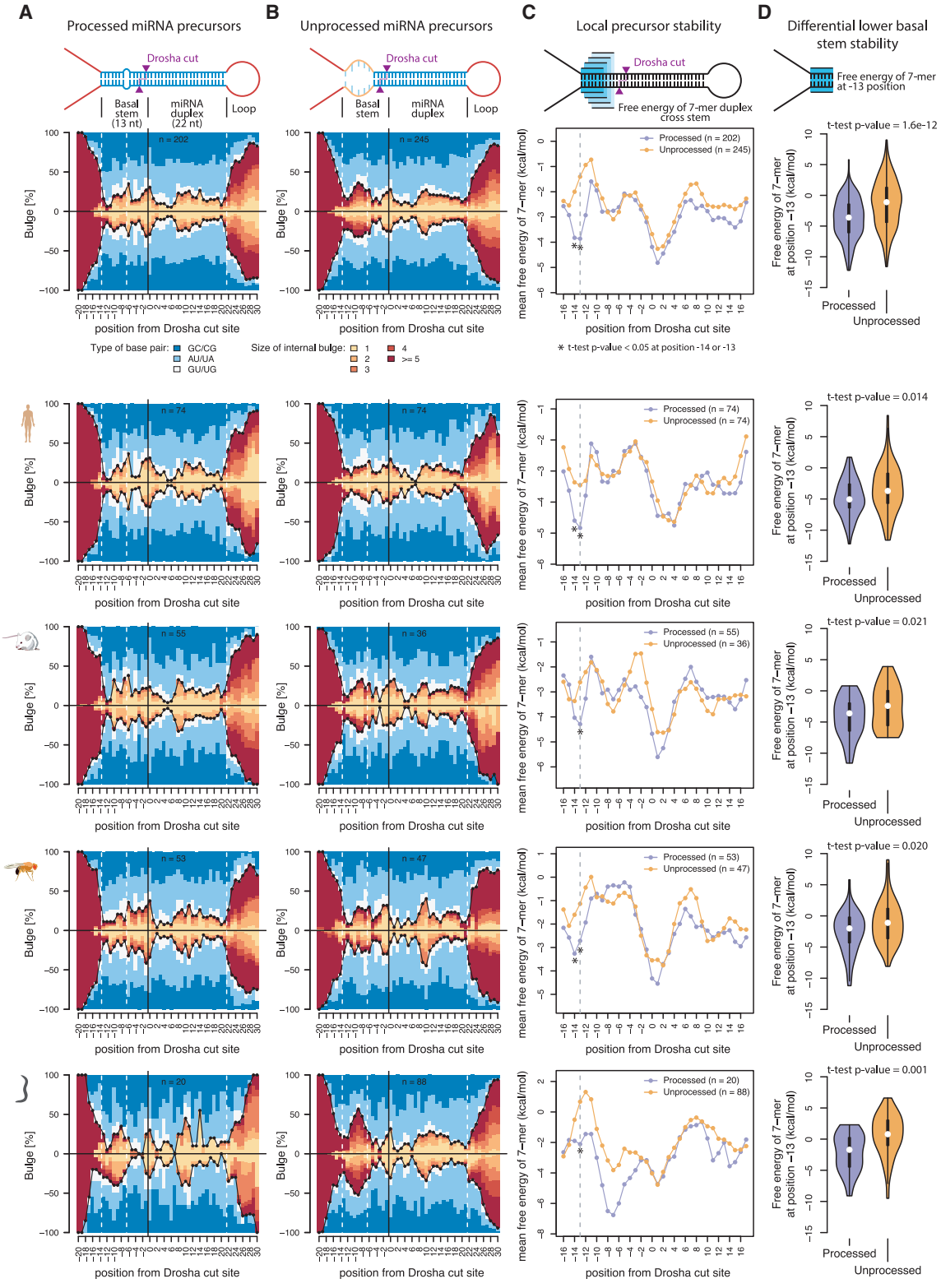
replicate experiments, when we compare processing in human HEK293T versus mouse NIH 3T3 cells, we find that 16% (28/176) of the tested precursors are processed more efficiently in one of the two cell types (Figure 2A, right). For instance, mir-872 is specific to the Glires animal group (rodents and lagomorphs) and is more efficiently processed in the mouse cell line (Figure 2A, right). Surprisingly, the human mir-130a also appears to be more efficiently processed in mouse cells than in human cells. Since this precursor appears to be more efficiently processed in other human cell lines (Figure S2), this could be due to some specific blocking of this precursor in the HEK293T cells. Based on the difference in percentages between the replicate experiment (3%, above) and the between cell type experiment (16%), we estimate that the biogenesis of 10%–15% of mammalian miRNA precursors is substantially influenced by cell-specific factors. We consider this estimate to be a higher bound since we here change both the species and the cell type. In summary, we demonstrate that MapToCleave can profile cell type-dependent processing of miRNA precursors.

miRNA biogenesis is functionally deeply conserved in animals

Having verified that MapToCleave recapitulates miRNA biogenesis in its natural cellular context, we next studied the processing of 709 non-human miRNA precursors in human HEK293T cells to evaluate species-specific features of miRNA biogenesis. The precursors originate from species ranging from mouse, fruit fly, nematode, planarian, sea anemone, animal sponge, slime mold, and single-cell green algae to thale cress (*Arabidopsis Thaliana)* (Figure 2B). Given the evidence that miRNAs originated through convergent evolution in plants and animals (Axtell et al., 2011), it would be expected that phylogeny strongly affects biogenesis. Indeed, we find that the precursors from species that are more closely related to humans are more likely to be processed. The percentages of human, mouse, and fruit fly precursors that are processed in human cells are comparable (ranging from 45% to 58%). In contrast, the percentage of nematode, planarian, and sea anemone precursors that are processed in human cells is low (ranging from 4% to 18%). Sea anemone is the species most distant from humans in which we detect processing above trace levels, spanning a gap of >600 million years of evolution. This suggests that miRNA processing is deeply conserved, while the substrate preference varies in species as a function of phylogenetic distance. We find that essentially no animal sponge, slime mold, green algae, or plant precursors are cleaved in human cells. miRNA biogenesis in animal sponges has previously been reported to be very different from other animals (Grimson et al., 2008), while slime mold and plant species do not have Drosha, which is a key biogenesis enzyme in animals (Avesson et al., 2012; Bologna and Voinnet, 2014; Bråte et al., 2018).

We next investigated whether the observed processing efficiency can be explained by the known sequence motifs, which

Structural features of miRNA precursors processed by MapToCleave assay



(legend on next page)

have been reported to facilitate mammalian miRNA processing. As has been previously reported, nematodes lack sequence motifs that are found in other animals (Auyeung et al., 2013), including planarians and sea anemone. However, there is no significant absence of sequence motifs in the precursors that are not processed in human cells (Figure 2B, green fields), suggesting that the low rate of processing has another explanation. Investigating the structures of the processed versus unprocessed planarian and sea anemone precursors, we found that the former had a tendency toward relatively structured and stable lower basal stems, defined as the first 7 nucleotides of the double-stranded stem structure (Figure 2B, in purple; Figure S3). This corresponds to positions −13 to −7 relative to the Drosha processing site. This apparent importance of the lower basal stem suggests that it is worth revisiting the influence of structural features on miRNA biogenesis.

Processed miRNA precursors have stable lower basal stems

We developed a new graphical representation to study the lower basal stem in more detail (Figure 3A). The "dumbbell" heat plots show the structure of miRNA precursors, with the single-stranded region to the left and the apical loop to the right and the 5′ strand on top and the 3′ strand below. The color code indicates CG base pairing (dark blue), AU base pairing (light blue), GU base pairs (white), or bulges of mismatched nucleotides of increasing size (yellow to red). When summing precursors over humans, mice, fruit flies, and nematodes, the most striking difference between the processed (Figure 3A, top) and the unprocessed miRNA precursors (Figure 3B, top) is at the lower basal stem from position −13 to −7 relative to the Drosha cleavage site (indicated with dotted white lines). Specifically, the precursors that are processed in our MapToCleave assay have fewer and smaller bulges in the lower basal stem (Figure 3A, top) than do the precursors that are not processed (Figure 3B, top). This difference is observed in the ΔG minimum free energy estimates (Figure 3C, top) and is statistically significant (Figure 3D, top, p = 1.6e−12). We observe the same tendency when human (p = 0.014), mouse (p = 0.021), fruit fly (p = 0.020), and nematode (p = 0.001) precursors are studied separately, covering >600 million years of evolution. This tendency also holds true for the MapToCleave precursors tested in mouse cell lines (Figure S4). To further support our findings, we re-analyzed miRNA precursor processing data from a previous study, in which Drosha cleavage efficiency of >50,000 sequence variants of three distinct primary miRNAs was tested *in vitro* in a lysate-containing Microprocessor (Fang and Bartel, 2015). By comparing the local structure profile of the variants with high, medium, and low cleavage efficiency, we find that introducing a bulge at the basal stem has a more detrimental effect on Drosha processing compared to bulges in other regions (Figure S5). In summary, we show that processed precursors have significantly more stable lower basal stem structures, from nematodes to humans.

Lower basal stem stability predicts miRNA expression levels *in vivo*

To test whether the stable lower basal stem is a robust biological feature for miRNA processing rather than an artifact resulting from our MapToCleave screening system, we reanalyzed public small RNA sequence data from various animals. These data are from tissues and therefore represent *in vivo* expression, completely independent of our screening system. Specifically, we took advantage of the recently released second version of the manually curated microRNA gene database (MirGeneDB; Fromm et al., 2020) and analyzed miRNA expression data composed of 191 tissue types from 20 species belonging to four clades: mammals, fruit flies, nematodes, and lophotrochozoans. We averaged miRNA expression over tissues within a species and then compared the mostly highly and lowly expressed miRNAs within a given clade. By comparing the structure profile between the highly and lowly expressed miRNA precursors, we find that the lower basal stem is consistently observed to be more stable in the highly expressed miRNA precursors in mammals (p = 5e−5; Figures 4A–4D, top row), fruit flies (p = 0.0084; Figures 4A–4D, second row), and lophotrochozoans (p = 0.013; Figures 4A–4D, fourth row). We do not observe the tendency in nematodes (p = 0.21; Figures 4A–4D, third row). Note that nematode precursors have slightly longer basal stems (Warf et al., 2011), which in turn shifts the location of their lower basal stem (around from position −16 to −10) by around 3 nt away from the Drosha cleavage site relative to the lower basal stem in other species (from position −13 to −7). Interestingly, the lower basal stem is also more stable in ancient miRNAs than in more recently emerged miRNAs (Figure S6). These findings support the idea that the stable lower basal stem is not an artifact of our RNA structure screening system but is rather a naturally occurring and deeply conserved biological feature for miRNA processing.

Chromatin-associated primary miRNA profiles support importance of lower basal stem

Previous studies indicate that miRNA primary transcripts may stably associate with chromatin (Pawlicki and Steitz, 2008). To study whether precursors with stable and unstable lower basal stems give rise to different primary miRNA profiles as a result of processing, we reanalyzed sequenced primary miRNA transcripts associated with chromatin from a study by Conrad et al. (2014). In this previous experiment, the amount of intact versus cleaved primary miRNA transcripts was used to estimate processing efficiency.

