Supplemental Table 1. Features calculated for building the novel miRNA detection model.

Feature
Sequence count in the cluster
Read count
The proportion of reads that are an exact match to the cluster sequence in the cluster
5’ Unstable length of the cluster
3’ Unstable length of the cluster
Proportion of genomic nucleotide at the positions -3, -2, -1 relative to the 5' end and +1, +2, +3, +4, +5, +6 relative to the 3' end of the stable range of the cluster sequences
Proportion of adenine (A) at the positions -3, -2, -1 relative to the 5' end and +1, +2, +3, +4, +5, +6 relative to the 3' end of the stable range of the cluster sequences
Proportion of thymine (T) at the positions -3, -2, -1 relative to the 5' end and +1, +2, +3, +4, +5, +6 relative to the 3' end of the stable range of the cluster sequences
Proportion of cytosine (C) at the positions -3, -2, -1 relative to the 5' end and +1, +2, +3, +4, +5, +6 relative to the 3' end of the stable range of the cluster sequences
Proportion of guanine (G) at the positions -3, -2, -1 relative to the 5' end and +1, +2, +3, +4, +5, +6 relative to the 3' end of the stable range of the cluster sequences
The arm where the stable range of sequences were located at the precursor hairpin structure
Distance between stable range of sequences and terminal loop
Number of bindings in the stable range of sequences divided by its length
Number of hairpin loops in the precursor
Number of bindings in the precursor
Number of interior loops in the precursor
Number of nucleotides in apical loop of the precursor
Length of stem in the precursor
Minimum free energy (MFE) of the precursor
Number of bindings in the stable range of sequences
Whether UGU or UGUG motif exists in the apical loop of the precursor
Whether there is another stable range of sequences located at the other arm of the precursor

Supplemental Table 2. 12 rat samples for evaluating the human and mouse novel miRNA predictive models.

Tissue Type	Sequence Read Archive (SRA) References
Lung	SRR3498192, SRR3498196, SRR3498182
Liver	SRR3498150, SRR3498148, SRR3498148
Kidney	SRR3498125, SRR3498118, SRR3498130
Heart	SRR3498075, SRR3498102, SRR3498103

Supplemental Table 3. All options to call in miRge 2.0.

Usage: miRge2.0.py annotate [-h] [<args>]

Annotate is just to find known miRNAs and other RNA species.

optional arguments:

-h, --help	show this help message and exit
-s	[sample <required> [sample <required> ...]] two options: 1. A file where each row represents one sample name; 2. *.fastq, *fastq, ...
-o	<dir> the directory of the outputs (default: current working directory)
-d	<string required> the miRNA database (default: miRBase. miRGeneDB is optional)
-pb	<dir required> the path to the system's bowtie binary
-lib	<dir required> the path to the miRge libraries
-sp	<string required> the species of comparison. Built in libraries are “human”, “mouse”, “fruitfly”, “nematode”, “rat” and other libraries can be built and called here “zebrafish” (novel miRNA detection is confined in human and mouse)
-ps	<dir required> the path to the system's samtools binary
-pr	<dir required> the path to the system's rnafold binary
-ad	<string> the adapter needed to be removed, which could be illumina, ion or a defined sequence (default: none)
-phred64	phred64 format (default: 64)
-spikeIn	switch to annotate spike-ins if the bowtie index files are located at the path of bowtie's index files (default: off)
-tcf	switch to write trimmed and collapsed fasta file (default: off)
-di	switch to calculate isomir entropy (default: off)
-cpu	<int> the number of processors to use for trimming, qc, and alignment (default: 1)
-ai	switch to calculate A to I editing (default: off)
-gff	switch to output results in gff format (default: off)
--version	show program's version number and exit

Usage: miRge2.0.py predict [-h] [<args>]

Predict will perform annotation as above AND novel miRNA prediction.

optional arguments:

-h, --help show this help message and exit

-s [sample <required> [sample <required> ...]]

Two options: 1. A file where each row represents one sample name; 2. *.fastq, *fastq, ...