Figure 3. Processed miRNA precursors have more stable lower basal stem structures

(A and B) Detailed structure profile of processed precursors (A) and unprocessed precursors (B). The "dumbbell" plots show the structure of miRNA precursors, with the single-stranded region to the left and the apical loop to the right, and the 5′ strand on top and the 3′ strand below. The color code indicates CG base pairing (dark blue), AU base pairing (light blue), GU base pairs (white), or bulges of mismatched nucleotides of increasing size (yellow to red). The Drosha cleavage site at the 5′ strand is at position zero, and the two white vertical lines to the left indicate the position of the lower basal stem.

(C) Thermodynamic stability profiles of processed and unprocessed precursors. The estimated minimum free energy (ΔG in kilocalories per mole) for RNA duplex was calculated by a rolling 7-nt window through the given precursor stem loop. Lower minimum free energy indicates more stable structures.

(D) Minimum free energy distribution of the lower basal stem, represented by the 7-mer window at position −13, of processed and unprocessed precursors.

CellPress
OPEN ACCESS

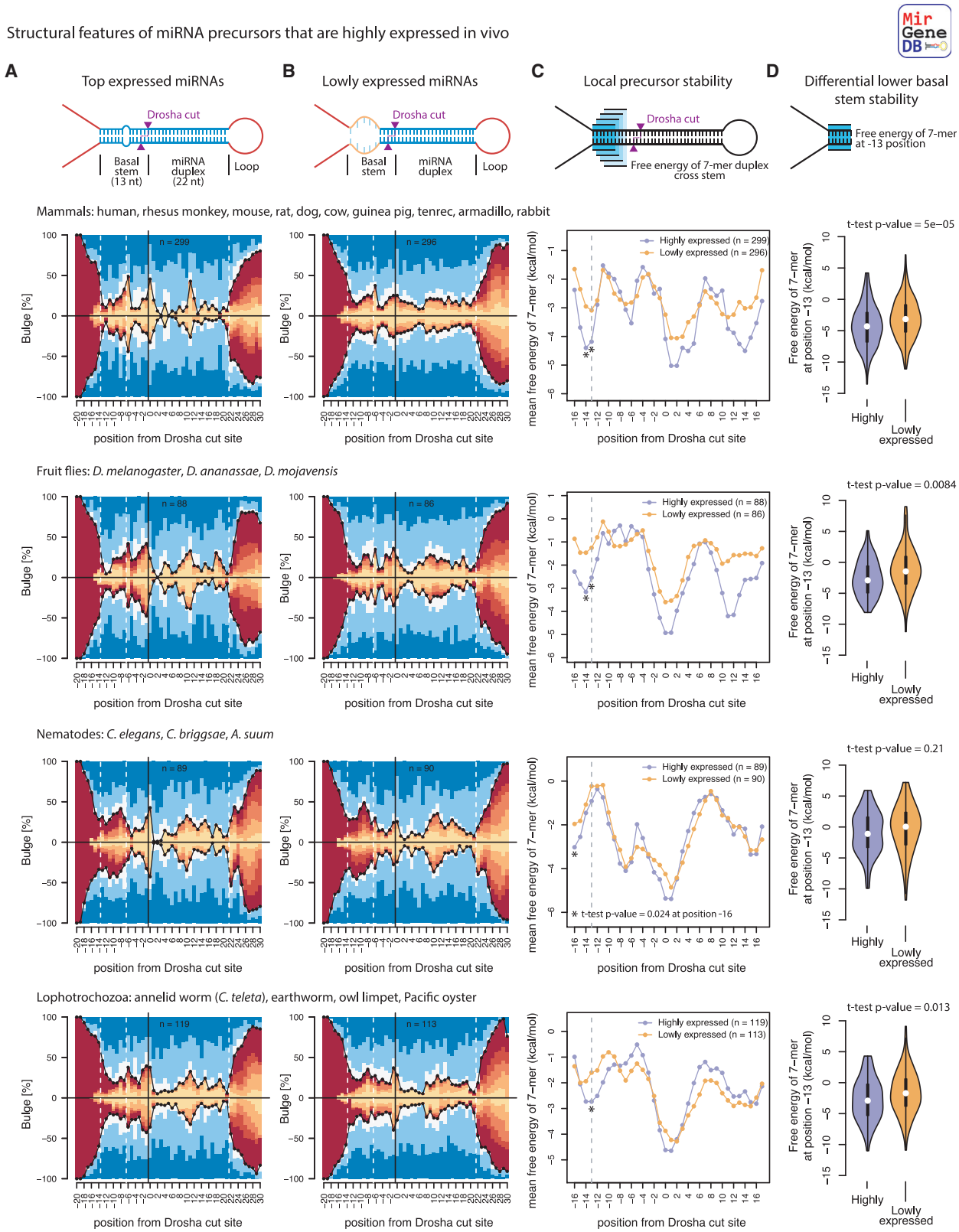


Figure 4. miRNAs with high *in vivo* expression have more stable lower basal stems

(A–D) Similar to Figure 3, but the plots were generated based on the miRNAs with highest and lowest expression in animal tissues according to MirGeneDB.

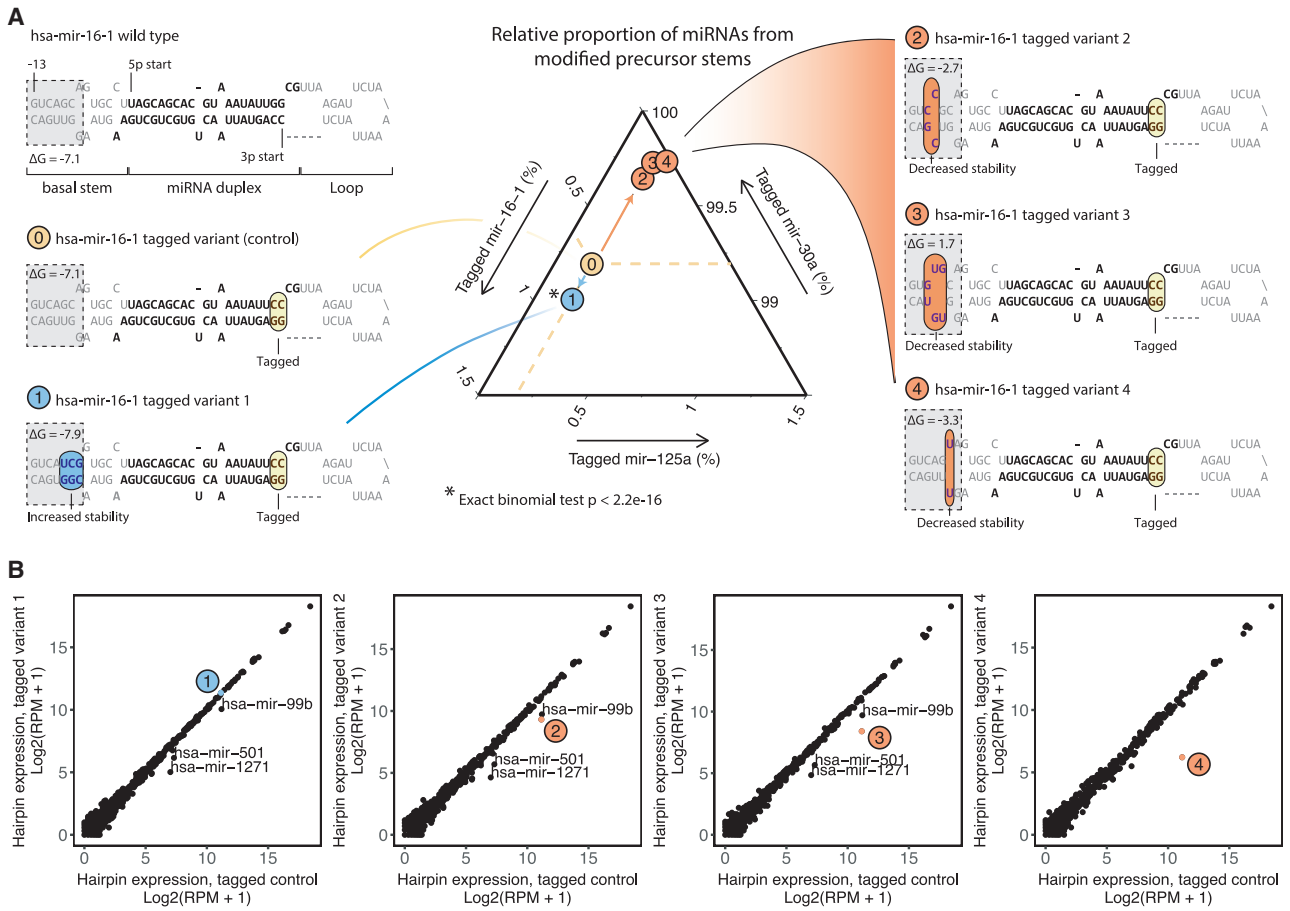


Figure 5. Design of miRNA precursors with improved or impaired processing

(A) (Left and right panels) Design of hsa-mir-16-1 variants with increased or decreased lower basal stem stability. All the variants are tagged by swapping 2 nucleotides at the 3′ end of the stem to distinguish them from endogenous miRNAs in sequencing. (Middle panel) Relative proportion of miRNAs from the tagged hairpin stem of hsa-mir-16-1, hsa-mir-30a, and hsa-mir-125a, as measured by small RNA sequencing.

(B) Scatterplots showing hairpin expression measured by summing 5p and 3p miRNAs in the mock or transfected cells.

We compared the structure profile between the most efficiently processed and the least efficiently processed miRNAs identified by the study (Figures S7A and S7B). As expected, the efficiently processed miRNA precursors have a more stable lower basal stem compared to the non-efficiently processed miRNAs, although this tendency is only significant when counting from position −14 and not from position −13 (Figures S7C and S7D). Again, this indicates the importance of the lower basal stem stability as a biological feature for miRNA processing.

Design of miRNA precursors with improved or impaired processing capacity

Hairpin RNAs are widely used in RNA interference experiments and also for therapeutic treatments (Beg et al., 2017; Janssen et al., 2013; Sahu et al., 2019). We next investigated whether it is possible to tune precursor design by modifying the lower basal stem regions. We designed four variants of mir-16, one of which should stabilize the lower basal stem and improve processing (variant 1) and three that should destabilize the lower basal stem and impair processing (variants 2–4; Figure 5A, left and right panels). In each experiment we co-transfected with equimolar abundances of mir-30a and mir-125a for normalization, and all transfected miRNAs were additionally modified (tagged) in the mature region to discern them from endogenous miRNAs (STAR Methods). We found that stabilizing the lower basal stem indeed improved expression subtly, while destabilizing the stem substantially reduced it (Figures 5A, middle, and 5B; Figure S8). Interestingly, endogenous mir-99b, mir-501, and mir-1271 were consistently reduced in the transfection experiments (Figure 5B). These miRNAs may be part of the same regulatory networks as the three transfected miRNAs and may be repressed through negative feedback loops. The influence of stability of the lower basal stem can also be observed in the designed variants of mir-30a (Figure S9). In summary, we show that hairpin design can be tuned by stabilizing or destabilizing the lower basal stem.

The GHG motif predicts processing better as a structure than as a sequence feature

Having focused on processing efficiency, we next investigated processing precision, measured as the percentage of sequenced

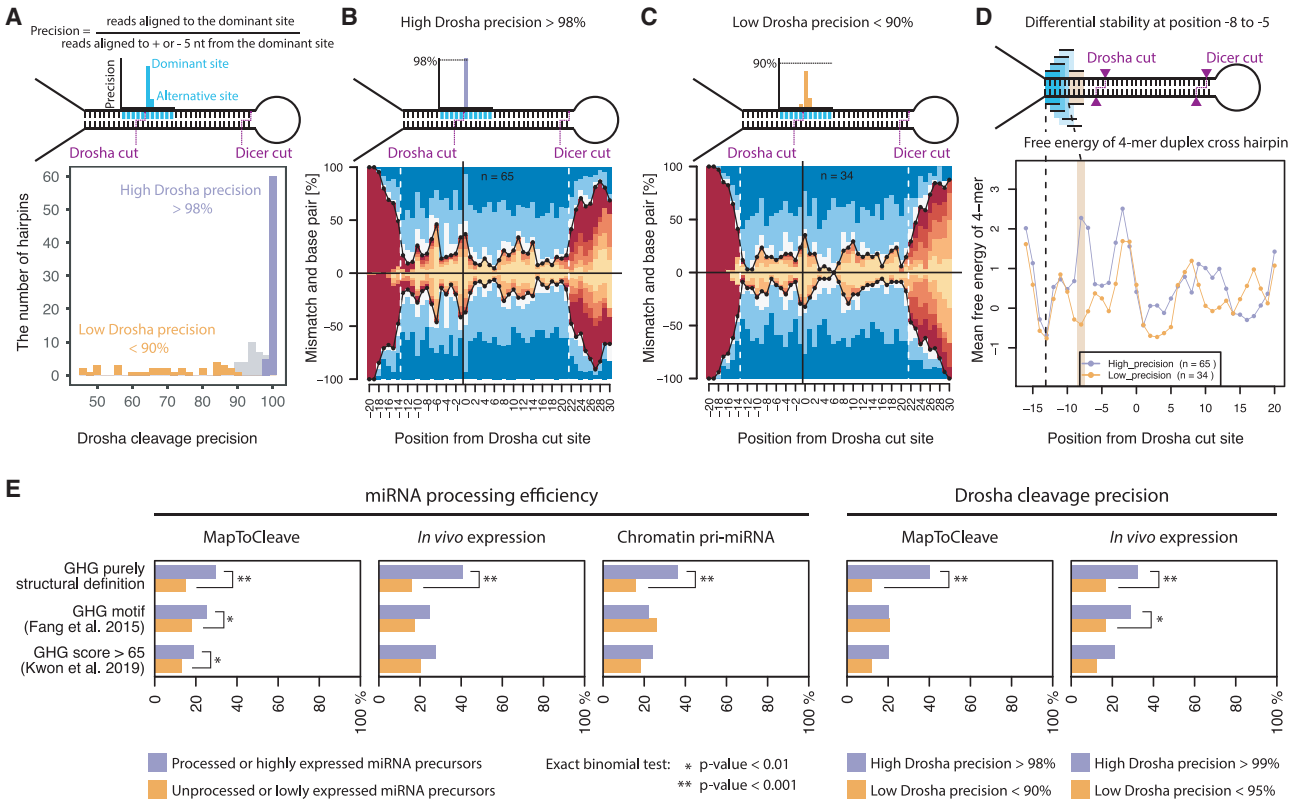


Figure 6. Influence of GHG feature on miRNA processing

(A) Histogram showing the Drosha cleavage precision of the processed precursors calculated by the equation on the top panel.

(B) Detailed structure profile of precursors with high Drosha cleavage precision (>98% of reads from dominant cleavage site).

(C) Same as (B) but using precursors with low Drosha cleavage precision (<90% of reads from dominant cleavage site).

(D) Thermodynamic stability profile of the processed precursors with high and low Drosha cleavage precision. The free energy (ΔG in kilocalories per mole) was calculated by a rolling 4-nt window through the given precursor stem loop. The orange bar shows position −7.

(E) The GHG motif predicts processing better when defined as a structural rather than sequence motif. miRNA precursors tested in our study were divided into the ones that are efficiently processed and highly expressed versus the ones that are unprocessed and had low expression. It was then tested how many miRNA precursors in the two groups contained the GHG motif, according to three different definitions. The "GHG motif" (Fang and Bartel, 2015) and the "GHG score >65" (Kwon et al., 2019) are defined by both structure and sequence features, while the "GHG structure" is a purely structure feature. For the purpose of this analysis, the MapToCleave data from HEK293T cells, MirGeneDB miRNA *in vivo* expression atlas of human tissues, and chromatin-associated primary miRNA data from Conrad et al. (2014) were used.

miRNAs that map exactly to the consensus cut site (Figure 6A). We find that the precursors with high Drosha precision (>98%) tend to have a small bulge of 1 or 2 nt that overlap with position −6, while the precursors that are processed with low precision (<90%) rarely have a bulge at this position (Figures 6B and 6C). This tendency for a bulge is clearly visible as an unstable region (Figure 6D). It is well established that the GHG motif, located from nucleotides −7 to −5 from the Drosha cleavage site, can facilitate processing efficiency and precision of miRNA precursors (Fang and Bartel, 2015; Kwon et al., 2019). However, it is debated whether the motif is functionally more a sequence motif or a structural motif. Given the clear bulge that we see in precisely processed hairpins (Figure 6B), we propose the purely structural definition that a precursor has the GHG structure motif if it has a bulge composed of 1 or 2 nt that overlap with position −6 (counted from the 5′ stand). We find that the structural definition better predicts processing efficiency in our MapToCleave assay and also better predicts miRNA expression *in vivo* (Figure 6E, left) than does the sequence definition—that a precursor has a GHG motif if the positions −7 to −5 relative to the Drosha cleavage site consist of an unmatched nucleotide other than guanosine that is flanked by two base-paired guanosines (definition by Fang and Bartel, 2015). The same holds true for miRNA processing efficiency estimated from chromatin-associated miRNA primary transcripts from Conrad et al. (2014) (Figure 6E, middle). We also find that the structural GHG definition better predicts processing precision in the MapToCleave assay or *in vivo* (Figure 6E, right). In summary, we find that the GHG motif better predicts miRNA processing efficiency and precision when defined only by its structure.

Relative importance of structures and sequence motifs for miRNA biogenesis

To understand the relative importance of known and novel sequence and structure features for miRNA biogenesis, we estimated how well each feature correlates with miRNA processing,

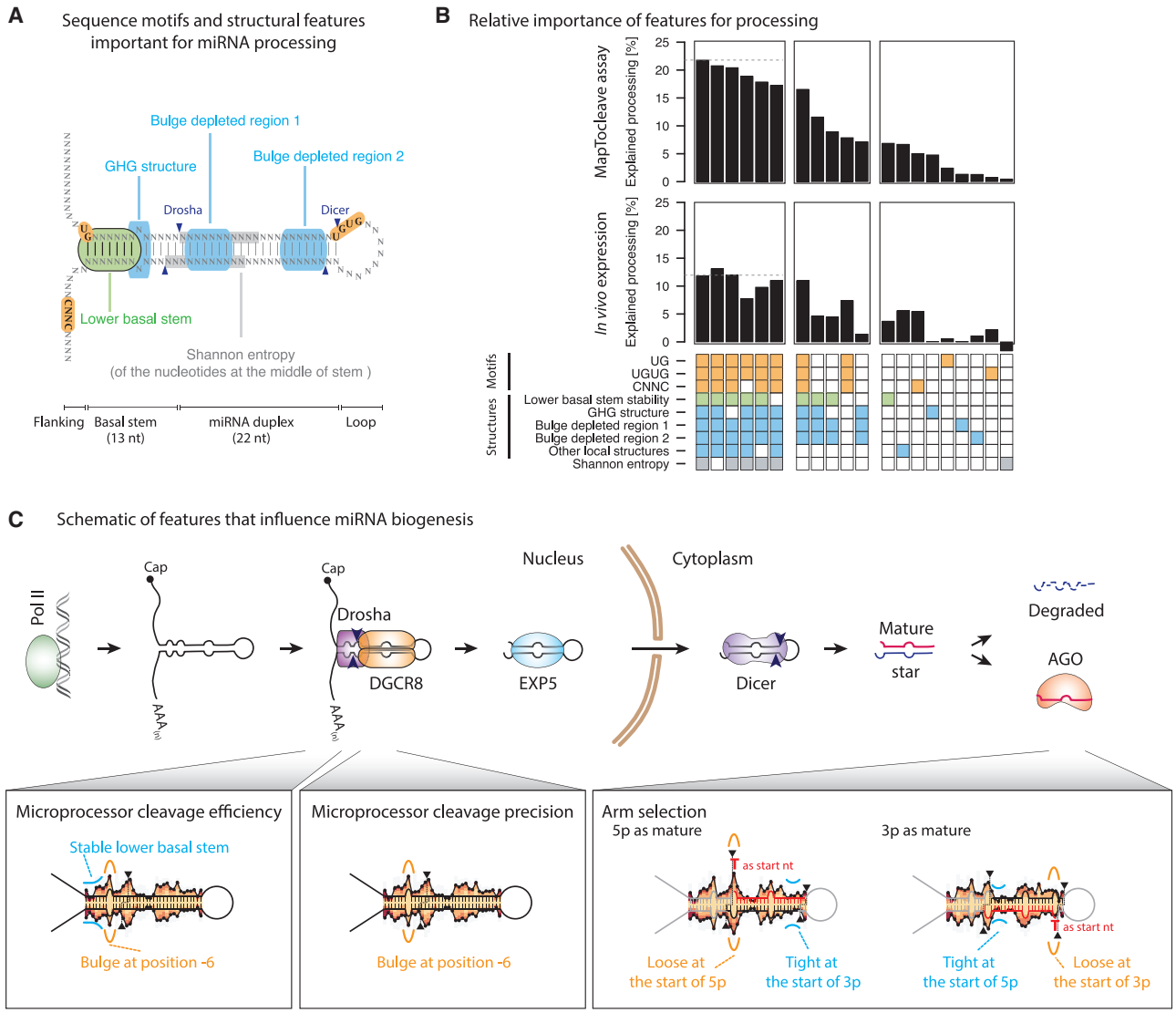


Figure 7. Relative importance of known and novel features for miRNA processing and expression

(A) Schematic of miRNA precursor stem showing location and type of known and newly identified features for miRNA processing efficiency.

(B) Feature importance estimated by adjusted R-squared value of the linear regression model with miRNA processing efficiency (MapToCleave data) or with mean miRNA RPM of human tissues (*in vivo* expression data from MirGeneDB) as the outcome variable and a given feature (or features) as the explanatory variable.

(C) Schematic of features that influence miRNA biogenesis. The background structure profile in the panels of Microprocessor cleavage efficiency and precision shows the presence of bulges in the MapToCleave-processed precursors. The color code is the same as in Figure 3A. In the panel on arm selection, the background structure profiles on the left and right show, respectively, the presence of bulges in the 5p arm- and the 3p-arm-selected MapToCleave-processed precursors.

as measured by MapToCleave, and miRNA *in vivo* expression, as collected in MirGeneDB (Figure 7A). Specifically, we applied linear regression to measure feature importance by the adjusted R-squared value, which reflects the amount of data variance of miRNA processing efficiency that is explained by the model built on the feature (STAR Methods). Intriguingly, the lower basal stem stability is ranked as the most important individual feature using MapToCleave data and the second most important using *in vivo* data (in green, Figure 7B), suggesting it is at least as important for processing as are the well-studied sequence motifs. We find that Shannon entropy (Rice et al., 2020) explains little of *in vivo* processing (in gray, Figure 7B), but does contribute to processing in our cleavage assay, although to a lesser extent than the lower basal stem stability (Figure 7B). Interestingly, two bulge-depleted regions of the precursors also contribute (in blue), consistent with previous results (Roden et al., 2017), as does the stability of other local structures along the miRNA stem that have only been investigated in a few studies (Li et al., 2020a; Nguyen et al., 2020). Overall, the combined structural features explain more of the miRNA processing (16.5%) than

do the combined sequence features composed of CNNC, UG, and UGUG (7.9%). The structural features explain comparable data variance of the *in vivo* expression (6.7%) to the sequence features (7.4%). In summary, we provide evidence that local structural precursor features are at least as important as the well-studied sequence motifs for miRNA processing.