-o <dir> the directory of the outputs (default: current working directory)

-d <string required> the miRNA database (default: miRBase. miRGeneDB is optional)

-pb <dir required> the path to the system's bowtie binary

-lib <dir required> the path to the miRge libraries

-sp <string required> the species of comparison. Built in libraries are “human”, “mouse”, “fruitfly”, “nematode”, “rat” and other libraries can be built and called here (novel miRNA detection is confined to human and mouse)

-ps <dir required> the path to the system's samtools binary

-pr <dir required> the path to the system's rnafold binary

-ad <string> the adapter need to be removed which could be illumina, ion or a defined sequence (default: none)

-phred64 phred64 format (default: 64)

-spikeIn switch to annotate spike-ins if the bowtie index files are located at the path of bowtie's index files (default: off)

-tcf switch to write trimmed and collapsed fasta file (default: off)

-di switch to calculate isomir entropy (default: off)

-cpu <int> the number of processors to use for trimming, qc, and alignment (default: 1)

-ai switch to calculate A to I editing (default: off)

-gff switch to output results in gff format (default: off)

-ws <file> the file containing the overall samples to analyze for novel miRNA prediction. No header, just a list of *.fastq file names in a column. Name of file can be to your choosing (e.g. filestochecknovel.txt)

-minl <int> the minimum length of the retained reads for novel miRNA detection (default: 16)

-maxl <int> the maximum length of the retained reads for novel miRNA detection (default: 25)

-cc <int> the maximum read count of the retained reads for novel miRNA detection (default: 2)

-ml <int> the maximum number of mapping loci for the retained reads for novel miRNA detection (default: 3)

-sl <int> the seed length when invoking Bowtie for novel miRNA detection (default: 25)

-olc <int> the length of overlapped sequence when joining reads into longer sequences based on the coordinate on the genome for novel miRNA detection (default: 14)

-clc <int> the maximum length of the clustered sequences for novel miRNA detection (default: 30)

--version show program's version number and exit

Supplemental Table 4. Top 12 features in the mouse novel miRNA predictive model. Hairpin structural features are labeled in bold/italics, while read compositional features are not.

Rank	Feature name	Description of the feature
1	***count_bindings_in_miRNA***	Number of bindings in the stable range of sequences
2	***pair_state_No***	Whether there is another stable range of sequences located at the other arm of precursor
3	***mFE***	Minimum free energy (MFE) of the precursor
4	***hairpin_count***	Number of hairpin loops in the precursor
5	***distanceToloop***	Distance between the stable range of sequences and the terminal loop
6	***stem_length***	Stem length of the precursor
7	exactMatchRatio	The proportion of reads that are an exact match to the cluster sequence in the cluster
8	headUnstableLength	5’ unstable length of the cluster
9	***percentage_PairedInMiRNA***	Number of bindings in the stable range of sequences divided by its length
10	***pair_state_Yes***	Whether there is another stable range of sequences located at the other arm of precursor
11	***binding_count***	Number of bindings in the precursor hairpin
12	***interiorLoopCount***	Number of interior loops in the precursor

Supplemental Table 5. Predictive results for 19 mouse RNA-seq datasets for novel miRNA detection.

Cell Type	SRA Reference	AUC	Precision	Recall	MCC
Smooth muscle cell	DRR041741	0.976	0.996	0.926	0.921
Erythroblast	DRR041759	0.976	1	0.917	0.926
Erythroblast	DRR041761	0.978	0.996	0.918	0.922
Erythroblast	DRR041762	0.977	0.996	0.916	0.929
Erythroblast	DRR041764	0.974	1	0.909	0.924
Erythroblast	DRR041765	0.976	0.991	0.909	0.922
Erythroblast	DRR041767	0.98	0.996	0.906	0.922
Erythroblast	DRR041768	0.972	0.995	0.902	0.912
Erythroblast	DRR041770	0.977	0.986	0.896	0.898
Embryo E17.5	DRR041745	0.991	0.996	0.958	0.957
Embryo E17.5	DRR041751	0.989	0.995	0.956	0.948
Embryo E12	DRR041771	0.983	0.988	0.937	0.939
Embryo E12	DRR041773	0.982	0.99	0.935	0.931
Embryo E15	DRR041774	0.984	0.992	0.926	0.93
Embryo E15	DRR041775	0.981	0.992	0.926	0.918
Embryo E18	DRR041777	0.982	0.985	0.933	0.939
Embryo E18	DRR041779	0.971	0.98	0.922	0.927
Neonate N09	DRR041780	0.982	0.995	0.947	0.951
Neonate N09	DRR041782	0.974	0.977	0.925	0.923
Mean		0.979	0.992	0.924	0.928
Std dev		0.005	0.006	0.017	0.014