MapToCleave recovers two rules of miRNA arm selection

Two rules have been proposed to determine which precursor arm gets selected as the guide miRNA and which gets degraded as a by-product of biogenesis (Czech et al., 2009; Khvorova et al., 2003; Okamura et al., 2009; Schwarz et al., 2003). According to the thermodynamics stability rule, the miRNA duplex end that is less stable is easier to open, and the arm whose 5′ end (so-called "5p" arms) is at this end will be selected. According to the nucleotide rule, the arm with U and A as the first nucleotide is more likely to be selected as the guide miRNA compared to the arm with G and C. We divided the processed MapToCleave precursors into four groups depending on their preference for arm selection and investigated their distinct structural features (Figure S10A). We find that precursors that have a strong 5p arm bias have a strong tendency for a bulge at the Drosha cleavage site (position 0), which would make the duplex end less stable, as predicted by the thermodynamic rule (Figure S10B). Interestingly, for the precursors with a strong 3p bias, this bulge tends to be located at position −1, just outside of the duplex (Figure S10A). The precursors that have a 3p arm bias also tend to have more bulges toward the 3′ end of the duplex (Figure S10A), resulting in less stability in that end (Figure S10B). Furthermore, precursors with extreme 5p and 3p arm usage have the highest local free energy at the 5′ and 3′ ends, respectively, of the miRNA duplex (Figure S10B), and they also have the highest proportions of U and A, respectively, as the start nucleotide (Figure S10C). These two rules of arm selection are identified by MapToCleave, suggesting that the method is able to capture features that impact different steps of miRNA biogenesis (Figure 7C).

DISCUSSION

In this study, we have systematically surveyed features of miRNA biogenesis through the use of our high-throughput screening method MapToCleave. This allows us to test processing of thousands of distinct RNA structures in one experiment, recapitulating miRNA biogenesis in the natural context of living cells with protein cofactors, cellular compartments, and more. We find that most of the tested human, mouse, and fruit fly miRNA precursors are efficiently processed in human HEK293T cells, while precursors of nematodes, planarians, and non-bilaterian animals are inefficiently processed, and precursors of organisms that lack Drosha are not processed above trace levels (Figure 2B). Surprisingly, the miRNA precursors that are not processed in our MapToCleave assay specifically tend to have unstable lower basal stems, defined as positions −13 to −7 relative to the Drosha cleavage site (Figure 3). Applying public data of *in vivo* expression of curated miRNA complements of 20 animal species from MirGeneDB, we find that highly expressed miRNA precursors tend to have stable lower basal stems, while lowly expressed precursors tend to have unstable lower basal stems, indicating that the stability of this region tunes miRNA expression (Figure 4). We find that a structural definition of the GHG motif better predicts precursor cleavage efficiency and precision than does a sequence definition (Figure 6E), consistent with recent cryo-EM studies of Drosha substrate recognition (Jin et al., 2020; Partin et al., 2020). Comparing the relative importance of precursor features, we find that novel structural features explain MapToCleave processing efficiency and *in vivo* miRNA expression as well as or better than sequence motifs (Figure 7B). We find that lower basal stem stability in itself explains ~7% of processing efficiency, more than each of the individual known sequence motifs. Lastly, we recover and confirm known features of miRNA biogenesis, including the rules that determine miRNA strand selection (Figure S10; Figure 7C).

It may seem surprising that Shannon entropy explains little of in *vivo* miRNA processing (Figure 7B), in contrast to findings in a recent in *vitro* large-scale screening study (Rice et al., 2020). This may in part be explained by the complexity of living cells, but it may also be explained by the definition of miRNA precursors. The previous screening study used miRBase annotations, which contain many young miRNA genes as well as false-positive annotations (Fromm et al., 2020). In contrast, our study uses MirGeneDB2 annotations, which are carefully curated. Thus, Shannon entropy may be a good measure for distinguishing genuine miRNAs from evolving genes or false positives (Figure S11), while lower basal stem stability distinguishes genuine miRNAs that are highly or lowly expressed in tissues.

It is well established that the length and stability of the ~35-nt miRNA stem is important for processing (Fang and Bartel, 2015; Roden et al., 2017), and the contribution of the lower stem (positions −13 to −1 from the Drosha cleavage site) has been shown before in *in vitro* assays (Auyeung et al., 2013; Han et al., 2006; Zeng et al., 2005). Here, we provide evidence that the first 7 nucleotides of the lower stem (positions −13 to −7) are of particular importance relative to other individual sequence and structure features for miRNA expression in cells and in tissues (Figure 7B). We argue that this relates to Drosha recognition and binding, rather than simply defining the single-stranded to double-stranded transition, since the stability of the full 7 nucleotides is critical and predicts processing much better than do shorter regions close to the single-stranded to double-stranded transition site (data not shown).

It may seem counterintuitive that the lower basal stem tunes miRNA expression, since a given precursor only gives rise to a single miRNA guide. However, there is evidence that many miRNA primary transcripts are not cleaved but rather remain relatively stable in the chromatin (Pawlicki and Steitz, 2008). Specifically, sequencing of RNAs in the chromatin allowed Conrad et al. (2014) to assign processing indexes to miRNA primary transcripts and to find that many had intermediate levels of processing. If the lower basal stem facilitates efficient precursor processing, it would result in higher expression of the resulting mature miRNA, as we observe in the *in vivo* MirGeneDB data from 20 animal species.

Surprisingly, in our MapToCleave assay, we found that only ~50% of the bona fide human miRNA precursors were processed in HEK293T cells. We estimate that ~5% of the tested

precursors appear to be unprocessed because the exogenous expression is masked by high endogenous expression. We further estimate that ~9% of the tested precursors may not have been cleaved because they are normally clustered with other precursors that may facilitate their biogenesis (Fang and Bartel, 2020; Hutter et al., 2020; Kretov et al., 2020; Shang et al., 2020). The remaining unprocessed precursors tend to have unstable lower basal stems (Figure 3A), which means they may be outcompeted for Drosha processing by the precursors that have more stable lower basal stems or may have other structural features that facilitate interactions with Microprocessor. We did not find any depletion of the known sequence motifs in the unprocessed precursors (Figure 2B). Finally, it is possible that some of the precursors may depend on biogenesis cofactors that are absent in HEK293T cells. This again highlights the advantage of studying miRNA biogenesis in a cellular system.

Interestingly, we find that known and novel precursor features overall explain less of the miRNA *in vivo* expression (13%) than they explain the miRNA processing (22%). This is what we expected, since MapToCleave comprises a well-controlled experiment in a human cell line, whereas the human MirGeneDB data comprise miRNA expression of various tissues that are affected by more layers of regulation of miRNA biogenesis as well as by the technical effects of heterogeneous data. Even with our new features, the current model of miRNA biogenesis has a relatively limited information content and is still far from explaining the specificity of miRNA biogenesis. The optimal structure profile and the known sequence motifs together only explain ~22% of data variance of miRNA processing in MapToCleave (Figure 7B). Of the remaining ~78% data variance, MapToCleave DNA construct copy number for each precursor explains ~14%, consistent with previous findings that primary miRNA transcription explains a substantial fraction of its final expression (de Rie et al., 2017). Besides data noise of experimental techniques, this points to more global factors, including, for example, RNA tertiary structure (Chaulk et al., 2011), global RNA structure (Rouleau et al., 2018), nuclear localization of precursors and biogenesis proteins, and biogenesis cofactors binding outside the local vicinity of the precursors (Nussbacher and Yeo, 2018; Treiber et al., 2017). Our results suggest that local features may only explain part of miRNA precursor selection and processing efficiency, and that a full model of miRNA biogenesis may also need to include global factors as critical components.

Limitations of the study

This study focuses on a single human cell line, HEK293T, and it is uncertain to what degree conclusions can be extended to other cell types. In particular, other cell types may contain cell-specific factors that facilitate or inhibit processing of specific miRNAs. However, we find overall good agreement between processing in human and mouse cells, with some notable differences (Figure 2A). A further limitation is that we are profiling miRNA biogenesis in its entirety and cannot unravel the contributions of individual biogenesis steps. For instance, we could not assign the contribution of the GHG motif to Microprocessor cleavage as opposed to nuclear export or Dicer processing, if Microprocessor activity had not already been studied *in vitro*. Finally, our tissue data can be confounded by transcription levels, which can differ from one tissue to another and which cannot easily be corrected for, and this may make our *in vivo* analyses more noisy. The widely used *in vitro* methods in contrast have the advantage of specifically profiling Microprocessor activity without confounding factors (e.g., Han et al., 2006; Auyeung et al., 2013; Fang and Bartel, 2015; (Li et al., 2020a); (Li et al., 2020b); Rice et al., 2020). These methods however may be limited by lack of cellular context and cofactors and it is not certain whether molecular concentrations reflect physiological levels. The two approaches seem complementary, and importantly the findings from our in-cell and in-tissue approach recover and converge with main findings of previous *in vitro* studies (Figure 7B).

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - MapToCleave library design
  - MapToCleave hairpin library
  - MapToCleave non-human miRNA hairpins
  - MapToCleave sequence cloning
  - Cell culture and MapToCleave transfections
  - MapToCleave DNA and small RNA sequencing
  - MapToCleave sequence data quality control and preprocessing
  - Estimating processing efficiency of MapToCleave sequences
  - Profiling cell type-dependent processing of MapToCleave sequences
  - Identifying structural features and sequence motifs in MapToCleave hairpins
  - Profiling local free energy of MapToCleave hairpins
  - Profiling Shannon entropy of MapToCleave hairpins
  - MirGeneDB miRNA expression analysis
  - Design of miRNA precursors with improved or impaired processing
  - Profiling Drosha cleavage precision of MapToCleave hairpins
  - Identifying the presence of the GHG feature using different definitions
  - Estimating the relative importance of features for miRNA processing and expression
  - Estimating the contribution of miRNA clustering to unprocessed MapToCleave precursors
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2021.110015.

ACKNOWLEDGMENTS

This work was supported by the following sources: ERC starting grant 758397, "miRCell"; Swedish Research Council (VR) grant 2015-04611, "MapToCleave"; and funding from the Strategic Research Area (SFO) program of the Swedish Research Council through Stockholm University. R.J. is supported by Science Foundation Ireland through Future Research Leaders award 18/FRL/6194. C.A. was supported by the Ministerio de Economía y Competitividad and FEDER funds under reference numbers BIO2011-26205 and BIO2015-70777-P and Secretaria d'Universitats i Investigació del Departament d'Economia i Coneixement de la Generalitat de Catalunya under award number 2014 SGR 1319. A.J.H. was funded as a Marie Curie Post-doctoral Fellow supported by the European Commission 7th Framework Program under grant agreement no. 330133. We thank Roderic Guigó, Xavier Estivill, and Joakim Lundeberg for support and advice. The computations were enabled by resources in a project (SNIC 2017/7-297) provided by the Swedish National Infrastructure for Computing (SNIC) at UPPMAX, partially funded by the Swedish Research Council through grant agreement no. 2018-05973.

AUTHOR CONTRIBUTIONS

The study was conceptualized by M.R.F., I.B., and R.J. The MapToCleave library was designed by M.R.F. and R.J. and generated by A.J.H., D.B., and C.A. K.T. and M.A. prepared two sequencing libraries. All other experimental work was performed by I.B. Computational analyses were performed by W.K. under supervision of M.R.F., with analysis contributions from B.F. and E.H. The manuscript was written by W.K. and M.R.F., with contributions from all authors.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: March 8, 2021
Revised: September 17, 2021
Accepted: October 27, 2021
Published: November 16, 2021

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.

Avesson, L., Reimegård, J., Wagner, E.G., and Söderbom, F. (2012). MicroRNAs in Amoebozoa: Deep sequencing of the small RNA population in the social amoeba *Dictyostelium discoideum* reveals developmentally regulated microRNAs. RNA *18*, 1771–1782.

Axtell, M.J., Westholm, J.O., and Lai, E.C. (2011). Vive la différence: Biogenesis and evolution of microRNAs in plants and animals. Genome Biol. *12*, 221.

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

Bartel, D.P. (2018). Metazoan MicroRNAs. Cell *173*, 20–51.

Beg, M.S., Brenner, A.J., Sachdev, J., Borad, M., Kang, Y.-K., Stoudemire, J., Smith, S., Bader, A.G., Kim, S., and Hong, D.S. (2017). Phase I study of MRX34, a liposomal miR-34a mimic, administered twice weekly in patients with advanced solid tumors. Invest. New Drugs *35*, 180–188.

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.

Bernstein, E., Caudy, A.A., Hammond, S.M., and Hannon, G.J. (2001). Role for a bidentate ribonuclease in the initiation step of RNA interference. Nature *409*, 363–366.

Bernstein, E., Kim, S.Y., Carmell, M.A., Murchison, E.P., Alcorn, H., Li, M.Z., Mills, A.A., Elledge, S.J., Anderson, K.V., and Hannon, G.J. (2003). Dicer is essential for mouse development. Nat. Genet. *35*, 215–217.

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

Bologna, N.G., and Voinnet, O. (2014). The diversity, biogenesis, and activities of endogenous silencing small RNAs in *Arabidopsis*. Annu. Rev. Plant Biol. *65*, 473–503.

Bråte, J., Neumann, R.S., Fromm, B., Haraldsen, A.A.B., Tarver, J.E., Suga, H., Donoghue, P.C.J., Peterson, K.J., Ruiz-Trillo, I., Grini, P.E., and Shalchian-Tabrizi, K. (2018). Unicellular origin of the animal microRNA machinery. Curr. Biol. *28*, 3288–3295.e5.

Cai, X., Hagedorn, C.H., and Cullen, B.R. (2004). Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. RNA *10*, 1957–1966.

Chaulk, S.G., Thede, G.L., Kent, O.A., Xu, Z., Gesner, E.M., Veldhoen, R.A., Khanna, S.K., Goping, I.S., MacMillan, A.M., Mendell, J.T., et al. (2011). Role of pri-miRNA tertiary structure in miR-17∼92 miRNA biogenesis. RNA Biol. *8*, 1105–1114.

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

Conrad, T., Marsico, A., Gehre, M., and Orom, U.A. (2014). Microprocessor activity controls differential miRNA biogenesis in vivo. Cell Rep. *9*, 542–554.

Czech, B., Zhou, R., Erlich, Y., Brennecke, J., Binari, R., Villalta, C., Gordon, A., Perrimon, N., and Hannon, G.J. (2009). Hierarchical rules for Argonaute loading in *Drosophila*. Mol. Cell *36*, 445–456.

de Rie, D., Abugessaisa, I., Alam, T., Arner, E., Arner, P., Ashoor, H., Åström, G., Babina, M., Bertin, N., Burroughs, A.M., et al.; FANTOM Consortium (2017). An integrated expression atlas of miRNAs and their promoters in human and mouse. Nat. Biotechnol. *35*, 872–878.

Esteller, M. (2011). Non-coding RNAs in human disease. Nat. Rev. Genet. *12*, 861–874.

Fang, W., and Bartel, D.P. (2015). The menu of features that define primary microRNAs and enable de novo design of microRNA genes. Mol. Cell *60*, 131–145.

Fang, W., and Bartel, D.P. (2020). MicroRNA clustering assists processing of suboptimal microRNA hairpins through the action of the ERH protein. Mol. Cell *78*, 289–302.e6.

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

Friedländer, M.R., Mackowiak, S.D., Li, N., Chen, W., and Rajewsky, N. (2012). miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic Acids Res. *40*, 37–52.

Friedländer, M.R., Lizano, E., Houben, A.J., Bezdan, D., Báñez-Coronel, M., Kudla, G., Mateu-Huertas, E., Kagerbauer, B., González, J., Chen, K.C., et al. (2014). Evidence for the biogenesis of more than 1,000 novel human microRNAs. Genome Biol. *15*, R57.

Friedman, R.C., Farh, K.K.-H., Burge, C.B., and Bartel, D.P. (2009). Most mammalian mRNAs are conserved targets of microRNAs. Genome Res. *19*, 92–105.

Fromm, B., Domanska, D., Høye, E., Ovchinnikov, V., Kang, W., Aparicio-Puerta, E., Johansen, M., Flatmark, K., Mathelier, A., Hovig, E., et al. (2020). MirGeneDB 2.0: The metazoan microRNA complement. Nucleic Acids Res. *48* (*D1*), D132–D141.

Garcia-Martin, J.A., and Clote, P. (2015). RNA thermodynamic structural entropy. PLoS ONE *10*, e0137859.

Gibson, D.G., Young, L., Chuang, R.-Y., Venter, J.C., Hutchison, C.A., 3rd, and Smith, H.O. (2009). Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat. Methods *6*, 343–345.

Giraldez, A.J., Cinalli, R.M., Glasner, M.E., Enright, A.J., Thomson, J.M., Baskerville, S., Hammond, S.M., Bartel, D.P., and Schier, A.F. (2005). MicroRNAs regulate brain morphogenesis in zebrafish. Science *308*, 833–838.

Grimson, A., Srivastava, M., Fahey, B., Woodcroft, B.J., Chiang, H.R., King, N., Degnan, B.M., Rokhsar, D.S., and Bartel, D.P. (2008). Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. Nature *455*, 1193–1197.

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., 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.

Hutter, K., Lohmüller, M., Jukic, A., Eichin, F., Avci, S., Labi, V., Szabo, T.G., Hoser, S.M., Hüttenhofer, A., Villunger, A., and Herzog, S. (2020). SAFB2 enables the processing of suboptimal stem-loop structures in clustered primary miRNA transcripts. Mol. Cell *78*, 876–889.e6.

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.

Iwasaki, S., Kobayashi, M., Yoda, M., Sakaguchi, Y., Katsuma, S., Suzuki, T., and Tomari, Y. (2010). Hsc70/Hsp90 chaperone machinery mediates ATP-dependent RISC loading of small RNA duplexes. Mol. Cell *39*, 292–299.

Janssen, H.L.A., Reesink, H.W., Lawitz, E.J., Zeuzem, S., Rodriguez-Torres, M., Patel, K., van der Meer, A.J., Patick, A.K., Chen, A., Zhou, Y., et al. (2013). Treatment of HCV infection by targeting microRNA. N. Engl. J. Med. *368*, 1685–1694.

Jin, W., Wang, J., Liu, C.-P., Wang, H.-W., and Xu, R.-M. (2020). Structural basis for pri-miRNA recognition by Drosha. Mol. Cell *78*, 423–433.e5.

Kang, W., Eldfjell, Y., Fromm, B., Estivill, X., Biryukova, I., and Friedländer, M.R. (2018). miRTrace reveals the organismal origins of microRNA sequencing data. Genome Biol. *19*, 213.

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

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

Knight, S.W., and Bass, B.L. (2001). A role for the RNase III enzyme DCR-1 in RNA interference and germ line development in *Caenorhabditis elegans*. Science *293*, 2269–2271.

Kozomara, A., and Griffiths-Jones, S. (2014). miRBase: Annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res. *42*, D68–D73.

Kretov, D.A., Walawalkar, I.A., Mora-Martin, A., Shafik, A.M., Moxon, S., and Cifuentes, D. (2020). Ago2-dependent processing allows miR-451 to evade the global microRNA turnover elicited during erythropoiesis. Mol. Cell *78*, 317–328.e6.

Kwon, S.C., Baek, S.C., Choi, Y.-G., Yang, J., Lee, Y.-S., Woo, J.-S., and Kim, V.N. (2019). Molecular basis for the single-nucleotide precision of primary microRNA processing. Mol. Cell *73*, 505–518.e5.

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

Lee, Y., Jeon, K., Lee, J.-T., Kim, S., and Kim, V.N. (2002). MicroRNA maturation: Stepwise processing and subcellular localization. EMBO J. *21*, 4663–4670.

Lee, Y., Kim, M., Han, J., Yeom, K.-H., Lee, S., Baek, S.H., and Kim, V.N. (2004). MicroRNA genes are transcribed by RNA polymerase II. EMBO J. *23*, 4051–4060.

Li, S., Nguyen, T.D., Nguyen, T.L., and Nguyen, T.A. (2020a). Mismatched and wobble base pairs govern primary microRNA processing by human Microprocessor. Nat. Commun. *11*, 1926.

Li, S., Le, T.N.-Y., Nguyen, T.D., Trinh, T.A., and Nguyen, T.A. (2020b). Bulges control pri-miRNA processing in a position and strand-dependent manner. RNA Biol. Published online December 31, 2020. https://doi.org/10.1080/15476286.2020.1868139.

Lim, L.P., Lau, N.C., Garrett-Engele, P., Grimson, A., Schelter, J.M., Castle, J., Bartel, D.P., Linsley, P.S., and Johnson, J.M. (2005). Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. Nature *433*, 769–773.

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

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

Lund, E., Güttinger, S., Calado, A., Dahlberg, J.E., and Kutay, U. (2004). Nuclear export of microRNA precursors. Science *303*, 95–98.

Nguyen, T.L., Nguyen, T.D., Bao, S., Li, S., and Nguyen, T.A. (2020). The internal loops in the lower stem of primary microRNA transcripts facilitate single cleavage of human Microprocessor. Nucleic Acids Res. *48*, 2579–2593.

Nussbacher, J.K., and Yeo, G.W. (2018). Systematic discovery of RNA binding proteins that regulate microRNA levels. Mol. Cell *69*, 1005–1016.e7.

Okada, C., Yamashita, E., Lee, S.J., Shibata, S., Katahira, J., Nakagawa, A., Yoneda, Y., and Tsukihara, T. (2009). A high-resolution structure of the pre-microRNA nuclear export machinery. Science *326*, 1275–1279.

Okamura, K., Liu, N., and Lai, E.C. (2009). Distinct mechanisms for microRNA strand selection by *Drosophila* Argonautes. Mol. Cell *36*, 431–444.

Partin, A.C., Zhang, K., Jeong, B.-C., Herrell, E., Li, S., Chiu, W., and Nam, Y. (2020). Cryo-EM structures of human Drosha and DGCR8 in complex with primary microRNA. Mol. Cell *78*, 411–422.e4.

Pawlicki, J.M., and Steitz, J.A. (2008). Primary microRNA transcript retention at sites of transcription leads to enhanced microRNA production. J. Cell Biol. *182*, 61–76.

Reuter, J.S., and Mathews, D.H. (2010). RNAstructure: Software for RNA secondary structure prediction and analysis. BMC Bioinformatics *11*, 129.

Rice, G.M., Shivashankar, V., Ma, E.J., Baryza, J.L., and Nutiu, R. (2020). Functional atlas of primary miRNA maturation by the Microprocessor. Mol. Cell *80*, 892–902.e4.

Roden, C., Gaillard, J., Kanoria, S., Rennie, W., Barish, S., Cheng, J., Pan, W., Liu, J., Cotsapas, C., Ding, Y., and Lu, J. (2017). Novel determinants of mammalian primary microRNA processing revealed by systematic evaluation of hairpin-containing transcripts and human genetic variation. Genome Res. *27*, 374–384.

Rouleau, S.G., Garant, J.M., Bolduc, F., Bisaillon, M., and Perreault, J.P. (2018). G-quadruplexes influence pri-microRNA processing. RNA Biol. *15*, 198–206.

Sahu, S.S., Dey, S., Nabinger, S.C., Jiang, G., Bates, A., Tanaka, H., Liu, Y., and Kota, J. (2019). The role and therapeutic potential of miRNAs in colorectal liver metastasis. Sci. Rep. *9*, 15803.

Schütze, T., Rubelt, F., Repkow, J., Greiner, N., Erdmann, V.A., Lehrach, H., Konthur, Z., and Glökler, J. (2011). A streamlined protocol for emulsion polymerase chain reaction and subsequent purification. Anal Biochem. *410*, 155–157. https://doi.org/10.1016/j.ab.2010.11.029.

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.

Shang, R., Baek, S.C., Kim, K., Kim, B., Kim, V.N., and Lai, E.C. (2020). Genomic clustering facilitates nuclear processing of suboptimal pri-miRNA loci. Mol. Cell *78*, 303–316.e4.

Treiber, T., Treiber, N., Plessmann, U., Harlander, S., Daiß, J.-L., Eichner, N., Lehmann, G., Schall, K., Urlaub, H., and Meister, G. (2017). A compendium of RNA-binding proteins that regulate microRNA biogenesis. Mol. Cell *66*, 270–284.e13.

Warf, M.B., Johnson, W.E., and Bass, B.L. (2011). Improved annotation of *C. elegans* microRNAs by deep sequencing reveals structures associated with processing by Drosha and Dicer. RNA *17*, 563–577.

Yi, R., Qin, Y., Macara, I.G., and Cullen, B.R. (2003). Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. Genes Dev. *17*, 3011–3016.

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.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and virus strains		
NEB DH 5-alpha *E. coli*	NEB	C2987H
One Shot TOP10 Chemically Competent *E. coli*	Invitrogen	C404006
Chemicals, peptides, and recombinant proteins		
PrimeStar GXL DNA polymerase	Takara-Clontech	R050A
T4 RNA ligase 2, deletion mut	NordicBiolabs/Lucigen	LR2D11310K
Superscript II and III	Invitrogen	18064071, 18080085
Novex TBE PAAG, 6%	Invitrogen	EC6265BOX
Costar spin-X(R) centrifuge tube filters	Sigma	CLS8162
GlycoBlue Coprecipitant	Ambion/ Invitrogen	AM9515
Sodium Acetate Solution (3 M), pH 5.2	Thermo Scientific	R1181
Critical commercial assays		
Gibson Assembly Master Mix	NEB	E2611S, E2611L
Lipofectamine 3000	Invitrogen	L3000008
Trizol	Ambion/ Invitrogen	15596026
Quick-RNA Microprep Kit	ZYMO RESEARCH	R1050/R1054
Agilent RNA 6000 Nano Kit	Agilent	5067-1511
Agilent High Sensitivity DNA Kit	Agilent	5067-4626
Qubit RNA Broad-Range assay	Invitrogen	Q10211
Qubit dsDNA High Sensitivity assay	Invitrogen	Q32854
Qubit dsDNA Broad-Range assay	Invitrogen	Q32853
TruSeq Small RNA Library Prep Kit	Illumina	RS-200-0012, RS-200-0024, RS-200-0036, RS-200-0048
NextSeq 500/550 High Output v2 kit (75 cycles)	Illumina	FC-404-2205
NextSeq 500/550 Mid Output v2 kit (150 cycles)	Illumina	FC-404-2001
NextSeq 500/550 High Output v2 kit (150 cycles)	Illumina	FC-404-2002
Deposited data		
Raw and processed MapToCleave sequencing data	This paper; GEO	GEO: GSE169020
Experimental validation of stability of lower basal stem tuning of microRNA processing	This paper; GEO	GEO: GSE169020
MapToCleave library. Related to STAR Methods.	This paper; Mendeley Data	Mendeley Data: https://doi.org/10.17632/4zb54wsfxz.1
Count table of MapToCleave hairpins tested in HEK and NIH 3T3 cells. Related to Figure 1.	This paper; Mendeley Data	Mendeley Data: https://doi.org/10.17632/6xtgkhsbds.1
MapToCleave processed miRNA precursors. Related to Figure 2B.	This paper; Mendeley Data	Mendeley Data: https://doi.org/10.17632/4xzrxbzdv8.1
MapToCleave unprocessed miRNA precursors. Related to Figure 2B.	This paper; Mendeley Data	Mendeley Data: https://doi.org/10.17632/msm34n43j4.1
Count table of MapToCleave miRNA precursors for cell type specific processing. Related to Figure 2A.	This paper; Mendeley Data	Mendeley Data: https://doi.org/10.17632/t5jdrzfwjs.1
Detailed feature profile of MapToCleave processed and unprocessed miRNA precursors. Related to Figure 3.	This paper; Figshare	figshare: https://doi.org/10.17044/scilifelab.15134739
Local free energy in 7-mer of MapToCleave processed and unprocessed miRNA precursors. Related to Figures 3C and 3D.	This paper; Mendeley Data	Mendeley Data: http://dx.doi.org/10.17632/gzmj7tctgs.1

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
MapToCleave feature importance estimation. Related to Figure 7B.	This paper; Mendeley Data	Mendeley Data: https://doi.org/10.17632/sgjv7tm2x3.1
Detailed feature profile of *in vivo* highly and lowly expressed miRNA precursors. Related to Figure 4.	This paper; Figshare	figshare: https://doi.org/10.17044/scilifelab.15134862
Modified miRNA precursors for experimental validation. Related to Figure 5.	This paper; Mendeley Data	Mendeley Data: https://doi.org/10.17632/zk257mkcvb.1
Count table for experimental validation. Related to Figure 5.	This paper; Mendeley Data	Mendeley Data: https://doi.org/10.17632/xkvpw9drvr.1
Drosha cleavage precision of MapToCleave processed miRNA precursors. Related to Figures 1H and 6A.	This paper; Mendeley Data	Mendeley Data: https://doi.org/10.17632/z6zf48nvct.1
GHG feature profiling with multiple definitions. Related to Figure 6E.	This paper; Figshare	figshare: https://doi.org/10.17044/scilifelab.15144339
In vivo expression feature importance estimation. Related to Figure 7B.	This paper; Mendeley Data	Mendeley Data: https://doi.org/10.17632/cgnggdp2by.1
MirGeneDB	Fromm et al., 2020	https://mirgenedb.org/
miRBase release 21	Kozomara and Griffiths-Jones, 2014	https://www.mirbase.org/
Experimental models: Cell lines		
HEK293T	SciLifeLab	N/A
NIH 3T3	SciLifeLab	N/A
MEF	SciLifeLab	N/A
HeLa	Stockholm University	N/A
Oligonucleotides		
MapToCleave sequences	This paper	Mendeley Data: https://doi.org/10.17632/4zb54wsfxz.1
Sequences used for experimental validation of stability of lower basal stem tuning of microRNA processing	This paper	Mendeley Data: https://doi.org/10.17632/zk257mkcvb.1
Oligonucleotides used for targeted DNA library construction and custom NGS	This paper	Table S1
Recombinant DNA		
MapToCleave library	This paper	N/A
Control and tagged pAH-C5-hsa-mir-125a	This paper	N/A
Control and tagged pAH-C5-hsa-mir-16-1	This paper	N/A
Control and tagged pAH-C5-hsa-mir-30a	This paper	N/A
Software and algorithms		
MapToCleave analysis pipeline	This paper	Zenodo: https://doi.org/10.5281/zenodo.5519203 or GitHub: https://github.com/wenjingk/MapToCleave
miRTrace version 1.0.1	Kang et al., 2018	https://github.com/friedlanderlab/mirtrace
bowtie version 1.1.2	Langmead et al., 2009	http://bowtie-bio.sourceforge.net/index.shtml
DESeq2 version 1.22.2	Love et al., 2014	https://bioconductor.org/packages/release/bioc/html/DESeq2.html
miRDeep2 version 2.0.0.8	Friedländer et al., 2012	https://www.mdc-berlin.de/content/mirdeep2-documentation
ViennaRNA Package	Lorenz et al., 2011	https://www.tbi.univie.ac.at/RNA/
RNAstructure version 6.2	Reuter and Mathews, 2010	https://rna.urmc.rochester.edu/RNAstructure.html
R version 3.5.3	The R Foundation	https://www.r-project.org/
Python version 3.6	Python Software Foundation	https://www.python.org/

RESOURCE AVAILABILITY

Lead contact

Further information for resources and reagents should be directed to and will be fulfilled by the lead contact Marc Friedländer (marc.friedlander@scilifelab.se).

Materials availability

All reagents generated in this study are available from the lead contact without restriction.

Data and code availability

- MapToCleave sequencing data and experimental validation data of stability of lower basal stem tuning of microRNA processing have been deposited at Gene Expression Omnibus (GEO) with accession number GSE169020. The supplemental datasets "Detailed feature profile of MapToCleave processed and unprocessed miRNA precursors" (https://doi.org/10.17044/scilifelab.15134739), "Detailed feature profile of *in vivo* highly and lowly expressed miRNA precursors" (https://doi.org/10.17044/scilifelab.15134862), and "GHG feature profiling with multiple definitions" (https://doi.org/10.17044/scilifelab.15144339) have been deposited on figshare, and the other supplemental datasets have been deposited on Mendeley Data. All the supplemental datasets are publicly available. DOIs are listed in the key resources table.
- The original code has been deposited on Zenodo and is publicly available. DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Human HEK293T (female) and HeLa (female), murine MEF (male and female mixed) and NIH 3T3 (male) were cultured in DMEM (Sigma-Aldrich, D6429) supplemented with 10%–15% heat-inactivated fetal bovine serum (GIBCO, 10500064) and penicillin-streptomycin (GIBCO, 15140122) under standard conditions. All cell lines were maintained at 37°C and 5% CO2. All cell lines in culture were routinely tested for mycoplasma using a qPCR-based test (Eurofins Genomics) and are mycoplasma negative.

METHOD DETAILS

MapToCleave library design

We selected 12,472 sequences to include in our MapToCleave library; please refer to Mendeley data (https://doi.org/10.17632/4zb54wsfxz.1). These include known human miRNA precursors that are not highly expressed in HEK293T and HeLa cells (test sequences) and sequences that are not predicted to form any hairpin structures (negative controls). The sequences also included novel predicted human miRNAs (Friedländer et al., 2014), hairpin-forming human sequences from exonic, intronic and intergenic regions, and known miRNAs from non-human animal, sponge, mold or plant species ('cross-species miRNAs'). STAR Methods Tables ''MapToCleave hairpin library'' and ''MapToCleave non-human miRNA hairpins'' give an overview of all sequences.

One expression library was generated and used for all experiments (see below). It was tested in human cell cultures HEK293T and mouse cell culture NIH 3T3. For the HEK293T cell, two concentrations of library (1 times and 10 times the concentration) and mock transfections were performed. For the NIH 3T3 cell, 1 times (1x) the concentration and mock transfections were performed. Four replicates were transfected for each condition. An Illumina small RNA TruSeq library was prepared for each transfection and sequenced in replicates (see below).

MapToCleave hairpin library

Type of sequence	Control	Number of sequences
Human known miRNAs	Positive control	921
Human sequence not forming hairpins	Negative control	1500
Random sequence not forming hairpins	Negative control	1500
Human predicted novel miRNAs		2469
Human predicted novel miRNAs (low-confidence)		2628
Human exonic hairpins		1000
Human intergenic and intronic hairpins		1000
Known miRNAs from non-human species		1454

MapToCleave non-human miRNA hairpins

Species	Organism	Number of sequences
M. musculus	Mouse	498
D. melanogaster	Fruit fly	221
C. elegans	Nematode	214
S. mediterranea	Planaria	137
N. vectensis	Sea anemone	48
A. queenslandica	Animal sponge	5
D. discoideum	Slime mold	2
C. reinhardtii	Single-cell algae	41
A. thaliana	Eudicot plant	288

MapToCleave sequence cloning

To generate the expression library, 118 nucleotides in length pre-miRNA hairpin, hairpin-forming and non-forming sequences, and in addition adjacent 5′- and 3′- PCR-adaptor sequences containing recognition sites for *XhoI* and EcoRI, were chemically synthesized (CustomArrayInc, US). In total, 12,472 ssDNA oligonucleotide sequences were converted into dsDNA and amplified using emulsion PCR, as previously described (Schütze et al., 2011), with the following PCR-adaptor primers: FRW 5′-AGGGATAACAGGGT AATCTCGAG-3′ and REV 5′-CTACCCGGTAGAATTGAAAGAATTC-3′. The PCR product was inserted into modified pMSCV-LMP-tomato vector (OpenBiosystems), pAH_c5 digested by *XhoI* and EcoRI (Thermo Scientific) using Gibson assembly cloning (Gibson et al., 2009). The resulting plasmids were transformed into *E. coli* DH10β cells by electroporation. Approximately 880,000 individual colonies were mixed and prepped in order to generate hairpin library pools. The hairpin library sequences were confirmed by the 150-bp single-end sequencing on an Illumina NextSeq 500 instrument.

Cell culture and MapToCleave transfections

HEK293T and NIH 3T3 were cultured in DMEM (Sigma-Aldrich) supplemented with 10%–15% heat-inactivated fetal bovine serum (GIBCO) and penicillin-streptomycin (GIBCO) under standard conditions. For the standard assays, HEK293T cells grown in 12-well plates were transiently transfected with either 1.8 μg of pAH_c5 (mock control) or an incremental amount of hairpin library (1.8 μg or 18 μg) using Lipofectamine 3000 (Invitrogen) following the manufacturer's protocol. Transfection efficiency was confirmed by tomato expression. Transient expression in murine cells was performed in 10 cm dishes. NIH 3T3 cells were transiently transfected with either 26 μg of pAH_c5 or 26 μg of the hairpin library. The NIH 3T3 tomato-positive cells were sorted on a BD Influx (BD Bioscience). 48 h after transfection, total RNA was extracted. miRNA expression was assessed by small RNA sequencing, and transfection efficiency was assayed by DNA sequencing.

MapToCleave DNA and small RNA sequencing

Total RNA and DNA were isolated using TRIzol reagent (Ambion). RNA integrity was estimated with a Bioanalyzer instrument using an RNA 6000 Nano kit (Agilent Technologies). 1 μg of HEK293T total RNA was used for standard small RNA library preparation using TruSeq small RNA kit v2 (Illumina). 1 μg of NIH 3T3 total RNA was used for preparation of small RNA libraries. The small RNA cDNA libraries were PCR-amplified in 15 cycles; the 75-bp single-end sequencing was carried out on a NextSeq500 (Illumina). DNA libraries were prepared in two steps: i) the DNA isolated from the HEK293T and NIH 3T3 transfected cells was pre-amplified using custom primers compatible with Illumina TruSeq adaptor sequences, ii) pre-amplified DNA fragments were multiplexed by PCR using Illumina TruSeq universal forward and single-indexed reverse PCR primers. The DNA libraries were resolved on a 6% Novex TBE gel (Invitrogen), and the 280 – 300 bp fraction was isolated. The 150-bp single-end sequencing was carried out using Illumina NextSeq500.

MapToCleave sequence data quality control and pre-processing

The small RNA sequencing data were quality control (QC) checked using miRTrace v 1.0.1 (Kang et al., 2018) qc mode. Human (miRTrace option -s hsa) and mouse (-s mmu) databases were used as references for HEK293T and NIH 3T3 samples, respectively. Illumina 3′ adaptor sequence TGGAATTCT was provided for 3′ adaptor trimming. The QC qualified reads of each sample were then aligned to the prepared reference sequences, which contained the sequences from the MapToCleave library and the endogenous human (for HEK293T sample) or mouse (for NIH 3T3 sample) miRNA hairpin sequences that are excluded from the MapToCleave library but are included in miRbase v 21 (Kozomara and Griffiths-Jones, 2014). The alignment was performed using bowtie v 1.1.2 (Langmead et al., 2009) without allowing mismatches (-v 0). The reads that were uniquely mapped to

the forward strand (-m 1) of the reference sequences were considered for the hairpin expression measurement; please refer to Mendeley data (https://doi.org/10.17632/6xtgkhsbds.1). The RNA counts were then normalized to reads per million (RPM) for each sample using the equations: *hairpin counts*/*total counts of human hairpins*$\times 10^6$ for the HEK293T samples and *hairpin counts*/*total counts of mouse hairpins*$\times 10^6$ for the NIH 3T3 samples. The DNA sequencing data were also aligned to the reference sequences in the same way as the small RNA sequencing data except for allowing one mismatch (-v 1) and considering only reads that were > 100 nucleotides in length; please refer to Mendeley data (https://doi.org/10.17632/6xtgkhsbds.1). The DNA counts were normalized to reads per million (RPM) for each sample using the formula *hairpin counts*/*total counts of hairpins*$\times 10^6$.

Estimating processing efficiency of MapToCleave sequences

To identify the hairpins that are differentially expressed (DE) in the cells transfected with the MapToCleave library compared to the mock cells, we applied DESeq2 v 1.22.2 (Love et al., 2014) Wald significance tests with the cutoffs: p value < 0.01, adjusted p value < 0.05 and absolute log2 fold-change > 1. A hairpin is defined as efficiently processed if the expression vector is successfully transfected with DNA RPM > = 5 and the hairpin is differentially expressed in the transfection cells. If the processed or unprocessed hairpins are annotated in MirGeneDB or miRBase (when MirGeneDB annotation is not available), we call them processed or unprocessed miRNA precursors; please refer to Mendeley data (https://doi.org/10.17632/4xzrxbzdv8.1 and https://doi.org/10.17632/msm34n43j4.1). The processing efficiency is represented by the fold-change of mean hairpin expression (RPM) between the transfection cells and the mock cells. This analysis relates to Figures 1 and 2B. In Figure 2B, the purple heatmap, the lower basal stem is defined as stable if the free energy of the duplex at −13 to −7 is lower than −2.3, −1.8, −1.3, −2.0, −1.9 and −4.2 (kcal/mol) respectively for human, mouse, fruit fly, nematode, planarian and sea anemone miRNA precursors. The processing efficiency is used in Figure 7B.

Profiling cell type-dependent processing of MapToCleave sequences

We measured the processing efficiency of MapToCleave hairpins in HEK293T cells transfected with 1x or 10x of the concentration of MapToCleave library and in the NIH 3T3 cells transfected with 1x the concentration of MapToCleave library in the following way. We applied quantifier.pl script from miRDeep2 v 2.0.0.8 to quantify the read counts of miRNAs, without allowing mismatches (-g 0) and with considering all aligned reads, including multiple aligned reads. The miRBase human (hsa), mouse (mmu), *Drosophila* (dme) and nematode (cel) miRNA hairpins were used as the precursor reference (-p) and the corresponding 5p and 3p arm sequences were used as the mature reference (-m). The read counts (see Mendeley data, https://doi.org/10.17632/t5jdrzfwjs.1) were then normalized to RPM using the following formulas: *miRNA counts*/*total counts of human miRNAs*$\times 10^6$ for the HEK293T samples and *miRNA counts*/*total counts of mouse miRNAs*$\times 10^6$ for the NIH 3T3 samples. The expression of miRNA hairpins was calculated by summing up the RPM expression of the corresponding arm sequences. The processing efficiency is represented by the difference of mean hairpin expression (RPM) between the transfection cells and the mock cells. The MapToCleave miRNA precursors that are processed in either the HEK293T or NIH 3T3 cells were used to generate Figure 2A. In Figure 2A, the HEK replicate 1 represents HEK293T cells with the 1x transfection concentration. The HEK replicate 2 represents HEK293T cells with the 10x transfection concentration.

Identifying structural features and sequence motifs in MapToCleave hairpins

The MapToCleave hairpin sequences were folded using RNAfold v 2.4.2 from the ViennaRNA Package (Lorenz et al., 2011) with the default setting but forced open if the flanking sequences are 19 nucleotides away from the 5′ side of Drosha cleavage site and 17 nucleotides away from the 3′ side of Drosha cleavage site. Each hairpin can be divided into four parts, comprising the flanking region, basal stem, miRNA duplex and apical loop, depending on the single and double stranded structure changes and Drosha cleavage site (illustrated by the example hairpin below). Since Drosha cleavage is the critical entry point for canonical miRNA biogenesis, we defined the coordinates along the hairpin stem loop relative to the Drosha cleavage site at the 5′ strand. For example, in the hairpin stem loop of hsa-mir-371a (Figure S12), the nucleotides away from the Drosha cleavage site to the apical loop are counted from 0 to 28 and 0 to 27 for the 5′ and 3′ strand. The nucleotides away from the Drosha cleavage site to the flanking sequences are counted from −1 to −30 and −1 to −31 for the 5′ and 3′ strand respectively. The 5p and 3p arm sequences at the 5′ and 3′ strand are colored in red and blue respectively. In this coordinate system, all the features were counted according to the relative distance from the Drosha cleavage site.

To determine the Drosha cleavage site of each hairpin in the HEK293T and NIH 3T3 cells, the QC qualified small RNA reads from all HEK293T or NIH 3T3 samples were pooled together and then aligned to the reference sequences in the same way as the previous alignment for MapToCleave sequence pre-processing. The mapping profile of each hairpin generated by the perfectly, uniquely and forward mapped reads was used to determine the Drosha cleavage site, which is identified by the 5′ end of the most abundant reads that aligned to the 5′ strand of the hairpin stem. For the hairpins with no reads aligned to the expected region, we used miRBase annotation to infer the Drosha cleavage site, which is located at the 5′ end of the 5p arm.

For each hairpin, we parsed the secondary structure to profile base pairing and mismatch information of each nucleotide at the 5′ strand and at the 3′ strand of the hairpin. If the nucleotide is base paired, we record the type of base pair, including A-U, U-A, C-G, G-C, G-U or U-G; if the nucleotide is unpaired and present as part of a bulge, we record the features of the bulge, including its size, symmetry and nucleotide content. We also checked if the known sequence motifs that influence miRNA processing are present at the

expected position. For example, the GU motif at the basal junction is expected to be located at position −15 to −12 of the 5′strand; the UGU or GUG motif at the apical junction is expected to be located at position + 19 to +27 of the 5′ strand; the CNNC motif is expected to be located at position −24 to −17 of the 3′ strand. The detailed feature profile of MapToCleave processed and unprocessed miRNA precursors are available in the figshare data (https://doi.org/10.17044/scilifelab.15134739). This analysis relates to Figures 3A and 3B.

Profiling local free energy of MapToCleave hairpins

To quantify local structure changes in terms of tightness or looseness across the hairpin, we measured the free energy of local regions. We first extracted local duplex segments by sliding a 7-mer window over the hairpin stem loop predicted by RNAfold from the flanking region to the apical loop. Since the size of the sliding window represents the number of nucleotides extracted from the 5′ strand, each segment always has 7 nucleotides from the 5′ stand but can have different numbers of nucleotides from the 3′ stand depending on the presence of bulges. For each segment, we then calculated the free energy using RNAeval v 2.4.2 from the ViennaRNA Package; please refer to Mendeley data (https://doi.org/10.17632/gzmj7tctgs.1). This analysis relates to Figures 3C and 3D.

Profiling Shannon entropy of MapToCleave hairpins

We calculated the structural positional entropy of each hairpin using the same approach as in the study by Rice et al. (Rice et al., 2020) (see Mendeley data, https://doi.org/10.17632/sgjv7tm2x3.1). For a given RNA sequence $\boldsymbol{a} = a_1, \ldots, a_i, a_j, \ldots, a_n$ and for $1 \leq i \leq j \leq n$, $p_{i,j}$ the probability of pairing for nucleotides i and j given the whole set of secondary structures is calculated using RNAstructure v 6.2 (Reuter and Mathews, 2010) by running partition function followed by ProbabilityPlot with -t option for text file output. We define the positional base pairing probability distribution at fixed position $1 \leq i \leq n$ by

$$p^*_{i,j} = \begin{cases} p_{i,j} & \text{if } i < j \\ p_{j,i} & \text{if } i > j \\ p_{i,j} = 0 & \text{if } i = j \end{cases}$$

The positional Shannon entropy H_i at nucleotide i, which provides a measure of local structural certainty, is calculated by $H_i = -\sum_{j=1}^{n} p^*_{i,j} log_{10} p^*_{i,j}$, with the convention for $p^*_{i,j} = 0$ that $\lim_{p \to 0} p log_{10} p = 0$. Low values of positional Shannon entropy at nucleotide i indicate the strong agreement among low energy structures. Therefore, the well-defined regions have low Shannon entropy. It should be noted that a given nucleotide does not pair with itself. Therefore, to calculate Shannon entropy, one needs to provide an additional assumption on $i = j$ case. The simplest way is to assume the $p_{i=j} = 0$, as described earlier (Rice et al., 2020). Alternatively, one can consider $p_{i=j} = 1 - \sum_{i \neq j} p^*_{i,j}$, which is used in mountain.pl script of ViennaRNA package and described in the section "Positional entropy" from Garcia-Martin and Clote (Garcia-Martin and Clote, 2015). This analysis relates to Figure 7B.

MirGeneDB miRNA expression analysis

Using the MirGeneDB miRNA expression atlas of tissues from twenty species that are grouped into four clades, namely, mammals, fruit flies, nematodes and lophotrochozoans (Fromm et al., 2020), we calculated the mean RPM of miRNA precursors by species. The miRNAs were then sorted according to the mean RPM. The top and bottom 30 miRNAs of each species were selected as highly and lowly expressed miRNAs and pooled by clade. This analysis relates to Figure 4. All the MirGeneDB miRNAs used in the study were processed in the same way as the MapToCleave hairpins to profile structure features, sequence motifs and Shannon entropy. MirGeneDB annotation was used to identify the Drosha cleavage site, which is important to define the coordinate of the hairpin stem. The detailed feature profile of *in vivo* highly and lowly expressed miRNA precursors in 20 species are available in the figshare data (https://doi.org/10.17044/scilifelab.15134862).

Design of miRNA precursors with improved or impaired processing

Hsa-mir-30a, hsa-mir-16 and hsa-mir-125a wild-type, tagged and mutant hairpins with 5′- and 3′-flanking adaptor sequences were synthesized (GeneArt, Thermofisher). The hairpins were subcloned into modified pMSCV-LMP-tomato vector (OpenBiosystems) using Gibson assembly cloning (NEB). HEK293T cells were transfected using Lipofectamin-3000 (Invitrogen). The hairpins were tested in triples; 0.6 μg of each hairpin was used for transfection in two biological replicates. The tested hairpin oligonucleotide sequences and hairpin combinatorial sets used in the lower basal stem stability validation are available in our Mendeley data (https://doi.org/10.17632/zk257mkcvb.1) and in Table S2. After 48 hours, total RNA was extracted using a Zymo quick-RNA microprep kit (Zymo Research). miRNA expression was assayed through small RNA sequencing using a TruSeq small RNA kit v2 (Illumina). We first checked the quality of the small RNA sequencing data generated from the validation experiments using miRTrace v 1.0.1 qc mode with option–species hsa–adaptor TGGAATTCT. The QC qualified reads were used to quantify expression of the human miRNA hairpins and the modified hairpins (Mendeley data, https://doi.org/10.17632/zk257mkcvb.1) using miRDeep2 (Friedländer et al., 2012) quantifier.pl with allowing 0 mismatches (-g 0) and allowing multi-mapping. The human miRNA hairpin sequences downloaded from miRBase v21 and the modified hairpin sequences were used as the precursor reference (-p). The corresponding human miRNA

sequences from miRBase v21 and the manually curated miRNA sequences of the modified hairpins (Mendeley data, https://doi.org/10.17632/zk257mkcvb.1) are used as the mature reference (-m). The hairpin sequence counts (Mendeley data, https://doi.org/10.17632/xkvpw9drvr.1) are measured by summing up the number of small RNA reads that mapped to an interval 2 nucleotides upstream and 5 nucleotides downstream of the 5p or/and 3p miRNA sequence from the hairpin. The hairpin counts were further normalized to RPM using the formula: $hairpin\ counts / total\ counts\ of\ hairpins \times 10^6$. This analysis relates to Figure 5.

Profiling Drosha cleavage precision of MapToCleave hairpins

The QC-passed reads from four replicates of 1x HEK transfection samples were pooled together and then aligned to the sequences of the MapToCleave library using bowtie v 1.1.2 without allowing mismatches (-v 0). The reads that were uniquely mapped to the forward strand of the reference sequences were used for the analysis; please refer to Mendeley data (https://doi.org/10.17632/z6zf48nvct.1). The Drosha cleavage precision was calculated by the number of reads aligned to the dominant cut site divided by the number of reads aligned in the region ranging from 5 nucleotides upstream to 5 nucleotides downstream of the dominant cut site. This analysis relates to Figure 6A.

Identifying the presence of the GHG feature using different definitions

The GHG feature defined by Fang et al., 2015 (Fang and Bartel, 2015) is located at position −7 to −5 of the 5′ strand and meets the following requirements: C-G or U-G pair at position −7, C*U, U*C, G*A or A*C mismatch or U-A, G-C or A-U pair at position −6, A-U, U-A, G-C or C-G pair at position −5. The GHG feature defined by Kwon et al. (Kwon et al., 2019) is based on the cleavage scores reported by Fang et al. and requires the cleavage score to be higher than 65. In the study by Fang et al., Microprocessor cleavage of the hairpin variants that were generated by randomizing the duplex segment at position −7 to −5 of three template pri-miRNAs was tested. Since all the other regions are the same except for the mutated duplex segment, the cleavage score indicates the influence of the duplex segment for miRNA processing. By matching the duplex segments at position −7 to −5 of MapToCleave hairpins to the duplex segments at position −7 to −5 of hairpin variants tested by Fang et al., the mGHG scores downloaded from Table S1 of Kwon et al. (2019) are assigned to MapToCleave duplex segments if they are matching. In this way, we can identify the GHG feature defined by Kwon et al. using MapToCleave hairpins. Our definition of the GHG feature requires a bulge composed of one or two nucleotides at position −6 counted from the 5′ side of the Drosha cleavage site. The highly and lowly expressed miRNA precursors in the plot "*in vivo* expression" in the panel "miRNA processing efficiency" of Figure 6E are represented by the top 70 and bottom 70 bona fide human miRNAs, which are ranked in descending order by mean RPM expression across human tissues from MirGeneDB data. The highly and lowly processed miRNA precursors in the plot "Chromatin pri-miRNA" in the panel "miRNA processing efficiency" of Figure 6E are represented by the top 50 and bottom 50 bona fide human miRNAs, which are ranked in descending order by pri-miRNA processing efficiency as indicated by deltaMPI values in HeLa cells from Conrad et al. (2014). The related datasets are available in our figshare data https://doi.org/10.17044/scilifelab.15144339). This analysis relates to Figure 6E.

Estimating the relative importance of features for miRNA processing and expression

To estimate the importance of individual features and the various combinations of features for miRNA processing, we fitted a linear regression model with the feature (or features) as the explanatory variable and miRNA processing efficiency as the outcome variable using R language lm() function. The importance of the feature (or features) is reflected by the adjusted R squared value of the model built on the feature (or features). The presence of CNNC at 3′ flanking sequence, UG at basal junction and UGUG at apical junction of each miRNA precursor was used to define the sequence motif features. The local free energy of a 4-mer window at each position of the hairpin stem was treated as a structural element, some of which were further combined to represent the structural features of interest. For example, the stable lower basal stem feature that is located from position −13 to −7 is represented by five 4-mer windows beginning from position −14, −13, −12, −11 and −10. The two bulge-depleted regions identified by Roden et al. (2017) are represented by three 4-mer windows beginning from position 1, 2 and 3 and three 4-mer windows beginning from position 14, 15 and 16. The other unrecognized region is represented by the 4-mer windows left over from the above-mentioned structural features and the GHG feature defined by the 4-mer windows at position −8 and −7. The GHG feature is defined by the presence of a bulge at position −6 of the 5′ strand with maximum two mismatches on either side of the bulge. We performed feature importance estimation using both MapToCleave and MirGeneDB data. For MapToCleave, we used the data of processed and unprocessed miRNA precursors to fit the linear regression model, where the outcome variable is represented by log2 fold-change of the mean hairpin expression (RPM) between the transfected cells and the mock HEK293T cells; please refer to Mendeley data (https://doi.org/10.17632/sgjv7tm2x3.1). For MirGeneDB, we used the miRNA expression atlas of human tissues to fit the model, where the outcome variable is represented by log2 of mean RPM of the tissues; please refer to Mendeley data (https://doi.org/10.17632/cgnggdp2by.1). These analyses relate to Figure 7B.

Estimating the contribution of miRNA clustering to unprocessed MapToCleave precursors

We calculated the percentage of processed MapToCleave human miRNA precursors that are localized to within 30 kb of each other. The same calculation was applied for the unprocessed human miRNA precursors. The contribution of miRNA clustering to miRNA processing is estimated by calculating the percentage of clustered, processed human miRNA precursors minus the percentage of clustered, unprocessed human miRNA precursors.

QUANTIFICATION AND STATISTICAL ANALYSIS

All the statistical tests are performed using R v3.5.3. The unpaired and two-tailed Student's t test t.test() was used to test whether the mean stability of lower basal stem is equal between the processed and the unprocessed MapToCleave precursors (Figure 3). The two-tailed exact binomial test binom.test() was used to compare the relative proportions of miRNAs from hsa-mir-16-1 variant 1 and hsa-mir-16-1 wild-type (Figure 5). The exact binomial test was also used in Figure 6 to compare the GHG proportions between the precursors with high and low processing efficiency and the GHG proportions between the precursors with high and low Drosha cleavage precision.